

ВОЗМОЖНОСТИ КЛАСТЕРНОГО И ФАКТОРНОГО АНАЛИЗОВ ДЛЯ ИССЛЕДОВАНИЯ СОСТОЯНИЯ СИСТЕМЫ «ОКРУЖАЮЩАЯ СРЕДА – ИНФЕКЦИИ»

В. В. Губарев¹, В. Б. Локтев², Е. Н. Наумова³, В. Е. Хиценко¹

¹Новосибирский государственный технический университет
Новосибирск, Россия,

E-mail: gubarev@vt.cs.nstu.ru, khits@is.cs.nstu.ru

²ФГУН Государственный научный центр вирусологии
и биотехнологии «Вектор»

Кольцово, Новосибирская область, 630559, Россия,

E-mail: valeryloktev@gmail.com

³National Center for Environmental Assessment, Cincinnati, USA

E-mail: elena.naumova@tufts.edu

Показаны возможности методов корреляционного, кластерного и факторного анализа для исследования инфекционной обстановки в крупном городе на примере анализа острых кишечных заболеваний в г. Барнауле в 2007 и 2008 гг. в их связи с водной и климатической обстановкой. Выделены наиболее опасные кластеры и фактор риска заболеваемости.

Ключевые слова: климат, качество воды, инфекции, корреляционный анализ, метод *k*-средних, факторный анализ, мониторинг динамики.

ВВЕДЕНИЕ. ПОСТАНОВКА ЗАДАЧИ

В последнее время все большее внимание уделяется изменению климата, влиянию этого изменения на окружающую среду и, как следствие, на заболевания, связанные с состоянием окружающей среды. Возникает задача поиска компактного описания для конкретного региона, которое позволяло бы совместно характеризовать систему «ОКРУЖАЮЩАЯ СРЕДА – ИНФЕКЦИИ» (ОСИ), категоризовать ее состояния, мониторировать динамику и прогнозировать будущие состояния. Варианты такого описания на конкретном примере и являются объектом данного исследования.

В работе предпринята попытка исследования возможных статистических закономерностей и связей острых кишечных инфекций в г. Барнауле с погодными условиями и качеством воды в открытых водоисточниках этого города по ежедневным данным 2007 и 2008 гг., построения такого описания состояния ОСИ, которое было бы более компактным, чем исходная совокупность показателей, но достаточно информативным для исследований. Используется оригинальный банк данных CliWaDIn (Climate, Water, Diseases, Infectiones) [1].

Сравнительно полно в исследуемом периоде представлены следующие ежедневные данные: число заболеваний в день (человеко-день) – X_1 ; мутность воды (мг/л) – X_2 ; температура воды (°C) – X_3 ; щелочность (мг-экв/л) – X_4 ; температура воздуха (°C) – X_5 ; влажность воздуха (%) – X_6 .

Предварительные корреляционный и регрессионный анализы позволили выявить допускающие логическую интерпретацию значимые связи между отдельными заболеваниями и показателями состояния воды и воздуха. Был обнаружен факт наличия нелинейного характера связей между некоторыми парами признаков, позволяющий предположить изменение направления и тесноты связей в разных областях пространства признаков.

В связи с этим было проведено многоэтапное изучение связей признаков и средств обобщенного описания ОСИ. В настоящей работе изложена часть результатов первых стадий исследования различных вариантов совместного описания инфекционных заболеваний, передаваемых преимущественно водным путем, с одной стороны, и состояний погоды и воды, как элементов окружающей среды, с другой, рассматриваемых как единая система ОСИ.

Конкретно в работе рассматриваются два метода решения данной задачи. Первый – кластеризация исходных данных как инструмент выделения важных паттернов для компактного, в том числе визуального, представления состояния объекта. Второй – переход к новой совокупности показателей состояния ОСИ меньшей размерности.

КЛАСТЕРНЫЙ АНАЛИЗ

Для выявления паттернов состояния ОСИ используется метод k -средних [2]. С его помощью разбиваем все N дней года ($N = 365$ в 2007 г. и $N = 366$ в 2008 г.) на k кластеров. В связи с различием измерительных шкал далее используются стандартизованные значения (z -значения) этих шести признаков. $Z = (X - x_{\text{ср}})/s_x$, где $x_{\text{ср}}$ – среднее арифметическое, s_x – стандартное отклонение признака.

