УДК 519.2

doi: 10.25750/1995-4301-2019-2-143-148

Статистические методы анализа взаимосвязи качества атмосферного воздуха и состояния здоровья детского населения Кировской области

© 2019. С. И. Калинин, д. п. н., профессор, С. И. Торопова, ассистент,
Вятский государственный университет,
610000, Россия, г. Киров, ул. Московская, д. 36,
e-mail: kalinin_gu@mail.ru, svetori82@mail.ru

Настоящее исследование реализовано с целью оценки вклада загрязнения атмосферного воздуха в формирование детской заболеваемости на территории Кировской области. На основе оценки областных показателей установлено наличие статистически значимых связей между выбросами в атмосферу углеводородов, летучих органических соединений и первичной заболеваемостью детского населения области болезнями органов дыхания. В результате анализа статистических показателей на уровне муниципальных образований обосновано наличие статистически значимого вклада выбросов загрязняющих веществ от стационарных источников загрязнения в возникновение заболеваний детского населения региона острыми инфекциями верхних дыхательных путей. Математическое моделирование осуществлялось с помощью регрессионного анализа по панельным данным, представляющего возможность анализировать индивидуальные отличия изучаемых объектов, чего нельзя сделать в рамках стандартных регрессионных моделей. Определено, что регрессионная модель с фиксированными эффектами обеспечивает получение обоснованного варианта моделирования.

Ключевые слова: корреляционно-регрессионный анализ, регрессионные модели, панельные данные, кластерный анализ.

Statistical methods for analyzing the correlatio between air quality and the state of children's health in the Kirov region

© 2019. S. I. Kalinin ORCID: 0000-0001-5439-9414**, S. I. Toropova** ORCID: 0000-0003-0533-5654**,**
Vyatka State University,
36, Moskovskay St., Kirov, Russia, 610000,
e-mail: kalinin_gu@mail.ru, svetori82@mail.ru

The aim of this study is to assess the contribution of atmospheric air pollution to the formation of childhood morbidity in the Kirov region. Statistically significant interaction of air emissions of hydrocarbons, volatile organic compounds and the primary incidence of respiratory diseases among the children of the region has been established on the base of the regional indicators assessment. The analysis of statistical indicators at the municipal level showed the presence of a statistically significant contribution of pollutant emissions from stationary pollution sources to the occurrence of diseases of the children of the region by acute upper respiratory tract infections. Mathematical modeling was carried out with the help of regression analysis based on panel data, which makes it possible taking into account the spatial and temporal structure of the analyzed statistical data, what cannot be done using standard regression models. The multidimensional classification of the regions of the Kirov region in terms of the level of atmospheric pollution by means of cluster analysis into four groups preceded the implementation of the panel data analysis. Using Fisher, Breusch-Pagan and Hausman tests, it is determined that the regression model with fixed effects provides a justified version of the simulation.

A similar research devoted to the correlation between the prevalence of diseases of the respiratory system, digestion, skin and subcutaneous tissue, the genitourinary system, infectious and parasitic diseases, the musculoskeletal system and connective tissue, the eye and its adnexa in the adult population of the Kirov region and the amount of pollutants released into the atmosphere doesn't show statistically significant interrelations.

Keywords: correlation and regression analysis, regression models from panel data, cluster analysis.

Согласно классификации Всемирной организации здравоохранения, второе место среди факторов, оказывающих влияние на процесс формирования здоровья населения, после социально-экономических показателей принадлежит экологическим факторам, или состоянию окружающей среды [1]. Имеется обширное количество серьёзных исследований по экологии и медицине, посвящённых изучению данной зависимости, существенное место среди которых занимают работы по анализу экологозависимых заболеваний, в том числе болезней органов дыхания (БОД) [1–6]. Особую обеспокоенность авторов вызывает заболеваемость детского населения как индикатора экологически неблагополучного состояния воздушной среды [2–4, 6].

В соответствии с официальными данными [7] наиболее частой причиной первичной заболеваемости всех категорий населения региона, в том числе детского, на протяжении нескольких последних десятилетий являются БОД. Следовательно, проблема влияния качества атмосферного воздуха на заболеваемость населения представляется актуальной для жителей области.

Одним из важнейших инструментов изучения рассматриваемых зависимостей выступает математический аппарат, в частности математическое моделирование и статистический анализ данных. В связи с тем, что многие характеристики окружающей среды и показатели здоровья населения являются непрерывными количественными переменными, для их оценки применяются методы корреляционно-регрессионного анализа [3].

