Kamil Gala[1]

Taryfikacja *a priori* w ubezpieczeniach komunikacyjnych z uwzględnieniem zależności przestrzennej

1. Wstęp

Standardową metodą oceny ryzyka ubezpieczeniowego w masowych ubezpieczeniach indywidualnych jest wykorzystanie metod statystycznych i modeli predykcyjnych. Szczególnie w ubezpieczeniach komunikacyjnych odpowiedzialności cywilnej posiadaczy pojazdów mechanicznych (OC p.p.m.) oraz autocasco (AC) składka ustalana jest na podstawie obserwowalnych cech ubezpieczonego i jego pojazdu – jest to tzw. taryfikacja *a priori*[2]. Do typowych zmiennych taryfowych można zaliczyć wiek kierowcy, rodzaj pojazdu czy też moc i pojemność silnika. W praktyce można również spotkać regionalne różnicowanie składki, podyktowane przestrzennym zróżnicowaniem ryzyka ubezpieczeniowego[3]. Zagadnienia związane z analizą przestrzennych aspektów ryzyka ubezpieczeniowego były poruszane również w literaturze aktuarialnej, m.in. w pracach M. Boskova i R.J. Verralla[4] oraz N. Brouhnsa i innych[5]. Autorzy wykorzystali w nich modele oparte na statystyce bayesowskiej. Odpowiednia analiza geograficznych aspektów ryzyka ubezpieczeniowego oraz zastosowanie metod statystyki przestrzennej wydają się więc drogą do bardziej efektywnej taryfikacji i lepszego dopasowania składki do rzeczywistego ryzyka.

---

1 Ubezpieczeniowy Fundusz Gwarancyjny.

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

3 Por. np. M. Denuit, X. Maréchal, S. Pitrebois, J. Walhin, *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*, Wiley, New York 2007 oraz N. Brouhns, M. Denuit, B. Masuy, R. Verrall, *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 2002, s. 1–26.

4 M. Boskov, R.J. Verrall, *Premium rating by geographical area using spatial models*, „ASTIN Bulletin" 1994, iss. 24, s. 131–143.

5 N. Brouhns, M. Denuit, B. Masuy, R. Verrall, op.cit.

W pracy K. Gali[6] przedstawiono wyniki analizy empirycznej danych pochodzących z bazy Ośrodka Informacji Ubezpieczeniowego Funduszu Gwarancyjnego (OI UFG). Wyniki te wskazują na występowanie zauważalnej autokorelacji przestrzennej między częstością szkód obserwowaną w różnych powiatach. Do analizy efektów przestrzennych zostały wykorzystane modele z czynnikami wielopoziomowymi[7]. Niniejsza praca stanowi kontynuację tych badań, rozszerzającą zakres modeli stosowanych do analizy efektów przestrzennych. W dalszej części artykułu przedstawiono uogólnione modele liniowe, w których przestrzenne efekty losowe modelowane są za pomocą uogólnionego modelu Bühlmanna-Strauba, uwzględniającego korelację między nieobserwowanymi zmiennymi losowymi.

2. Opis zagadnienia

Dane przestrzenne można zdefiniować jako dane dotyczące zjawisk zachodzących w przyjętym układzie współrzędnych oraz podzielić na trzy kategorie[8]:

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

W kontekście ubezpieczeń najczęściej wykorzystywane i najłatwiej dostępne są dane obszarowe, uzyskiwane poprzez agregację danych indywidualnych, np. na podstawie adresu zamieszkania ubezpieczonego.

W niniejszej pracy rozważane będą zagadnienia dotyczące danych obszarowych. W przypadku takich danych istotnym elementem analizy jest określenie, w jaki sposób mierzyć odległość między regionami i w jaki sposób definiować sąsiedztwo. Można wskazać wiele sposobów definiowania macierzy sąsiedztwa[9], jednak w niniejszej pracy rozważane są dwa z nich:

---

[6] K. Gala, *Taryfikacja a priori z uwzględnieniem efektów przestrzennych*, „Śląski Przegląd Statystyczny" 2017, nr 15(21), s. 99–113.

