С. Н. Белугин

КЛАСТЕРНЫЙ АНАЛИЗ МУЛЬТИПОЛЛЮТАНТНЫХ ВЫБРОСОВ В АТМОСФЕРУ И ПЕРВИЧНАЯ ЗАБОЛЕВАЕМОСТЬ ДЕТЕЙ В БЕЛАРУСИ, 2005–2013 гг.

УО «Белорусский государственный медицинский университет»

В исследовании применен кластерный анализ для классификации профилей мультиполлютантных выбросов в атмосферу (ПМПВА, n = 63) в Беларуси. Для деления ПМПВА на кластеры метод PAM (partitioning around medoids) был предпочтен методу k-средних (k-means) и алгоритмам иерархического анализа. Основываясь на выделенных кластерах мультиполлютантного загрязнения воздуха удалось разделить данные первичной детской заболеваемости в областях Беларуси (с 2005 по 2013 г.) на четыре группы. С помощью дисперсионного анализа выявлена зависимость первичной заболеваемости по болезням кожи, нервной системы, психическим расстройствам и болезням органов кровообращения среди детей от особенностей мультиполлютантного загрязнения воздуха.

Ключевые слова: *мультиполлютантные выбросы в атмосферу, первичная детская заболеваемость, кластер.*

S. N. Belugin

CLUSTER ANALYSIS OF MULTIPOLLUTANT AIR EJECTIONS AND PRIMARY CHILDREN MORBIDITY IN BELARUS, 2005–2013 yy.

Nine years of regional multipollutant air ejections data reported in Belarus were classified on their annual multipollutant profiles (n = 63). Classification of annual multipollutant profiles was performed using partitioning around medoids (PAM) clustering. The optimal solution by PAM contained four distinct groups of annual multipollutant profiles. Mean annual air ejections of 12 pollutants (tons per year) and their ratios were computed in order to characterize the differences between clusters. The values of annual primary children morbidity (2005–2013 yy.) were partitioned for four groups. This allowed us to link the characteristics of each cluster to health effects and show that primary morbidity of kids with diseases of skin, nervous system, psychic and cardiovascular disorders were associated with distinct cluster ofmultipollutant profiles.

Key words: *multipollutant air ejections, primary children morbidity, cluster.*

Эпидемиологические исследования указывают на негативные эффекты от взвешенных частиц, оксидов азота, оксидов серы, летучих органических соединений и других поллютантов воздуха на здоровье взрослых и детей [9, 11, 12]. Вместе с тем, интерпретация негативных эффектов на здоровье со стороны каждого из поллютантов всегда осложнена из-за присутствия в воздухе многокомпонентной смеси загрязнителей.

На территории Республики Беларусь каждый год в воздух выбрасываются тысячи тонн смесей поллютантов от стационарных и мобильных источников, что отражено в ежегодных сводках НАН Беларуси и Министерства природных ресурсов и охраны окружающей среды [1]. При имеющихся регионарных особенностях источников выбросов в атмосферу формируются различные профили мультиполлютантных выбросов в атмосферу (ПМПВА). Оценка качества воздуха не только на основании уровней отдельных поллютантов, но и их количественных соотношений по данным мультиполлютантных выбросов в атмосферу, актуальна и крайне важна, в частности,

при анализе ассоциированности состояния здоровья человека с особенностями многокомпонентного состава загрязнений в воздухе. Для развития методов оценки связи заболеваемости с мультиполлютантным загрязнением воздуха существует необходимость классифицирования многообразных ПМПВА и формирования интегративной шкалы качества воздуха по мультиполлютантным критериям.

Кластерный анализ совсем недавно был применен для идентификации и описания мультиполлютантных профилей загрязнения воздуха [4, 13]. В особенности, в первых работах для кластерного анализа использованы данные ежедневного мониторинга воздуха в Бостоне и был выбран метод *k*-средних для выделения в группы дней с однотипными мультиполлютантными профилями загрязнения воздуха. В нашем исследовании кластерному анализу подвергались ПМПВА по данным ежегодных выбросов в атмосферу от стационарных и мобильных источников для каждой из шести областей Беларуси и г. Минска за период с 2005 по 2013 г.

Задачей настоящего исследования было объединение при помощи кластерного анализа многомерные объекты – ПМПВА – в кластеры. Целью исследования было применение кластерного критерия определения типа ПМПВА для объединения данных заболеваемости в Беларуси в группы и последующего анализа ассоциированности негативных последствий на здоровье с многокомпонентным составом загрязненного воздуха.

