TARYFIKACJA *A PRIORI* Z UWZGLĘDNIENIEM EFEKTÓW PRZESTRZENNYCH

Kamil Gala
Ubezpieczeniowy Fundusz Gwarancyjny
e-mail: kgala@ufg.pl

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY
Nr 15(21)

ISSN 1644-6739
e-ISSN 2449-9765

DOI: 10.15611/sps.2017.15.05
JEL Classification: G22 Insurance; Insurance Companies; Actuarial Studies

Streszczenie: Artykuł poświęcony jest metodom aktuarialnej taryfikacji *a priori* w ubezpieczeniach komunikacyjnych, w których jednym z czynników taryfikacyjnych jest adres zamieszkania ubezpieczonego. Podział na klasy taryfowe związane z miejscem zamieszkania jest często stosowany w praktyce ubezpieczeń komunikacyjnych, co jest podyktowane m.in. zróżnicowanymi warunkami drogowymi w zależności od miejsca użytkowania pojazdu. Celem pracy jest identyfikacja i opis przestrzennego zróżnicowania ryzyka ubezpieczeniowego w ubezpieczeniach komunikacyjnych, a także określenie, czy wykorzystanie metod statystyki przestrzennej pozwala na zwiększenie efektywności standardowych modeli taryfikacyjnych i w konsekwencji na lepszą ocenę ryzyka ubezpieczeniowego. W pracy rozważane są modele z klasy uogólnionych modeli liniowych oraz ich modyfikacje mające na celu uwzględnienie zagadnień typowych dla statystyki przestrzennej, szczególnie autokorelacji przestrzennej oraz wygładzania przestrzennego związanego z nierówną ekspozycją poszczególnych obszarów geograficznych na ryzyko.

Słowa kluczowe: ubezpieczenia komunikacyjne, uogólnione modele liniowe, statystyka przestrzenna, efekty przestrzenne.

1. Wstęp

Standardową praktyką rynkową w ubezpieczeniach komunikacyjnych, głównie ubezpieczeniu odpowiedzialności cywilnej posiadaczy pojazdów mechanicznych (OC p.p.m.) oraz ubezpieczeniu *autocasco* (AC), jest ustalanie składki na podstawie obserwowalnych cech ubezpieczonego i jego pojazdu. Jest to tzw. taryfikacja *a priori* [Ostasiewicz (red.) 2004]. Często stosowaną praktyką w tym zakresie jest różnicowanie składki ze względu na region geograficzny, w którym użytkowany jest pojazd [Denuit i in. 2007; Brouhns i in. 2002]. Wynika to z przestrzennego zróżnicowania ryzyka ubezpieczeniowego, które z kolei związane jest z szeregiem czynników ekonomicznych, społecznych i technicznych.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 15(21)

Z powyższych względów pożądany jest rozwój metod taryfikacji, które pozwalają w precyzyjny sposób analizować wpływ lokalizacji na ryzyko ubezpieczeniowe. Niniejsza praca poświęcona jest metodom statystycznym pozwalającym na identyfikację zależności przestrzennych, a następnie na wykorzystanie tej informacji w procesie taryfikacji *a priori*.

2. Analiza danych przestrzennych

2.1. Definicja i rodzaje efektów przestrzennych

W niniejszej pracy **dane przestrzenne** zostały zdefiniowane jako dane dotyczące zjawisk zachodzących w przyjętym układzie współrzędnych. Dane przestrzenne można podzielić według typu informacji na trzy kategorie [Suchecki (red.) 2010]:

- **dane punktowe** – pokazujące wartości zmiennych zlokalizowanych w konkretnych punktach przestrzeni,
- **dane powierzchniowe** – cechujące się ciągłą zmiennością (np. temperatura, ciśnienie atmosferyczne),
- **dane obszarowe** – dotyczące zmiennych obserwowanych dla obiektów w postaci fragmentów powierzchni (np. jednostek podziału administracyjnego).

W ubezpieczeniach najczęściej dostępne są dane obszarowe, uzyskiwane przez agregację danych indywidualnych, np. na podstawie adresu zamieszkania ubezpieczonego. W dobie nowoczesnych technologii (np. lokalizatorów GPS) można spodziewać się, że zwiększy się również dostępność danych punktowych, np. w postaci współrzędnych geograficznych miejsca zajścia zdarzenia ubezpieczeniowego.

W dalszej części pracy rozważone zostały metody analizy danych obszarowych. Kluczowym zagadnieniem w analizie takich danych jest definicja sąsiedztwa i odległości, które określają charakter i siłę interakcji przestrzennych. W niniejszym opracowaniu rozważane są dwa rodzaje macierzy sąsiedztwa:

- **macierz binarna** – $\boldsymbol{D}_{bin} = \left[d_{ij}^{bin}\right]_{i=1,\dots n, j=1,\dots n}$, gdzie $d_{ij}^{bin} = 1$, jeśli obszary i oraz j mają wspólną granicę, i $d_{ij}^{bin} = 0$ w przeciwnym przypadku,
- **macierz odległości oparta na centroidach** – $\boldsymbol{D}_{centr} = \left[d_{ij}^{centr}\right]_{i=1,\dots n, j=1,\dots n}$, gdzie d_{ij}^{centr} jest równe odległości (w kilometrach) między geograficznymi środkami obszarów i oraz

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 15(21)

j, jeśli obszary mają wspólną granicę, oraz równe 0 w przeciwnym przypadku.