[7] E. Ohlsson, B. Johansson, *Non-Life Insurance Pricing with Generalized Linear Models*, Springer-Verlag, Berlin Heidelberg 2010, s. 71–100.

[8] B. Suchecki (red.), *Ekonometria przestrzenna. Metody i modele analizy danych przestrzennych*, C.H. Beck, Warszawa 2010, s. 38–39.

[9] Ibidem, s. 105–107.

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

Na podstawie macierzy odległości wyznaczana jest macierz wag przestrzennych. Wagi te wykorzystywane są do konstrukcji miar autokorelacji przestrzennej, a także przy budowie modeli statystycznych. W niniejszej pracy przyjęto macierz wag $\boldsymbol{W} = \left[w_{ij} \right]_{i=1,..,n,\ j=1,..n}$ taką, że:

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

Dodatkowo przyjęto, że macierz wag jest wystandaryzowana wierszami, a elementy standaryzowanej macierzy wag $\boldsymbol{W}^*$ obliczane są według wzoru:

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

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

3. Opis modelu

3.1. Model Bühlmanna-Strauba ze skorelowanymi efektami losowymi

Niech X_{ij} będzie zmienną losową interpretowaną jako pewna statystyka szkodowa (np. częstość szkód lub średnia wartość szkody) dla obserwacji j w jednostce badania i, gdzie $i = 1,\ldots, I$ oraz $j = 1,\ldots, n_i$, I jest liczbą jednostek badania, a n_i oznacza liczbę obserwacji w jednostce badania i. Niech w_{ij} będzie znaną wagą związaną ze zmienną losową X_{ij}. W standardowym modelu Bühlmanna-Strauba przyjmowane są następujące założenia[10]:

[10] H. Bühlmann, A. Gisler, *A Course in Credibility Theory and its Applications*, Springer-Verlag, Berlin Heidelberg 2005, s. 79.

- (BS1) Jednostka badania i cechuje się indywidualnym parametrem ryzyka θ_i, będącym realizacją zmiennej losowej Θ_i.
- (BS2) Zmienne $\left\{X_{ij} : j = 1,\ldots, n_i\right\}$ są warunkowo niezależne przy ustalonym Θ_i oraz

$$\mathbb{E}\left(X_{ij}|\Theta_i\right) = \mu\left(\Theta_i\right),$$

$$Var\left(X_{ij}|\Theta_i\right) = \frac{\sigma^2\left(\Theta_i\right)}{w_{ij}},$$

  gdzie $\mu\left(\Theta_i\right)$ i $\sigma^2\left(\Theta_i\right)$ są rzeczywistymi funkcjami zmiennej losowej Θ_i.
- (BS3) Pary $\left(\Theta_1, \boldsymbol{X}_1\right)$, $\left(\Theta_2, \boldsymbol{X}_2\right)$, … są niezależne (gdzie $\boldsymbol{X}_i = \left(X_{i1}, X_{i2},\ldots, X_{in_i}\right)$), a zmienne Θ_1, Θ_2,… są niezależne i mają ten sam rozkład.

Niech $\mu := \mathbb{E}\left(\mu\left(\Theta_i\right)\right)$, $\sigma^2 := \mathbb{E}\left(\sigma^2\left(\Theta_i\right)\right)$ oraz $\tau^2 := Var\left(\mu\left(\Theta_i\right)\right)$. Konsekwencją tych założeń są następujące wzory:

$$\mathbb{E}\left(X_{ij}\right) = \mathbb{E}\left(\mathbb{E}\left(X_{ij}|\Theta_i\right)\right) = \mathbb{E}\left(\mu\left(\Theta_i\right)\right) = \mu, \tag{1}$$

$$Var\left(X_{ij}\right) = \mathbb{E}\left(Var\left(X_{ij}|\Theta_i\right)\right) + Var\left(\mathbb{E}\left(X_{ij}|\Theta_i\right)\right) = \frac{\sigma^2}{w_{ij}} + \tau^2, \tag{2}$$

$$Cov\left(X_{ij}, X_{rs}\right) = \mathbb{E}\left(Cov\left(X_{ij}, X_{rs}|\Theta_i, \Theta_r\right)\right) + Cov\left(\mathbb{E}\left(X_{ij}|\Theta_i, \Theta_r\right), \mathbb{E}\left(X_{rs}|\Theta_i, \Theta_r\right)\right) =$$
$$= \begin{cases} 0 & \text{dla } i \neq r \\ \tau^2 & \text{dla } i = r \text{ i } j \neq s \\ \dfrac{\sigma^2}{w_{ij}} + \tau^2 & \text{dla } i = r \text{ i } j = s \end{cases} \tag{3}$$