Не менее важной задачей, решаемой методами математической статистки, является задача классификации объектов и их группировка по следующим причинам. Во-первых, существуют многочисленные исследования, авторы которых делают обоснованный вывод о том, что оценка взаимосвязи факторов окружающей среды и состояния здоровья населения должна осуществляться с учётом конкретной территории [3–5]. В тех случаях, когда исследование охватывает, например, федеральный округ, целесообразно распределение территорий на группы, характеризующиеся однотипными показателями. Во-вторых, одной из существенных проблем, возникающих в подобных исследованиях, является необходимость учёта одновременного воздействия огромных массивов факторов разной силы, интенсивности и природы. В связи с тем, что анализ модели, содержащей десятки или сотни переменных, затруднителен, целесообразно уменьшить размерность множества исходных данных и оставить из них приоритетные группы.

Обзор научных исследований свидетельствует о том, что большинство из них может быть отнесено к одному из двух направлений: изучение временных рядов (статистических показателей определённой территории в разные моменты времени) или пространственных совокупностей (информации в фиксированный момент времени на различных территориях). Практически отсутствуют работы, реализующие одновременно анализ по времени и в пространстве. Можно выделить исследование [3], в котором описано несколько типов регрессионных моделей, отражающих влияние факторов воздушного бассейна на распространение БОД населения в зависимости от биоклиматических зон Приморского края за 2000–2013 гг.

Построение одной модели, совмещающей в себе пространственные и временные ряды и сочетающей достоинства каждого из этих видов данных, предусматривает аппарат регрессионного анализа панельных данных. Вследствие своей специальной структуры данные этого типа обеспечивают построение более адекватных и содержательных математических моделей с целью изучения причинно-следственной связи между факторами [9–11].

Цель настоящего исследования – установить наличие статистически значимого вклада загрязнителей атмосферного воздуха Кировской области в формирование заболеваемости детского населения региона БОД на основе моделей панельных данных, учитывающих пространственную и временную структуру имеющихся результатов мониторинга.

Объекты и методы

Объектами исследования являлись статистические данные, опубликованные на официальных сайтах правительства Кировской области, Роспотребнадзора, Росстата и Кировстата за 2002–2017 гг.

На уровне Кировской области для установления возможной связи качества атмосферного воздуха и заболеваемости детского населения БОД в качестве исследуемых функций были составлены и проанализированы регрессионные модели с помощью программы MS Excel. Целью исследования на уровне районов Кировской области явилось

изучение вклада отдельных загрязнителей в формирование заболеваемости детского населения острыми инфекциями верхних дыхательных путей с использованием моделей панельных данных, параметры которых были оценены посредством программы Gretl.

Принято выделять три типа регрессионных моделей по панельным данным: объединённая модель (модель сквозной регрессии, pooled model), модель с фиксированными эффектами (fixed effects model, FEM), модель со случайными эффектами (random effects model, REM) [9–11]. Представим данные модели.

Объединённая модель есть обычная линейная модель регрессии, она фактически не учитывает панельную структуру данных, в частности индивидуальные различия изучаемых объектов. Учёт упоминаемых индивидуальных различий предусматривает модель с фиксированными эффектами за счёт включения в уравнение регрессии N фиктивных переменных, где N – количество объектов исследования. Считается целесообразным использование модели с фиксированными эффектами, если выбирается уникальный набор N регионов страны, N муниципальных образований определённой административной территории и т. п. [12]. Модель регрессии со случайными эффектами основана на предположении о случайном выборе N объектов из некоторой генеральной совокупности.

Панельный анализ данных представлен в публикациях, посвящённых современным исследованиям в области медицины и экологии. Например, целью работы [12] явилась оценка социально-экономических и экологических факторов на региональные демографические процессы в РФ. В нём обоснована адекватность модели с фиксированными эффектами. В исследовании [13] представлен анализ обычных регрессионных моделей и регрессионных моделей по панельным данным, рассмотрены их сходства и различия, недостатки и преимущества. В данном источнике приведены примеры панельных исследований из отечественной и зарубежной литературы по медицине.

В настоящем исследовании реализации панельного анализа данных предшествовала многомерная классификация районов Кировской области по уровню загрязнения атмосферы средствами кластерного анализа с применением пакета прикладных программ Statistica 12.0 методом k-средних.

Результаты и обсуждение

В процессе реализации поставленной цели исследования формулировалась последовательность взаимосвязанных задач, решение которых осуществлялось поэтапно.

