

W artykule przedstawiono wyniki retrospektywnej analizy ekologicznej, dotyczącej zachorowalności na raka stercza w relacji do sytuacji etnicznej w województwie opolskim. W pracy zostało wykorzystanych kilka znanych metod statystycznych. Detekcję skupień zachorowań przeprowadzono metodą Tango [10]. Estymatory relatywnego ryzyka wyznaczono metodą empiryczną bayesowską Clayтона i Kaldora [2]. Do wyznaczenia współczynnika korelacji przestrzennej zastosowano statystykę Morana [5]. Modelowanie poissonowskie zmiennych przeprowadzono za pomocą techniki komputerowej wykorzystującej połączone własności łańcucha Markowa i symulacji Monte Carlo – znane pod nazwą próbkowania Gibbsa [8, 9]. Przedstawione tabelarycznie i graficznie wyniki wskazują na prawdopodobny związek pomiędzy rozkładem etnicznym i zachorowalnością na raka stercza w województwie opolskim.

Słowa kluczowe: rak stercza, grupa etniczna, skupisko zachorowań, korelacja przestrzenna, statystyka bayesowska.

Rak stercza w relacji do sytuacji etnicznej w województwie opolskim

Prostate Cancer Incidence in Relation to Ethnic Situation in Opole Province, Poland

Andrzej Tukiendorf

Technical University, Opole, Poland

INTRODUCTION

The striking variations in the risk of disease in different geographical locations, and in different ethnic groups in the same locality, have long been important in generating hypotheses about causation. In studying the etiology of a disease, one valuable approach is to compare morbidity rates in different areas and to see if any differences can be related to variation in some other measurable characteristics of those areas.

In recent years there has been a growing medical and public concern about prostate cancer incidence. Some extensive epidemiological debates on this topic appeared in many recognized world scientific journals. In the province of Opole, Poland, the morbidity of prostate cancer reaches 5% of total cancer incidence in males. The amount is similar to the level for the entire country [1]. However, the considerable differences in incidence have been observed throughout the region. This fact predisposes to investigate their possible demographic background due to the very specific ethnic situation that appeared in Opole province after World War II.

This paper deals with a retrospective ecological analysis of prostate cancer incidence in relation to ethnic situation in the province of Opole. The paper begins with historical facts, assumed groups' definitions, and information on materials. In the next section, the methodology of the research is presented concerning procedures of geographical data aggregation, detecting spatial disease clusters, computing empirical Bayes estimates of relative risks, variables' spatial correlation, and their fully Bayesian modeling. Finally, the results and the discussion are given.

HISTORICAL FACTS AND ETHNIC DEFINITIONS

After World War II, Poland became a very ethnically homogenous country. Under the terms of international agreements, all Germans were forced to leave the western territories newly regained by Poland. Simultaneously, the process of resettling Polish repatriates from the former eastern and central Poland began. Although the withdrawal of Germans was widespread, they did not leave their

This paper presents results of the retrospective ecological analysis of prostate cancer incidence in relation to ethnic situation in the province of Opole, Poland. A few well-known methodological procedures were applied in the study. Detecting spatial disease clusters followed Tango's (1995) [10] method. Estimates of relative risks of cancer morbidity were calculated via Clayton and Kaldor's (1987) [2] empirical Bayes approach. To measure spatial correlation between neoplastic incidence and ethnic density, Moran's (1950) [5] statistic was used. The Poisson modeling of the variables was performed using Markov Chain Monte Carlo technique - Gibbs sampling (Smith and Roberts 1993 [8], Spiegelhalter et al. 1996 [9]). The results presented in tables and graphs provide evidence of possible association between ethnic pattern and prostate cancer incidence in the analyzed province.

Key words: prostate cancer, ethnic group, disease cluster, spatial correlation, Bayesian statistics.

lands entirely. This fact led to a specific ethnic situation in the province of Opole. According to the demographic statistics, the estimated percentage of the autochthons within the province was up to 53% in the early 50s, 44% in the mid 60s, and 38% in the late 70s [2].

For the research purpose, after Balaryn [2] the following definitions of the autochthons and the repatriates were adopted:

- ▶ the autochthons are persons born within the former German Reich borders before it became a part of Poland in 1945 and inhabiting the present territory of Opole province after the World War II, and their descendants;
- ▶ the repatriates are all Polish immigrants mainly from the former Poland's eastern and central territories inhabiting the present territory of Opole province after the World War II, and their descendants.

Particularly, the autochthons are concentrated mainly in the eastern part of the region, while the repatriates resettled the surrounding western territories. Figure 1 clearly exposes the ethnic pattern in the region.

The map in Figure 1 indicates a low degree of integration of the ethnic groups across the region (a lack of geographic units with half and half shares of the populations). Certainly, this fact does not exclude the mixing (plus biological) of these two groups after the war. However, a large percentage of the population living in the region still identifies themselves with one or the other ethnic group. This argues for the remaining distinction between these populations.

It is to add that the information survey on the autochthons within the province of Opole was politically motivated. Former governments tried to monitor such ethnically mixed regions and the inhabitants were regularly inquired about their national identities.

MATERIALS

The data concerning all the registered cases of prostate cancer (185, ICD-9) in the years 1985-1994 originated from the Regional Cancer Registry (RCR) in Opole. During the study period 665 cases were confirmed. It is of note that since 01.01.1985 RCR in Opole has provided, following Finnish Cancer Registry in Helsinki, Finland, the active program of cancer registration. This form of activity with supplementary 'follow-back' and 'follow-up' verifications significantly improved the reliability of gathered information, making RCR in Opole a contributor of *International Agency for Research on Cancer and International Association of Cancer Registries* [3].

The population numbers in the age groups were taken from the National Census of 1988 that concerned 494,385 males.

The data regarding the ethnic situation in the region originated from the Silesian Institute in Opole [2] providing information of 178,045 male autochthons inhabiting the region (the remaining male population was assumed as the repatriates). To calculate the percentages of the autochthonic population in geographic units, the supplemental information from the National Census of 1978 was taken into account.