Материалы и методы

В настоящей работе использованы отчетные сведения о выбросах в атмосферу с 2005 по 2013 г. на территориях шести областей Беларуси и г. Минска [1]. По стационарным источникам использованы данные выбросов в атмосферу (тыс. тонн в год): угарного газа (с*CO*), оксидов азота (с*NOx*), диоксида серы (сSO_2), метановых углеводородов без летучих органических соединений (сМУВ), неметановых летучих органических соединений (сНМЛОС), взвешенных частиц (с*PM, particulate matter*). По мобильным источникам использованы данные выбросов в атмосферу (тыс. тонн в год): угарного газа (м*CO*), оксидов азота (м*NOx*), диоксида серы (мSO_2), летучих органических соединений (мЛОС), взвешенных частиц (м*PM, particulate matter*), бензопирена (мБП, тонн в год). Таким образом, сформирована выборка из 63 многомерных (12 параметров) объектов ПМПВА, каждый из которых соответствовал определенному году (n = 9) и региону (n = 7).

Абсолютные значения 12 параметров выбросов в атмосферу были преобразованы в z-значения:

$$z = (X - \mu)/\sigma,$$

где X – значение параметра; μ – среднее значение; σ – среднеквадратичное отклонение. Для преобразования в z-значения использована программа *SPSS v.*13.0.

Данные первичной заболеваемости детей по болезням глаз, кожи, дыхательной системы, органов кровообращения, нервной системы и психическим расстройствам взяты из ежегодных статистических отчетов Министерства здравоохранения Республики Беларусь о состоянии здоровья населения по областям за период с 2005 г. по 2013 г. [2, 3].

Использование кластерного анализа предполагало распределение ПМПВА по кластерам, каждый из которых объединял бы ПМПВА с однотипным соотношением 12 параметров. На выборке из 63 объектов ПМПВА протестированы на валидность методы кластерного анализа: *PAM* (partitioning *around medoids*) [8], *k*-средних [6] и иерархические алгоритмы «*centroid*» и «*complete*». На основании расчета *R*-квадратичного индекса (R^2) [10] и индекса *Davies-Bouldin* (*DB*) [5] производился выбор метода для кластерного анализа и выбор оптимального количества кластеров для разбиения выборки из 63 объектов ПМПВА. Для расчета *R*-квадратичного и *Davies-Bouldin* индексов, а также для разбиения 63 объектов ПМПВА на кластеры использована программа *CVAP v.*3.7 (*Cluster Validity Analysis Platform* версия 3.7) на базе *Matlab v.*7.0.1 [7].

R-квадратичный индекс рассчитывался по отношению *SSB/TSS*, где *SSB* (*sum of squares between*) – сумма квадратов отклонений между группами, и *TSS* (*total sum of squares*) – общая сумма квадратов отклонней между группами и внутригрупповых отклонений. Минимальное значение (0) *R*-квадратичного индекса отражает абсолютное подобие групп, максимальное значение (1) – полное отличие между выделенными группами.

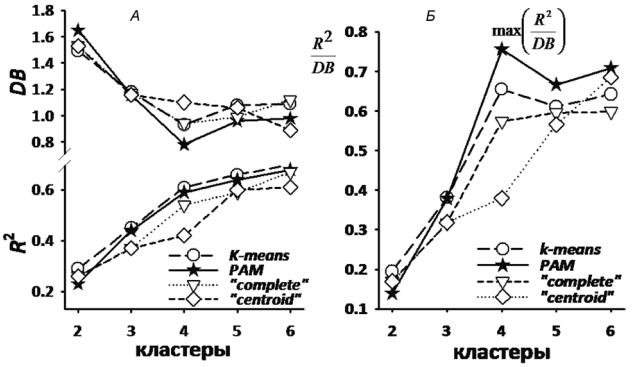
Davies-Bouldin индекс [5]:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max\left(\frac{S_i + S_j}{M_{ij}}\right),$$

где *n* – количество кластеров; S_i – среднее расстояние от центра в кластере *i*; S_j – среднее расстояние от центра в кластере *j*; M_{ij} – расстояние между центрами кластеров *i* и *j*. Минимальные значения *Davies-Bouldin* индекса свидетельствуют о сравнительной компактности групп вокруг кластеробразующих центров при сравнительной удаленности последних друг от друга.

