УДК 51-76 DOI: 10.23885/2500-395X EDN LNFFYE

COMBINATORIAL AGGREGATIONS AND ARIMA MODELLING ANALYSIS IN ENVIRONMENTAL SPATIO-TEMPORAL EPIDEMIOLOGY

***S.V. Venevsky*[1], *S.K. Pinaev*[2, 3], *L. Tian*[4], *P. Gong*[4], *O.P. Gretsova*[5]**

[1] Federal Research Centre the Southern Scientific Centre of the Russian Academy of Sciences, Rostov-on-Don
[2] Khabarovsk Federal Research Center of the Russian Academy of Sciences, Khabarovsk
[3] Far Eastern State Medical University, Khabarovsk
[4] The University of Hong Kong, Hong Kong
[5] National Medical Research Radiological Centre of the Ministry of Health of the Russian Federation, Moscow
sergvene@gmail.com, pinaev@mail.ru, linweit@hku.hk, penggong@hku.hk

Absctract. Finding spatio-temporal disease clusters and pointing to their environmental drivers, even with correlation sense, require sometimes sophisticated statistical and GIS methods manifested in a new branch of EE, namely spatio-temporal epidemiology. Two key problems of spatio-temporal epidemiology is (1) selection of scale of analysis for aggregated data and (2) choosing a method for spatio-temporal pattern identification. We advocate using of linear correlation analysis and complete combinatorial analysis to determine possible spatial correlations between environmental drivers and disease characteristics for solution of problem 1. We advocate using of ARIMA models for identification of spatio-temporal patterns of dependencies disease characteristics by environmental driver (i.e. sets of spatial units at a fixed time interval with significant statistically correlations) for solution of problem 2.

So, identification of spatio-temporal patterns for statistical associations between number of fires (NF) as environmental driver and cancer incidence (CI) as disease characteristics using general approach and selection of best scale of analysis for aggregated data were done for seven administrative units (AUs) of Far Eastern Federal District (FEFD) of Russia for the period years 1992–2019. It was found that one fifth of AU-CI combinations has correlation between NF and CI. Three blood cancers (leukemia, non-Hodgkin lymphoma, and Hodgkin lymphoma) had the strongest statistical associations with wildfire. We also have seen two best scales of analysis of aggregated data; the first scale depicts typical type of forest (20 % of FFED); the second scale depicts typical climate (75 % of FFED).

Keywords: Spatio-temporal ecological epidemiology, combinatorial linear correlation analysis, ARIMA models, Russian Far East, fires, cancer.

1. Spatio-temporal epidemiology problems. Environmental epidemiology (EE) is a discipline which historically focuses particularly at disease clusters [1]. Traditional questions "Who is ill, when and where" in EE are oriented to finding a spatio-temporal clusters for defined population groups and some disease (type of diseases). The three questions are combined with a question "Why?" or "What is environmental (natural or man-made) driver for the presence of certain spatio-temporal clusters?" It is anticipated that finding a biologically explainable causal relationship between environmental driver and diseases can be quite unreachable. However, even visible correlation relationship between environmental driver and disease can be an important signal for public health and environmental managers, according to precautionary principle in EE (when some driver is strongly suspected to cause harm, one should not wait until a proof arrives, but rather take an action).

Finding spatio-temporal disease clusters and pointing to their environmental drivers, even with correlation sense, require sometimes sophisticated statistical and GIS methods [2] manifested in a new branch of EE, namely spatio-temporal epidemiology [3]. Two key problems of spatio-temporal epidemiology is (1) selection of scale of analysis for aggregated data and (2) choosing a method for spatio-temporal pattern identification.

2. General problems in understanding dynamics of environmental driver and response disease in EE. Spatio-temporal epidemiology uses localized registries of disease incidence and/or disease mortality for a wide spectrum of population groups (e.g. age cohorts, age classes, gender, ethnicity etc.). These localized time series of disease characteristics, obtained for temporal analysis may be long or short, and may have gaps. Very often the time series have seasonality, or some other cyclic features and trends,