METHODS

In the analysis, a complete census list of 944 municipalities within the province was considered. Based upon geographical coordinates of their centroids, all the demographic and epidemiological information associated with these points was aggregated into 10 km x 10 km square areas - a new representation set. The aggregation of data had in purpose to avoid the size and shape irregularities of the administrative units only and to achieve the geographical information

homogeneity. However, the epidemiological analysis conducted afterwards would have been possible without this technique as well.

To detect spatial disease clustering and clusters' locations, Tango's method was adopted [4]. In its classical approach, a test index C and a test statistic T to measure general and focused clustering of rare diseases are defined as follows:

if O_i and P_i denote the observed numbers of cases and the individuals 'at risk' respectively in the i -th area, for $i = 1, \dots, I$, then

$$C = (r - p)'A(r - p)$$

and

$$T = \{C - E(C)\} / \{\text{var}(C)\}^{1/2} \sim \chi^2$$

with degrees of freedom adjusted by skewness (may not be integer), where

$$r' = (O_1, O_2, \dots, O_I) / O, \{O = O_1 + O_2 + \dots + O_I\}$$

denotes a vector of observed relative frequencies in I areas with parameter vector

$$p' = (p_1, p_2, \dots, p_I), \{p_i = P_i / P, P = P_1 + P_2 + \dots + P_I\}.$$

Furthermore,

$$A(\lambda) = a_{ij}(\lambda), \text{ for } i, j = 1, 2, \dots, I,$$

is an $I \times I$ matrix that measures closeness between areas with a constant scale parameter λ [4]. In the presented analysis the following non-linear exponential decay model was assumed:

$$a_{ij}(\lambda) = \exp\{-4(d_{ij}/\lambda)^2\}$$

where d_{ij} represents a distance between the i -th and j -th squares. To detect spatial disease clusters, a value of $\lambda = 20$ (km) was taken into account, which means that any pair of cases far apart beyond this distance cannot be considered as clusters. Particularly, the choice of λ is associated with the reasonable 'prior' knowledge about the mode of disease clusters. Large λ will give a test sensitive to a large cluster and small λ to a small cluster, however, a previous study of Tango [4] showed that the resulting C and T do not change drastically by choosing the constant λ . The candidate clusters were indicated by the percentage of the i -th area's contribution in the total C index as $>1\%$. In case of no clustering, that means that the observed numbers of cases in I regions are independent Poisson variables with the means proportional to the populations 'at risk', the percentage of the i -th area equals $1/103 \cdot 100\%$ in the total C index. Otherwise, different contributions of

areas to C are detected, and the most extreme values can be considered as locations of likely clusters [4]. The ratios of observed to expected cases in such locations that are smaller or larger the unity indicate an appropriate type of clusters of low or high incidence, respectively.

A parametric empirical Bayes approach was used to estimate *Standardized Morbidity Ratios* (SMR = observed/expected cases) of prostate cancer in the province. Following the basic method proposed by Clayton and Kaldor (1987) [5], the relative risks θ_i are following a gamma distribution with scale parameter α and shape parameter μ . Since conditional on θ_i , the observed cases of disease O_i are Poisson variates with expectation $\theta_i E_i$. Thus, the posterior expectation of θ_i conditional on O_i is

$$E(\theta_i | O_i, E_i; \alpha, \nu) = \frac{O_i + \nu}{E_i + \alpha}.$$

The estimation of α and ν can be accomplished via maximum likelihood. The log-likelihood based on the distribution of O_i and E_i is

$$L(\alpha, \nu) = \sum_i \left[\log \frac{\Gamma(O_i + \nu)}{\Gamma(\nu)} + \nu \log(\alpha) - (O_i + \nu) \log(E_i + \alpha) \right]$$

(Clayton and Kaldor 1987 [5]).

Numerically, the solution for the maximum likelihood estimates relies on finding the minimum of the negative log-likelihood.

A simple estimation of the expected number of cases - E_i was carried out conditionally on the set of observed cases - O_i and the reference populations 'at risk' (the conditional approach). Given a total number of cases for each age category (O_{+1}, \dots, O_{+K}), $O_{+k} = O_{1k} + \dots + O_{Ik}$, the conditional expected number of cases E_i is derived from the following equation:

$$E_i = \sum_{k=1}^K O_{+k} \left(\frac{P_{ik}}{\sum_{j=1}^I P_{jk}} \right), \quad (i = 1, 2, \dots, I; k = 1, 2, \dots, K).$$

Taking spatial interactions into account between the SMRs and the percentages of ethnic population, Moran's I statistic [6] is equal to

$$I = \frac{\frac{n}{W} \sum_{ij} w_{ij} z_i z_j}{\sum_{ij} z_i^2},$$

where

$$W = \sum_{ij} w_{ij},$$

and w_{ij} is the weight for the relationship between the observations $i, j = 1, 2, \dots, l$, while

$$z_i = SMR_i - \overline{SMR}$$

is the centered variate obtained from SMR_i [6]. In the presented study (similarly to Cressie [7], Kaluzny *et al.* [8])

$$w_{ij} = (\min\{d_{ij} : i = 1, 2, \dots, l\} / d_{ij}) (X_i X_j)^{1/2}$$

where d_{ij} denotes a distance between the i -th and the j -th squares, and X_i, X_j are the percentages of autochthonic population within areas i, j . In this research, two squares were considered to be neighbors if their centers are within 20 kilometers of each other (maximum distance between cases).