Исходя из предпочтительности наибольшего значения R-квадратичного индексаили наименьшего значения *Davies-Bouldin* индекса, в настоящем исследовании введено отношение $\frac{R^2}{DB}$ (отношение *R*-квадратичного индекса к *Davies-Bouldin* индексу), максимальное значение которого ($\max\left(\frac{R^2}{DB}\right)$) мы рассматривали в качестве критерия предпочтительности метода и оптимальности выбранного количества кластеров (рисунок).

Для сравнения чувствительности методов кластерного анализа вся выборка из 63 объектовПМПВА разбивались при помощи *PAM* и *k*-средних на 4 кластера по



Индексы валидности методов кластерного анализа при объединении ПМПВА в кластеры: А – *R*-квадратичный индекс (R^2) и индекс *Davies-Bouldin* (*DB*); Б – Максимальное отношение $\frac{R^2}{DB}$ свидетельствует о предпочтительности метода и оптимальном количестве кластеров

данным: 1) шести областей и г. Минска, и 2)только шести областей. Оценивали количество несоответствий ПМПВА кластеру при применении методов на выборке содержащей данные только по шести областям, т. е. уменьшенной на 14,2%. Метод *k*-средних при уменьшении выборки ПМПВА на 9 (на 14,2%, от 63) дал разбиение на четыре кластера с 7 ошибками (12,9%). Метод *PAM* при уменьшении выборки ПМПВА на 14,2% дал четыре кластера с 1-й ошибкой (1,8%), что явилось дополнительным подтверждением предпочтительности метода *PAM* перед методом *k*-средних.

Статистическая обработка данных осуществлялась в программе SigmaStat for Windows v.3.10. Использован дисперсионный анализ *One Way ANOVA*. Достоверность различий между средними значениями согласно *t*-теста Стьюдента принималась при $P < 0.05$.

Результаты и обсуждение

Значения *R*-квадратичного индекса (R^2) и *Davies-Bouldin* (*DB*) индекса рассчитаны для условий разбиения на 2–6 кластеров выборки из 63 объектов ПМПВА (рисунок, *А*). Наибольшее отношение $\frac{R^2}{DB}$ выявлено при разбиении всей выборки ПМПВА методом *PAM* на четыре кластера (рисунок, Б). Исходя из наименьшего значения *Davies-Bouldin* индекса и наибольшего отношения $\frac{R^2}{DB}$ метод *PAM* оказался более предпочтителен для классифицирования ПМПВА по сравнению с методом *k*-средних и алгоритмами иерархического анализа.

Выбор метода *PAM* отражает особенность выборки данных ПМПВА (n = 63, каждый ПМПВА по 12 параметров), которая невелика по сравнению с выборками в исследованиях других авторов [4, 13]. Метод *PAM* предпочтителен при относительно небольшом массиве данных и менее подвержен влиянию со стороны крайних величин (*outliers*), тогда как для реализации метода *k*-средних требуются большие массивы данных.

Разбиение выборки ПМПВА на четыре кластера позволило таблично представить четыре типа мультиполлютантных профилей (№ 1, № 2, № 3, № 4) в соответствии с годом и областью (табл. 1). В кластер № 1 вошли ПМПВА следующих областей: Брестской – 2005–2013 гг., Гомельской – 2008, 2011, 2012 гг., Гродненской – 2005–2011 гг., Минской – 2011–2013 гг., а также г. Минска – 2005–2013 гг. В кластер № 2 вошли ПМПВА Витебской области – 2005–2013 гг. и Гомельской области – 2005–2007 гг., 2009, 2010 гг. В кластер № 3 вошли ПМПВА областей: Могилевской – 2005–2013 гг., Гомельской – 2013 г. и Гродненской – 2012–2013 гг. В кластер № 4 вошли ПМПВА Минской области – 2005–2010 гг.

Рассчитаны средние значения и стандартные ошибки 12 химических параметров для каждого кластера (табл. 2). С помощью дисперсионного анализа (*One Way ANOVA*) выявлены значимые различия между кластерами по всем 12 химическим поллютантам.

Для каждого кластера рассчитаны показатели, характеризующие источники выбросов – отношения значений выбросов поллютантов от стационарных источников к соответствующим значениям от мобильных источников, а также отношение сМУВ к сНМЛОС (табл. 3).