Oprócz powyższych możliwe są inne definicje, uwzględniające np. sąsiedztwo wyższych rzędów lub odległość ekonomiczną między obszarami (np. czas podróży między głównymi miastami).

2.2. Autokorelacja przestrzenna

Jednym z głównych zagadnień związanych z analizą danych przestrzennych jest **autokorelacja przestrzenna**, która oznacza stopień skorelowania wartości zmiennej obserwowanej w danej lokalizacji z wartością tej samej zmiennej w innych lokalizacjach [Suchecki (red.) 2010]. Autokorelacja dodatnia oznacza tendencję do występowania przestrzennych skupień wysokich lub niskich wartości zmiennej, natomiast w przypadku autokorelacji ujemnej wysokie wartości badanej zmiennej sąsiadują z niskimi.

W analizie autokorelacji przestrzennej dużą rolę odgrywają tzw. **macierze wag przestrzennych**, definiowane na podstawie macierzy odległości i wykorzystywane do konstrukcji miar interakcji przestrzennych. W niniejszej pracy przyjęto macierz wag (dla obu definicji odległości) $\boldsymbol{W} = \left[w_{ij}\right]_{i=1,\ldots,n,j=1,\ldots n}$ taką, że:

$$w_{ij} = \begin{cases} 0 & \text{jeśli } d_{ij} = 0 \\ 1/d_{ij} & \text{jeśli } d_{ij} > 0 \end{cases}.$$

Dodatkowo macierz wag może być standaryzowana wierszami – elementy standaryzowanej macierzy wag $\boldsymbol{W}^*$ obliczane są wtedy według wzoru

$$w_{ij}^* = \frac{w_{ij}}{\sum_j w_{ij}}.$$

W dalszej części pracy przez "macierz wag" rozumiana będzie macierz wag standaryzowana wierszami.

Omówione teraz zostaną miary autokorelacji przestrzennej dla danych obszarowych. Zostały przyjęte następujące oznaczenia:

- n – liczba jednostek terytorialnych,
- $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ – suma wag przestrzennych,
- x_i – wartość analizowanej zmiennej w i-tej jednostce terytorialnej,
- $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ – średnia wartość analizowanej zmiennej we wszystkich jednostkach terytorialnych.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 15(21)

Pierwszą z omawianych miar jest **współczynnik *I* Morana**, zdefiniowany wzorem:

$$I = \frac{n}{S_0} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Statystyka I wykorzystywana jest do badania, czy sąsiadujące ze sobą obszary są podobne do siebie bardziej, niż wynikałoby to ze stochastycznego charakteru analizowanego zjawiska. Statystyka ta przyjmuje wartości z przedziału $[-1,1]$, przy czym jej wartość oczekiwana jest równa

$$\mathbb{E}(I) = -\frac{1}{n-1}.$$

Wartości statystyki bliskie $\mathbb{E}(I)$ świadczą o braku autokorelacji przestrzennej, wartości większe od $\mathbb{E}(I)$ wskazują na autokorelację dodatnią, natomiast mniejsze niż $\mathbb{E}(I)$ – na autokorelację ujemną.

Kolejną miarą autokorelacji przestrzennej jest **współczynnik *C* Geary'ego**, zdefiniowany następującym wzorem:

$$C = \frac{n-1}{2S_0} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Wartości tak zdefiniowanej statystyki należą do przedziału $[0,2]$, przy czym wartość współczynnika bliska 1 świadczy o braku autokorelacji przestrzennej, wartość z przedziału $[0,1)$ oznacza dodatnią autokorelację, natomiast wartość z przedziału $(1,2]$ – ujemną.

Powyższe miary autokorelacji pozwalają ocenić autokorelację w sposób syntetyczny, za pomocą jednej wartości liczbowej wspólnej dla wszystkich jednostek terytorialnych. Nie pozwalają jednak na ustalenie, które jednostki terytorialne w największym stopniu oddziałują na swoich sąsiadów. W celu rozwiązania tego problemu zostały zdefiniowane lokalne wskaźniki autokorelacji przestrzennej (*Local Indicators of Spatial* Autocorrelation – LISA) [Anselin 1995].

Lokalna statystyka Morana dla i-tej jednostki terytorialnej zdefiniowana jest następująco:

$$I_i = \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x}),$$

natomiast lokalna statystyka Geary'ego dana jest wzorem:

$$C_i = \sum_{j=1}^{n} w_{ij}(x_i - x_j)^2.$$

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 15(21)

Lokalne wskaźniki autokorelacji pozwalają określić wkład i-tej jednostki terytorialnej do globalnego wskaźnika oraz zidentyfikować skupienia jednostek o wysokich lub niskich wartościach badanej zmiennej.