The fully Bayesian ecological analysis was based on the random-effects Poisson regression - the conditional autoregressive (CAR) model using Markov Chain Monte Carlo (MCMC) technique, known as Gibbs sampling [9], Smith and Roberts [10]. Following Spiegelhalter *et al.* [9], the model assumes that if O_i and E_i are the observed and the expected cancer cases in the i -th area respectively, then

$$O_i \sim \text{Poisson}(\mu_i)$$

$$\log \mu_i = \log E_i + \alpha X_i + b_i$$

$$b_i \sim \text{Normal}(\bar{b}_i, n_i)$$

$$\bar{b}_i = 1/n_i \sum_{j \in \text{neighbors}(i)} b_j$$

where α is given independent 'noninformative' prior, n_i is the number of neighbors of i and X_i is an ecological covariate (percentage of autochthons). In

the Bayesian approach, in addition to specifying the model for the observed data (usually in the form of a probability distribution - likelihood), a vector of unknown parameters (random quantity) is supposed as well, having its own prior distribution and hierarchical vector of hyperparameters. Inference concerning these unknown parameters is then based on its posterior distribution expressed in the convenient shorthand as proportional to the likelihood times the prior. The key thing is that in prior-to-posterior analysis, the prior always informs the posterior. The only question is how much? In this model, the authors [9] propose a noninformative prior for α that is insensitive to a choice of a very different prior yielding the same posterior. Moreover, the model builds the full joint distribution from complete conditional distributions for each b_i (given all others) that induce the spatial correlations [9].

The basic philosophy behind MCMC is an iterative simulation of Markov chain whose equilibrium distribution is the desired distribution. Instead of calculating exact estimates, this technique generates a stream of simulated values for each quantity of interest [10]. In the study, the computation based on a burn-in of 1,000 iterations and following 10,000 samples provided convincing evidence of converged distributions of the streams of values during the Gibbs sampling. Because an assessment of convergence in the MCMC technique is difficult, without an intuitive understanding of the shape of the posterior, one can never tell whether it has sampled from all the modes of the posterior. The key thing to convergence is: if one runs it longer, will the outcome change? If it has truly converged, then the answer is no. One good way to test convergence is using the multiple chains options in BUGS which technique was used in the presented study. Because the chains yielded the same answer, the convergence of the streams of output values was stated in the study.

The resulting SMRs achieved from the empirical and fully Bayesian approaches were exposed in color coded maps. Additionally, to show how the SMRs (Poisson regression) change with the X s - a linear fit of the variables was plotted.

The following licensed computer programs were used in the research: MapInfo 4.5, S-PLUS 4.0, S+SPATIALSTATS 1.0, and WinBUGS 1.3.

The presented methodological approaches are quite easy to use and can be successfully adopted in other types of ecological analyses [11]. For users who are interested in applying these methods, the program codes can be provided by the author upon request.

RESULTS

The adopted set of methods was used to assess the relation between prostate cancer (185, ICD-9) incidence and the ethnic situation in the province of Opole.

The estimates of Tango's test for detecting spatial disease clustering [4] are set in Table 1.

From the results in Table 1 it can be ascertained that standardizing distribution of the C index gives the high amount of the T statistic at a very low value of p . That supports the clustering evidence of the disease morbidity in the region. The accompanying geographical analysis of the clusters' locations is presented in Figure 2.

The map in Figure 2 perfectly demonstrates that the clusters of low incidence are concentrated in the 'autochthonic' areas, while the clusters of the high disease incidence are dispersed in the territories inhabited by the repatriates (compare with Figure 1).

The spatial distribution of empirical Bayes estimates of the relative risks [5] of prostate cancer is presented in Figure 3.

From the fairly mosaic pattern of the $SMRs$ presented in Figure 3 it is seen that the lowest cancer incidence is characteristic for the 'autochthonic' areas, while the highest cancer rates are observed in the territories populated by the repatriates (compare with Figure 1).

The Moran's spatial correlation [6] estimates of prostate cancer incidence with respect to ethnic densities are set in Table 2.

The results given in Table 2 testify that $SMRs$ are spatially dependent series and tend to have neighboring values similar to one another when referred to Xs . In case of chronic diseases, this correlation can be thought of as arising indirectly through the influence on the disease of the ecological covariate [12].

Such a spatially correlated process supports the modeling of the variables via spatial regression models. The posterior estimates of the α slope parameter (Poisson regression) for the spatial CAR model in the fully Bayesian approach [9] are given in Table 3.

The posterior estimates in Table 3 prove evidently the negative influence of the X ecological variable (% of autochthons) on the SMR level. The map of the spatially modeled $SMRs$ of prostate cancer incidence via autochthonic densities is presented in Figure 4.

Similarly, the map in Figure 4 clearly shows that the lowest cancer risk areas overlap the territories

populated by the autochthons, while the highest cancer risk areas are represented by the repatriates (compare with Figure 1). A supplemental picture of an impact of changes of the $SMRs$ in the Xs (% of autochthons) is given in Figure 5.

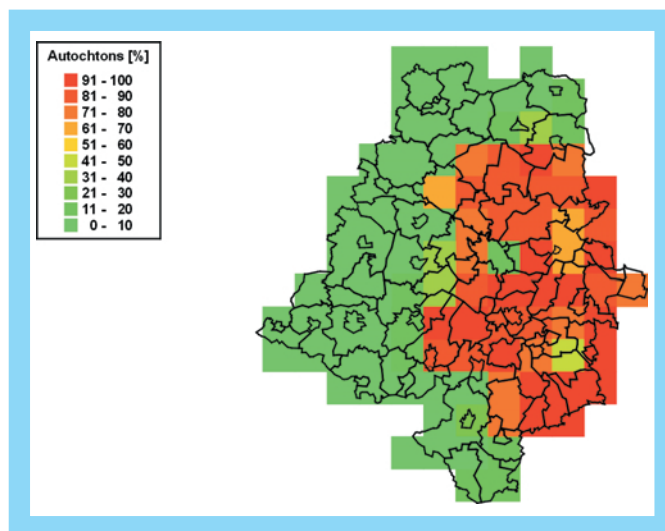


Fig. 1. Percentages of Autochthons (males) in Opole Province (1977)