В качестве критерия качества кластеризации принят максимум значения статистики Фишера – отношения среднего квадрата межклассовых отклонений (между центрами кластеров) к среднему квадрату внутриклассовых отклонений от центра, характеризующих вариабельность признаков между и вариабельность внутри кластеров соответственно. Чем больше это отношение, тем «очевиднее» разделение точек дней, тем отчетливее границы кластеров в пространстве признаков.

Центры кластеров для 2007 – 2008 гг.

Признаки	Номер кластера			
	1	2	3	4
X_1 – число заболевших в день	8,58/10,12	12,97/14,69	14,38/12,80	6,87/7,70
X_2 – мутность воды	22,68/22,96	56,07/51,56	4,02/2,88	4,49/11,86
X_3 – температура воды	18,14/18,50	9,24/4,82	0,19/0,10	1,96/5,47
X_4 – щелочность	1,53/1,39	1,49/1,65	2,73/2,77	2,57/1,72
X_5 – температура воздуха	16,51/17,09	11,33/6,27	-7,75/-13,62	6,20/1,32
X_6 – влажность воздуха	66,35/64,96	53,54/55,49	72,86/78,02	76,21/77,91
Количество дней в кластере	138/133	44/46	63/110	120/77

По этому критерию наилучшим оказалось разбиение дней года на 4 кластера, приблизительно соответствующих сезонам года. В таблице сведены результаты.

По признаку заболеваемости X_1 наиболее опасен в 2007 г. третий кластер, а в 2008 г. – второй, хотя оба достаточно близки по данному признаку. Нумерация кластеров произвольна, но согласована по принципу близости их центров в различных годах.

Корреляционный анализ внутри каждого кластера подтвердил факт изменения характера, направления и тесноты связи между признаками в отдельных диапазонах их значений, т. е. в кластерах (паттернах) состояния ОСИ. Например, в 2008 г. коэффициент корреляции Спирмена между мутностью воды и заболеваемостью изменяется от $r_{12}=0,193$ в первом кластере до $r_{12}=-0,306$ во втором, а в остальных кластерах эта пара признаков не коррелирует. Щелочность, температура и влажность воздуха влияют на заболеваемость только в третьем кластере ($r_{14}=0,388$; $r_{15}=0,356$; $r_{16}=-0,337$). Характерно, что частные корреляции на этих парах при исключении всех остальных признаков все сохраняют значимость кроме r_{12} во втором кластере.

ФАКТОРНЫЙ АНАЛИЗ

Значимая линейная корреляция многих пар признаков погодной и водной обстановки свидетельствует о перспективности факторного анализа, как еще одного способа сокращения размерности пространства признаков на линейаризованном уровне. Факторный анализ массива данных (день-признак) включает: а) определение количества факторов, равного числу собственных значений корреляционной матрицы, ответственных за основную долю суммарной дисперсии (вариативности) признаков; б) определение линейной комбинации факторов, удовлетворительно описывающей признаки, т. е. проекций признаков в ортогональной системе факторов и в) определение выражения факторов в виде линейной комбинации признаков, объясняющей их корреляцию и допускающей содержательную интерпретацию каждого фактора, как обобщенного показателя. По методу главных компонент [2] было выявлено два фактора Y_1 и Y_2 , найдены координаты стандартизованных признаков в ортогональной системе факторов. Они оказались близкими в двух исследуемых периодах.

Найденная в итоге зависимость факторов от признаков для 2007 г. такова:

$$Y_1 = -0,15048z_1 + 0,14323z_2 + 0,30781z_3 - 0,29381z_4 + 0,29393z_5 - 0,09455z_6,$$

$$Y_2 = 0,73363z_1 + 0,35606z_2 - 0,20604z_3 + 0,01129z_4 - 0,05115z_5 - 0,33325z_6.$$

Для 2008 г. эти выражения достаточно похожи:

$$Y_1 = -0,19528z_1 + 0,12645z_2 + 0,30233z_3 - 0,31773z_4 + 0,29473z_5 - 0,08909z_6,$$

$$Y_2 = 0,66666z_1 + 0,34724z_2 - 0,13944z_3 + 0,09279z_4 + 0,01527z_5 - 0,42983z_6.$$

На рисунке показана динамика изменений факторов. Видно, что первый фактор Y_1 носит сезонный характер. Вторым фактором Y_2 , судя по силе влияния на него признака заболеваемости z_1 , можно назвать инфекционным риском. Он практически не отличается в рассматриваемых периодах и отчетливо показывает повышение риска заболеваемости ранней весной.

