УДК519.2

© 2022 г. **М.З. Ермолицкая,** канд. биол. наук
(Институт автоматики и процессов управления ДВО РАН, Владивосток)

ВЫЯВЛЕНИЕ ВЗАИМОСВЯЗИ МЕЖДУ ПОКАЗАТЕЛЯМИ КАЧЕСТВА ЖИЗНИ НАСЕЛЕНИЯ И ЗАБОЛЕВАЕМОСТЬЮ РАКОМ МОЛОЧНОЙ ЖЕЛЕЗЫ НА ТЕРРИТОРИИ ПРИМОРСКОГО КРАЯ

Проведен статистический анализ данных, который позволил выявить взаимосвязи между показателями качества жизни населения и заболеваемостью раком молочной железы на территории Приморского края. Значимые показатели сгруппированы в главные компоненты и могут быть использованы при построении прогнозных моделей, результаты которых, в свою очередь, могут быть пригодны при обосновании управленческих решений на региональном уровне.
Ключевые слова: статистический анализ данных, анализ главных компонент, регрессионный анализ, заболеваемость раком молочной железы.

DOI: 10.22250/18142400_2022_72_2_13

Введение

Статистические методы обработки данных широко используются в разных сферах жизнедеятельности человека, в том числе и при решении задач в области здравоохранения и медицины. Эффективность проводимых исследований во многом зависит от выбора адекватного метода статистического анализа, что, в первую очередь, связано с необходимостью определения типов исходных данных. Для установления взаимосвязи между количественными показателями и построением прогнозных моделей активно применяется корреляционно-регрессионный анализ. Этот статистический метод позволяет определить и оценить тесноту и направление связи между показателями. Так как на практике медицинские данные, как правило, не подчиняются нормальному закону распределения, для вычисления коэффициентов корреляции используют методы Кендалла и Спирмена. При этом значимость

коэффициентов корреляции, значение которых по модулю меньше 0.5, проверяют с помощью критерия Стьюдента. Для моделирования данных и исследования их свойств применяют регрессионный анализ. Регрессионный анализ входит в раздел математической статистики и машинного обучения и используется для выявления скрытых взаимосвязей в данных [1, 2].

Актуальность данного исследования заключается в том, что заболеваемость раком молочной железы как в России в целом, так и в Приморском крае в частности неуклонно растет и занимает лидирующее место среди онкологических заболеваний женского населения. В многочисленных исследованиях подтверждается, что, наряду с генетическими предрасположенностями, на заболеваемость РМЖ и выживаемость пациентов оказывают влияние показатели качества жизни населения: социально-экономические, санитарно-гигиенические, образ жизни, питание и т.п. [3 – 9]. Знания о факторах риска и их связи с заболеваемостью РМЖ имеют важное значение для общественного здравоохранения, особенно при выработке стратегии по укреплению здоровья и увеличению продолжительности жизни населения.

Целью данного исследования является выявление значимых показателей качества жизни населения, влияющих на заболеваемость раком молочной железы, на территории Приморского края с помощью методов статистического анализа.

Статистическая обработка и анализ данных

В качестве исходных данных были взяты стандартизованные показатели заболеваемости раком молочной железы (РМЖ) на территории Приморского края за период с 2007 г. по 2019 г. В качестве внешних факторов были выбраны следующие три группы показателей:

социально-экономические индикаторы уровня жизни населения – доход (среднедушевые денежные доходы в месяц, руб.), прожиточный уровень (величина прожиточного уровня в среднем на душу населения, руб. в месяц), коэффициент Джини (индекс концентрации доходов), объем платных услуг населению в расчете на душу населения, жилье (введено в действие общей площади жилых домов и общежитий, тыс. $м^2$), ВРП (валовый региональный продукт на душу населения), ИЧР (индекс человеческого развития), индекс образования, уровень безработицы (по методологии МОТ), индекс потребительских цен (в разах к декабрю предыдущего года), численность врачей на 10 тыс. чел. населения, численность среднего медицинского персонала на 10 тыс. чел. населения, обеспеченность больничными койками на 10 тыс. населения;