При дисперсионном анализе данных первичной детской заболеваемости в соответствии с выделенным кластером выявлены достоверные различия по болезням органов кровообращения и психическим расстройствам. Максимальный уровень первичных проявлений болезней органов кровообращения, а также болезней кожи представлен в кластере № 4 (табл. 4). Максимальное количество первичных случаев проявления психических расстройств среди детей приходится на кластер № 1. Минимальная первичная заболеваемость детей по болезням кожи и психическим расстройствам приходится на кластер № 2. Минимальная первичная заболеваемость детей по органам кровообращения приходится на кластер № 3.

Каждый из четырех кластеров – четырех типов ПМПВА – имеет характерные особенности по количественному соотношению 12 параметров. Кластер № 1 в сравнении с кластерами № 2 и № 3 характеризуется более высокими значениями выбросов в атмосферу от мобильных источников. В кластере № 1 самый высокий уровень выбросов от мобильных источников относительно выбросов от стационарных источников. Первичная заболеваемость среди детей отнесенная к кластеру № 1 имеет максимальные значения по болезням глаз, дыхательной системы, психическим расстройствам и болезням нервной системы. Первичная заболеваемость среди детей отнесенная к кластеру № 1 превышает также по болезням кожи и органов кровообращения первичную заболеваемость в кластерах № 2 и № 3.

Кластер № 2 характеризуется максимальными значениями выбросов в атмосферу от стационарных источников оксидов азота, оксидов серы, неметановых летучих органических соединений. В кластере № 2 отмечается максимальное отношение выбросов оксидов серы и азота от стационарных источников к выбросам оксидов серы и азота (соответственно) от мобильных источников, и наименьшее отношение выбросов метановых углеводородов к неметановым летучим органическим соединениям от стационарных источников (табл. 3).

Таблица 1. **Классификация типов ПМПВА в соответствии с номером кластера**

	Брестская область	Витебская область	Могилевская область	Гомельская область	Гродненская область	Минская область	г. Минск
2013 г.	1	2	3	3	3	1	1
2012 г.	1	2	3	1	3	1	1
2011 г.	1	2	3	1	1	1	1
2010 г.	1	2	3	2	1	4	1
2009 г.	1	2	3	2	1	4	1
2008 г.	1	2	3	1	1	4	1
2007 г.	1	2	3	2	1	4	1
2006 г.	1	2	3	2	1	4	1
2005 г.	1	2	3	2	1	4	1

Таблица 2. **Химическая характеристика кластеров**

Вид источника	Поллютант		Кластер № 1, n = 31	Кластер № 2, n = 14	Кластер № 3, n = 12	Кластер № 4, n = 6	*P**
Стационарный источник	*CO* тыс. тонн/год	*M* ± *SE*	11,12 ± 0,54	15,76 ± 0,88	9,6 ± 0,77	26,76 ± 2,29	<0,001
Стационарный источник	МУВ без ЛОС тыс. тонн/год	*M* ± *SE*	7,64 ± 1,21	5,64 ± 1,23	13,15 ± 1,86	10,05 ± 1,0	<0,015
Стационарный источник	*NOx* тыс. тонн/год	*M* ± *SE*	6,74 ± 0,35	14,66 ± 0,62	8,48 ± 0,63	12,71 ± 0,98	<0,001
Стационарный источник	SO_2 тыс. тонн/год	*M* ± *SE*	5,01 ± 1,14	28,05 ± 2,45	4,48 ± 1,65	13,85 ± 3,95	<0,001
Стационарный источник	*PM* тыс. тонн/год	*M* ± *SE*	5,17 ± 0,3	6,65 ± 0,29	6,32 ± 0,2	12,08 ± 0,46	<0,001
Стационарный источник	НМЛОС тыс. тонн/год	*M* ± *SE*	4,78 ± 0,81	28,33 ± 2,17	5,15 ± 0,89	7,66 ± 0,94	<0,001
Мобильный источник	*PM* тыс. тонн/год	*M* ± *SE*	4,91 ± 0,11	4,48 ± 0,09	3,3 ± 0,16	6,06 ± 0,22	<0,001
Мобильный источник	*CO* тыс. тонн/год	*M* ± *SE*	109,27 ± 3,46	89,38 ± 3,2	67,54 ± 2,93	140,91 ± 6,63	<0,001
Мобильный источник	SO_2 тыс. тонн/год	*M* ± *SE*	0,3 ± 0,02	0,21 ± 0,02	0,12 ± 0,02	0,3 ± 0,04	<0,001
Мобильный источник	*NOx* тыс. тонн/год	*M* ± *SE*	16,27 ± 0,34	13,88 ± 0,24	10,72 ± 0,52	19,16 ± 0,64	<0,001
Мобильный источник	УВ тыс. тонн/год	*M* ± *SE*	31,75 ± 0,71	26,51 ± 0,56	20,47 ± 0,88	39,05 ± 1,46	<0,001
Мобильный источник	Бензопирен тонн/год	*M* ± *SE*	126,14 ± 3,61	113,43 ± 3,56	90,23 ± 3,36	157,5 ± 13,14	<0,001

