



УДК 519.8

© 2019 г. А.А. Яковлев

(Дальневосточный федеральный университет, Владивосток),

А.И. Абакумов, д-р физ.-мат. наук

(Институт автоматизации и процессов управления ДВО РАН, Владивосток)

АНАЛИЗ ДАННЫХ О ТУБЕРКУЛЕЗЕ И ВЛИЯНИИ НА ЗАБОЛЕВАНИЕ УСЛОВИЙ ЖИЗНИ В ПРИМОРСКОМ КРАЕ

По данным о туберкулезе среди населения и условиях жизни в Приморском крае в 2002 – 2015 гг. проведено исследование связей и причин возникновения и течения туберкулеза. Установлено, что все представленные характеристики условий жизни можно разбить на две основные группы: позитивные и негативные. Характеристики условий жизни хорошо группируются вдоль главных компонент и слаженно воздействуют на характеристики заболевания. Выявлен временной лаг в 5 – 7 лет воздействия условий жизни на туберкулез.

Ключевые слова: кластерный анализ, корреляция, главные компоненты.

DOI: 10.22250/isu.2019.61.42-53

Введение

Проблемы динамики распространения туберкулеза и качества медицинской помощи в лечении этой болезни являются объектом постоянного внимания управленцев в области здравоохранения [1]. При исследовании закономерностей и связей этих процессов с демографическими, социальными и экономическими факторами используются методы математического моделирования [2, 3].

В этом направлении нами сделана попытка анализа данных о динамике ряда социальных, экономических и демографических показателей в связи с заболеваемостью туберкулезом и его распространенностью в Приморском крае в период 2002 – 2015 гг. [4 – 7]. Аналитический обзор [1] является совместным изданием Министерства здравоохранения РФ и ряда медицинских учреждений. В обзоре представлен анализ показателей о туберкулезе, обсуждено их значение для оценки эпидемиологической ситуации и качества

противотуберкулезной помощи в РФ в 2007 – 2013 гг., рассмотрена ее динамика за последние 10 – 15 лет. Дано сравнение ситуации по туберкулезу в РФ и странах мира. Общие отчеты ВОЗ [8] содержат информацию о туберкулезе в странах мира за несколько лет. В данном источнике опубликованы стратегии ликвидации туберкулеза, проведен анализ причин возникновения заболевания. Данные анализируются по многим параметрам, в том числе по гендерному признаку, по детскому туберкулезу, по сочетанию с ВИЧ, по лекарственно-устойчивому туберкулезу. В [4] мы рассматривали данные по Приморскому краю, добавив данные из местных источников. Анализ данных с точки зрения эффективности медицинских учреждений дан в работе [5], мы же сосредоточились на анализе влияния среды обитания.

Данные и методы их анализа

Исходные данные [4] представлены матрицей $X = (X_{ij})_{i,j=1}^{n,m}$, где n – обозначает количество показателей; m – количество данных у каждого показателя. Количество данных одинаково у всех показателей и равно числу рассматриваемых лет, $m = 14$ (за 2002 – 2015 гг.). В табл. 1 показаны основные характеристики населения и условий его жизни, разбитые на три группы.

Таблица 1

№	Показатели	Примечание
1	Распространенность туберкулеза на 100 тыс. чел.	Охват населения туберкулезом
2	Заболеваемость туберкулезом на 100 тыс. чел.	
3	Распространенность больных активной фазой туберкулеза на 100 тыс. чел.	
4	Вылечившиеся больные, %	
5	Обеспеченность фтизиатрами на 10 тыс. чел.	Демография и медицинское обеспечение
6	Численность населения	
7	Валовый региональный продукт (ВРП) к ценам 2001 г.	Характеристики условий жизни (социально-экономические факторы)
8	Уровень дохода населения к ценам 2001 г.	
9	Обеспеченность жилплощадью на душу населения	
10	Оборот общественного питания, млн. руб.	
11	Обеспеченность мясом на душу населения	
12	Обеспеченность молоком на душу населения	
13	Обеспеченность яйцами на душу населения	
14	Обеспеченность водопроводом, %	
15	Обеспеченность канализацией, %	
16	Число жителей, живущих ниже черты бедности, на 1 тыс. чел.	
17	Уровень безработицы, %	
18	Смертность на 100 тыс. чел.	
19	Потребление водки на душу населения	
20	Потребление наркотиков на 100 тыс. чел.	
21	Вредные выбросы	
22	Потребление вина на душу населения	
23	Потребление пива на душу населения	

Данные проанализированы статистическими и близкими к статистическим методами на предмет взаимосвязей и взаимовлияния. Стремясь избавиться от разнобоя в единицах измерения и масштабах показателей табл. 1, выполним центрирование и нормировку данных, т.е. каждый показатель x заменен на $\frac{x - x_{cp}}{s_x}$, где $x_{cp} = \frac{1}{m} \sum_j x_j$, $s_x^2 = \frac{1}{m-1} \sum_j (x_j - x_{cp})^2$. В этом случае единицы измерения конкретных показателей не влияют на результаты расчетов. Статистические характеристики показывают согласованность динамик различных показателей, а не их абсолютных значений [9].