Jak widać, model Bühlmanna-Strauba tylko w ograniczonym stopniu jest w stanie uchwycić korelację między różnymi jednostkami badania – kowariancja obserwacji z tej samej grupy jest równa τ^2, natomiast obserwacje z różnych grup są nieskorelowane. Jeśli celem jest uchwycenie korelacji między grupami (jak to ma miejsce w przypadku analiz przestrzennych), to model musi być zmodyfikowany. W tym celu założenie (BS3) zostanie zastąpione założeniem (BS3a):

- (BS3a) Zmienne $\Theta_1, \Theta_2,\ldots, \Theta_I$ są skorelowane i przyjmujemy

$$Cov\left(\mu\left(\Theta_i\right), \mu\left(\Theta_r\right)\right) =: \sigma_{ir}^2 \text{ dla } i, r = 1,\ldots, I.$$

W rezultacie zmianie ulega wzór na kowariancję X_{ij} oraz X_{rs}:

$$Cov\left(X_{ij},X_{rs}\right)=\mathbb{E}\left(Cov\left(X_{ij},X_{rs}|\Theta_i,\Theta_r\right)\right)+Cov\left(\mathbb{E}\left(X_{ij}|\Theta_i,\Theta_r\right),\mathbb{E}\left(X_{rs}|\Theta_i,\Theta_r\right)\right)=$$
$$=\begin{cases}\sigma_{ir}^2 & \text{dla } i\neq r\\ \tau^2 & \text{dla } i=r \text{ i } j\neq s\\ \dfrac{\sigma^2}{w_{ij}}+\tau^2 & \text{dla } i=r \text{ i } j=s\end{cases}. \tag{4}$$

Macierz wariancji-kowariancji wektora $\left(\mu\left(\Theta_1\right),\ldots,\mu\left(\Theta_I\right)\right)$ oznaczana będzie przez Σ. W uogólnionym modelu B-S macierz ta nie jest diagonalna, dzięki czemu możliwe jest uwzględnienie w analizie korelacji między obserwacjami pochodzącymi z różnych grup.

Kolejnym krokiem w analizie jest znalezienie liniowego predyktora bayesowskiego zmiennej losowej $\mu\left(\Theta_i\right)$ bazującego na danych złożonych z zestawu wektorów $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_I$. W niniejszej pracy rozważony zostanie przypadek niejednorodny, zatem celem jest minimalizacja wyrażenia[11]:

$$V=\mathbb{E}\left(\mu\left(\Theta_i\right)-a_0-\sum_{k=1}^{I}\sum_{l=1}^{n_k}a_{kl}X_{kl}\right)^2$$

ze względu na wartość parametrów $a_0,a_{11},\ldots,a_{In_I}$. Z warunku $\dfrac{\partial V}{\partial a_0}=0$ wynika, że:

$$\hat{a}_0=\mathbb{E}\left(\mu\left(\Theta_i\right)\right)-\sum_{k=1}^{I}\sum_{l=1}^{n_k}\hat{a}_{kl}\mathbb{E}\left(X_{kl}\right)=\mu\cdot\left(1-\sum_{k=1}^{I}\sum_{l=1}^{n_k}\hat{a}_{kl}\right),$$

z kolei warunek $\dfrac{\partial V}{\partial a_{rs}}=0$ prowadzi do równoważnego warunku:

$$Cov\left(\mu\left(\Theta_i\right),X_{rs}\right)=\sum_{k=1}^{I}\sum_{l=1}^{n_k}\hat{a}_{kl}\cdot Cov\left(X_{kl},X_{rs}\right),$$

gdzie $r=1,\ldots,I$ oraz $s=1,\ldots,n_r$. Jest to taki sam układ równań normalnych, jak w standardowym modelu Bühlmanna-Strauba. Wyznaczenie i podstawienie poszczególnych wartości prowadzi do wzoru:

---

[11] Ibidem, s. 62. Wskazane zadanie minimalizacji jest sformułowane tak samo, jak w modelu Bühlmanna i Bühlmanna-Strauba, natomiast modyfikacja struktury stochastycznej modelu wpływa na postać rozwiązania.

$$\sum_{k=1}^{I}\sigma_{kr}\cdot\hat{a}_{k\bullet}+\frac{\hat{a}_{r\bullet}}{w_{r\bullet}}\sigma^2=\sigma_{ir}, \tag{5}$$

gdzie $\hat{a}_{k\bullet}=\sum_{l=1}^{n_k}\hat{a}_{kl}$ i $w_{r\bullet}=\sum_{l=1}^{n_r}w_{rl}$. Powyższy wzór wynika z tego, że współczynniki $\hat{a}_{kl}$ spełniające układ równań normalnych związane są zależnością

$$\frac{\hat{a}_{kl}}{w_{kl}}=\frac{\hat{a}_{k\bullet}}{w_{k\bullet}}$$

dla $l=1,\ldots,n_k$. Relacja ta pozwala zredukować problem do układu I równań liniowych z I niewiadomymi, co oznacza, że rozwiązanie tego układu można przedstawić w postaci macierzowej. Należy przy tym zwrócić uwagę na dwa fakty:

- lewa strona wzoru (5) nie zależy od i,
- prawa strona wzoru (5) zależy od i, a wektorem wyrazów wolnych w tym układzie jest i-ty wiersz (lub, równoważnie, i-ta kolumna) macierzy $\boldsymbol{\Sigma}$.

Powyższe obserwacje pozwalają przedstawić rozwiązanie wszystkich I równań w postaci macierzowej:

$$\hat{a}_0^{(i)}=\mu\cdot\left(1-\sum_{k=1}^{I}\hat{a}_{k\bullet}^{(i)}\right)$$

$$\hat{\boldsymbol{a}}=\left(\boldsymbol{\Sigma}+\sigma^2\cdot diag\left(\boldsymbol{w}^{-1}\right)\right)^{-1}\cdot\boldsymbol{\Sigma}, \tag{6}$$

gdzie: $\boldsymbol{w}^{-1}=\left(w_{1\bullet}^{-1},w_{2\bullet}^{-1},\ldots,w_{I\bullet}^{-1}\right)$ jest wektorem odwrotności sum wag, $diag\left(\boldsymbol{x}\right)$ oznacza macierz diagonalną posiadającą wektor $\boldsymbol{x}$ na głównej przekątnej, natomiast $\hat{\boldsymbol{a}}=\left(\hat{\boldsymbol{a}}^{(1)},\ldots,\hat{\boldsymbol{a}}^{(I)}\right)$ jest macierzą, w której i-ta kolumna $\hat{\boldsymbol{a}}^{(i)}=\left(\hat{a}_{1\bullet}^{(i)},\ldots,\hat{a}_{I\bullet}^{(i)}\right)'$ jest rozwiązaniem układu równań (5). Indeks górny $\left(i\right)$ został wprowadzony w celu podkreślenia, że rozwiązanie rozważanego układu równań może być różne dla każdej jednostki badania.

Warto wskazać wybrane własności otrzymanego rozwiązania:

- Jeśli $\boldsymbol{\Sigma}=diag\left(\left(\tau^2,\ldots,\tau^2\right)\right)$ (efekty losowe są nieskorelowane i mają tę samą wariancję równą τ^2), to dla ustalonego i otrzymujemy:

$$\hat{\boldsymbol{a}}=diag\left(\left(\frac{\tau^2}{\tau^2+\frac{\sigma^2}{w_{1\bullet}}},\ldots,\frac{\tau^2}{\tau^2+\frac{\sigma^2}{w_{I\bullet}}}\right)\right),$$

co daje standardowy model Bühlmanna-Strauba.