Wyniki analizy empirycznej

Dane pochodzące z bazy danych Ośrodka Informacji Ubezpieczeniowego Funduszu Gwarancyjnego (dalej: OI UFG) pozwalają na przeprowadzenie badań empirycznych. Zakres danych gromadzonych w tej bazie określony jest w art. 102 Ustawy z dnia 22 maja 2003 r. o ubezpieczeniach obowiązkowych, Ubezpieczeniowym Funduszu Gwarancyjnym i Polskim Biurze Ubezpieczycieli Komunikacyjnych (Dz.U. 2013, poz. 392 tj.) i obejmuje informacje o zawartych umowach ubezpieczenia OC p.p.m. i AC, szkodach powodujących odpowiedzialność zakładu ubezpieczeń z tytułu tych umów oraz wypłaconych odszkodowaniach lub odmowach wypłaty. Baza ta jest obowiązkowo zasilana przez zakłady ubezpieczeń prowadzące w Polsce działalność w zakresie OC p.p.m. i we wrześniu 2016 r. zawierała ponad 350 milionów rekordów.

Na podstawie danych pochodzących z bazy OI UFG przygotowany został zbiór danych zawierający częstość szkód z tytułu umów ubezpieczenia OC p.p.m. i AC w podziale na powiaty. Do analizy wybrano umowy zawarte w 2014 r. o rocznym okresie ochrony – w takim przypadku częstość szkód zdefiniowana jest jako iloraz, w którym mianownikiem jest liczba zawartych umów, a licznikiem liczba szkód związanych z tymi umowami. W analizie uwzględniono tylko umowy indywidualne (inne niż flotowe), w których wśród ubezpieczonych wskazano osobę fizyczną, natomiast każda umowa została przypisana do powiatu na podstawie adresu najstarszego ubezpieczonego[1].

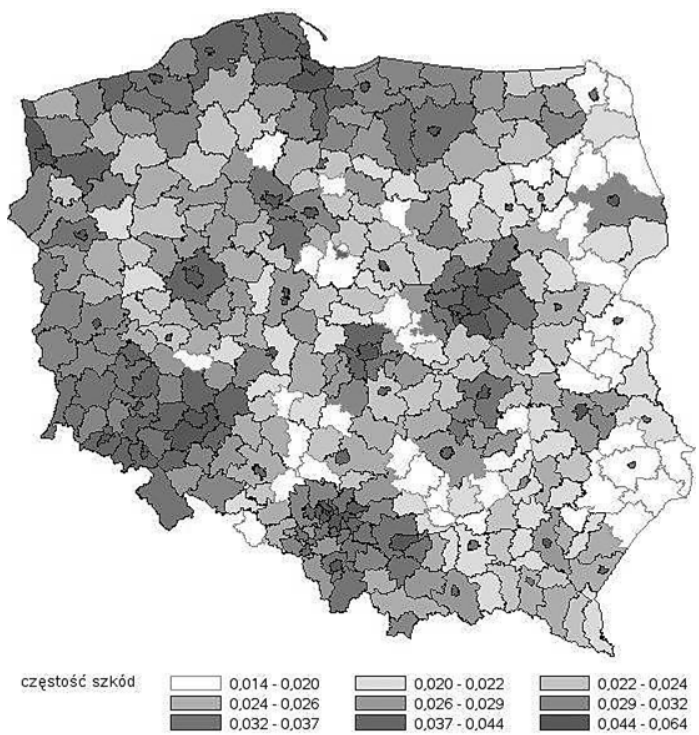
Oprócz informacji o liczbie szkód, liczbie umów oraz powiecie w analizowanym zbiorze znalazła się informacja o płci i wieku najstarszego ubezpieczonego, rodzaju i marce pojazdu, a także wskazanie, czy w umowie występuje wielu ubezpieczonych, czy występuje podmiot zagraniczny i czy któryś z ubezpieczonych jest osobą prawną. Dodatkowo uwzględniono zmienne na poziomie powiatu – wskazanie, czy powiat jest miastem na prawach powiatu, czy jest miastem

[1] Metoda przypisania umowy do powiatu została ustalona ekspercko, na podstawie doświadczeń z analizą danych zgromadzonych w bazie OI UFG. Możliwe są również inne podejścia, np. wybór adresu najmłodszego ubezpieczonego.

wojewódzkim oraz czy jest miastem powyżej 500 tys. mieszkańców. Wszystkie obliczenia zostały wykonane z wykorzystaniem środowiska *R*.