Количественные значения силы взаимосвязей данных получены вычислением выборочной ковариационной матрицы. Для нормированных и центрированных данных ковариационная матрица совпадает с корреляционной. Ковариационную матрицу для этих данных обозначаем $K = (k_{ij})_{i,j=1}^n$, $k_{ij} = \text{cov}(x_i, x_j)$, в случае центрированных и нормированных данных ее элементы вычисляются по формуле $k_{ij} = \frac{1}{m-1} (X_i \cdot X_j)$, где X_i – строка матрицы X , символ " \cdot " обозначает скалярное произведение. Матрица K симметрична и неотрицательно определена [10].

Для исследования взаимосвязей показателей применен метод кластерного анализа. Методы кластерного анализа позволяют решать следующие задачи: проведение классификации объектов с учетом признаков, отражающих сущность, природу объектов; проверку выдвигаемых предположений о наличии некоторой структуры в изучаемой совокупности объектов; построение новых классификаций для слабоизученных явлений, когда необходимо установить наличие связей внутри совокупности.

Большинство методов кластерного анализа являются эвристическими и представляют собой довольно простые процедуры, что позволяет свести к минимуму ошибки при трактовке результатов. Однако необходимо иметь в виду, что кластерные методы размещают объекты по группам, которые могут существенно различаться по составу при использовании различных методов кластеризации. В общем случае смысл классификации объектов заключается в том, чтобы всю совокупность объектов разбить на сравнительно небольшое (заранее известное или нет) однородных в определенном смысле групп или классов. В ряде случаев возникает необходимость использовать алгоритмы, на каждом шаге которых обсчитывается лишь небольшая часть исходных наблюдений. Одним из таких методов является метод k -средних.

Пусть $x = (x_i)_{i=1}^n$ – набор переменных, представлен векторами центри-

рованных и нормированных экспериментальных данных в виде матрицы $X = (X_{ij})_{i,j=1}^{n,m}$, где n – число переменных (наблюдений, показателей); m – длина наборов экспериментальных данных, имеющих выборочные средние 0 и выборочные дисперсии 1. Метод k -средних разделяет m наблюдений на k групп (или кластеров) ($k \leq n$): $S = (S_i)_{i=1}^k$, чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:

$$\sum_{i=1}^k \sum_{X_j \in S_i} \rho(X_j, C_i) \rightarrow \min_S,$$

где C_i – центр кластера S_i ; X_j – строка матрицы X . Под расстоянием $\rho(X_j, C_i)$ понимается евклидово расстояние [99].

Применен также метод главных компонент. Выделены ведущие показатели, играющие основную роль в динамике всей совокупности данных.

Метод главных компонент используется по следующей схеме. Выборочную ковариационную матрицу K приводим к диагональному виду ортогональным преобразованием подобия $\Lambda = U^{-1}KU$, матрица U ортогональна, $U^{-1} = U^T$, ее столбцы являются собственными векторами для неотрицательных собственных чисел матрицы K , стоящих по диагонали в матрице Λ . Верхний индекс “Т” у матрицы традиционно обозначает действие транспонирования. Замена $y = U^T x$ приводит нас к главным компонентам. Матрица Λ представляет собой выборочную ковариационную матрицу выборки $Y = U^T X$. Через y_k обозначим вектор из первых k компонент вектора y , это и есть выбранные нами главные компоненты. Соответственно изображение данных в подпространстве первых k собственных векторов представляется столбцами матрицы $Y_k = (U^T X)_k$. Индекс k означает выбор первых k строк матриц [99]. При вычислениях использовались пакеты SPSS, STATISTICA [1111, 1212].

Анализ данных о социально-экономических характеристиках условий жизни

Выполнен анализ взаимовлияний между показателями, характеризующими условия жизни (в табл. 1 это показатели номеров 7 – 23). Однако значения ковариационной (корреляционной) матрицы приведены в табл. 2 частично. Поскольку эта матрица симметричная, то под ее диагональю указаны лишь значения нижней треугольной матрицы, так как ее транспонирование определяет значения (в табл. 2 они опущены) верхней треугольной матрицы.