На первом этапе моделирования зависимостей «Загрязнение атмосферного воздуха – здоровье детского населения» проводился научно обоснованный отбор факторов в модель. Были учтены имеющиеся в открытом доступе статистические данные, ряд исследований [1, 5, 6], наличие тесной связи факторов с зависимыми переменными и слабой коррелированности друг с другом. В результате в качестве исследуемых функций были отобраны y_1 – первичная заболеваемость БОД детского населения Кировской области (на 1000 чел.) и y_2 – заболеваемость детского населения районов области острыми инфекциями верхних дыхательных путей (на 100 тыс. чел.); в число независимых переменных были включены следующие показатели: x – количество выброшенных в атмосферу загрязняющих веществ от стационарных источников загрязнения (тыс. тонн), из них x_1 – газообразные и жидкие вещества (тыс. тонн), x_2, x_3 – аналогичные показатели по углеводородам (без летучих органических соединений) и летучие органические соединения (ЛОС) соответственно.

Статистические данные, соответствующие числовым значениям переменных y_1, x, x_1, x_2, x_3, были представлены графически (рис.). Сравнение построенных графиков послужило основанием предположить, что имеет место линейная корреляционная зависимость между заболеваемостью детского населения БОД и выбросами загрязняющих веществ в атмосферу Кировской области. В дальнейшем сформулированная гипотеза обосновывается аналитически.

На втором этапе моделирования применение аппарата корреляционно-регрессионного анализа к моделированию зависимости заболеваемости детского населения Кировской области БОД обеспечило установление двух статистически значимых переменных x_2 и x_3. Линейные модели парной регрессии имеют вид: $\hat{y}_1 = 1074{,}62 + 25{,}44x_2$ и $\hat{y}_1 = 950{,}67 + 111{,}92x_3$. Характеристики данных уравнений статистически значимы на уровне значимости $\alpha = 0{,}001$.

Известно, что корректное использование построенных регрессионных моделей, в частности прогнозирование на их основе, осуществляется при условии установления их адекватности моделируемому объекту [8].