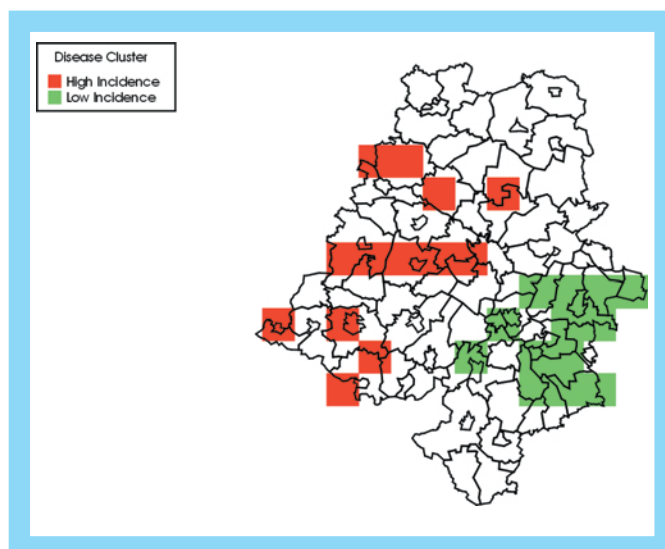


Fig. 2. Disease Clusters of Prostate Cancer (185, ICD-9) Incidence in Opole Province (1985-1994)

Tab. 1. Tango's Disease Clustering Test

C index	0.0047
T statistic	8.3649
p -value	3.6 e-6
d.f. (adjusted)	6.3855

Tab. 2. Moran's Spatial Correlation Estimates

spatial correlation	0.1956
standard error	0.0607
normal statistic	3.3850
normal p -value (2-sided)	7.1 e-4

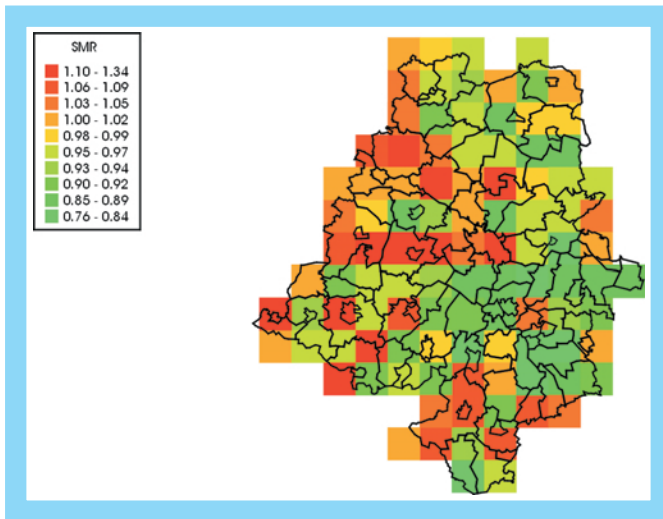


Fig. 3. SMRs of Prostate Cancer (185, ICD-9) in Opole Province (1985-1994) - empirical Bayes approach

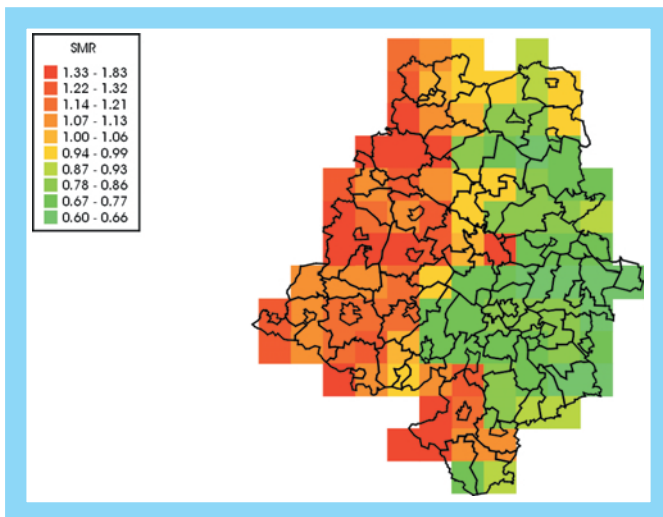


Fig. 4. SMRs of Prostate Cancer (185, ICD-9) in Opole Province (1985-1994) - fully Bayes approach

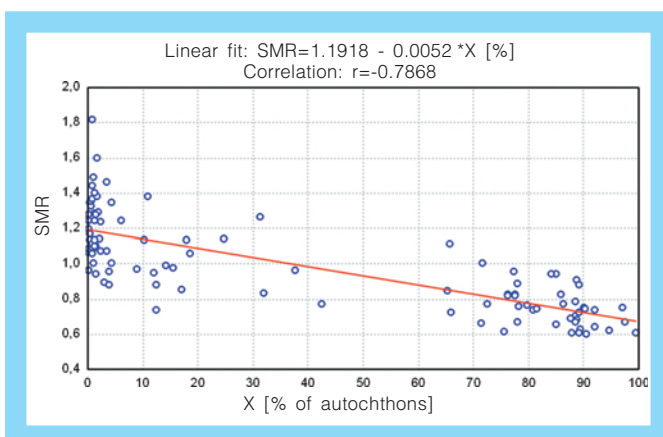


Fig. 5. SMRs of Prostate Cancer (185, ICD-9) vs. Xs (% of autochthons) - fully Bayes approach

Tab. 3. Posterior Means of α Slope Parameter - Poisson regression

mean	-5.261e-3
standard deviation	2.099e-3
2.5 % credible interval	-9.351e-3
97.5 % credible interval	-1.036e-3

This simple linear fit of the variables plotted in Figure 5 shows that the risk of prostate cancer disease in the repatriates is about twice higher than in the autochthons.

CONCLUSIONS

The presented retrospective ecological analysis concerned the relationship between cancer incidence of prostate cancer (185, ICD-9) and the ethnic situation in the province of Opole. Despite the deficiency of prior knowledge whether the population differences might have a real influence on the disease, the obtained results in the presented methodological approaches testify an obvious similarity between the ethnic and the health patterns in the region. In the aspect of the prostate neoplastic incidence among the two studied groups, the repatriates suffer from cancer about twice more intensively than the neighboring autochthons. This fact may lead to a certain speculation over the etiology of the disease that is related to differences in, e.g., individual habits, occupational factors, diet or alcohol/tobacco consumption, etc., among these populations. It is of note that in the most common local opinion, the autochthons take care of their widely understood hygiene much more than the repatriates. Certainly, at this stage of research nothing can be ascertained but the epidemiological problem can be signaled only. However, in the Author's opinion, these results shall indicate the proper direction of some more expanded scientific investigations, such as case-control or longitudinal studies (if carried out in the future), to determine the reason of difference in prostate risk cancer in relation to ethnical pattern in the region.