although the trends cannot be simply explained by biological, physical or any other natural reasons. Thus, time series of disease incidence and/or disease mortality hardly can be represented by stationary stochastic approximations with constant moments (e.g. mean and standard deviation) over time. On the other hand, disease characteristics should have some autoregressive features (i.e. their recent values would depend on previous historic values because of some natural reasons, like genetic, demographic or (sometimes) registration methods). Accounting for historical dependence of present values in the time series from the past, we should assume that their residual stochastic component should also be related to previous values of residuals, but perhaps differently weighted to describe some events which may influence their values (e.g. changes in registration system, or other). Potential environmental drivers of diseases, either climatic or anthropogenic many of which can be reasonably simulated, also have a stochastic component. Interpreted as stochastic time series, values of environmental drivers have often autoregressive behavior due to their physical and/or chemical nature and demonstrate non-stationarity related to change in trends or seasonality or other reasons. Meanwhile, time series of fixed diseases characteristics for different spatial units may be correlated or not, depending on size of the units. If spatial units are large and not-connected by demographic processes (like migration), it can be assumed that their time series, representing disease characteristics, are independent. Environmental drivers for large spatial units may be also independent, if the units are not connected by some joint ecological or physical processes. Moreover, possible statistical relations between environmental drivers and disease characteristics may be also seen only at some certain spatial and temporal scales.

3. Mathematical formulation of selection of best scale of analysis for aggregated data in EE. Let us have $Y_{jik}(t)$ registered disease characteristics for the time period $t = [t_{year1}, t_{year2}]$, where j – is an index of spatial geographical unit $j = [1 : l]$, i – is an index of population group $i = [1 : m]$, k – is an index of an analyzed disease $k = [1 : n]$. We also have a suspect environmental driver $X_j(t)$ in each of spatial unit for the same period.

How many spatial units should be taken for an analysis to maximize strength of statistical relation between $X_j(t)$ and $Y_{jik}(t)$ and what are these units? Thus we need to find

$$A_{best} = \sum\nolimits_{J_g = j_s, j_u \dots j_z}^{w} A(j_g),$$

where A_{best} is an area representing best scale, $A(g)$ is an area of spatial unit with an index g; s, $u \dots z$ are indexes of spatial units, which maximize a value of functional reflecting strength of statistical relations

$$F_{J_g} =_{j_s, j_u \dots j_z} (X_{J_g}(t);\ Y_{J_g ik}(t))$$

for all $i = [1 : m]$ and $k = [1 : n]$ over entire analysis interval $t = [t_{year1}, t_{year2}]$.

4. Mathematical formulation for spatio-temporal pattern identification in EE. The generalized task can be formulated as following: to find all statistical functions (or models)

$$M_{J_g = j_s, j_u \dots j_z; i_h = i_a, j_b \dots i_c; k_f = k_d; k_e \dots k_f} (X_{J_g}(t);$$
$$Y_{J_g\, i_h k}(t)),$$

describing statistical dependence of diseases with indexes $d, e \dots f$ for population group with indexes a, $b \dots c$ in spatial units with indexes $s, u \dots z$ from suspected driver with a confidence interval α, preliminary set. As a matter of fact, this task is usually done for each of disease i and for each of population group k (i.e. for one type of disease and for one population group in sets of spatial units). Ideally, fulfilling of mentioned generalized task will have intersection with task of finding of the best spatial scale.

5. Suggested method for selection of best scale of analysis for aggregated data in EE. Significant body of analysis methods are designed in environmental epidemiology to determine disease clusters and their impacting environmental drivers. Detection of Spatial Autocorrelation (SA) for neighboring areas, using of Moran I spatial statistics or Getis-Ord Gi statistics, spatial analysis learning machines etc. (see for example review of [4]). Such GIS based correlation studies can be complex and difficult to explain for public health and environmental managers. We advocate using of linear correlation analysis to determine possible spatial correlations between environmental drivers and disease characteristics. These correlations do not explain casual relationships, but rather give an information on strength and directions of relationships between environmental drivers and disease characteristics. In this approach the functional $F_{lg} =_{js,\ ju \dots jz} (X_{lg}(t);\ Y_{lgik}(t))$, will be a maximum coefficient of correlation

$$\max_{J_g, i_h, k_e} R_{J_g = j_s, j_u \ldots j_z; i_h = i_a, j_b \ldots i_c; k_f = k_d; k_e \ldots k_f} (\bar{X}_{J_q}(t); \bar{Y}_{J_g i_h k_e}(t)),$$