Таблица 2

1																
0,96	1															
0,95	0,99	1														
0,88	0,94	0,97	1													
0,97	0,99	0,99	0,95	1												
0,92	0,98	0,94	0,85	0,95	1											
0,95	1,00	0,99	0,94	0,99	0,97	1										
0,91	0,86	0,84	0,77	0,85	0,83	0,84	1									
0,78	0,66	0,60	0,48	0,66	0,69	0,65	0,83	1								
-0,89	-0,93	-0,89	-0,76	-0,88	-0,96	-0,92	-0,78	-0,65	1							
-0,54	-0,56	-0,57	-0,60	-0,56	-0,54	-0,52	-0,57	-0,32	0,47	1						
-0,76	-0,76	-0,79	-0,78	-0,82	-0,72	-0,75	-0,64	-0,53	0,57	0,48	1					
-0,86	-0,91	-0,91	-0,87	-0,91	-0,89	-0,93	-0,72	-0,63	0,82	0,34	0,70	1				
-0,84	-0,93	-0,95	-0,97	-0,93	-0,87	-0,94	-0,78	-0,52	0,76	0,55	0,76	0,91	1			
-0,90	-0,95	-0,93	-0,88	-0,92	-0,93	-0,94	-0,79	-0,53	0,91	0,57	0,62	0,82	0,85	1		
0,85	0,86	0,79	0,64	0,82	0,91	0,85	0,78	0,75	-0,94	-0,30	-0,54	-0,76	-0,66	-0,81	1	
0,76	0,83	0,75	0,62	0,78	0,92	0,82	0,71	0,66	-0,92	-0,41	-0,43	-0,76	-0,67	-0,81	0,92	1

Выделяются три группы показателей данных об условиях жизни, сравнительно тесно коррелирующих между собой (табл. 3).

Первая и вторая группы имеют отрицательные корреляции между собой. Из содержательного смысла показателей первую группу естественно назвать позитивной, а вторую – негативной. Неожиданностью является выделение двухэлементной третьей группы (слабый алкоголь), имеющей тесные положительные корреляции внутри и положительные корреляции с позитивной группой при отчетливых отрицательных корреляциях с негативной группой.

Таблица 3

Наименование группы	Показатели
Позитивная группа	ВРП к ценам 2001 г. Уровень дохода населения к ценам 2001 г. Обеспеченность жилплощадью на душу населения Оборот общественного питания, млн. руб. Обеспеченность мясом на душу населения Обеспеченность молоком на душу населения Обеспеченность яйцами на душу населения Обеспеченность водопроводом Обеспеченность канализацией
Негативная группа	Число жителей, живущих ниже черты бедности, на 1 тыс. чел. Уровень безработицы, % Смертность на 100 тыс. чел. Потребление водки на душу населения Потребление наркотиков на 100 тыс. чел. Вредные выбросы
Слабый алкоголь	Потребление вина на душу населения Потребление пива на душу населения

Далее проведена кластеризация данных. В методе k -средних число кластеров задается, мы варьировали это число от четырех до шести. При разбиении на четыре кластера эти кластеры фактически совпадают с указанными выше тремя группами. Если в табл. 3 из позитивной группы выделить «водоснабжение» (показатели 14 – 15) в отдельный кластер, то получаем 4-кластерное разбиение (табл. 4).

Таблица 4

№	Показатели	Характеристика кластеров
1	ВРП к ценам 2001 г. Уровень дохода населения к ценам 2001 г. Обеспеченность жилплощадью на душу населения Оборот общественного питания, млн. руб. Обеспеченность мясом на душу населения Обеспеченность молоком на душу населения Обеспеченность яйцами на душу населения	Позитивный
2	Обеспеченность водопроводом Обеспеченность канализацией	Водоснабжение
3	Число жителей, живущих ниже черты бедности, на 1 тыс. чел. Уровень безработицы, % Смертность на 100 тыс. чел. Потребление водки на душу населения Потребление наркотиков на 100 тыс. чел. Вредные выбросы	Негативный
4	Потребление вина на душу населения Потребление пива на душу населения	Слабый алкоголь

При разбиении на 5 кластеров дополнительно выделяется «оборот общественного питания» (показатель 10, табл. 1), а при разбиении на 6 кластеров «общественное питание» возвращается в первый кластер, при этом третий кластер разбивается на три новых кластера. Это подвигло авторов остановиться на 4-кластерном разбиении, придав кластерам названия по аналогии с названиями групп табл. 3 и табл. 4.