WSTĘP

Wyraźne zróżnicowanie geograficzne i etniczne zachorowań, od dawna leżało u podstaw poszukiwania przyczyn chorób. W etiologii chorób, jedną z ważnych technik poznawczych jest porównanie współczynników zachorowalności w różnych regionach i sprawdzenie, czy jakiegokolwiek zmiany w ich poziomach mogą być związane ze zmiennością innych mierzalnych charakterystyk tych obszarów.

W ostatnich latach obserwuje się wzrost zainteresowania medycznego i publicznego zachorowaniami na raka prostaty. Przeprowadzono szereg debat naukowych w wielu uznanych światowych czasopiśmie naukowych. W województwie opolskim zachorowalność na nowotwory prostaty sięga 5 proc. wszystkich przypadków zachorowań na raka u mężczyzn i wartość ta jest podobna do poziomu krajowego [1]. Zauważa się jednak wyraźne zróżnicowanie wewnątrz regionu. Fakt ten predysponuje do po-

szukiwania tła demograficznego, w związku z bardzo specyficzną sytuacją etniczną, która wystąpiła na Opolszczyźnie po II wojnie światowej.

W niniejszym artykule zanalizowano zachorowalność na raka prostaty w województwie w aspekcie etycznym. Na początku podano przyjęte definicje grup etnicznych oraz informacje o materiałach. Następnie przedstawiono metodykę badań dotyczącą procedur geograficznej agregacji danych, detekcji przestrzennych skupień zachorowań, wyznaczania empirycznych estymatorów bayesowskich, przestrzennej korelacji zmiennych i ich pełnego modelowania bayesowskiego. Na końcu podano wyniki oraz ich dyskusję.

FAKTY HISTORYCZNE I DEFINICJE GRUP ETNICZNYCH

Po II wojnie światowej Polska stała się krajem o wysoce jednorodnej strukturze etnicznej. Na podstawie konwencji międzynarodowych wszyscy Niemcy mieli obowiązek opuszczenia Polski w nowych warunkach terytorialnych kraju. Równocześnie rozpoczął się proces zasiedlania opuszczonych terenów polskimi repatriantami z byłej wschodniej i centralnej Polski. Chociaż uchodźstwo Niemców było powszechne, nie opuścili jednak tych stron w zupełności. Ten fakt doprowadził do bardzo specyficznej sytuacji na Opolszczyźnie. Powołując się na statystykę demograficzną, szacowany procentowy udział autochtonów w województwie kształtował się na poziomie 53 proc. we wczesnych latach 50., 44 proc. w połowie 60. i 38 proc. w późnych 70. [2].

Do celów badawczych, za [2], zaadaptowano następujące definicje autochtonów i repatriantów:

- za autochtonów uznano osoby urodzone wewnątrz granic byłej Trzeciej Rzeszy, zanim ta stała się częścią Polski w 1945 r., i zamieszkujących obecne terytorium województwa opolskiego po II wojnie światowej oraz ich potomków;
- za repatriantów uznano wszystkich polskich imigrantów, głównie z byłej wschodniej i centralnej Polski, zamieszkujących obecne terytorium województwa opolskiego po II wojnie światowej oraz ich potomków [2].

Autochtoni skupieni byli głównie we wschodniej i środkowej części regionu, zaś repatrianci zasiedlili otaczające zachodnie rubieże. Sytuację etyczną w regionie przedstawia jasno ryc. 1. (fig. 1.).

Mapa na ryc. 1. (fig. 1.) wskazuje niski poziom zintegrowania się grup etnicznych na przestrzeni regionu (mało jest jednostek geograficznych o połowicznych udziałach populacji). Oczywiście, ten fakt nie wyklucza wymieszania (w tym biologicznego) tych dwóch grup po wojnie. Niemniej duży udział

procentowy populacji żyjącej w regionie wciąż utożsamia się z jedną bądź drugą grupą etniczną. To uzasadnia stałą odmienność tych populacji.

Należy dodać, że zbieranie informacji o autochtonach w województwie opolskim było motywowane politycznie. Poprzednie rządy starały się monitorować tego typu etnicznie wymieszane regiony, a ich mieszkańcy byli regularnie wypytywani o przynależność narodowościową.

MATERIAŁY

Dane dotyczące wszystkich zarejestrowanych przypadków nowotworów złośliwych prostaty (185, IX M. Kl. Ch.) w latach 1985–1994 pochodziły z Wojewódzkiego Rejestru Nowotworów (WRN) w Opolu. W tym czasie potwierdzonych zostało 665 przypadków raka. Należy podkreślić, że od 1.01.1985 r. WRN w Opolu prowadzi, po konsultacjach z Fińskim Rejestrem Nowotworów w Helsinkach, program aktywnej rejestracji nowotworów. Ta forma działalności, z uzupełniającymi retro- i prospektywnymi weryfikacjami, istotnie poprawiła wiarygodność zbieranych informacji, czyniąc WRN w Opolu współpracownikiem Międzynarodowej Agencji Badań nad Rakim oraz Międzynarodowego Stowarzyszenia Rejestrów Nowotworowych [3].

Liczyby ludności w grupach wiekowych zostały zaczerpnięte z Narodowego Spisu Powszechnego (1988) i dotyczyły 494,385 mężczyzn.

Dane na temat ludności autochtonicznej zamieszkującej region uzyskano w Instytucie Śląskim w Opolu i obejmowały w tej grupie etnicznej liczbę 178 045 mężczyzn [2] (za liczbę repatriantów w regionie uznano pozostałych mężczyzn w zbiorze). Obliczenia procentowego udziału mężczyzn autochtonicznych w jednostkach geograficznych oparto na danych z Narodowego Spisu Powszechnego (1978).