- W ogólnym przypadku (macierz $\boldsymbol{\Sigma}$ nie jest diagonalna, więc występuje korelacja między efektami losowymi) macierz $\hat{\boldsymbol{a}}$ nie jest diagonalna. Oznacza to, że predyktor liniowy

$$\widehat{\mu(\Theta_i)} = \hat{a}_0^{(i)} + \sum_{k=1}^{I}\sum_{l=1}^{n_k}\hat{a}_{kl}^{(i)}X_{kl}$$

wykorzystuje wszystkie dane, a nie tylko obserwacje dla *i*-tej jednostki badania, jak w przypadku standardowego modelu Bühlmanna-Strauba.
Dla *i*-tej jednostki badania można napisać:

$$\widehat{\mu(\Theta_i)} = \hat{a}_0^{(i)} + \sum_{k=1}^{I}\sum_{l=1}^{n_k}\hat{a}_{kl}^{(i)}X_{kl} = \mu\cdot\left(1-\sum_{k=1}^{I}\sum_{l=1}^{n_k}\hat{a}_{kl}^{(i)}\right) + \sum_{k=1}^{I}\sum_{l=1}^{n_k}\hat{a}_{kl}^{(i)}X_{kl} =$$

$$= \mu\cdot\left(1-\sum_{k=1}^{I}\hat{a}_{k\bullet}^{(i)}\right) + \sum_{k=1}^{I}\hat{a}_{k\bullet}^{(i)}\tilde{X}_k, \tag{7}$$

gdzie $\tilde{X}_k = \dfrac{1}{w_{k\bullet}}\sum_{l=1}^{n_k}w_{kl}X_{kl}$ jest średnią ważoną obserwacji dla *i*-tej jednostki badania. Postać predyktora jest więc analogiczna do modelu B-S – stanowi on kombinację liniową wartości oczekiwanej *a priori* oraz średnich dla poszczególnych jednostek badania.
- Wagi (tzn. współczynniki przy μ oraz $\tilde{X}_1,\ldots,\tilde{X}_I$) w powyższym wzorze sumują się do 1 dla każdego *i*. Należy jednak zwrócić uwagę, że poszczególne współczynniki $\hat{a}_{k\bullet}^{(i)}$ nie muszą zawierać się w przedziale $(0,1)$.

Do wyznaczenia wektora $\hat{\boldsymbol{a}}$ konieczna jest znajomość parametrów $\boldsymbol{w}$, μ, σ^2 oraz $\boldsymbol{\Sigma}$. Wektor wag $\boldsymbol{w}$ jest znany, natomiast pozostałe wartości muszą zostać oszacowane na podstawie danych.

Wartość μ może zostać przyjęta *a priori* na podstawie średniej dla całej populacji[12]. Z kolei parametr σ^2 pełni taką samą funkcję, jak w standardowym modelu Bühlmanna-Strauba. Można więc zastosować następujący wzór:

$$\widehat{\sigma^2} = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{n_i-1}\sum_{j=1}^{n_i}w_{ij}\cdot\left(X_{ij}-X_i\right)^2\right).$$

[12] W przypadku standardowego modelu Bühlmanna-Strauba estymator parametru μ otrzymywany w modelu jednorodnym różni się od średniej z całej populacji. W związku z tym kierunkiem dalszych badań może być uogólnienie jednorodnej wersji modelu Bühlmanna-Strauba.

Powyższy estymator jest średnią arytmetyczną indywidualnych estymatorów wariancji dla i-tej jednostki badania, S_i:

$$S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} w_{ij} \cdot \left(X_{ij} - X_i \right)^2 .$$

W standardowym modelu Bühlmanna-Strauba powyższy estymator jest nieobciążony i zgodny[13], w związku z czym pojawia się pytanie, czy tak jest również w przypadku modelu uogólnionego, ponieważ $\widehat{\sigma^2}$ jest funkcją skorelowanych zmiennych losowych. Łatwo sprawdzić, że korelacja ta nie wpływa na nieobciążoność estymatora, natomiast korzystając z nierówności Czebyszewa można napisać:

$$\mathbb{P}\left(\left| \widehat{\sigma^2} - \sigma^2 \right| > \varepsilon \right) \leq \frac{Var\left(\widehat{\sigma^2} \right)}{\varepsilon^2},$$

$$Var\left(\widehat{\sigma^2} \right) = \frac{1}{I^2} Var\left(\sum_{i=1}^{I} S_i \right) = \frac{1}{I^2} \cdot \left(\sum_{i=1}^{I} Var\left(S_i \right) + 2 \sum_{i<j} Cov\left(S_i, S_j \right) \right).$$

Z powyższego wzoru wynika, że do zgodności estymatora wystarczy, żeby $Var\left(S_i \right)$ były ograniczone, a także aby każda jednostka miała ograniczoną z góry liczbę sąsiadów. Założenia te pozwalają kontrolować poszczególne składniki sumy, gdy I dąży do nieskończoności. Wydaje się, że założenia te powinny być spełnione w praktyce.