Применение метода главных компонент приводит к собственным числам матрицы K ковариаций. Первые 10 собственных чисел приведены в табл. 5. Остальные 7 собственных чисел (по величине $\leq 0,01$) являются не отрицательными. Первые 4 собственных числа обеспечивают более 95% общей дисперсии данных. Уже первое собственное число «забирает» более 80% общей дисперсии данных.

Таблица 5

	1	2	3	4	5	6	7	8	9	10
Собственное число	13,74	1,16	0,74	0,67	0,29	0,20	0,09	0,05	0,03	0,02
% общей дисперсии	80,82	6,81	4,34	3,95	1,70	1,19	0,51	0,30	0,19	0,12
% дисперсии накопленный	80,82	87,64	91,98	95,93	97,63	98,82	99,33	99,63	99,82	99,94

Нормированные собственные векторы первых 4 собственных чисел (табл. 6) указывают направления в пространстве R^n , вдоль которых с наибольшим размахом распределены исходные данные.

Таблица 6

№ показателя	Собственный вектор			
	U ₁	U ₂	U ₃	U ₄
1	0,262	-0,025	-0,102	0,142
2	0,269	0,038	0,065	-0,043
3	0,264	0,137	0,104	0,007
4	0,248	0,331	0,133	0,006
5	0,267	0,097	0,058	0,065
6	0,264	-0,099	0,036	-0,116
7	0,267	0,031	0,123	-0,014
8	0,239	-0,075	-0,371	0,178
9	0,193	-0,373	-0,447	0,479
10	-0,250	0,225	-0,053	0,252
11	-0,154	-0,407	0,672	0,439
12	-0,206	-0,324	0,012	-0,470
13	-0,246	0,005	-0,320	-0,152
14	-0,248	-0,261	-0,178	-0,040
15	-0,253	-0,028	-0,093	0,289
16	0,235	-0,416	0,037	-0,046
17	0,227	-0,378	0,016	-0,334

Наиболее информативно направление, указываемое собственным вектором первого собственного числа. Компоненты этого вектора для первого, позитивного, а также второго и четвертого кластеров (табл. 4) имеют один знак, а для третьего, негативного, противоположный. Здесь так же, как и в ковариационной матрице, негативный кластер проявляется собственными свойствами.

При этом все показатели примерно одинаково «участвуют» компонентами в первом собственном векторе, он аккумулирует информацию обо всех показателях в равной степени. Остальные собственные векторы имеют менее выраженные свойства.

В данных выделяются четыре первых главных компоненты (табл. 7), которые представляют собой проекции данных на соответствующие собственные векторы.

Главные компоненты представляют собой проекции векторов показателей каждого года на собственные векторы. Первая главная компонента «собирает» наиболее значимые величины этих проекций. С 2002 г. по 2007 г. проекции вектора показателей в целом убывали, а с 2008 г. начался убыстряющийся рост.

Таблица 7

Год сбора информации	Y1	Y2	Y3	Y4
2002	-6,14	0,73	-0,44	1,35
2003	-5,38	1,54	-0,21	0,75
2004	-4,48	-0,21	1,57	-0,47
2005	-3,35	-0,44	-0,36	-1,30
2006	-1,61	-0,47	-0,59	-0,85
2007	-0,46	-0,16	-0,78	-1,02
2008	0,21	-0,62	-0,40	-0,67
2009	0,60	-1,28	1,14	0,33
2010	1,57	-1,47	1,05	0,83
2011	2,81	-1,33	-0,71	0,93
2012	3,47	-0,29	-1,40	0,67
2013	3,71	0,91	0,01	-0,21
2014	4,08	1,43	0,27	-0,20
2015	4,97	1,66	0,85	-0,14

Если вспомнить, что компоненты первого собственного вектора положительны для всех кластеров, кроме негативного (табл. 6), то это означает, что условия жизни с 2002 г. по 2007 г. в целом ухудшались, а с 2008 г. стали улучшаться. А поскольку первой компоненте соответствует 80% всей информации, то этот вывод приобретает достаточную весомость. Более того, если бы удалось собирать данные соответственно направлению собственного вектора первого собственного числа (каждый такт времени по одному числу), то мы сразу получали бы 80% информации о вариациях данных этого такта времени.