социально-гигиенические показатели (гигиена, загрязнение окружающей среды): объем используемой свежей воды (млн. $м^3$); сброс загрязненных сточных вод (млн. $м^3$); качество питьевой воды из водопроводной сети и нецентрализованного водоснабжения по санитарно-химическим и микробиологическим показателям (в % нестандартных проб); наличие водопровода (удельный вес общей площади жилищного фонда края с водопроводом на конец года, в %) и водоотведения (удельный вес общей площади жилищного фонда края с водоотведением на конец года, в %); выбросы в атмосферу (выбросы в атмосферу ЗВ от стационарных источников, тыс. тонн); уловленные и обезвреженные загрязняющие вещества (тыс. тонн); доля неблагоприятных проб почвы по санитарно-химическим, микробиологическим и гигиеническим показателям в %; доля неблагоприятных проб почвы селитебной территории по санитарно-химическим показателям в %;

потребление продуктов питания на душу населения в год: мясо и мясопродукты, кг, молоко и молокопродукты, хлебные продукты, картофель, овощи, фрукты и ягоды, сахар, масло растительное, рыба и рыбопродукты, яйца и яйцепродукты (шт.); доля расходов домашних хозяйств на покупку табачной продукции, %; потребление алкогольных напитков и пива на душу населения, л; стоимость условного (минимального) набора продуктов питания (на конец декабря, руб.).

Всего рассматривали 41 показатель качества жизни населения, с учетом загрязнения среды обитания. Численные значения показателей были получены из материалов официального сайта Федеральной службы государственной статистики и Федеральной службы по надзору в сфере защиты прав потребителей и благополучия человека по Приморскому краю [10, 11].

Сбор данных проводился в программе MS Excel 2010, обработка и анализ исходных данных осуществлялись в программе RStudio (Version 1.0.153). Проверка нулевых гипотез проводилась методом Р. Фишера на заданном уровне значимости 0.05 [12, 13].

Для проверки на нормальность распределения исходных данных использовали критерий Шапиро – Уилка. Статистика критерия вычислялась по следующей формуле:

$$U = n_1 * n_2 + \frac{n_x * (n_x + 1)}{2} - T_x, \qquad (1)$$

где n_1 – количество элементов в первой выборке; n_2 – количество элементов во второй выборке; T_x – бóльшая из двух ранговых сумм, соответствующая выборке с n_x элементами.

Построение матриц корреляций осуществляли методом Спирмена (так

как показатель заболеваемости не подчиняется нормальному закону распределения).

Формула расчета коэффициента корреляции Спирмена:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \quad (2)$$

где d_i – разность между рангами по показателям x и y; n – количество признаков.

Отбор значимых показателей, взаимосвязанных с заболеваемостью РМЖ, осуществляли с помощью простой линейной регрессии. Уравнение регрессии имеет следующий вид:

$$y = \beta_0 + \beta_1 * x, \quad (3)$$

где β_0 – свободный коэффициент регрессии; β_1 – коэффициент регрессии при показателе x.

Для проверки гипотезы об отсутствии линейной зависимости между показателями использовали критерий Фишера:

$$F = \frac{n-k-1}{k} \frac{\sum(\hat{y} - \bar{y})^2}{\sum(\hat{y} - y_i)^2}, \quad (4)$$

где $\hat{y}$ и $\bar{y}$ – предсказанное значение и среднее значение зависимого показателя; n – объем выборки; k – количество показателей в уравнении.

Для проверки значимости коэффициентов регрессии применяли критерий Стьюдента:

$$t_\beta = \frac{\beta}{S_\beta}, \quad (5)$$

где S_β – стандартная ошибка коэффициента регрессии β.

В табл. 1 представлены полученные коэффициенты регрессии при независимом показателе и соответствующий уровень их значимости.

Таблица 1

	Показатели	Коэффициенты регрессии / уровень значимости (p-value)
1	2	3
Социально-экономические показатели	Доход	0.0011 / 3.78e-05
	Прожиточный уровень	0.0031 / 1.74e-05
	Коэффициент Джини	85.102 / 1.8e-09
	Платные услуги населению в расчете на душу населения	4.776e-04 /3e-05
	Индекс потребительских цен	0.3059 / 7.12e-10
	Стоимость условного набора продуктов питания	0.0073 / 4.52e-06