W przypadku parametru τ^2 przyjęty zostanie ten sam estymator, co w standardowym modelu Bühlmanna-Strauba[14]:

$$\widehat{\tau^2} = c \cdot \left\{ T - \frac{I \widehat{\sigma^2}}{w_{\bullet\bullet}} \right\},$$

gdzie $T = \frac{I}{I-1} \sum_{i=1}^{I} \frac{w_{i\bullet}}{w_{\bullet\bullet}} \cdot \left(X_i - \bar{X} \right)^2$ oraz $c = \frac{I-1}{I} \cdot \left(\sum_{i=1}^{I} \frac{w_{i\bullet}}{w_{\bullet\bullet}} \cdot \left(1 - \frac{w_{i\bullet}}{w_{\bullet\bullet}} \right) \right)^{-1}$, natomiast $w_{\bullet\bullet} = \sum_{i,j} w_{ij}$. Również w tym przypadku korelacja między zmiennymi nie powoduje obciążenia estymatora, natomiast kwestia zgodności tego estymatora może być przedmiotem dalszych badań.

[13] H. Bühlmann, A. Gisler, op.cit., s. 93.
[14] Ibidem, s. 95.

Problem estymacji pozostałych elementów macierzy $\mathbf{\Sigma}$ nie będzie tu rozważany całościowo, w kolejnym punkcie zostanie jedynie przestawiony szczególny przypadek przyjęty na potrzeby analizy autokorelacji przestrzennej.

3.2. Uogólniony model liniowy ze skorelowanymi efektami losowymi

Jednymi z podstawowych narzędzi w taryfikacji *a priori* są modele należące do klasy uogólnionych modeli liniowych (UML, ang. *generalized linear models* – GLM), wprowadzonej przez J.A. Neldera i R. Wedderburna[15]. W modelach tych zakłada się, że zmienna objaśniana Y ma rozkład należący do tzw. rodziny wykładniczej rozkładów o funkcji gęstości (lub funkcji prawdopodobieństwa w przypadku rozkładów dyskretnych), danej wzorem:

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

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

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

$$\eta_i=\beta_0+\beta_1X_{i1}+\ldots+\beta_kX_{ik},$$

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 $g\left(\mu_i\right)=\eta_i$, gdzie g jest tzw. funkcją wiążącą (ang. *link function*). Taka definicja modelu pozwala na estymację parametrów $\beta_0,\ldots,\beta_k$ oraz ψ za pomocą metody największej wiarygodności.

Podstawową metodą uwzględnienia efektów przestrzennych w UML jest wprowadzenie zmiennych objaśniających zdefiniowanych na poziomie jednostki terytorialnej (np. gęstość sieci drogowej w powiacie) lub wskazujących

[15] J.A. Nelder, R.W.M. Wedderburn, *Generalized Linear Models*, „Journal of the Royal Statistical Society" 1972, Series A (General), vol. 135, s. 370–384.

na konkretną jednostkę terytorialną[16]. Ta ostatnia metoda nie sprawdzi się jednak przy próbie oszacowania indywidualnych efektów dla poszczególnych jednostek ze względu na ich dużą liczbę (np. 380 powiatów) i prawdopodobną niewielką liczbę obserwacji w części z nich. W tej sytuacji mogą zostać zastosowane modele z czynnikami wielopoziomowymi (ang. *multi-level factors*, MLF), opisane np. w pracy E. Ohlssona i B. Johanssona[17]. W modelu tym zmienna kategoryczna o wielu poziomach traktowana jest jako efekt losowy, a do estymacji parametrów jej rozkładu wykorzystywane są metody teorii wiarygodności (ang. *credibility theory*).