METODY

W analizie wzięto pod uwagę kompletną spisową listę 944 miejscowości, występujących w województwie opolskim. Na podstawie współrzędnych geograficznych ich środków, wszystkie informacje demograficzne i epidemiologiczne przypisane tym punktom zostały zagregowane do 10 km x 10 km² powierzchni – nowej reprezentacji zbioru danych. Agregacja miała na celu uniknięcie wielkościowych i kształtnych nieregularności jednostek administracyjnych oraz osiągnięcie geograficznej jednorodności danych. Jakkolwiek analiza epidemiologiczna byłaby możliwa do przeprowadzenia bez stosowania tej techniki.

Do detekcji zjawiska skupienia zachorowań i ich lokalizacji wykorzystano metodę Tango [4]. W jej klasycznym ujęciu, indeks testowy C i statystyka testo-

wa T do pomiaru ogólnego i specyficznego zjawiska skupienia rzadkich chorób są definiowane następująco:

jeśli O_i i P_i określają odpowiednio liczby obserwowanych przypadków raka i populacji *narażonej* w i -tej powierzchni, dla $i = 1, \dots, I$, to

$$C = (r - p)'A(r - p)$$

i

$$T = \{C - E(C)\} / \{\text{var}(C)\}^{1/2} \sim \chi^2$$

ze stopniami swobody wyznaczonymi przez skośność (nie muszą być całkowite), gdzie

$$r^i = (O_1, O_2, \dots, O_I) / O, \{O = O_1 + O_2 + \dots + O_I\}$$

określa wektor obserwowanych względnych częstości w I powierzchniach z wektorem parametrów

$$p^i = (p_1, p_2, \dots, p_I), \{p_i = P_i / P, P = P_1 + P_2 + \dots + P_I\}.$$

Dodatkowo

$$A(\lambda) = a_{ij}(\lambda), \text{ for } i, j = 1, 2, \dots, I,$$

jest $I \times I$ macierzą, która określa bliskość pomiędzy powierzchniami, ze stałym parametrem skali λ [4]. W prezentowanej analizie został przyjęty następujący nieliniowy wykładniczy zanikający model

$$a_{ij}(\lambda) = \exp\{-4(d_{ij}/\lambda)^2\}$$

gdzie d_{ij} reprezentuje odległość pomiędzy i -tym i j -tym kwadratem. Do określenia przestrzennych skupień zachorowań, pod uwagę brana była wartość $\lambda = 20$ (km), która oznacza, że każda para przypadków raka odległych od siebie poza tę odległość nie mogła być uznana za skupienie. W szczególności wybór λ jest związany z sensowną wiedzą *a priori* o wielkości skupień zachorowań. Duże wartości λ czynią test wrażliwym na duże skupienia, zaś małe wartości λ na małe skupienia, jednakże doświadczenia Tango [4] pokazują, że otrzymywane C i T nie zmieniają się drastycznie w zależności od wyboru stałej λ . Kandydaci na skupienie wskazywani byli poprzez procentowy udział i -tej powierzchni w całkowitym indeksie C jako >1 proc. W przypadku braku zjawiska skupienia, które oznacza, że obserwowane liczby przypadków w I powierzchniach są niezależnymi zmiennymi poissonowskimi, ze średnimi proporcjonalnymi do populacji *narażonej*, udział procentowy i -tej powierzchni jest równy $1/103 \cdot 100$ proc. w całkowitym indeksie C . W innym przypadku, różne udziały powierzchni w C byłyby widoczne, a najbardziej ekstremalne wartości mogą być uważane za najbardziej prawdopodobne skupienia [4]. Stosunki obserwowanych do oczekiwanych przypadków dla takich lokalizacji, większe lub mniejsze od jedności, wskazują właściwy typ skupień odpowiednio o niskiej lub wysokiej zachorowalności.

Do estymacji standaryzowanego współczynnika zachorowalności ($SWZ = \text{obserwowane/oczekiwane}$ przypadki) raka prostaty w województwie wykorzystano parametryczną empiryczną metodę bayesowską. W podstawowym ujęciu metody proponowanej przez [2], relatywne ryzyka θ podlegają własnościom rozkładu gamma z parametrem skali α i parametrem kształtu ν . Zakładając warunek dla θ_i , można przyjąć, że obserwowane przypadki chorób O_i są zmiennymi o rozkładzie Poissona z parametrem $\theta_i E_i$. Wtedy oczekiwany rozkład *a posteriori* θ_i , pod warunkiem O_i można zapisać jako

$$E(\theta_i | O_i, E_i; \alpha, \nu) = \frac{O_i + \nu}{E_i + \alpha}.$$

Estymację α i ν można osiągnąć metodą określania najwyższej wiarygodności. Logarytm prawdopodobieństwa na bazie O_i i E_i jest równy

$$L(\alpha, \nu) = \sum_i \left[\log \frac{\Gamma(O_i + \nu)}{\Gamma(\nu)} + \nu \log(\alpha) - (O_i + \nu) \log(E_i + \alpha) \right] [2].$$

Numerycznie, estymacja parametrów funkcji najwyższej wiarygodności polega na znalezieniu minimalnej sumy ujemnego logarytmu prawdopodobieństwa.