1	2	3
	ВРП на душу населения	9.174e-05 / 3.06e-05
	ИЧР	39.450 / 5.87e-09
	Индекс образования	35.44 / 2.58e-09
	Уровень безработицы	4.6485 / 8.36e-11
	Жилье	0.0604 / 5.84e-07
	Численность врачей	0.6285 / 1.54e-09
	Численность среднего медперсонала	0.3780 / 5.82e-10
	Обеспеченность больничными койками	0.3266 / 2.36e-08
Социально-гигиенические показатели Вода	Объем использованной воды	0.0533 / 8.49e-09
	Сброс сточных вод	0.1077 / 2.42e-11
	Водопроводная вода (санитарно-химические показатели)	1.2178 / 1.65e-07
	Водопроводная вода (микробиологические показатели)	3.5678 / 2.97e-11
	Нецентрализ. водоснабжение (санитарно-химические показатели)	1.0080 / 3.26e-06
	Нецентрализ. водоснабжение (микробиологические показатели)	0.6957 / 1.53e-08
	Наличие водопровода	0.4355 / 2.01e-09
	Наличие водоотведения	0.4437 / 1.52e-09
Атмосфера	Выбросы в атмосферу	0.1583 / 4.83e-12
	Уловленные выбросы	0.01633 / 4.64e-11
Загрязнение почвы	На селитебной территории по санитарно-химическим показателям	0.6878 / 1.43e-07
	По санитарно-химическим показателям	0.6603 / 4.52e-08
	По микробиологическим показателям	1.0920 / 6.15e-08
	По гигиеническим показателям	10.610 / 5.26e-06
Потребление продуктов питания на душу населения в год	Доля расходов на покупку табачной продукции	7.298 / 0.00106
	Потребление алкоголя на душу населения	3.759 / 1.07e-09
	Мясо	0.0402 / 1.87e-07
	Молоко	0.0156 / 1.24e-08
	Яйца	0.0117 / 8.44e-08
	Хлебные продукты	0.0193 / 1.63e-09
	Картофель	0.2748 / 3.29e-09
	Овощи	0.3041 / 1.16e-09
	Фрукты, ягоды	0.4524 / 9.09e-08
	Сахар	0.8142 / 1.6e-09
	Масло растительное	2.479 / 2.479
	Рыба и рыбопродукты	1.0343 / 1.69e-08

В результате получилось, что все взятые нами показатели по отдельности взаимосвязаны с заболеваемостью РМЖ. Среди социально-экономиче-

ских показателей наибольшее влияние на заболеваемость РМЖ оказывают коэффициент Джини, индекс человеческого развития (ИЧР), индекс образования и уровень безработицы, т.е. те показатели, которые характеризуют основные параметры человеческого потенциала с учетом неравномерности распределения потребления и доходов в обществе.

В группе социально-гигиенических показателей наибольшее влияние на заболеваемость оказывают загрязнение почвы и качество водопроводной воды; в группе потребления продуктов питания – расходы на покупку табачной продукции, потребление алкоголя, растительного масла и рыбопродуктов.

Для уменьшения размерности данных, но без значительной потери информации, использовали метод главных компонент (разложение Карунена – Лоева, PCA), предварительно проведя стандартизацию исходных данных:

$$X' = \frac{x - \bar{x}}{\sigma(x)}, \qquad (6)$$

где $\sigma(x)$ – стандартное отклонение.

Основная идея метода главных компонент заключается в поиске в исходном пространстве гиперплоскости заданной размерности с последующим проектированием на нее исходной выборки. При этом выбирается та гиперплоскость, ошибка проектирования данных на которую является минимальной в смысле суммы квадратов отклонений. Результатом PCA-преобразования матрицы наблюдений X' размерностью $n \times m$ является матрица факторных значений T размерностью $n \times u$, содержащая проекции исходных точек выборки X' в новом u-мерном базисе. Матрица P размерности $u \times m$ содержит факторные нагрузки (loadings) – коэффициенты пропорциональности, обеспечивающие пересчет данных из m-мерного пространства исходных переменных в u-мерное пространство главных компонент. Связь между этими матрицами выглядит следующим образом:

$$X'P \to T. \qquad (7)$$

При этом имеет место соотношение:

$$T'T = P^t CP = \Lambda, \qquad (8)$$

где $\Lambda = diag\{\lambda_1, \ldots, \lambda_m\}$ – матрица собственных значений; $C = X'^t X'$ – корреляционная матрица.

Дисперсии столбцов матрицы T соответствуют собственным значениям λj матрицы C, являющимися диагональными элементами матрицы Λ.