Przez Y_{ijt} została oznaczona 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)[18]:

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

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

W celu uwzględnienia korelacji między różnymi regionami wymagana jest modyfikacja założeń standardowego modelu z czynnikiem wielopoziomowym w zakresie struktury stochastycznej wektorów losowych $\left(V_j,\ Y_{1j1},\ Y_{1j2},\ldots,Y_{2j1},\ Y_{2j2},\ldots\right)$ dla $j = 1,\ldots,J$ oraz ich rozkładu łącznego. Przyjęte zostały następujące założenia:

- zmienne V_j ($j = 1,\ldots,J$) mają jednakowy rozkład z parametrami $\mathbb{E}\left(V_j\right) = \mu > 0$ oraz $Var\left(V_j\right) = \tau^2 > 0$. Macierz wariancji-kowariancji wektora $\boldsymbol{V} = \left(V_1,\ldots,V_J\right)$ jest równa $\boldsymbol{\Sigma}$ (macierz ta nie musi być diagonalna);
- dla każdego *j* zmienne Y_{ijt} są niezależne warunkowo względem V_j, ze średnią γ_iV_j i wariancją spełniającą warunek $\mathbb{E}\left(Var\left(Y_{ijt}|\,V_j\right)\right) = \dfrac{\gamma_i\sigma^2}{w_{ijt}}$.

[16] Zbliżone podejście zostało przedstawione np. w pracy J. Lemaire, S.C. Park, K.C. Wang, *The use of annual mileage as a rating variable*, „ASTIN Bulletin" 2016, vol. 46, iss. 1, s. 39–69, w której użyto wskaźników odpowiadających różnym regionom kraju.

[17] E. Ohlsson, B. Johansson, *Non-Life Insurance Pricing with Generalized Linear Models*, Springer-Verlag, Berlin Heidelberg 2010, s. 71–99.

[18] Ibidem, s. 74.

Jeśli wartość γ_i jest znana, to rozważenie w miejsce Y_{ijt} oraz w_{ijt} zmiennych $\tilde{Y}_{ijt} = \frac{Y_{ijt}}{\gamma_i}$ oraz $\tilde{w}_{ijt} = w_{ijt}\gamma_i$ pozwala na zastosowanie uogólnionego modelu Bühlmanna-Strauba. Predyktor dla zmiennej U_i dany jest wtedy wzorem (7).

Należy zwrócić uwagę, że parametry UML są estymowane przy założeniu, że wartości U_j są ustalone, z kolei estymatory $\hat{U}_j$ zależą od γ_i. Do estymacji parametrów takiego modelu można zastosować procedurę iteracyjną[19]:

1. Przyjmij $U_j = 1$ dla $j = 1,\ldots,J$.
2. Oszacuj parametry UML, przyjmując U_j jako zmienną określającą przesunięcie (ang. *offset*).
3. Wyznacz estymatory $\hat{\sigma}^2$ oraz $\hat{\Sigma}$ w uogólnionym modelu Bühlmanna-Strauba:
   - $\widehat{\sigma^2} = \frac{\Sigma_j \left(n_j - 1\right)\widehat{\sigma_j^2}}{\Sigma_j \left(n_j - 1\right)}$, gdzie $\widehat{\sigma_j^2} = \frac{1}{n_j - 1}\sum_{i,t} \tilde{w}_{ijt}\left(\tilde{Y}_{ijt} - \overline{\tilde{Y}}_{\bullet j\bullet}\right)^2$,
   - $\widehat{\tau^2} = \frac{\sum_j \tilde{w}_{\bullet j\bullet}\left(\overline{\tilde{Y}}_{\bullet j\bullet} - \overline{\tilde{Y}}_{\bullet\bullet\bullet}\right)^2 - \left(J-1\right)\widehat{\sigma^2}}{\tilde{w}_{\bullet\bullet\bullet} - \sum_j \tilde{w}_{\bullet j\bullet}^2 / \tilde{w}_{\bullet\bullet\bullet}}$,
   - $\hat{\rho} = \widehat{\mathrm{Corr}}\left(\mathrm{X},\mathrm{Y}\right)$, gdzie jako realizacje wektora $\left(X,Y\right)$ przyjęto pary $\left(\overline{\tilde{Y}}_{\bullet i\bullet}, \overline{\tilde{Y}}_{\bullet j\bullet}\right)$ takie, że regiony *i* oraz *j* sąsiadują ze sobą (symboliczny zapis $i \sim j$), a *Corr* oznacza współczynnik korelacji Pearsona,
   - $\widehat{\sigma_{ir}} = \hat{\rho}\cdot\sqrt{s_i^2 \cdot s_r^2}$, jeśli $i \sim r$ oraz $\widehat{\sigma_{ir}} = 0$ w przeciwnym przypadku, gdzie $s_i^2 = \widehat{\tau^2}\cdot\frac{\sum_{i,j}\tilde{w}_{ij}^2}{\tilde{w}_{i\bullet}^2} + \frac{\widehat{\sigma^2}}{\tilde{w}_{i\bullet}}$.
4. Wyznacz nowe wartości $\hat{U}_j$ dla $j = 1,\ldots,J$.
5. Powtarzaj punkty 2–4 do uzyskania zbieżności przy ustalonym kryterium stopu.