between environmental driver $\bar{X}_{J_g}(t)$ for a set of spatial units J_g averaged over the time period [t_{year1}, t_{year2}] and disease characteristic $\bar{Y}_{J_g i_h k_e}(t)$ averaged over the time period [t_{year1}, t_{year2}] for a set of spatial units J_g for set of population groups i_h and for set of diseases k_f. When analyzing a set of population groups (i.e. all population in groups with numbers i_a, i_b...i_c), or/and a set of diseases (i.e. all population in groups with numbers k_d, k_e... k_f) an average correlation for the sets should be estimated. This can be done by simple averaging, where average spatial correlation for the sets i_h and k_f, declared as the same sample, $R_{J_g;\, i_h;\, k_f}(\bar{X}_{J_g}(t); \bar{Y}_{J_g i_h k_f}(t))$ is calculated as $R_{J_g; i_h; k_e} = \sum_{i_a, i_b \ldots i_c} \sum_{k_d, k_e \ldots k_f} R_{J_g} \frac{1}{n(i_h) m(k_e)}$ – the sum of spatial correlations for each particular population group in the set i_h and each particular disease in the set k_f, divided by multiple of the total number of elements in the set of population groups $n(i_h)$ and in the set of diseases $m(k_e)$. It was shown, however, that such a simple averaging of correlations potentially may result in a bias of the estimate of averaged correlation coefficient. Using recommendation of [5] we suggest using of alternate method of averaging, when each R_{J_g} is first converted to Fishers' z: $Z_{J_g} = 0{,}5 \times ln \frac{1+R_{J_g}}{1-R_{J_g}}$, then all transformed R_{J_g} can be averaged and the result back-converted to $R_{J_g;\, i_h;\, k_e} = \frac{exp^{2Z_{J_g;\, i_h;\, k_e}} + 1}{exp^{2Z_{J_g;\, i_h;\, k_e}} - 1}$. It was demonstrated by [5] that z-transformation greatly reduces bias in averaging of correlations. The task of finding maximum spatial correlation between environmental driver and disease characteristics for all the spatial units would be equivalent to calculation of $R_{J_g;\, i_h;\, k_f}(\bar{X}_{J_g}(t); \bar{Y}_{J_g i_h k_f}(t))$ for each possible set J_g of combinations of spatial units with further identification of a set providing maximum spatial correlation. Number of calculations for fixed sets of population groups i_h and for set of diseases k_f will be equal to $C^l_3 + C^l_4 + \ldots C^l_{(l-1)} + C^l_1$, where C^n_m is a number of combinations of m – spatial units from l – total spatial units.

6. Suggested method for spatio-temporal pattern identification in EE. Spatio-temporal pattern analysis in EE is always based on time series analysis and forecasting. Time series analysis appeared almost hundred years ago as a branch of practical statistics, but developed particularly wide in the last thirty years [6]. Firstly, classical statistical models were developed. Subsequently, exponential smoothing techniques were suggested for refining time series analysis. Now days auto-regressive moving average models (ARIMA) are mostly applied. Some time series analysis approaches incorporate now Machine Learning.

We strongly advocate using of ARIMA models with combinatorial aggregations for identification of spatio-temporal patterns in EE. ARIMA models (h, d, q) for time series λ are combinations of a difference autoregressive model with a moving average model [6], which are expressed as:

$$\Delta^d \gamma(t) = \alpha_0 + \alpha_1 \Delta^d \gamma(t-1) + \alpha_2 \Delta^d \gamma(t-2) + \ldots + \alpha_h \Delta^d \gamma(t-h) + \in(t) + \theta_1 \in(t-1) + \theta_2 \in(t-2) + \ldots + \theta_g \in(t-q),$$

where $\Delta^d\gamma(t)$ is analyzed variable in time moment t, differenced d times, $\in(l)$ are normally distributed residuals in time moment l, $\alpha_1 \ldots \alpha_h$ are the coefficients of the autoregressive (AR) part of the model, $\theta_1 \ldots \theta_q$ are the coefficients of the moving average (MA) part, and α_0 is a constant. In an ARIMA (h, d, q) model the predictors (d differenced) are lagged h data points for the autoregressive part and q residuals are considered for the moving average part. When the task is to make an identification of spatio-temporal patterns reflecting relations of disease characteristics with a certain environmental driver one need to minimize influence of different regional co-factors of diseases by statistical analysis. We suggest to make a normalization to a maximum value in each of spatial units both for environmental driver $\check{X}_{J_g}(t) = \frac{X_{J_g}(t)}{\max X_{J_g}(t)}$ and for disease characteristic $\check{Y}_{J_g ik}(t) = \frac{Y_{J_g ik}(t)}{\max Y_{J_g ik}(t)}$. Such a normalization allows comparison of amplitudes, trends, number of cycles and autoregressive features of time series within spatial units or within sets of several spatial units. Identification of spatio-temporal patterns in a set of several spatial units assumes that one should analyze time series for environmental driver and disease characteristics, representative for the set. These time series obtained by unification of time series for spatial units from the set, which can be calculated in two ways:

1) by simple averaging of time series for each year:

$$g_{set}(t) = \frac{\sum_i^{n_{set}} g_i(t)}{n_{set}},$$

where g_{set} is the time series, representing either environmental driver or disease characteristic for the entire set, n_{set} is the number of spatial units in the set, g_i is the time series in the spatial unit i;

2) by averaging with weighting to population number in each spatial unit:

$$g_{set}(t) = \frac{\sum_i^{n_{set}} p_i(t) \times g_i(t)}{\sum_i^{n_{set}} p_i(t)},$$

where $p(t)_i$ is a population number and g_i is either disease characteristic in spatial unit i, or environmental driver (if physical sense allows such averaging for the driver).

Kruskal – Wallis test with an α = 0,05 [7] should be applied before unification of time series from different spatial units in order to see if samples presenting environmental driver and/or disease characteristics from one spatial unit dominate similar samples from another spatial unit (in this case samples are dependent and one from the tested spatial units should be taken out from further spatio-temporal analysis). The task of spatio-temporal pattern identification in EE, after defining of a rule for calculating of representative time series for environmental driver and disease characteristics for a set of aggregated spatial units, can be formulated as following: to find all statistically significant (usually with confidence interval α = 0,05 or α = 0,1) correlations for all spatial units and all possible aggregations of the spatial units between environmental driver (normalized) and disease characteristics (normalized) and identify spatial units in aggregations, which provide statistical associations, i.e. find all sets J_g with spatial units J_s, J_u ... J_z, for which absolute value of correlation coefficient for a set of population groups (i.e. all population in groups with numbers i_a, i_b ... i_c), or/and a set of diseases (i.e. all population in groups with numbers k_d, k_e ... k_f) is larger than a fixed threshold r_{min} (usually > 0,5) with certain confidence α (0,05 or 0,1):

$$abs[r_{J_g = j_s, j_u \ldots j_z;\ i_h = i_a,\ i_b \ldots i_c;\ k_e = k_d,\ k_e \ldots k_f}(\check{X}_{J_g}(t); \check{Y}_{J_g i_h k_e}(t))] > r_{min}$$

with p-value < α. We suggest that one conduct the correlation analysis for all combinatorial aggregations of spatial units, which total number is equal to $C^l_1 + C^l_2 + \ldots C^l_{(l-1)} + C^l_l$, where C^n_m is a number of combinations of m – spatial units from l – total spatial units and with two ways of unification of the time series within the spatial units. We advocate using of ARIMA models for identification of spatio-temporal patterns of dependencies disease characteristics by environmental driver (i.e. sets of spatial units at a fixed time interval with significant statistically correlations). In this approach environmental driver in a set of spatial units is firstly approximated by ARIMA model, possibly with backwards time lags, depicting memory in diseases development. Disease characteristics in the set are also approximated by ARIMA and afterwards the predictor and the response variable are cross-correlated. The Ljung-Box Q test [6] should be applied to the residuals in ARIMA models to ensure that the residuals series are white noise, which indicates the goodness of the resulting fit. Holm – Bonferroni method [8] with correcting (increasing) of p-value should be applied to control family-wise error of I type (accepting of false positive hypothesis on existing of significant correlation) during multiple calculations of ARIMA models. The final sets with found cross-correlations between ARIMA described environmental driver and disease characteristics are ranked by the value of cross-correlation coefficients in order to see spatio-temporal patterns with strongest relations between the predictor and response function.