Для оценки необходимого числа главных компонент воспользовались правилом Кеттелла с построением диаграммы «каменистой осыпи» и параллельным анализом с целью дополнительной оценки числа главных компо-

нент. Судя по рис. 1, для характеристики набора данных достаточно извлечь 3 – 4 главные компоненты, которые вместе объясняют 71% общей дисперсии исходных данных. При этом первая компонента объясняет 37%, вторая – 15%, третья – 13%, четвертая – 6%.

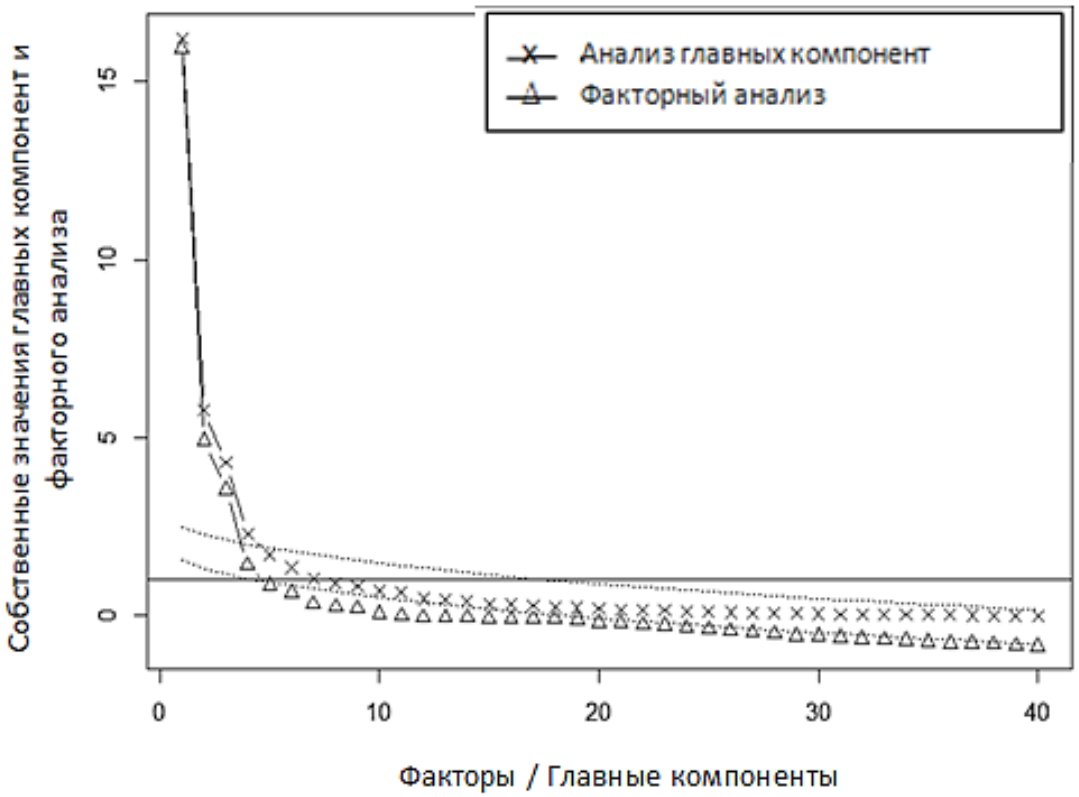

Рис.1. Диаграмма собственных значений с параллельным анализом.

При выделении трех главных компонент получаем следующий результат (табл. 2). В первую компоненту (PC1) вошли социально-экономические показатели качества жизни населения. PC1 положительно коррелирует с показателями: ВРП на душу населения (коэффициент корреляции –0,99), доход (0,98), платные услуги населению (0,98), ИЧР (0,98). Наименьшая связь имеется с показателями медицинского обслуживания населения. Вторая компонента (PC2) связана с социально-гигиеническими показателями (гигиена, загрязнение окружающей среды). Наибольшее влияние оказывают санитарно-химические показатели нецентрализованного водоснабжения (0,93), выбросы в атмосферу (–0,92); наименьшее – загрязнение почвы по гигиеническим показателям (0,04), наличие водоотведения (–0,07), микробиологические показатели нецентрализованного водоснабжения (0,10). В третью компоненту (PC3) вошли показатели потребления продуктов питания. Сильная положительная корреляция наблюдается с показателями – потребления молока (0,97), яиц (0,96), мяса (0,94); слабая – с потреблением хлебных продуктов (0,18). Значимость коэффициентов корреляции проверяли с помощью критерия Стьюдента.