Влияние условий жизни на распространение туберкулеза

Четыре показателя распространения туберкулеза (номера 1 – 4, табл. 1) характеризуют разные стороны заболевания. Для каждого из четырех показателей туберкулеза построим ковариационные функции относительно выделенных кластеров [11]. Это функции f_{ik} сдвига p каждого показателя туберкулеза x_i ($i = 1, \dots, 4$) относительно характеристик кластеров S_k ($k = 1, \dots, 4$) в табл. 4: $f_{ik}(p) = \text{cov}_p(x_i, c_k)$. Здесь c_k – показатель центра кластера k . Этот показатель заменяет показатели, вошедшие в кластер. Выборочный центр C_k кластера вычисляется как среднее арифметическое векторов X_j выборочных данных, вошедших в кластер. Индекс p указывает на сдвиг по компонентам векторов: $\text{cov}_p(X_i, C_k) = \frac{1}{m-p} \sum_{l=1}^{m-p} X_{i,l+p} C_{kl}$. В нашем случае переменная p изменяется по целым числам от 0 до 9. Ковариационные функции характери-

зуют влияние кластеров показателей одного года на уровень заболеваемости и состояние больных через p лет.

Ниже приведены графики ковариационных функций для трех показателей туберкулеза (рис. 1 – 3).

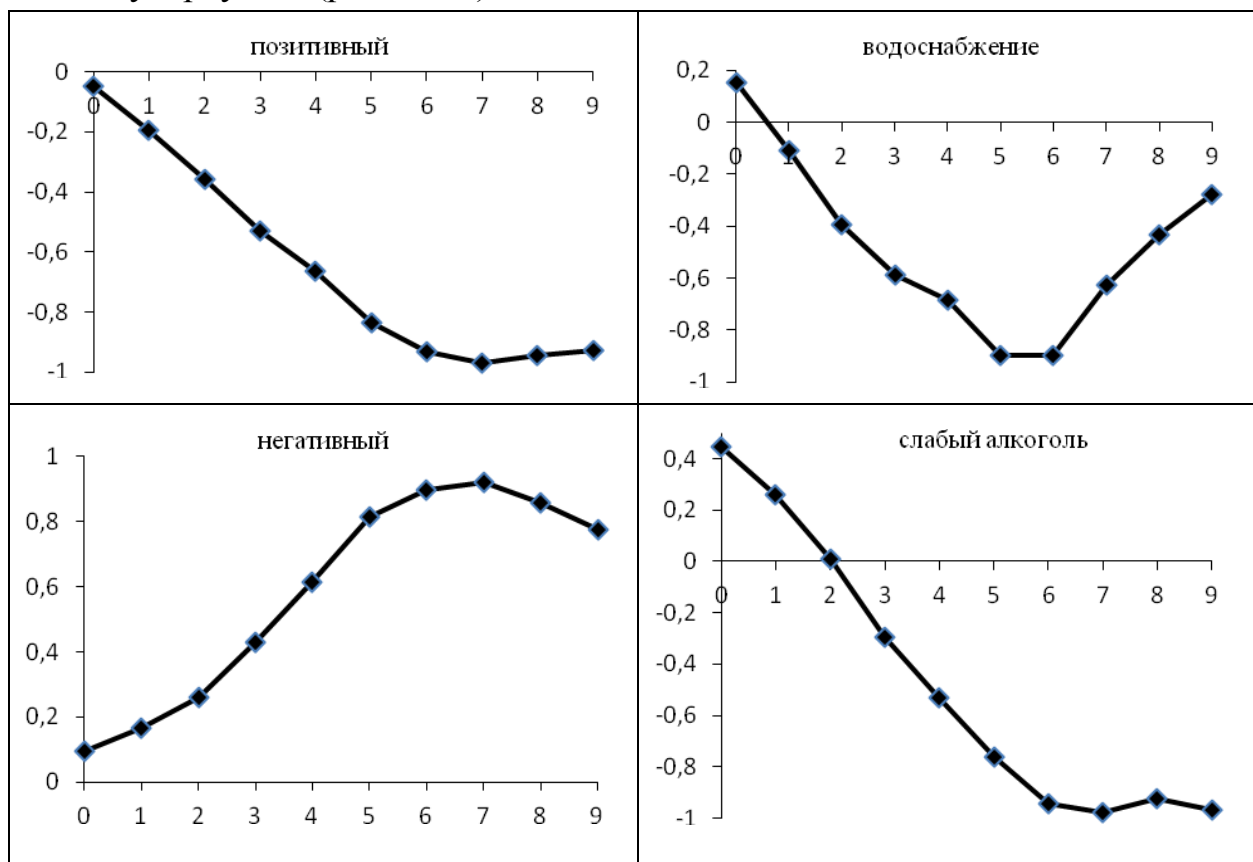


Рис. 1. Ковариационная функция зависимости заболеваемости туберкулезом.