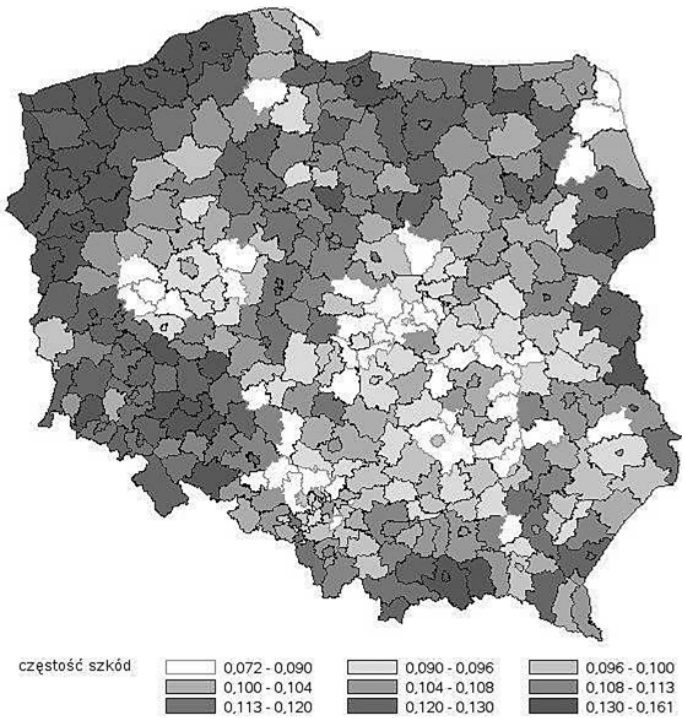
Na rysunku 1 przedstawiona została częstość szkód w ujęciu przestrzennym dla ubezpieczeń OC p.p.m., natomiast na rys. 2 – dla ubezpieczeń AC.

Dla obu rodzajów ubezpieczeń można dostrzec wyraźne skupiska wysokich lub niskich wartości. W przypadku OC p.p.m. wysoka częstość szkód jest obserwowana zwłaszcza w dużych aglomeracjach i terenach je otaczających (np. Warszawa, Górnośląski Okręg Przemysłowy), natomiast w przypadku AC wysokie wartości wskaźnika obserwowane są w zachodniej Polsce. W tabeli 1 przedstawione zostały wartości współczynników autokorelacji przestrzennej Morana i Geary'ego. Obliczenia przeprowadzono dla dwóch wersji macierzy wag – opartej na sąsiedztwie oraz opartej na odległości między środkami obszarów posiadających wspólną granicę.



Rys. 1. Częstość szkód OC p.p.m. w podziale na powiaty

Źródło: opracowanie własne.



Rys. 2. Częstość szkód AC p.p.m. w podziale na powiaty

Źródło: opracowanie własne.

Tabela 1. Wartości współczynników autokorelacji przestrzennej dla analizowanych danych

Rodzaj umowy	Rok	*I* Morana		*C* Geary'ego	
		binarna	centroidy	binarna	centroidy
AC	2013	0,42	0,43	0,53	0,53
	2014	0,59	0,61	0,40	0,38
OC	2013	0,45	0,47	0,43	0,47
	2014	0,45	0,47	0,43	0,47

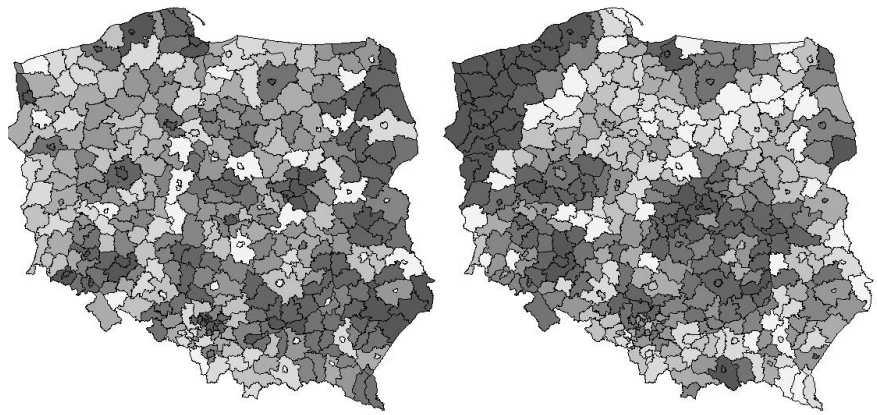
Źródło: opracowanie własne.

Uzyskane wyniki potwierdzają występowanie wyraźnej autokorelacji przestrzennej. Dodatkowo na rysunku 3 zostały przedstawione wartości lokalnej statystyki Morana przy zastosowaniu macierzy wag bazującej na sąsiedztwie.

Przedstawione wyniki wskazują, że w obu rodzajach ubezpieczeń występują skupiska obszarów o wysokiej wartości lokalnej wskaźnika

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 15(21)



Rys. 3. Lokalna statystyka Morana dla OC p.p.m. (po lewej) oraz AC (po prawej). Ciemniejszy kolor oznacza większą wartość statystyki

Źródło: opracowanie własne.

autokorelacji. Oznacza to, że obszary te istotnie oddziałują na swoich sąsiadów i wnoszą duży wkład do globalnego współczynnika autokorelacji.