Примечание: МУВ без ЛОС – метановые углеводороды без летучих органических соединений; *PM* (*particulate matter*) – взвешенные частицы; НМЛОС – неметановые летучие органические соединения; УВ – углеводороды. * – применен *One Way ANOVA*.

Таблица 3. **Отношения выбросов поллютантов от стационарных (с) источников к выбросам от мобильных (м) источников**

Поллютант	Средняя (*M*) ± стандартная ошибка (*SE*)	Кластер № 1, n = 31	Кластер № 2, n = 14	Кластер № 3, n = 12	Кластер 4, n = 6
с*CO*/м*CO*	*M* ± *SE*	0,1 ± 0,005	0,17 ± 0,007	0,14 ± 0,008	0,19 ± 0,01
с*PM*/м*PM*	*M* ± *SE*	1,0 ± 0,05	1,4 ± 0,06	1,96 ± 0,1	2,0 ± 0,1
сSO_2/мSO_2	*M* ± *SE*	19,2 ± 4,4	154,2 ± 26,9	41,4 ± 17,1	57,0 ± 22,4
с*NOx*/м*NOx*	*M* ± *SE*	0,42 ± 0,02	1,06 ± 0,05	0,8 ± 0,06	0,6 ± 0,06
сМУВ/мУВ	*M* ± *SE*	0,24 ± 0,04	0,21 ± 0,04	0,63 ± 0,06	0,2 ± 0,02
сНМЛОС/мУВ	*M* ± *SE*	0,15 ± 0,02	1,08 ± 0,09	0,24 ± 0,03	0,2 ± 0,02
сМУВ/сНМЛОС	*M* ± *SE*	2,22 ± 0,37	0,22 ± 0,05	2,75 ± 0,32	1,39 ± 0,17

Примечание: МУВ – метановые углеводороды без летучих органических соединений; *PM* (*particulate matter*) – взвешенные частицы; НМЛОС – неметановые летучие органические соединения; УВ – углеводороды.

Таблица 4. **Первичная заболеваемость детей**

Первичная заболеваемость детей (на 100 тыс.)	Средняя ± стандартная ошибка	Кластер № 1, n = 31	Кластер № 2, n = 14	Кластер № 3, n = 12	Кластер № 4, n = 6	*P* (*One Way ANOVA*)
Болезни глаз	*M* ± *SE*	5025,4 ± 472,0	3953,6 ± 138,1	4188,3 ± 249,7	4241,8 ± 144,3	0,294
Болезни органов дыхания	*M* ± *SE*	132121,4 ± 5893,9	120915,3 ± 3488,4	120089,0 ± 2693,0	127061,0 ± 3811,2	0,381
Болезни кожи	*M* ± *SE*	6217,2 ± 338,7	5641,7 ± 645,4	5896,6 ± 510,7	8207,6 ± 312,1	0,056
Психические расстройства	*M* ± *SE*	1793,6 ± 148,2	1051,3 ± 109,5	1385,1 ± 57,3	1489,3 ± 151,5	0,005
Болезни нервной системы	*M* ± *SE*	1016,4 ± 124,3	988,0 ± 72,9	699,0 ± 33,8	899,8 ± 31,7	0,328
Заболевания органов кровообращения	*M* ± *SE*	827,2 ± 86,4	470,6 ± 46,0	355,8 ± 24,0	1014,8 ± 36,3	<0,001

Кластер № 3 характеризуется максимальными значениями выбросов в атмосферу от стационарных источников метановых углеводородов без летучих органических соединений. В кластере № 3 отмечаются минимальные выбросы в атмосферу от мобильных источников. Первичная заболеваемость среди детей отнесенная к кластеру № 3 имеет минимальные значения по болезням нервной системы, которые достоверно отличаются от значений в кластере № 2 ($P = 0.002$) и кластере № 4 ($P = 0.001$) (табл. 4).