7. Case study: Cancer incidence and number of fires in Far Eastern Federal District of Russia – spatio-temporal pattern identification. Identification of spatio-temporal patterns for statistical associations between number of fires (NF) as environmental driver and cancer incidence (CI) as disease characteristics using general approach, described above (see 5), was done for seven administrative units (AUs) of Far Eastern Federal District (FEFD) of Russia (see [9]. We started our study knowing that temporal dynamics of cancer incidence was quite successfully described by ARIMA models in different countries like USA, Brazil, Switzerland [10] and relationship between cancer and wildfires was found in Canada [11]. In our study (see [9]) data on CI (persons with cancer for 100 000 persons) for five major cancer types in seven listed AUs in two age groups (children and teenagers 0–14 years and the entire population) for the 28-year period (1992–2019) were used. The algorithms of spatio-temporal patterns identification in our case study were realized in R language and using Excel (calculation files can be send by request) and both methods of unification of time series (simple averaging and averaging with weighting to population number) were applied. Approximately one fifth from all possible sets were found to have cross-correlations

between NF and CI. Three blood cancers (leukemia, non-Hodgkin lymphoma, and Hodgkin lymphoma) had the strongest statistical associations with wildfires [9].

8. Case study: Cancer incidence and number of fires in Far Eastern Federal District of Russia – selection of best scale of analysis for aggregated data. We made a spatial analysis of geographical coincidence between CI and NF for all sets of three to seven possible combinations of AUs of FEFD for the age population groups (children/teens 0–14 year and entire population) in order to find best scale of analysis of aggregated data. We found that distribution of maximum spatial correlation coefficient R^2 (averaged for all five cancer types) over number of AUs in a combination set has two peaks (see Fig. 1) for both age population groups (children/teens 0–14 year and entire population).

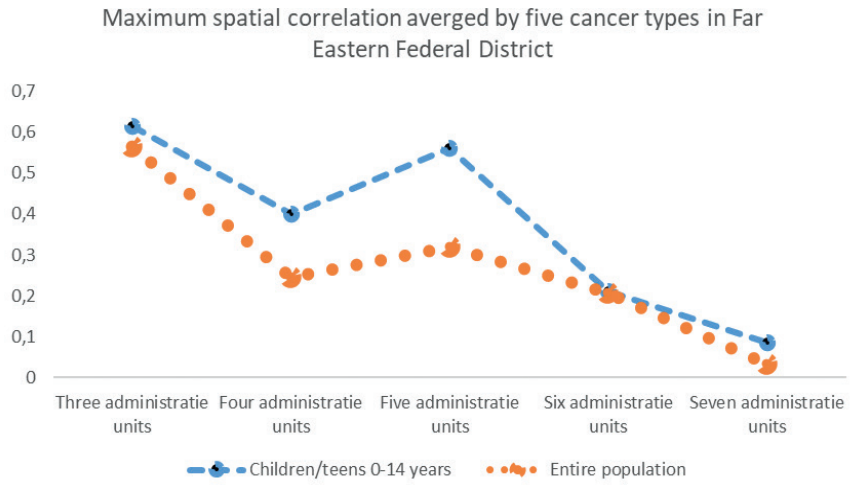
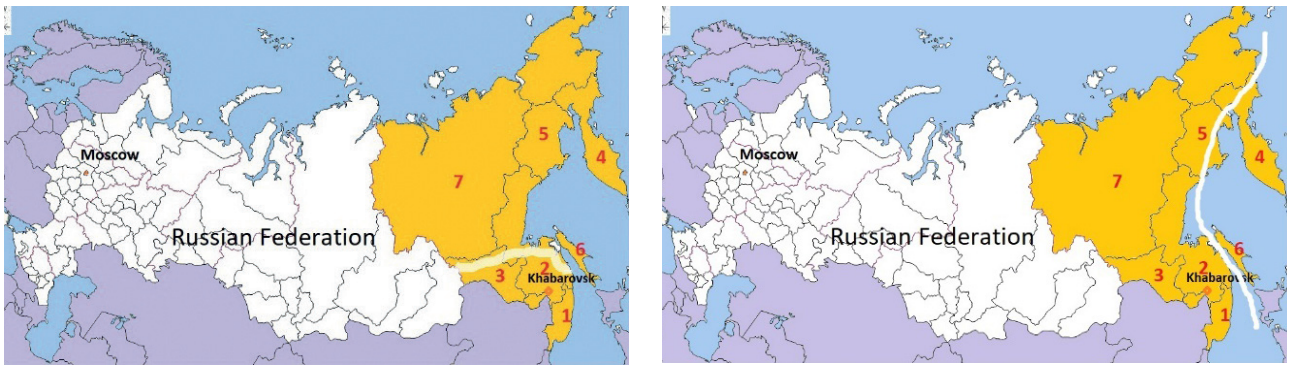


Fig. 1. Maximum spatial correlation coefficient R2 (averaged for all five cancer types) over number of administrative units

The first peak is observed for combination from the three AUs Primorskij Kraj – Khabarovskij Kraj – Amurskaja Oblast'. Thus, the first possible best scale of analysis is sum of areas for these administrative units (over 1 314 000 km^2, which is one fifth of the entire area). These three AUs have similar forest types (south taiga and temperate forests), which likely result in a similar smoke pollution (see Fig. 2a).