Таблица 2

Главные компоненты	Показатели	Факторные нагрузки (loading)	Общность (component communalities)	Коэффициенты для главных компонент (scores)
1	2	3	4	5
PC_1	Доход	0.3363	0.9609	1.6134
	Прожиточный уровень	0.3324	0.9140	1.5548
	Коэффициент Джини	0.0821	0.1561	–0.2483
	Платные услуги в расчете на душу населения	0.3289	0.9669	1.5987
	Индекс потребительских цен	–0.2274	0.3882	–3.1359
	Стоимость условного набора продуктов питания	0.3286	0.9476	1.5885
	ВРП на душу населения	0.3344	0.9796	1.6269
	ИЧР	0.3188	0.9573	1.5709
	Индекс образования	0.3300	0.9274	1.5619
	Уровень безработицы	–0.3201	0.8114	–3.9721
	Жилье	0.1140	0.2491	0.0588
	Численность врачей	–0.1803	0.1584	–2.5937
	Численность среднего медперсонала	–0.058	0.0093	–1.6095
	Обеспеченность больничными койками	0.1679	0.3599	0.3856
PC_2	Объем использованной воды	0.0618	0.0270	0.3017
	Сброс сточных вод	–0.3551	0.6242	–1.8614
	Водопроводная вода (санитарно-химические показатели)	0.3202	0.5298	1.6235
	Водопроводная вода (микробиологические показатели)	–0.3117	0.4824	–1.6389
	Нецентрализ. водоснабжение (санитарно-химические показатели)	0.4045	0.8620	2.0657
	Нецентрализ. водоснабжение (микробиологические показатели)	0.0364	0.0110	0.1669
	Наличие водопровода	0.3470	0.6176	1.7614
	Наличие водоотведения	–0.0349	0.0043	–0.2047
	Выбросы в атмосферу	–0.4094	0.8550	–2.1500
	Уловленные выбросы	–0.4018	0.8418	–2.1167
	Загрязнения почвы на селитебной территории по сани	0.1349	0.1110	0.6822

1	2	3	4	5
	тарно-химическим показателям			
	Загрязнение почвы по санитарно-химическим показателям	0.1407	0.1192	0.7111
	Загрязнение почвы по микробиологическим показателям	0.1288	0.0932	0.6409
	Загрязнение почвы по гигиеническим показателям	0.0079	0.0012	0.0183
PC3	Доля расходов на покупку табачной продукции	–0.1534	0.183	–0.4718
	Потребление алкоголя на душу населения	0.2677	0.336	2.2151
	Мясо	–0.3543	0.891	–1.7945
	Молоко	–0.3702	0.941	–1.8771
	Яйца	–0.3709	0.913	–1.8526
	Хлебные продукты	–0.0251	0.033	0.2537
	Картофель	0.2603	0.309	2.1576
	Овощи	0.3079	0.478	2.4940
	Фрукты, ягоды	–0.3315	0.818	–1.6706
	Сахар	–0.2524	0.447	–1.0995
	Масло растительное	0.3268	0.786	2.8536
	Рыба и рыбопродукты	–0.2470	0.556	–1.2079

Примечание. Общность (component communalities) – доля учтенной компонентой дисперсии каждой переменной.

При выделении четырех главных компонент происходит следующее: первая компонента, отвечающая за социально-экономические показатели качества жизни населения, делится в свою очередь на две главные компоненты таким образом: во вторую компоненту входят показатели (PC12) – коэффициент Джини, жилье, численность врачей, численность среднего медицинского персонала, обеспеченность больничными койками; в первую компоненту (PC11) – все остальные показатели. Показатели в PC_2 и PC_3 остаются прежними.

С помощью регрессии на главные компоненты были построены линейные модели. Вид регрессионного уравнения в матричной форме следующий:

$$y = Tb + e, \tag{9}$$

где T – матрица факторных значений; b – коэффициенты регрессии; e – ошибка.

Всего построено шесть моделей. Результаты регрессионного анализа

представлены в табл. 3, где приведены полученные коэффициенты регрессии и соответствующие им уровни значимости.