Кластер № 4 характеризуется максимальными значениями выбросов в атмосферу от мобильных источников по всем шести параметрам, и максимальными выбросами угарного газа с взвешенными частицами от стационарных источников. Первичная заболеваемость детей по болезням кожи в кластере № 4 достоверно выше, чем в кластерах № 1 ($P = 0.002$), № 2 ($P = 0.016$) и № 3 ($P = 0.008$) (табл. 4).

Учитывая наибольший относительный и абсолютный вклад в выбросы от мобильных источников, кластеры № 1 и № 4 характеризуются максимальным проявлением первичной детской заболеваемости по сравнению с первичной заболеваемостью соотносимой с кластерами № 2 и № 3.

Полученные результаты при помощи метода кластерного анализа предполагается использовать для более детального ретроспективного и проспективного анализа связи заболеваемости с определенными типами ПМПВА, а также для выявления поллютантов являющихся доминирующей причиной в негативных последствиях на здоровье. Примененный подход предполагает возможности оценки риска эколого-зависимых заболеваний в Беларуси по критериям тесно связанным с характером мультиполлютантного профиля выбросов в атмосферу, и исследовать эффекты на здоровье со стороны отдельных поллютантов с учетом типа ПМПВА в качестве модифицирующего условия.

Таким образом, с помощью кластерного анализа классифицированы основные типы ПМПВА для Беларуси. Определение типа ПМПВА рассматривается в качестве необходимого условия для оценки связи заболеваемости с многокомпонентным загрязнением воздуха.

Литература

1. *Состояние* природной среды Беларуси. Экологический бюллетень. Под ред. академика НАН Беларуси В. Ф. Логинова. 2005–2013 гг. – [Электронный ресурс]. – Режим доступа: http://www.minpriroda.gov.by/

2. *Здравоохранение* в Республике Беларусь. Официальные статистические сборники. РНМБ, Минск. 2010–2014 гг. – [Электронный ресурс]. – Режим доступа: http://minzdrav.gov.by/ru/static/numbers/zabolevaemost/.

3. *Первичная* заболеваемость населения Республики Беларусь отдельными болезнями в 2005–2012 гг. – [Электронный ресурс]. – Режим доступа: http://minzdrav.gov.by/ru/static/numbers/zabolevaemost/

4. *Austin,* E., Coull B., Thomas D., Koutrakis P. A framework for identifying distinct multipollutant profiles in air pollution data // Environ Int. – 2012. – Vol. 45. – P. 112–121.

5. *Davies,* D. L., Bouldin D. W. A cluster separation measure // IEEE Trans Pattern Anal Mach Intell. – 1979. – P. 224–227.

6. *Hartigan,* J., Wong M. A k-means clustering algorithm // J. R. Stat. Soc. C. – 1979. – Vol. 28. – P. 100–108.

7. *Kaijun,* W., Baijie W., Peng L. CVAP: Validation for cluster analyses // Data Science Journal. – 2009. – Vol. 8, № 20. – P. 88–93.

8. *Kaufman*, L., Rousseeuw P. Clustering by means of medoids // Statistical Data Analysis Based on the L1 Norm and Related Methods. North-Holland. – 1987. – P. 405–416.

9. Lee, B. J., Kim B., Lee K. Air pollution exposure and cardiovascular disease // Toxicol Res. – 2014. – Vol. 30, № 2. – P. 71–75.

10. *Sharma,* S. // Applied multivariate techniques. John Wiley & Sons, Inc..: New York, – 1996. – P. 493.

11. *Schwela,* D. Air pollution and health in urban areas // Rev. Environ. Health. – 2000. – Vol. 15, № 1. –2. – P. 13–42.

12. *Wang*, S., Zhang J., Zeng X., Zeng Y., Wang S., Shuyun C. Association of Traffic-Related Air Pollution with Children's Neurobehavioral Functions in Quanzhou, China// Environmental Health Perspectives. – 2009. – Vol.117, № 10. – P. 1612–1618.

13. *Zanobetti*, A., Austin E., Coull B. A., Schwartz J., Koutrakis P. Health effects of multi-pollutant profiles // Environ Int. – 2014. – Vol. 71. – P. 13–19.

Поступила 29.06.2015 г.