Мы не использовали показатель распространенности активной фазы заболевания, считая его в данном случае частично дублирующим общий показатель распространенности заболевания. По горизонтальной оси указан сдвиг p в ковариационной функции, а по вертикальной – значение функции.

Все три группы графиков показывают особенность поведения ковариационных функций при сдвиге 5, 6 и, особенно, 7 лет. С этим сдвигом позитивный кластер (табл. 4) оказывает наибольшее негативное влияние на заболеваемость и распространенность туберкулеза.

Негативный кластер (табл. 4) при этих же сдвигах в наибольшей мере «помогает» заболеваемости и распространенности туберкулеза. Два других малых по числу показателей кластера (табл. 4) играют менее отчетливые роли, но в целом оба проявляют позитивные воздействия на показатели заболевания. Интересно, что слабый алкоголь оказывается скорее полезным для больных, чем вредным.

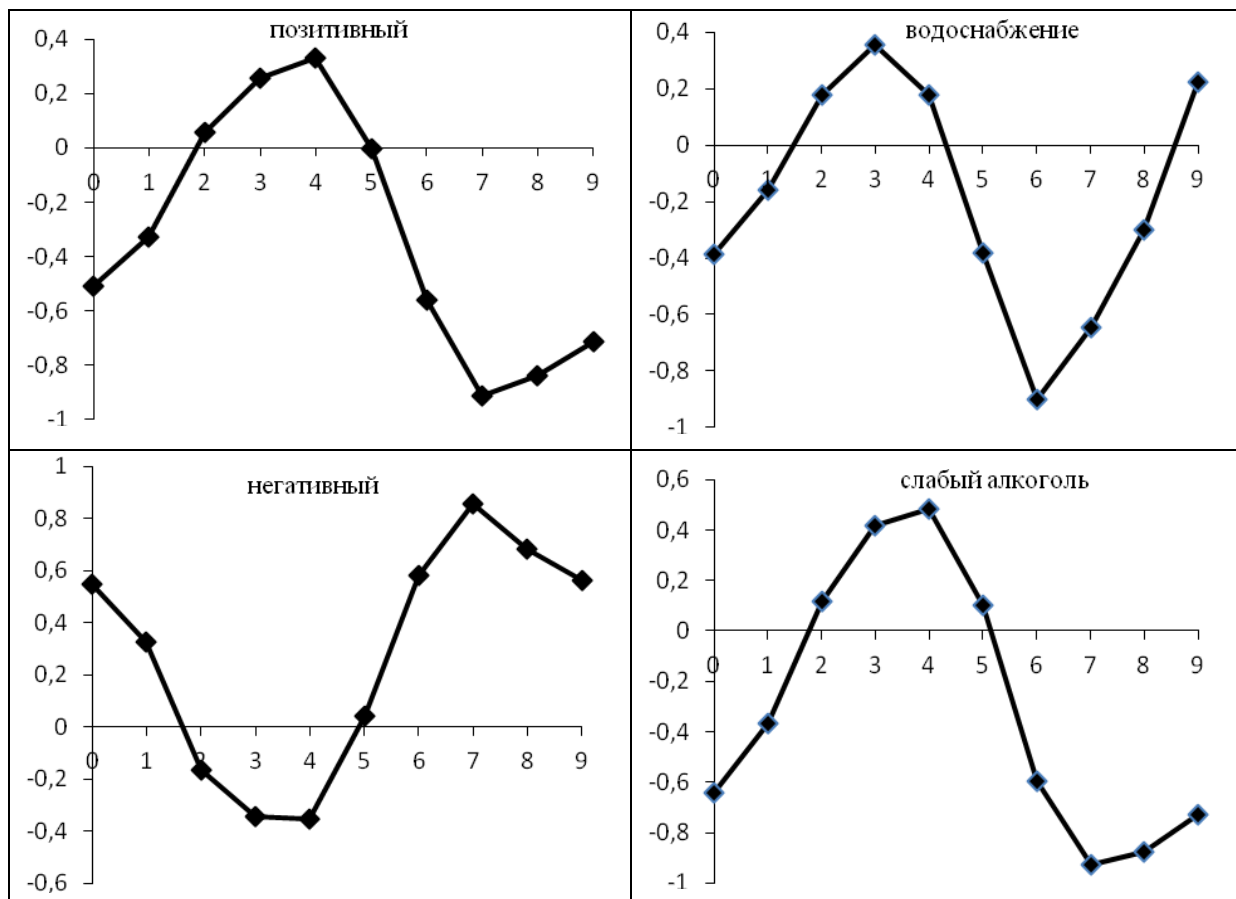


Рис. 2. Ковариационная функция зависимости распространенности туберкулеза.