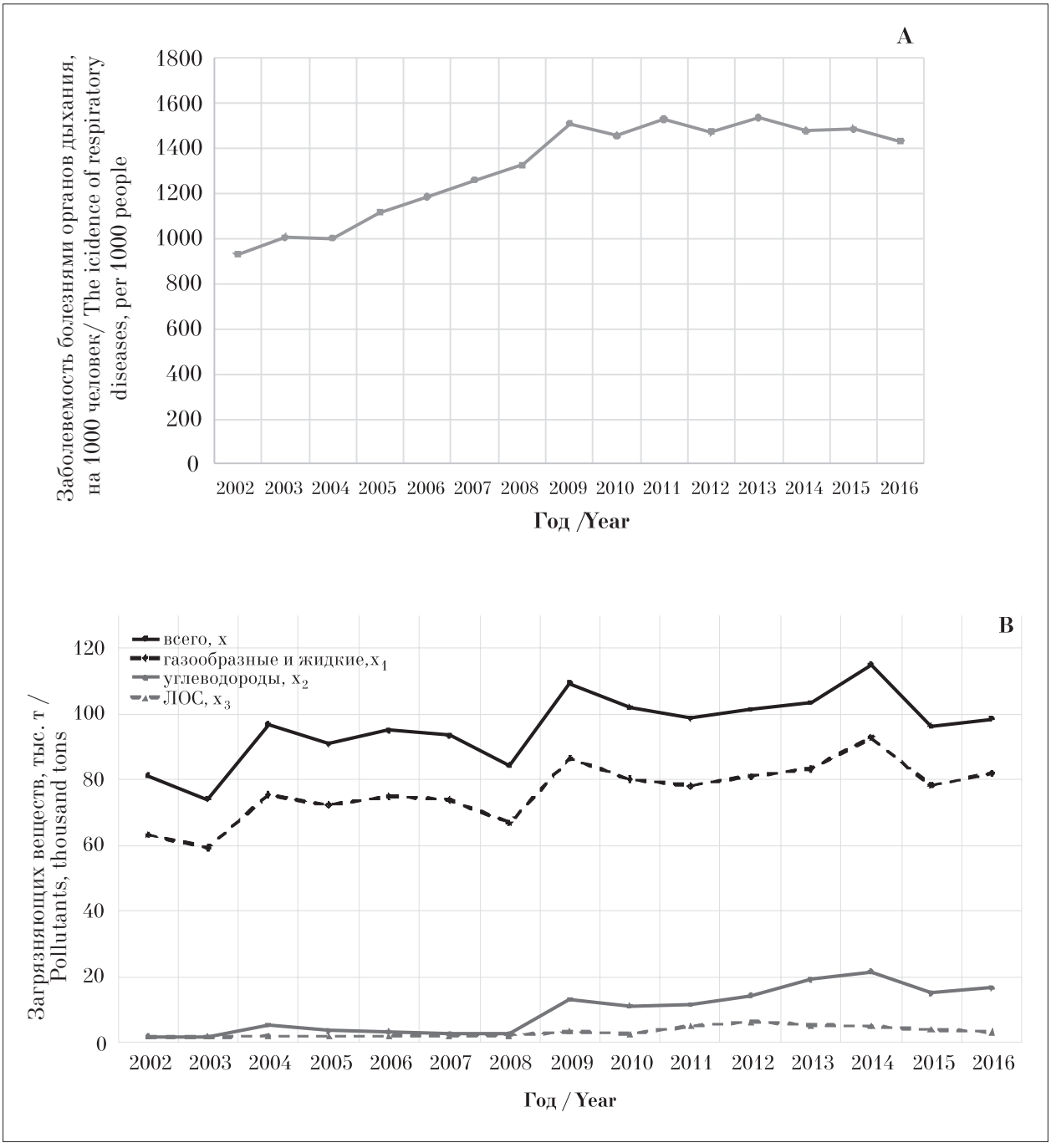

Рис. Динамика первичной заболеваемости БОД детского населения Кировской области (А) и количества выброшенных в атмосферу региона загрязняющих веществ от стационарных источников загрязнения (В) в 2002–2016 гг.

Fig. Dynamics of primary morbidity of children's population in the Kirov region (A) and the amount of pollutants emitted to the atmosphere from stationary sources of pollution (B) in 2002–2016

Дальнейшие рассуждения проведём с моделью $\hat{y}_1 = 1074{,}62 + 25{,}44x_2$. Её достоверность была обусловлена относительной ошибкой аппроксимации $\bar{A} = 8{,}53\%$, не превышающей 10%, коэффициентом детерминации $R^2 = 0{,}66$, объясняющим уравнением регрессии 66% дисперсии результативного признака, и соответствием случайных остатков нормальному закону распределения на основе коэффициентов асимметрии $A = -0{,}21$ и эксцесса $E = -1{,}02$, удовлетворяющих неравенствам $|A| < 1{,}5\sigma_A$ и $\left|E + \frac{6}{n+1}\right| < 1{,}5\sigma_E$,

где σ_A и σ_E – это среднее квадратическое отклонение асимметрии и эксцесса соответственно. Однако анализируемая модель имеет ряд недостатков, среди которых – не-

возможность учёта индивидуальных различий районов области по концентрации загрязняющих веществ и количеству источников загрязнения.

Как упоминалось выше, регрессионная модель с фиксированными эффектами обычно представляет наиболее значимый вариант моделирования для исследования регионов России или районов определённой области. В связи с тем, что указанная модель, содержащая одну независимую переменную и сорок фиктивных переменных (г. Киров и 39 районов Кировской области), является довольно громоздкой для исследования, перед анализом панельных данных целесообразно разбить рассматриваемые районы на группы со сходными значениями и уменьшить размерность конструируемой модели.

На третьем этапе моделирования на основании изучения статистики выбросов вредных веществ в атмосферу и количества стационарных источников загрязнения за 2008–2017 гг. был осуществлён кластерный анализ, в результате которого административные территории Кировской области были разбиты на четыре группы.

Кластер № 1 включает только г. Киров, поскольку все рассматриваемые показатели для областного центра превышают соответствующие показатели по районам области [7]; кластер № 2 – Верхнекамский, Зуевский, Кирово-Чепецкий, Кумёнский, Малмыжский, Нолинский, Омутнинский, Оричевский, Слободской и Советский районы; кластер № 3 – Афанасьевский, Белохолуницкий, Вятскополянский, Даровской, Верхошижемский, Лузский, Нагорский, Орловский, Уржумский, Фалёнский, Юрьянский и Яранский районы; остальные районы Кировской области были отнесены к кластеру № 4. Кластеры 2–4 упорядочены по убыванию указанных показателей и улучшению состояния атмосферного воздуха соответственно.

На четвёртом этапе моделирования с учётом реализованного кластерного анализа для построения моделей панельных данных были отобраны исследуемые показатели г. Кирова и трёх районов Кировской области по одному из оставшихся кластеров. Результаты регрессионного анализа по панельным данным следующие.