3. Taryfikacja *a priori* z uwzględnieniem efektów przestrzennych

3.1. Uogólnione modele liniowe z efektami przestrzennymi

Na potrzeby taryfikacji *a priori* często wykorzystuje się modele należące do klasy uogólnionych modeli liniowych (UML, *Generalized Linear Models – GLM*) i ich modyfikacje pozwalające na uwzględnienie efektów przestrzennych. W modelach tych zakłada się, że zmienna objaśniana $\boldsymbol{Y}$ ma rozkład należący do tzw. rodziny wykładniczej rozkładów o funkcji gęstości (lub funkcji prawdopodobieństwa w przypadku rozkładów dyskretnych) danej wzorem:

$$f_Y(y;\theta;\psi) = \exp\left(\frac{y\theta - b(\theta)}{\psi} + c(y;\psi)\right), y \in D_\psi,$$

gdzie θ i ψ to parametry rozkładu, $b\colon \mathbb{R} \to \mathbb{R}$ i $c\colon \mathbb{R}^2 \to \mathbb{R}$ to ustalone funkcje, a D_ψ jest nośnikiem rozkładu, który może zależeć od parametru ψ. Do tej rodziny należy wiele popularnych rozkładów, np. rozkład normalny, rozkład gamma i rozkład Poissona. Dla rozkładu należącego do rodziny wykładniczej wartość oczekiwana jest równa $\mu = \mathbb{E}(Y) = b'(\theta)$, gdzie b' oznacza pochodną funkcji b.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 15(21)

Kolejnym elementem modelu jest składnik systematyczny dany dla i-tej obserwacji wzorem:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \ \ldots + \beta_k X_{ik},$$

gdzie $\beta_0, \ldots, \beta_k$ to parametry, a X_{ir} jest wartością r-tej zmiennej objaśniającej dla i-tej obserwacji. Dla i-tej obserwacji parametry rozkładu zmiennej Y_i są związane ze zmiennymi objaśniającymi relacją $g(\mu_i) = \eta_i$, gdzie g jest tzw. funkcją wiążącą. Taka definicja modelu pozwala na estymację parametrów $\beta_0, \ldots, \beta_k$ oraz ψ za pomocą metody największej wiarygodności.

Najprostszą metodą uwzględnienia efektów przestrzennych w UML jest wprowadzenie zmiennych objaśniających zdefiniowanych na poziomie jednostki terytorialnej (np. dotyczących infrastruktury drogowej). W takim przypadku efekt przestrzenny jest zawarty w składniku systematycznym, a estymacja parametrów modelu przebiega w standardowy sposób.

Należy jednak zwrócić uwagę, że zastosowanie tej metody do oszacowania indywidualnej wartości oczekiwanej dla każdej jednostki terytorialnej może nie być dobrym podejściem. Przykładowo w przypadku Polski uwzględnienie województwa jako zmiennej objaśniającej oznacza potrzebę estymacji dodatkowych 15 parametrów (jedno województwo jest wartością referencyjną), natomiast w przypadku powiatów jest to już 379 parametrów. Można zatem oczekiwać, że estymatory tych parametrów zazwyczaj nie będą wiarygodne ze względu na niewystarczającą liczbę obserwacji dla poszczególnych poziomów zmiennej. Rozwiązaniem tego problemu mogą być modele z czynnikami wielopoziomowymi (Multi-Level Factors, MLF), opisane np. w pracy [Ohlsson, Johansson 2010]. W modelu tym zmienna kategoryczna o wielu poziomach traktowana jest jako efekt losowy, a do estymacji parametrów jej rozkładu wykorzystywane są metody teorii wiarygodności (*credibility theory*, por. np. [Bühlmann, Gisler 2005]).

W dalszej analizie model z czynnikiem wielopoziomowym zostanie wykorzystany do modelowania częstości szkód z tytułu zawartej umowy ubezpieczenia. Przez Y_{ijt} oznaczono częstość szkód dla i-tej umowy, stanowiącej obserwację t w regionie j ($t = 1, \ldots, n_j$), natomiast U_j oznacza efekt losowy dla regionu j ($j = 1, \ldots, J$). Zakłada się, że Y_{ijt} dla ustalonego U_j można opisać za pomocą UML z rozkładem Poissona i logarytmiczną funkcją wiążącą (model multiplikatywny):

$$\mathbb{E}\left(Y_{ijt} \middle| U_j\right) = \mu \gamma_1^i \gamma_2^i \cdot \ldots \gamma_R^i U_j = \gamma_i V_j,$$

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 15(21)

gdzie μ jest oczekiwaną częstością szkód dla bazowej grupy taryfowej, γ_k^i jest względną oczekiwaną częstością szkód dla k-tej zmiennej taryfowej $(k = 1, \dots, R)$ dla i-tej umowy, $\gamma_i = \gamma_1^i \gamma_2^i \cdot \dots \gamma_R^i U_j$ oraz $V_j = \mu U_j$.

Przyjmujemy dalej następujące założenia:

- wektory losowe $(V_j, Y_{1j1}, Y_{1j2}, \dots, Y_{2j1}, Y_{2j2}, \dots)$ są niezależne dla $j = 1, \dots, J$,
- zmienne V_j $(j = 1, \dots, J)$ są niezależne i mają jednakowy rozkład z parametrami $\mathbb{E}(V_j) = \mu > 0$ oraz $Var(V_j) = \tau^2 > 0$,
- dla każdego j zmienne Y_{ijt} są niezależne warunkowo względem V_j, ze średnią $\gamma_i V_j$ i wariancją spełniającą $\mathbb{E}\left(Var(Y_{ijt}|V_j)\right) = \frac{\gamma_i \sigma^2}{w_{ijt}}$.