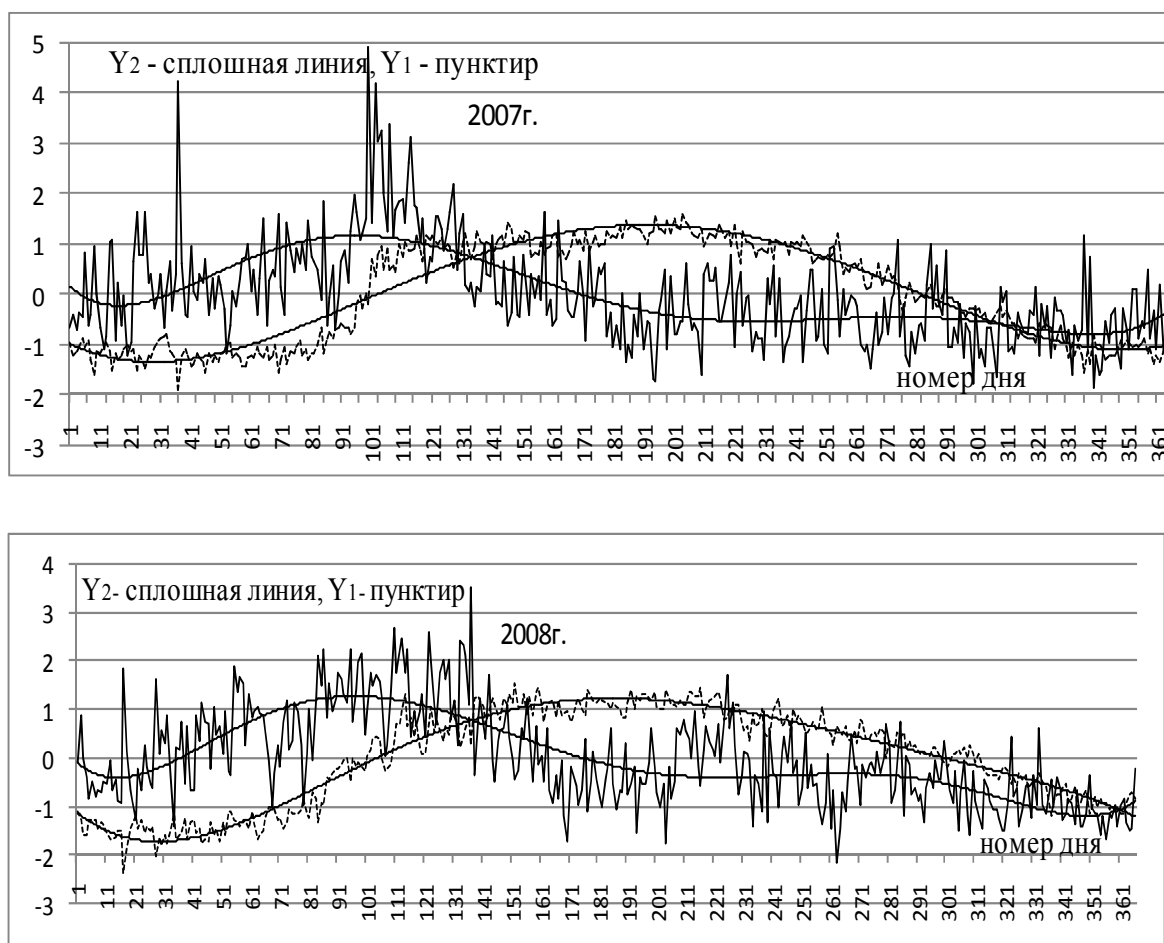


Рис.1. Динамика факторов Y_1 (пунктир) и Y_2 (сплошная линия) и полиномиальные тренды. На оси абсцисс дни года

Была предпринята попытка факторного анализа внутри массива дней, принадлежащих конкретному кластеру. Выяснилось существенное расхождение результатов, связанное с отмеченным выше различием корреляционных матриц в каждом кластере. В первом и четвертом кластерах 2008 г. было выявлено три фактора, во втором и третьем – два фактора. При этом выражения для факторов существенно отличаются.