Prostą estymację oczekiwanych przypadków chorób - E_i przeprowadzono warunkowo na bazie przypadków obserwowanych i populacji odniesienia (podejście warunkowe). Dla danej liczby wszystkich przypadków zachorowań dla każdej kategorii wiekowej (O_{+1}, \dots, O_{+K}) , $O_{+k} = O_{1k} + \dots + O_{Ik}$, warunkowa oczekiwana liczba przypadków jest wyznaczana z równania

$$E_i = \sum_{k=1}^K O_{+k} \left(\frac{P_{ik}}{\sum_{j=1}^I P_{jk}} \right), \quad (i = 1, 2, \dots, I; k = 1, 2, \dots, K).$$

Do analizy przestrzennych interakcji pomiędzy SWZ a udziałami procentowymi ludności etnicznej zaadoptowano statystykę I Moran'a [6]

$$I = \frac{\frac{n}{W} \sum_{ij} w_{ij} z_i z_j}{\sum_{ij} z_i^2},$$

gdzie

$$W = \sum_{i,j} w_{ij},$$

i w_{ij} jest wagą dla związku pomiędzy obserwacjami $i, j = 1, 2, \dots, I$, podczas gdy

$$z_i = SWZ_I - \overline{SWZ}$$

jest scentralizowaną zmienną uzyskaną z SWZ_I [6]. W niniejszym studium (podobnie do [7,8])

$$w_{ij} = (\min\{d_{ij} : i = 1, 2, \dots, I\} / d_{ij}) (X_i X_j)^{1/2}$$

gdzie d_{ij} oznacza odległość pomiędzy i -tym i j -tym kwadratem, zaś X_i, X_j są procentowymi udziałami ludności autochtonicznej w powierzchniach i, j . W badaniach dwa kwadraty były uznane za sąsiadów, jeśli ich środki znajdowały się w odległości do 20 km od siebie (maksymalny dystans pomiędzy przypadkami).

Pełna ekologiczna analiza bayesowska bazowała na regresji poissonowskiej efektów losowych – modelu warunkowej autoregresji przy użyciu techniki sprzęgającej własności łańcucha Markowa z metodą symulacji Monte Carlo, znanej pod nazwą próbkowania Gibbsa [9,10]. W założeniach modelu podanego przez [9] przyjmuje się, że jeżeli O_i i E_i są odpowiednio obserwowanymi i oczekiwanymi przypadkami zachorowań na raka w i -tej powierzchni, to

$$O_i \sim \text{Poisson}(\mu_i)$$

$$\log \mu_i = \log E_i + \alpha X_i + b_i$$

$$b_i \sim \text{Normal}(\bar{b}_i, n_i)$$

$$\bar{b}_i = 1/n_i \sum_{j \in \text{neighbors}(i)} b_j$$

gdzie α jest nieokreśloną informacją aprioryczną, n_i jest liczbą sąsiadów i , zaś X_i jest zmienną ekologiczną (procentem autochtonów). W ujęciu bayesowskim, jako dodatek do określenia modelu dla danych obserwowanych (zazwyczaj w formie rozkładu prawdopodobieństwa), rozważany jest również wektor nieznanych parametrów rozkładu (wielkość losowa), mających swój własny rozkład *a priori* i hierarchiczny wektor hiperparametrów. Wnioskowanie statystyczne bazuje na rozkładzie *a posteriori*, poglądowo wyrażonym jako proporcjonalny do rozkładu prawdopodobieństwa w krotkości rozkładu *a priori*. Kluczowym elementem analizy *a priori-a posteriori* jest, by informacja *a priori* zawsze informowała o *a posteriori* [9]. Kwestią sporną jest, w jakim stopniu? W przyjętym modelu [9] autorzy pro-

ponują nieokreśloną informację *a priori* dla α , co oznacza, że jakkolwiek jej wybór daje zawsze ten sam rozkład *a posteriori*. Model buduje pełny połączony rozkład z kompletnych rozkładów warunkowych dla każdego b_i (wziętych wszystkich jednocześnie), które indukują korelacje przestrzenne [9].

Podstawowa filozofia stojąca za techniką łańcucha Markowa Monte Carlo polega na iteracyjnej symulacji łańcucha Markowa, której rozkład równowagowy jest rozkładem oczekiwanym. Zamiast kalkulowania dokładnych wartości, technika ta generuje strumień symulowanych wartości dla każdej wielkości poszukiwanej [10]. W badaniach, symulacja oparta na rozruchowych 1 000 iteracjach i następujących 10 000 próbkach była przekonująca co do zbieżności rozkładów strumieni wartości w trakcie próbkowania Gibbsa.

Otrzymane SWZ z empirycznej i pełnej analizy bayesowskiej zostały przedstawione na kolorowych mapach. Zmiany SWZ (otrzymane w regresji poissonowskiej) w obrębie X z liniowym dopasowaniem przedstawiono dodatkowo na wykresie.

W badaniach wykorzystano następujące licencjonowane programy komputerowe: MapInfo 4.5, S-PLUS 4.0, S+SPATIALSTATS 1.0 i WinBUGS 1.3.

Prezentowane techniki obliczeń są łatwe w użyciu i mogą być wykorzystane w innych rodzajach analiz ekologicznych [11]. Dla zainteresowanych autor oferuje możliwość przesłania kodów programowych na życzenie.

WYNIKI

Adoptowany zestaw metod został wykorzystany do oceny relacji pomiędzy zachorowalnością na raka prostaty a sytuacją etniczną w województwie opolskim. Wyniki obliczeń testu Tango [4] do wykrywania zjawiska skupienia przestrzennego zachorowań zestawiono w tab. 1.

Z wyników przedstawionych w tab. 1. można stwierdzić, że standaryzacja rozkładu indeksu C daje wysoką wartość statystyki T przy niskiej wartości poziomu błędów p . Liczby te potwierdzają występowanie zjawiska skupienia zachorowań w regionie. Wyniki analizy geograficznej lokalizacji skupień przedstawiono na ryc. 2. (fig. 2.). Mapa na ryc. 2. (fig. 2.) precyzyjnie demonstrowuje, że skupienia niskiej zachorowalności występują w powierzchniach *autochtonicznych*, natomiast skupienia wysokiej zachorowalności są rozrzucone na terytoriach zamieszkałych przez repatriantów (porównaj z fig. 1.).

Przestrzenny rozkład empirycznych estymatorów bayesowskich relatywnego ryzyka [5] raka prostaty przedstawia ryc. 3. (fig. 3.)