Zastosowanie modelu Bühlmanna-Strauba do estymacji U_j prowadzi do wzoru

$$\widehat{U}_j = \tilde{z}_j \frac{\bar{\bar{\tilde{Y}}}_{.j.}}{\mu} + \left(1 - \tilde{z}_j\right),$$

gdzie: $\bar{\bar{\tilde{Y}}}_{.j.} = \frac{\sum_{i,t} \tilde{w}_{ijt} \tilde{Y}_{ijt}}{\tilde{w}_{.j.}}$, $\tilde{Y}_{ijt} = \frac{Y_{ijt}}{\gamma_i}$ oraz $\tilde{w}_{ijt} = w_{ijt}\gamma_i$, a wagi $\tilde{z}_j$ dane są wzorem

$$\tilde{z}_j = \frac{\tilde{w}_{.j.}}{\tilde{w}_{.j.} + \frac{\sigma^2}{\tau^2}}.$$

Parametry σ^2 oraz τ^2 najczęściej nie są znane i muszą zostać oszacowane. W niniejszej pracy wykorzystane zostały estymatory podane w pracy [Ohlsson, Johansson 2010].

Należy zwrócić uwagę, że w powyższej procedurze parametry UML są estymowane przy założeniu, że wartości U_j są ustalone, natomiast estymatory $\widehat{U}_j$ zależą od γ_i. Do estymacji parametrów tego modelu można zastosować metodę iteracyjną:

1. Przyjmij $U_j = 1$ dla $j = 1, \dots, J$.
2. Oszacuj parametry UML, przyjmując U_j jako zmienną określającą przesunięcie (*offset*).
3. Wyznacz estymatory $\hat{\sigma}^2$ oraz $\hat{\tau}^2$ w modelu Bühlmanna-Strauba.
4. Wyznacz nowe wartości $\widehat{U}_j$ dla $j = 1, \dots, J$.
5. Powtarzaj punkty 2-4 do uzyskania zbieżności.

Powyższa procedura pozwala uzyskać zarówno oszacowania parametrów dla poszczególnych zmiennych objaśniających, jak i wartości efektu losowego $\widehat{U}_j$ oraz odpowiadające im współczynniki wiarygodności, $\tilde{z}_j$.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY

Nr 15(21)

3.2. Estymacja parametrów modelu – przykład empiryczny

Zilustrujmy zastosowanie procedury opisanej w punkcie 3.1 przykładem empirycznym szacowania parametrów modelu opisującego liczbę szkód z tytułu umów ubezpieczenia *autocasco*. W analizie został wykorzystany zbiór danych opisany w punkcie 2.3. Pod rozwagę wzięto trzy modele:

- **model bazowy (model 1)** – uwzględniający podstawowe zmienne objaśniające z wyłączeniem zmiennych na poziomie powiatu,
- **model ze zmiennymi na poziomie powiatu (model 2)** – uwzględniający wszystkie dostępne zmienne objaśniające,
- **model z efektem losowym (model 3)** – model 2 rozszerzony o efekt losowy.

Na podstawie analizowanego zbioru utworzono losowo dwa rozłączne zbiory – zbiór uczący, liczący milion obserwacji, oraz zbiór walidacyjny, liczący 250 tys. obserwacji. Parametry poszczególnych modeli zostały oszacowane na podstawie zbioru uczącego. Zmienne objaśniające zostały wybrane za pomocą selekcji wstecznej na podstawie wartości kryterium informacyjnego Akaike'a (AIC). Następnie otrzymane modele zostały porównane za pomocą błędu średniokwadratowego (BŚK) na zbiorze walidacyjnym.

Podsumowanie wyników estymacji zostało przedstawione w tab. 2. Znak "-" oznacza, że wyższa kategoria zmiennej (lub poziom 'TAK' w przypadku zmiennych dychotomicznych) przekłada się na niższą oczekiwaną częstość szkód, natomiast znak "+" – na wyższą. W przypadku zmiennych nominalnych podano kategorię o największej i najmniejszej oczekiwanej częstości szkód.

Warto zwrócić uwagę, że w modelu 3 po uwzględnieniu efektu losowego zmienne objaśniające dotyczące powiatu okazały się nieistotne statystycznie. Jeśli chodzi o pozostałe zmienne, to ich statystyczna istotność oraz kierunek oddziaływania okazały się takie same dla wszystkich rozważanych modeli.

W tabeli 3 przedstawiono porównanie błędu średniokwadratowego poszczególnych modeli na zbiorze walidacyjnym.

Model z efektem przestrzennym cechuje się najmniejszym błędem średniokwadratowym na zbiorze walidacyjnym, przy czym różnica między modelami jest niewielka. Może to wynikać z faktu, że we wszystkich modelach występuje podobny lub ten sam zestaw zmiennych objaśniających, a model z efektem losowym nie wykorzystuje w pełni przestrzennej struktury danych.

Tabela 2. Podsumowanie wyników estymacji parametrów rozważanych modeli.