а *б*

Fig. 2. *а*) Three aggregated administrative units, providing best spatial correlation between cancer incidence and number of fires, are situated completely or partly to the south of border of southern taiga and temperate forest (white line) [12]; *b*) Five aggregated administrative units, providing second-best spatial correlation between cancer incidence and number of fires, are situated completely to the west of winter isobaric border 1016 millibar (white line), dividing territory with continental and marine climate [13]

Note. Far Eastern Federal District (FEFD) of Russian Federation in borders till the year 2018 (light grey) and administrative units used in the case study: *1* – Primorskij Kraj; *2* – Khabarovskij Kraj; *3* – Amurskaja Oblast; *4* – Kamchatskij Kraj; *5* – Magadan Oblast; *6* – Sakhalin Oblast. Khabarovsk is the largest city of FEFD (1,6 million inhabitants) and capital of FEFD till the year 2018 (Map of Russian Federation till the year 2018)

Adding to this set Respublika Sakha (central and northern taiga) decrease maximum R^2. Further addition of Magadan Oblast to the set Primorskij Kraj – Khabarovskij Kraj – Amurskaja Oblast' – Respublika Sakha, provides the second peak for maximum R^2 distribution at a five AUs set. Thus, the second possible best scale of analysis, which takes three fourths of the entire area (over 4 800 000 km^2). Geographic area which includes Primorskij Kraj – Khabarovskij Kraj – Amurskaja Oblast' – Respublika Sakha – Magadan Oblast' has continental climate with similar geostrophic pressure fields and prevailing winds in all seasons (see Fig. 2b), thus inducing homogeneity in distribution of smoke pollution. Addition to the area with homogeneous continental climate of AUs with maritime climate (Kamchatskij Kraj and/or Sakhalin oblast moves maximum R^2 to low values. Maximum R^2 for the age group "children/teens 0–14 years" was always smaller than maximum R2 for "entire population" at all sets of administrative unit combinations (from three to seven) (see Fig. 1).

9. Conclusion: What is learned from the case study for methodology? We saw that there were to best spatial scales for analysis of aggregated data. The first one took near 20 % of the entire area, but we found almost 40 % of spatio-temporal statistical associations between environmental driver and disease characteristics here (11 from the total 29 statistical associations found by ARIMA modelling), The second spatial scale constitutes 75 % of the entire area and got all the spatio-temporal statistical associations. Thus, we speculate that there is a trade-off in environmental epidemiology spatio-temporal analysis between finding of maximum number of spatio-temporal patterns at the best scale and finding all the patterns at the second-best scale.

Our findings in spatial correlations analysis need an EE process-oriented exposure modelling proof. Indeed, maximum spatial correlation between CI and NF in aggregated set "Primorskij Kraj – Khabarovskij Kraj – Amurskaja Oblast" may be explained as by carcinogenic features of smoke from burned southern taiga and temperate forests, so by largest population density in the entire Federal District here.

The study was supported from the Russian State Assignment of the Federal Research Centre of the Southern Scientific Centre of the Russian Academy of Sciences (SSC RAS) (122013100131-9).

References

1. Megan L. Environmental epidemiology. 2006: McGraw-Hill Education (UK).

2. Shaddick G., Zidek J.V., Schmidt A.M. 2023. *Spatio-Temporal Methods in Environmental Epidemiology with R.* CRC Press.

3. Meliker J.R., Sloan C.D. 2011. Spatio-temporal epidemiology: Principles and opportunities. *Spatial and spatio-temporal epidemiology.* 2(1): 1–9.

4. Mena C., et al. 2018. Spatial analysis for the epidemiological study of cardiovascular diseases: A systematic literature search. *Geospatial health.* 13(1).