Z lekko mozaikowego układu SWZ przedstawionego na ryc. 3. (fig. 3.) można wyczytać, że najniższe poziomy zachorowań są charakterystyczne dla powierzchni *autochtonicznych*, zaś najwyższe poziomy zachorowań na raka są obserwowane na terytoriach zaludnionych przez repatriantów (porównaj z fig. 1.).

Wyniki korelacji przestrzennej Morana [6] pomiędzy zachorowalnością na raka w odniesieniu do gęstości etnicznych zestawiono w tab. 2.

Wyniki potwierdzają, że SWZ są przestrzennie zależne i zmiernają do posiadania podobnych wartości sąsiedzkich, gdy odnoszone są do X . W przypadku chorób chronicznych, ta korelacja może być rozumiana jako pojawiająca się warunkowo w wyniku wpływu zmiennej ekologicznej na tę chorobę [12].

Taki przestrzennie skorelowany proces uzasadnia modelowanie zmiennych za pomocą modeli regresji przestrzennych. Wyniki analizy *a posteriori* dla parametru α (regresja poissonowska) dla modelu warunkowej autoregresji w metodzie bayesowskiej pełnej [9] są podane w tab. 3.

Wyniki *a posteriori* zawarte w tab. 3. dowodzą ewidentnie wpływu ekologicznej zmiennej X (proc. autochtonów) na poziom SWZ. Mapa przestrzennie wymodelowanych SWZ raka prostaty z udziałem gęstości autochtonicznych przedstawiona została na ryc. 4. (fig. 4.)

Podobnie jak wcześniej, mapa na ryc. 4. (fig. 4.) jasno pokazuje, że najniższe ryzyko zachorowań na raka nakłada się na terytoria zaludnione przez autochtonów, natomiast powierzchnie najwyższego ryzyka zachorowań na ten nowotwór złośliwy są charakterystyczne dla repatriantów. Komplementarny obraz wpływu X na zmiany SWZ przedstawia ryc. 5. (fig. 5.)

Proste liniowe dopasowanie zmiennych wyrysowane na ryc. 5. (fig. 5.) pokazuje, że ryzyko zachorowań na raka prostaty u repatriantów jest ok. dwukrotnie wyższe niż u autochtonów.

WNIOSKI

Przedstawiona retrospektywna analiza ekologiczna dotyczyła związku pomiędzy zachorowalnością na raka prostaty a sytuacją etniczną w województwie opolskim. Mimo braku pierwotnej wiedzy na temat tego, czy różnice populacyjne mogły mieć wpływ na tę chorobę, uzyskane za pomocą przyjętych metod wyniki potwierdzają oczywiste podobieństwo rozkładu etnicznego i zdrowotnego w regionie. W aspekcie zachorowalności na nowotwory złośliwe prostaty wśród dwóch badanych grup można stwierdzić, że repatrianci cierpią na raka mniej więcej 2-krotnie in-

tenywniej, niż sąsiadujący autochtoni. Ten fakt może wywołać określone spekulacje na temat etiologii choroby w odniesieniu do różnic w, na przykład, indywidualnych zwyczajach, czynnikach zawodowych, diety czy konsumpcji alkoholowej/tytoniowej itd., wśród tych populacji. Warto podkreślić, że w najbardziej powszechnej lokalnej opinii, autochtoni uznani są za znacznie bardziej dbających o szeroko rozumianą higienę niż repatrianci. Oczywiście, na tym etapie badań ten problem epidemiologiczny można jedynie zasygnalizować. Jednak w opinii autora, wyniki te mogą wskazywać właściwy kierunek dla dalszych bardziej zaawansowanych badań naukowych, pomocnych w określeniu przyczyny różnicy w ryzyku zachorowań na raka w relacji do układu etnicznego w województwie.

PIŚMIENNICTWO

1. Zatoński WA ed. *Nowotwory złośliwe w Polsce*. Warszawa (PL): Centrum Onkologii – Instytut im. M. Curie-Skłodowskiej 1993.
2. Balaryn J. *Szacunek ludności autochtonicznej województwa opolskiego według stanu w dniu 30 VI 1977 r.* Opole (PL): Instytut Śląski w Opolu, 1977.
3. Parkin DM, Muir CS, Whelan SL, Gao Y-T, Ferlay J, Powell J eds. *Cancer Incidence in Five Continents*. Lyon (F): IARC Scientific Publications 1992; VI, 120.
4. Tango T. *A class of tests for detecting „general” and „focused” clustering of rare disease*. *Statistics in Medicine* 1995; 14: 2323-34.
5. Clayton D, Kaldor J. *Empirical Bayes Estimates of Age-standardized Relative Risks for Use in Disease Mapping*. *Biometrics* 1987; 43: 671-81.
6. Moran PAP. *Notes on continuous stochastic phenomena*. *Biometrika* 1950; 37: 17-23.
7. Cressie NAC. *Statistics for Spatial Data*. New York (USA): Wiley 1993.
8. Kaluzny SP, Vega SC, Cardoso TP, Shelly AA. *S+SPATIAL-STATS 1.0 – User's Manual*. Seattle (USA): MathSoft, Inc. 1996.
9. Spiegelhalter D, Thomas A, Best N, Gilks W. *BUGS 0.5 Examples, Volume 2*. Cambridge (UK): Medical Research Council Biostatistics Unit 1996: 37-42, <http://www.mrc-bsu.cam.ac.uk/bugs/>.
10. Smith AFM, Roberts GO. *Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods*. *Journal of the Royal Statistical Society* 1993; B, 55: 3-23.
11. Tukiendorf A. *An ecological analysis of leukemia incidence around the highest ¹³⁷Cs concentration in Poland*. *Cancer Causes and Control* 2001; 12: 653-59.
12. Richardson S, Guihenneuc C, Lasserre V. *Spatial linear models with autocorrelated error structure*. *The Statistician* 1992; 41: 539-57.

ADRES DO KORESPONDENCJI

dr Andrzej Tukiendorf

Politechnika Opolska
ul. Mikołajczyka 5
45-271 Opole
antu@po.opole.pl