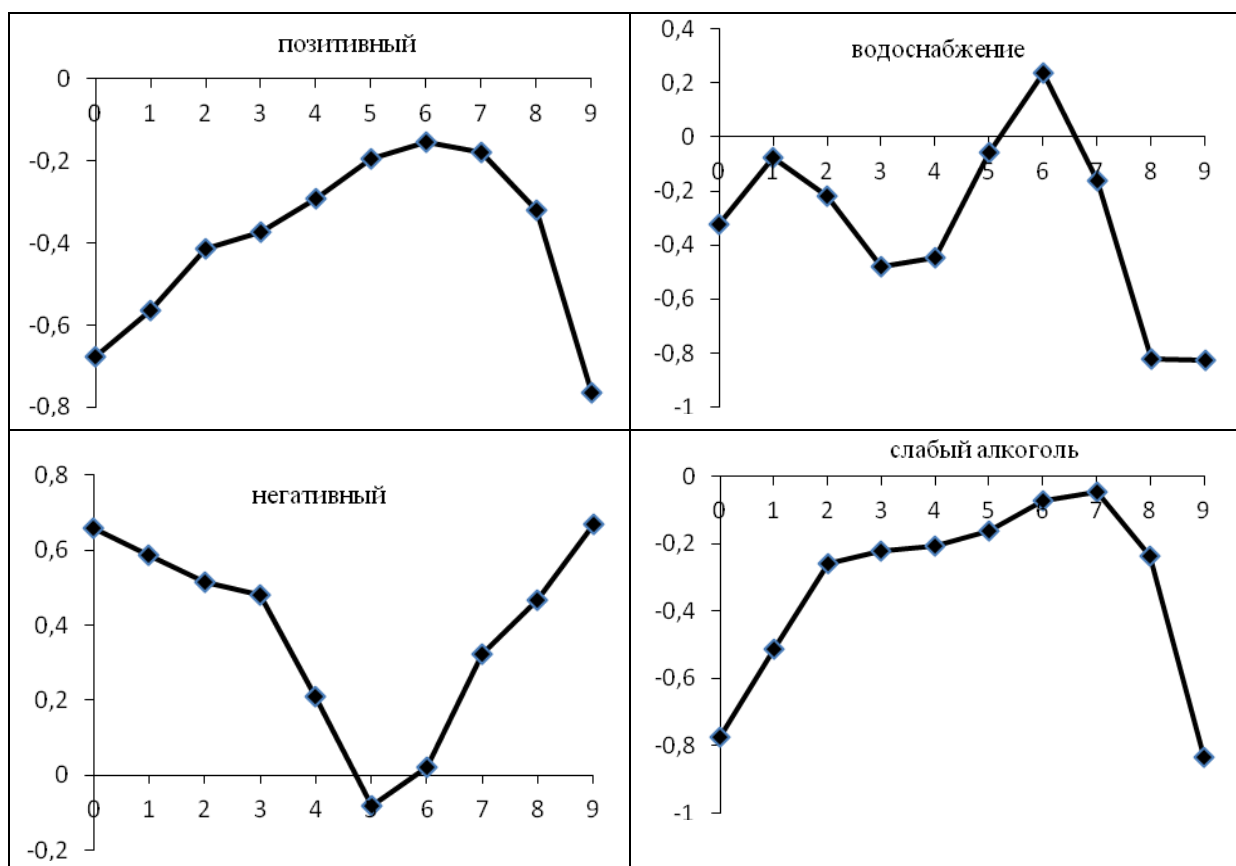


Рис. 3. Ковариационная функция зависимости выздоровевших.

Заключение

Короткие ряды данных (всего 14 лет), конечно, повлияли на результаты исследования, но некоторые закономерности выявились вполне четко.

Среди показателей об условиях жизни анализ ковариаций выявил большие группы позитивных и негативных показателей и две небольшие группы скорее позитивных показателей (табл. 3 и 4). Позитивные показатели благоприятно влияют на уменьшение распространения болезни и выздоровление, негативные действуют противоположным образом. Неожиданным оказался результат положительного влияния на больных слабого алкоголя (вино, пиво). Выделение главных компонент отчетливо подтвердило этот результат. С 2002 г. по 2007 г. позитивные показатели в целом убывали, а с 2008 г. начался убыстряющийся рост, т.е. с 2002 г. по 2007 г. условия жизни в целом ухудшались, а с 2008 г. стали улучшаться.