На рисунке 2 показано евклидово расстояние от центра текущего дня до центра наиболее опасного второго кластера для 2008 г., измеренное в стандартизованных z -признаках. Полиномиальный тренд явно показывает весеннее приближение к опасному кластеру. При этом отчетливо возрастает вариативность этой меры риска, что также может служить индикатором опасности инфекций. Следует отметить, что подобный подход полезен для ретроспективного анализа инфекционной обстановки с осторожными обобщениями и рекомендациями на следующий период.

Примечательно, что для 2008 г. кривая удаленности дневной обстановки от наиболее опасного кластера 2, показанная на рис. 2, в точности противоположна динамике фактора Y_2 , который отражает инфекционный риск, как степень близости к опасному кластеру.

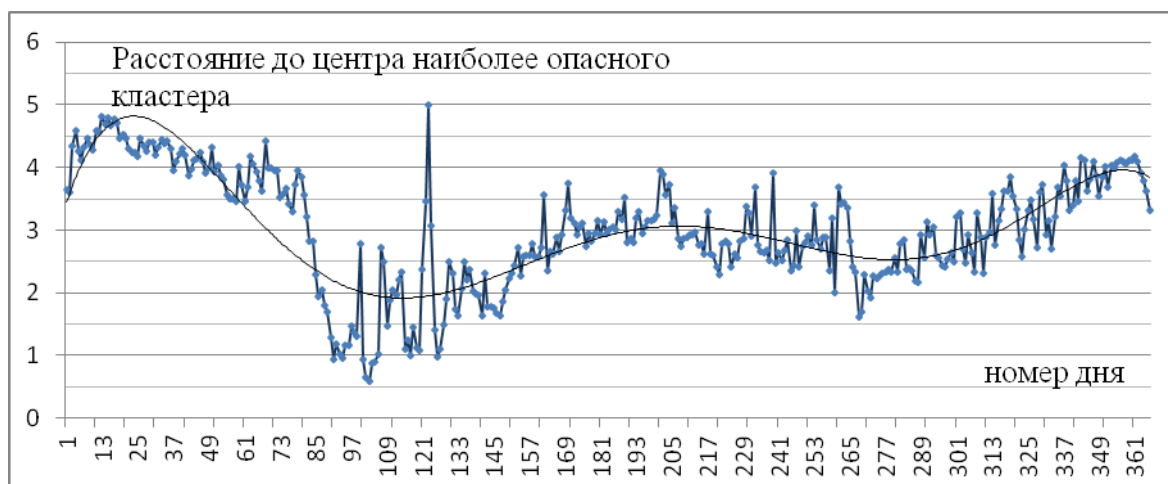


Рис.2. Евклидово расстояние дней 2008 г. до центра кластера № 2. Гладкая кривая – полиномиальный тренд

ЗАКЛЮЧЕНИЕ

Проведенный анализ данных инфекционной, водной и погодной обстановки показывает перспективность подхода, предлагаемого для разработки системы мониторинга заболеваемости и контроля уровня инфекционной угрозы для населения большого города по одному или нескольким факторам. Каждый день по признакам погоды и качества воды может быть отнесен к конкретному кластеру и, следовательно, ранжирован по степени опасности риска заражения острыми кишечными инфекциями, скажем, по близости к наиболее опасному из кластеров. Можно выявить несколько критических значений признаков, превышение которых должно вызывать тревогу у соответствующих специалистов и служб.

В дальнейшем предполагается более детальное исследование в рамках данных подходов на примере других городов и отдельных инфекций, а также развитие подходов с применением нелинейных корреляционного и регрессионного анализов, других методов классификации. Визуализации и интеллектуального анализа данных.

ЛИТЕРАТУРА

1. *Acseanova, V.* Climate and infectious diseases. Establishment of data repository for examining associations between, weather quality and infectious diseases / V. Acseanova [et al.]// NIAID-ISTC Conference on Bioinformatics Tools and Techniques for Allergy and Infectious Diseases Research, Moscow, June 07-09, 2010. P. 66.
2. *Айвазян, С. А.* Прикладная статистика: Классификация и снижение размерности / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. М.: Финансы и статистика, 1989. 472 с.