Объединённая модель имеет вид $\hat{y}_2 = 74565{,}17 + 2423{,}53x$, все параметры данной модели статистически значимы на уровне значимости $\alpha = 0{,}05$, $R^2 = 0{,}54$.

Уравнение $\hat{y}_2 = -188{,}06x + 157222f_1 + 71900{,}5f_2 + 39617{,}9f_3 + 117892f_4$ есть модель с фиксированными эффектами, в которой статистически значим на уровне значимости $\alpha = 0{,}05$ коэффициент при переменной x, остальные параметры статистически значимы на уровне значимости $\alpha = 0{,}001$, $R^2 = 0{,}92$.

Модель со случайными эффектами имеет вид: $\hat{y}^*_2 = 92012{,}8 + 361{,}04x$, где $y^*_{it} = y_{it} - \theta\, y_i$, $\hat{y}^*_{it}$ – преобразованное значение зависимой переменной для i-ой единицы совокупности в момент времени t ($i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$), y_{it} – аналогичное исходное значение функции, $\overline{y_i}$ – среднее значение по времени для каждого объекта наблюдения, $\theta = 0{,}88$ – параметр корректировки; для независимой переменной имеют место аналогичные соотношения. Угловой коэффициент статистически значим на уровне значимости $\alpha = 0{,}01$, свободный член – на уровне значимости $\alpha = 0{,}001$, $R^2 = 0{,}54$.

Оценка основных типов регрессионных моделей по панельным данным на адекватность с помощью тестов Фишера, Бреуша-Пагана и Хаусмана [8], показала, что модель с фиксированными эффектами наиболее достоверно описывает исходные данные.

Проведено аналогичное исследование, посвящённое изучению распространённости БОД, заболеваний органов пищеварения, кожи и подкожной клетчатки, мочеполовой системы, инфекционных и паразитарных заболеваний, костно-мышечной системы и соединительной ткани, глаза и его придаточного аппарата взрослого населения Кировской области от количества выброшенных в атмосферу загрязняющих веществ. Статистически значимые связи не выявлены.

Данный вывод согласуется с результатами, полученными другими исследователями, указывающими на социально-значимый характер заболеваемости органов дыхания взрослого населения Ханты-Мансийского автономного округа [2], Приморского края [3] и др.

Заключение

В результате исследования были установлены статистически значимые связи между количеством выбросов загрязняющих веществ в атмосферу Кировской области и показателем заболеваемости детского населения региона БОД, в частности, острыми инфекциями верхних дыхательных путей. Отличительной чертой данного исследования явилось построение модели с фиксированными эффектами, позволяющей учесть пространственную и временную структуру анализируемых стати-

стических данных, что нельзя сделать в рамках стандартных регрессионных моделей.

Литература

1. Семёнова Н.П. Состояние атмосферного воздуха и заболеваемость населения республики Саха (Якутия) // Экология человека. 2013. № 12. С. 14–19.

2. Аристархов А.Б., Козлова И.И., Кашапов Н.Г., Миняйло Л.А., Галиев А.Г. Использование методологии оценки риска при ведении социально-гигиенического мониторинга по атмосферному воздуху и связь здоровья населения с загрязнением атмосферы в г. Нижневартовске // Гигиена и санитария. 2015. № 94 (2). С. 10–12.

3. Кику П.Ф., Гельцер Б.И., Ярыгина М.В., Бениова С.Н., Горборукова Т.В., Морева В.Г., Шитер Н.С., Сабирова К.М., Мезенцева М.А. Эколого-гигиенические аспекты распространённости болезней органов дыхания у подростков и детей Приморского края // Гигиена и санитария. 2016. № 95 (8). С. 749–753.

4. Кочурова Л.В., Елисеев В.А. Множественность заболеваний у детей, проживающих в экологически неблагополучных регионах Сибири // Экология человека. 2011. № 11. С. 19–24.

5. Новиков С.М., Шашина Т.А., Додина Н.С., Кислицин В.А., Воробьёва Л.М., Горяев Д.В., Тихонова И.В., Куркатов С.В. Сравнительная оценка канцерогенных рисков здоровью населения при многосредовом воздействии химических веществ // Гигиена и санитария. 2015. № 94 (2). С. 88–92.

6. Тафеева Е.А., Иванов А.В., Титова А.А., Ахметзянова И.Ф. Мониторинг загрязнения атмосферного воздуха как фактор риска здоровью населения Казани // Гигиена и санитария. 2015. № 94 (3). С. 37–40.