Zmienna	Model 1	Model 2	Model 3
Wiek	-	-	-
Płeć	częstość szkód większa dla kobiet	częstość szkód większa dla kobiet	częstość szkód większa dla kobiet
Rodzaj pojazdu	największa częstość – samochody osobowe	największa częstość – samochody osobowe	największa częstość – samochody osobowe
	najmniejsza częstość – jednoślady	najmniejsza częstość – jednoślady	najmniejsza częstość – jednoślady
Marka pojazdu	największa częstość – Toyota	największa częstość – Toyota	największa częstość – Toyota
	najmniejsza częstość – Fiat	najmniejsza częstość – Fiat	najmniejsza częstość – Fiat
Czy os. Prawna	+	+	+
Czy współubezpieczeni	-	-	-
Miasto na prawach powiatu		+	nieistotne
Miasto pow. 500 tys. mieszkańców		+	nieistotne

Źródło: opracowanie własne.

Tabela 3. Błąd średniokwadratowy dla porównywanych modeli

Model	BŚK – zbiór uczący	BŚK – zbiór walidacyjny
Model 1	0,1272	0,1256
Model 2	0,1272	0,1256
Model 3	0,1270	0,1255

Źródło: opracowanie własne.

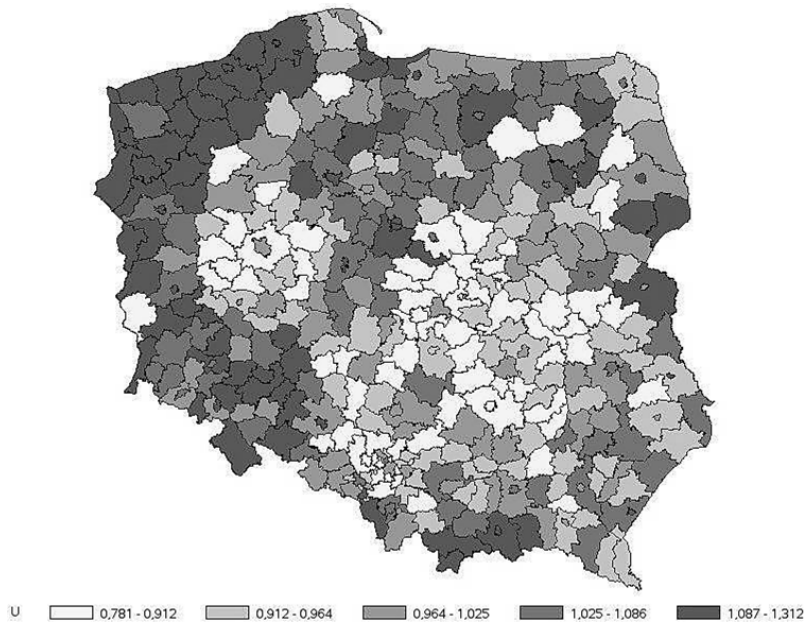
Na rysunkach 4 i 5 przedstawiono oszacowania efektów losowych w modelu 3, a także towarzyszące im współczynniki wiarygodności.

Rozkład przestrzenny efektów losowych jest zbliżony do tego przedstawionego na rys. 2. Oznacza to, że składnik systematyczny modelu uwzględniający dostępne zmienne objaśniające nie wyjaśnia przestrzennego zróżnicowania badanego zjawiska, które zostało uchwycone w efekcie losowym. Przedmiotem dalszych badań może być odpowiedź na pytanie, w jaki sposób najbardziej efektywnie wykorzystać uzyskane wyniki w procesie taryfikacji.

Na rysunku 5 zostały przedstawione współczynniki wiarygodności dla poszczególnych powiatów otrzymane w wyniku estymacji.

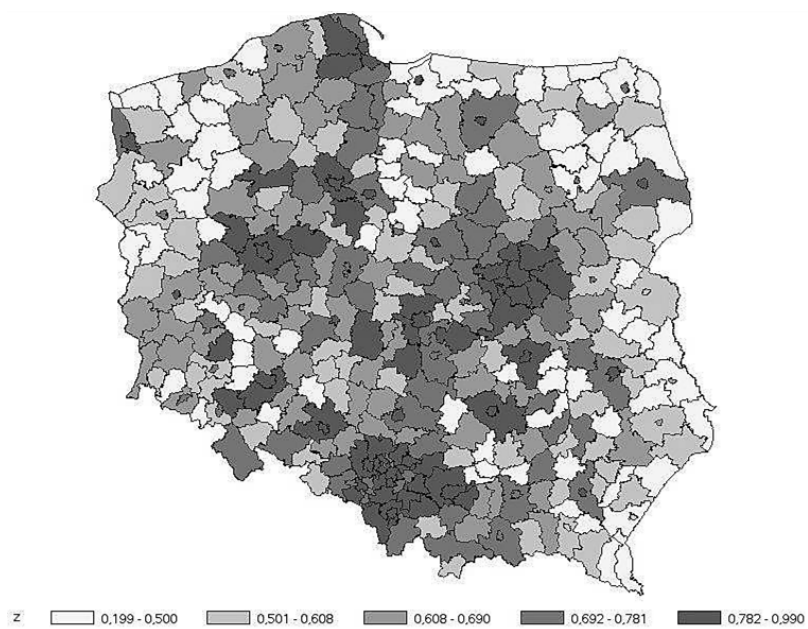
ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY

Nr 15(21)



Rys. 4. Oszacowania losowych efektów przestrzennych dla modelu liczby szkód AC

Źródło: opracowanie własne.