Таблица 3

Компоненты	Коэффициенты регрессии / уровень значимости (p-value)
PC_1	–0.0465 (0.004)
PC_{11}+PC_{12}	–0.036 (0.011)+0.427 (0.030)
PC_2	–0.119 (0.0002)
PC_3	0.931 (0.022)
PC_1+PC_2+PC_3	–0.028 (0.316) –0.108 (0.019) –0.058 (0.372)
PC_{11}+PC_{12}+ PC_2+PC_3	0.041(0.449) +0.928(0.175)+0.038(0.728)+0.216(0.289)

При использовании регрессионного анализа на главных компонентах не было получено качественной модели с учетом всех компонент, так как не все коэффициенты регрессии в уравнениях значимы. По отдельности каждая компонента значимо взаимосвязана с заболеваемостью, в совокупности компонент качественной модели на основе линейной регрессии получить не удалось. Для построения моделей необходимо использовать другие методы прогнозирования.

Заключение

Проведенный статистический анализ позволил выявить показатели, влияющие на заболеваемость раком молочной железы на территории Приморского края. С помощью анализа главных компонент выделены четыре компоненты, объясняющие 71% общей вариации исходных данных. Первые две компоненты содержат показатели, характеризующие социально-экономические условия жизни населения; вторая компонента связана с социально-гигиеническими показателями, с учетом загрязнения окружающей среды; третья компонента содержит показатели потребления продуктов питания. При этом показатель заболеваемости имеет обратную умеренную зависимость с первой и второй компонентами (коэффициенты корреляции равны –0.401 и –0.544 соответственно) и прямую умеренную зависимость с третьей компонентой (0.456).

На основе выделенных главных компонент не удалось построить качественной регрессионной модели. Для этого нужно выбрать другие методы прогнозирования (например, нейронные сети, случайный лес и др.).

ЛИТЕРАТУРА

1. *Шитиков В.К., Мастицкий С.Э.* Классификация, регрессия и другие алгоритмы DataMining с использованием R. 2017. Режим доступа: https://github.com/ranalytics/data-mining (дата обращения 26.03.2022).

2. *Мун С.А., Глушов А.Н., Штернис Т.А., Ларин С.А., Максимов С.А.* Регрессионный анализ в медико-биологических исследованиях. – Кемерово: КемГМА, 2012.
3. *Французова И.С.* Анализ факторов риска развития рака молочной железы // Международный научно-исследовательский журнал. – 2019. – №3 (81) – С. 68-74.
4. *Юдин С.В., Маслов Д.В.* Влияние антропогенных факторов на онкологическую заболеваемость населения Приморского края // Тихоокеанский медицинский журнал. – 2004. – №3(17). – С.46-49.
5. *Ермолицкая М.З., Кику П. Ф., Абакумов А.И.* Смертность от рака молочной железы в Приморском крае: анализ данных и моделирование // Здоровье населения и среда обитания. –2021. – №29(11). – С.16-22.
6. *Tyrer J., Duffy S.W., Cuzick J.* A breast cancer prediction model incorporating familial and personal risk factors // Stat. Med. – 2004. –№23. –1111-1130.
7. *Asghari Jafarabadi M., Iraji Z., Dolatkhah R., Jafari Koshki T.* Modeling the Factors Associated with Mortality in Patients with Breast Cancer: A Machine Learning Approach. URL: https://www.researchsquare.com/article/rs-57685/v1 (дата обращения 26.03.2022)
8. *Sternfeld B.* Physical activity and risk of recurrence and mortality in breast cancer survivors: findings from the LACE study // Cancer Epidemiology and Prevention Biomarkers. – 2009. – 18(1). – 87-95.
9. *Sekeroglu B., Tuncal K.* Prediction of cancer incidence rates for the European continent using machine learning models // Health Informatics Journal. – 2021. – 27(1). URL: https://journals.sagepub.com/doi/full/10.1177/1460458220983878 (дата обращения 26.03.2022).
10. Федеральная служба государственной статистики: официальный сайт: Режим доступа: https://rosstat.gov.ru (дата обращения: 26.03.2022).
11. Федеральная служба по надзору в сфере защиты прав потребителей и благополучия человека по Приморскому краю. Режим доступа: http://25.rospotrebnadzor.ru (дата обращения 26.03.2022).
12. *Кабаков Р.И.* R в действии. Анализ и визуализация данных в программе R. / пер. с англ. П.А. Волковой. – М.: ДМК Пресс, 2014.

Статья представлена к публикации членом редколлегии А.И. Абакумовым.

E-mail:
Ермолицкая Марина Захаровна – ermmz@mail.ru.