7. О состоянии санитарно-эпидемиологического благополучия населения в Кировской области в 2017 году: государственный доклад [Электронный ресурс] http://www.43.rospotrebnadzor.ru/documents/ gosregdoklad/publications/gosudarstvennyy-doklad-2017.pdf (Дата обращения: 01.05.2018).

8. Эконометрика: учебник для бакалавриата и магистратуры / Под ред. И.И. Елисеевой. М.: Издательство Юрайт, 2015. 449 с.

9. Mátyás L. The econometrics of panel data. Fundamentals and recent developments in theory and practice. Berlin: Springer, 2008. 954 p.

10. Stock J., Watson M. Introduction to econometrics. Pearson, Addison Wesley, 2010. 827 p.

11. Wooldridge J.M. Econometric analysis of cross section and panel data. Boston: The MIT Press, 2010. 1096 p.

12. Буркин М.М., Молчанова Е.В., Кручек М.М. Интегральная оценка влияния социально-экономических и экологических факторов на региональные демографические процессы // Экология человека. 2016. № 6. С. 39–46.

13. Холматова К.К., Гржибовский А.М. Панельные исследования и исследования тренда в медицине и общественном здравоохранении // Экология человека. 2016. № 10. С. 57–64.

References

1. Semenova N.P. Atmospheric air state and morbidity among population in republic of Sakha (Yakutia) // Ekologiya cheloveka. 2013. No. 12. P. 14–19 (in Russian).

2. Aristarkhov A.B., Kozlova I.I., Kashapov N.G., Minyaylo L.A., Galiyev A.G. The use of risk assessment methodology in the management of social-hygienic monitoring for ambient air and the relationship of population's health state with the air pollution in Nizhnevartovsk // Gigiyena i sanitariya. 2015. No. 94 (2). P. 10–12 (in Russian).

3. Kiku P.F., Gel'tser B.I., Yarygina M.V., Beniova S.N., Gorborukova T.V., Moreva V.G., Shiter N.S., Sabirova K.M., Mezentseva M.A. Ecological-hygienic aspects of the prevalence of respiratory diseases in adolescents and children of the Primorsky Krai // Gigiyena i sanitariya. 2016. No. 95 (8). P. 749–753 (in Russian). doi: 10.18821/0016-9900-2016-95-8-749-753

4. Kochurova L.V., Yeliseyev V.A. Disease multiplicity in children – indicator of environmentally neglected regions of Siberia // Ekologiya cheloveka. 2011. No. 11. P. 19–24 (in Russian).

5. Novikov S.M., Shashina T.A., Dodina N.S., Kislitsin V.A., Vorob'yeva L.M., Goryayev D.V., Tikhonova I.V., Kurkatov S.V. Comparative assessment of the multimedia cancer health risks caused by contamination of the Krasnoyarsk Krai regions' environment // Gigiyena i sanitariya. 2015. No. 94 (2). P. 88–92 (in Russian).

6. Tafeyeva Ye.A., Ivanov A.V., Titova A.A., Akhmetzyanova I.F. Air pollutions as a risk factor for the population health in Kazan city // Gigiyena i sanitariya. 2015. No. 94 (3). P. 37–40 (in Russian).

7. On the state of sanitary and epidemiological welfare of the population in the Kirov region in 2017: state report [Internet resource] http://www.43.rospotrebnadzor.ru/documents/ gosregdoklad/publications/gosudarstvennyy-doklad-2017.pdf (Accessed: 01.05.2018) (in Russian).

8. Econometrics: a textbook for undergraduate and graduate studies / Ed. I.I. Yeliseyeva. Moskva: Izdatel'stvo Yurayt, 2015. 449 p. (in Russian).

9. Mátyás L. The Econometrics of panel data. Fundamentals and recent developments in theory and practice. Berlin: Springer, 2008. 954 p.

10. Stock J., Watson M. Introduction to econometrics. Pearson, Addison Wesley, 2010. 827 p.

11. Wooldridge J.M. Econometric analysis of cross section and panel data. Boston: The MIT Press, 2010. 1096 p.

12. Burkin M.M., Molchanova Ye.V., Kruchek M.M. Integral criterion of the influence of social, economic and environmental factors on the regional demographic processes // Ekologiya cheloveka. 2016. No. 6. P. 39–46 (in Russian).

13. Kholmatova K.K., Grzhibovskiy A.M. Panel and trend studies in medicine and public health // Ekologiya cheloveka. 2016. No. 10. P. 57–64 (in Russian).