Rys. 5. Współczynniki wiarygodności dla modelu liczby szkód AC

Źródło: opracowanie własne.

Uzyskane wyniki wskazują, że powiaty odpowiadające dużym ośrodkom miejskim (szczególnie Warszawa, Górnośląski Okręg Przemysłowy, Poznań, Wrocław) są na tyle duże i wewnętrznie mało zróżnicowane, że efekty losowe mogą być szacowane praktycznie tylko na podstawie danych dotyczących tych powiatów, jednakże w wielu przypadkach wykorzystanie danych dotyczących całego kraju do estymacji efektów przestrzennych za pomocą metod bayesowskich jest uzasadnione.

4. Zakończenie

W niniejszej pracy przedstawiono podstawowe zagadnienia związane z analizą danych przestrzennych w kontekście modelowania liczby szkód w ubezpieczeniach komunikacyjnych OC p.p.m. i AC. Omówiono rodzaje danych przestrzennych, miary służące do badania autokorelacji przestrzennej oraz uogólniony model liniowy uwzględniający losowe efekty przestrzenne. Przeprowadzono również estymację parametrów omawianego modelu na podstawie danych pochodzących z bazy danych Ośrodka Informacji Ubezpieczeniowego Funduszu Gwarancyjnego.

Eksploracja danych przestrzennych wykazała, że częstość szkód w obu analizowanych rodzajach ubezpieczeń jest zróżnicowana między powiatami oraz występuje istotna autokorelacja przestrzenna. Estymacja parametrów uogólnionego modelu liniowego potwierdziła tę obserwację, ponieważ przestrzenne zróżnicowanie częstości szkód pozostało zauważalne również po uwzględnieniu w modelu szeregu zmiennych objaśniających. Uzyskane wyniki wskazują również na zasadność stosowania metod bayesowskich w estymacji efektów losowych na poziomie powiatu i pozwalają zidentyfikować obszary, dla których indywidualna estymacja częstości szkód nie jest właściwym podejściem.

Jako główny kierunek dalszych badań można wskazać uwzględnienie przestrzennej struktury danych w uogólnionym modelu liniowym z efektami losowymi, a także zagadnienie budowy taryfy uwzględniającej wymiar geograficzny. Interesujące może być również zbadanie wrażliwości wyników na zmianę założeń dotyczących przestrzennych interakcji (np. metody pomiaru odległości) oraz analiza modeli innych niż ten przedstawiony w niniejszej pracy.

ŚLĄSKI
PRZEGLĄD
STATYSTYCZNY

Nr 15(21)

Literatura

Anselin L., 1995, *Local Indicators of Spatial Association – LISA*, Geographical Analysis, vol. 27, no. 2.

Brouhns N., Denuit M., Masuy B., Verrall R., 2002, *Ratemaking by geographical area: A case study using the Boskov and Verrall model*, Discussion paper 0202, Publications of the Institut de statistique, Louvain-la-Neuve, s. 1-26.

Bühlmann H., Gisler A., 2005, *A Course in Credibility Theory and its Applications*, Springer-Verlag Berlin Heidelberg.

Denuit M., Maréchal X., Pitrebois S., Walhin J., 2007, *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*, Wiley, New York.

Ohlsson E., Johansson B., 2010, *Non-Life Insurance Pricing with Generalized Linear Models*, Springer-Verlag Berlin Heidelberg.

Ostasiewicz W. (red.), 2004, *Składki i ryzyko ubezpieczeniowe. Modelowanie stochastyczne*, Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu, Wrocław.

Suchecki B. (red.), 2010, *Ekonometria przestrzenna. Metody i modele analizy danych przestrzennych*, Wydawnictwo C.H. Beck, Warszawa.

Ustawa z dnia 22 maja 2003 r. o ubezpieczeniach obowiązkowych, Ubezpieczeniowym Funduszu Gwarancyjnym i Polskim Biurze Ubezpieczycieli Komunikacyjnych, Dz.U. 2013, poz. 392, t.j.

MODELLING SPATIAL EFFECTS IN THE A PRIORI RISK CLASSIFICATION

Summary: The standard market practice in automobile insurance is to include information on the place of residence of the insured as one of the rating factors. In such a situation geographic area is used as a proxy for various risk factors associated with this area, e.g. traffic intensity and commuting patterns. The subject of this paper is to analyze how actuarial risk classification models could be extended to take these spatial effects into account in the most effective way. To this end, a combination of generalized linear models (GLM) and spatial statistics methods (such as spatial autocorrelation analysis and spatial smoothing) are used.

Keywords: automobile insurance, generalized linear models, spatial statistics, spatial effects.