5. Corey D.M., Dunlap W.P., Burke M.J. 1998. Averaging correlations: Expected values and bias in combined Pearson rs and Fisher's z-transformations. *The Journal of general psychology.* 125(3): 245–261.

6. Montgomery D.C., Jennings C.L., Kulahci M. 2015. *Introduction to time series analysis and forecasting.* John Wiley & Sons.

7. *Kruskal W.H., Wallis W.A.* 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association.* 47(260): 583–621.

8. Holm S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics.* P. 65–70.

9. Pinaev S., et al. 2023. Possible links of wildfires with oncological diseases of children and adults in the Russian Far East. *Bulletin of Russian State Medical University.* 4: 21–31.

10. Bouzon Nagem Assad D., et al. 2024. Comparing the current short-term cancer incidence prediction models in Brazil with state-of-the-art time-series models. *Scientific Reports.* 14(1): 4566.

11. Korsiak J., et al. 2022. Long-term exposure to wildfires and cancer incidence in Canada: a population-based observational cohort study. *The Lancet Planetary Health.* 6(5): e400-e409.

12. Chen D., et al. 2015. *Russian Boreal Forest Disturbance Maps Derived from Landsat Imagery, 1984–2000.* ORNL DAAC.

13. Tarasov A.V., Rakhmanov R.S. 2023. *Marine Climate of Russian Coastal Territories. Public Health Aspects of Biological Adaption.* Springer Nature Switzerland AG.

КОМБИНАТОРНЫЕ АГРЕГАЦИИ И МОДЕЛЬ АНАЛИЗА ARIMA В ЭКОЛОГИЧЕСКОЙ ПРОСТРАНСТВЕННО-ВРЕМЕННОЙ ЭПИДЕМИОЛОГИИ

***С.В. Веневский*[1], *С.К. Пинаев*[2, 3] *Л. Тьян*[4], *П. Гонг*[4], *О.П. Грецова*[5]**

[1] Федеральный исследовательский центр Южный научный центр Российской академии наук, Ростов-на-Дону
[2] Хабаровский федеральный научный центр РАН, Хабаровск
[3] Дальневосточный медицинский институт, Хабаровск
[4] Университет Гонконга, Гонконг, Китай
[5] Национальный медико-радиологический центр им. П.А. Герцена, Москва
sergvene@gmail.com, pinaev@mail.ru, linweit@hku.hk, penggong@hku.hk

Аннотация. Обнаружение пространственно-временных кластеров заболеваний и поиск их экологических причин, даже корреляционных, иногда требуют сложных статистических и ГИС-методов, что проявляется в появлении новой отрасли экологической эпидемиологии, а именно: пространственно-временной экологической эпидемиологии. Двумя ключевыми проблемами пространственно-временной эпидемиологии являются: 1) выбор масштаба анализа агрегированных данных; 2) выбор метода выявления пространственно-временных закономерностей.

Для определения возможных пространственных корреляций между всеми экологическими факторами и характеристиками заболеваний и с целью решения проблемы 1 рекомендовано использовать линейный корреляционный анализ и полный комбинаторный перебор. Для выявления пространственно-временных закономерностей (решения проблемы 2) целесообразно использовать модели ARIMA с комбинаторными агрегациями.

Выявлены пространственно-временные закономерности статистических связей (проблема 2) между количеством пожаров (NF) как экологическим фактором и заболеваемостью раком (CI) как характеристикой заболевания с использованием модели ARIMA и определен наилучший масштаб анализа агрегированных данных (проблема 1). Исследование проводилось для семи административных единиц (AE) Дальневосточного федерального округа (ДВФО) России за 28-летний период (1992–2019 гг.).

Выявлено, что примерно одна пятая из всех возможных наборов AE–CI имеет взаимную корреляцию между NF и CI. Три вида рака крови (лейкемия, неходжкинская лимфома и лимфома Ходжкина) имели самую сильную статистическую связь с лесными пожарами. Обнаружено, что существуют два наилучших пространственных масштаба для анализа агрегированных данных: первый масштаб отражает тип доминантного леса (20 % территории ДВФО), а второй – тип доминантного климата (75 % территории ДВФО).

Ключевые слова: пространственно-временная экологическая эпидемиология, комбинаторный линейный корреляционный анализ, модели ARIMA, Дальний Восток России, пожары, рак.

Работа выполнена в рамках ГЗ ЮНЦ РАН (122013100131-9).