Выяснился временной сдвиг в 5 – 7 лет наибольшей отчетливости воздействий групп показателей (табл. 4) на показатели распространения туберкулеза (рис. 1 – 3). При сдвиге в 7 лет позитивный кластер оказывает наибольшее негативное влияние на заболеваемость и распространенность туберкулеза. Негативный кластер при этих же сдвигах в 5 – 7 лет в наибольшей мере «способствует» распространению туберкулеза.

Попытки найти взаимодействия на уровне регрессии, в том числе и с учетом уровня медицинского обслуживания, не удались. Можно предположить, что причины кроются в кратком временном промежутке и в двойной роли медицинского обслуживания. С одной стороны, это лечебные воздействия, а с другой, – эффективность выявления больных.

В целом применение математических средств к данным медицинской статистики дало новую информацию о динамике заболевания в зависимости от условий жизни населения. Появилась возможность разработки собственной математической модели динамики заболевания туберкулезом, что является следующим шагом нашей работы. Поскольку туберкулез относится к эпидемическим заболеваниям [8], то его моделирование следует вести в стиле подходов известной модели Бейли [1313] и многочисленных последующих публикаций. В то же время течение болезни связано с функционированием иммунной системы человека, поэтому следует использовать подходы Марчука [14,15].

ЛИТЕРАТУРА

1. Туберкулез в Российской Федерации, 2012-2014 гг. Аналитический обзор статистических показателей, используемых в Российской Федерации и в мире. – М., 2015.

2. *Романюха А.А.* Математические модели в иммунологии и эпидемиологии инфекционных заболеваний. – М.: БИНОМ. Лаборатория знаний, 2015.
3. *Абакумов А.И., Яковлев А.А.* Модель иммунного ответа на вирусное заражение // Информатика и системы управления. – 2018. – № 3. – С. 3-9.
4. Состояние здоровья населения и организация здравоохранения на территории Приморского края. Государственный доклад администрации Приморского края для Минздрава РФ. Электронный ресурс <https://www.primorsky.ru/authorities/executive-agencies/departments/health/statisticheskie-dannye-o-sostoyanii-i-dinamike-razvitiya-zdravookhraneniya-primorskogo-kraya/> (дата обращения: 15.09.2018)
5. *Волкова М.В., Шахгельдян К.И., Гельцер Б.И., Кривелевич Е.Б., Транковская Л.В., Ермолицкая М.З., Кучерова С.В.* Анализ кадрового ресурса системы здравоохранения Приморского края // Тихоокеанский медицинский журнал. – 2016. – № 3. – С. 52-56.
6. *Лавренюк В.В., Мотанова Л.Н.* Оценка ситуации по туберкулезу в Приморском крае за последние 14 лет (2003 – 2016) // Тихоокеанский медицинский журнал. – 2017. – № 4. – С. 74 – 76
7. *Мурашкина Г.С., Алексеева Т.В., Новикова Н.М. и др.* Эпидемическая ситуация по туберкулезу в Дальневосточном федеральном округе в 2003–2008 гг. // Туберкулез и болезни легких. – 2011. – № 1. – С. 10-16.
8. World Health Organization. https://www.who.int/tb/publications/global_report/ru/ (дата обращения: 22.08.2018)
9. *Ниворожкина Л.И., Арженовский С.В.* Многомерные статистические методы в экономике. Учебник. – М.: Издательско-торговая корпорация «Дашков и К»; Ростов н/Д: Наука-Спектр, 2008.
10. *Гантмахер Ф.Р.* Теория матриц. – М.: Наука, 1966.
11. <https://www.ibm.com/ru-ru/analytics/spss-statistics-software> (дата обращения 01.07.2019)
12. <http://statsoft.ru/> (дата обращения 01.07.2019)
13. *Bailey N.T.J.* The Mathematical Theory of Epidemics. – London. Griffin, 1957.
14. *Марчук Г.И.* Математические модели в иммунологии. – М.: Наука, 1985.
15. *Белых Л.Н., Марчук Г.И.* Качественный анализ простейшей математической модели инфекционного заболевания // Математическое моделирование в иммунологии и медицине. – 1982. – С. 5-27.

E-mail

Абакумов Александр Иванович – abakimov@iacp.dvo.ru;

Яковлев Анатолий Александрович – yakovlev_aal@dvfu.ru.