W niniejszej pracy w analizie empirycznej zostały przyjęte przedstawione wyżej estymatory parametrów strukturalnych, mogą to jednak być również inne estymatory o odpowiednich własnościach. W szczególności istotnym zagadnieniem jest wybór struktury autokorelacji przestrzennej i efektywna estymacja

[19] Przedstawiona procedura stanowi rozszerzenie metody estymacji modeli z czynnikiem wielopoziomowym przedstawionej w cytowanej pracy E. Ohlssona i B. Johanssona. Modyfikacji uległ punkt 3, w którym wymagana jest estymacja parametrów uogólnionego modelu Bühlmanna-Strauba.

parametrów definiujących tę strukturę, co może być dokonane na wiele sposobów w zależności od charakteru badanego zjawiska. W niniejszej pracy założono, że występuje wyłącznie korelacja między regionami sąsiadującymi ze sobą, a współczynnik korelacji jest stały. Założenie to ułatwia estymację, natomiast analiza innych struktur może być przedmiotem dalszych badań.

4. Wyniki analizy empirycznej

4.1. Opis zbioru danych

Źródłem danych wykorzystanych w analizie empirycznej jest baza danych OI UFG. 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 (tekst jedn.: Dz.U. 2018, poz. 473) 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 2017 r. zawierała blisko 400 mln rekordów.

Analizie statystycznej zostały poddane umowy ubezpieczenia OC p.p.m. i AC zawarte w 2015 r. Wykluczone z analizy zostały umowy obejmujące flotę pojazdów oraz takie, w których jako posiadacza pojazdu wskazano tylko osoby prawne (np. przedsiębiorstwo). Ostatecznie zbiór do analizy liczył około 15 mln obserwacji.

Dla każdej z badanych umów wyznaczono szereg charakterystyk, które mogą posłużyć do modelowania liczby szkód obciążających tę umowę. Lista zmiennych wykorzystanych w badaniu przedstawiona została w tabeli 1. Każda ze zmiennych z grup „dane osoby", „historia osoby" i „dane geograficzne" dostępna jest w dwóch wariantach, wyznaczonych dla najstarszego i najmłodszego posiadacza pojazdu[20].

[20] W przypadku jednego posiadacza pojazdu zmienne objaśniające w obu wariantach są identyczne. Ta sama sytuacja może również zaistnieć mimo występowania wielu posiadaczy pojazdu, jeśli ich dane są zgodne dla przyjętego poziomu agregacji. Z tego względu dwa warianty tej samej zmiennej nie powinny być analizowane jednocześnie ze względu na możliwą współliniowość zmiennych.

Tabela 1. Lista zmiennych występujących w zbiorze danych

Grupa	Zmienna
Zmienna objaśniana	Liczba szkód obciążających umowę ubezpieczenia
Dane umowy	Rodzaj umowy (OC p.p.m./AC)
	Data początku okresu obowiązywania umowy
	Data końca okresu obowiązywania umowy
	Ekspozycja na ryzyko[21]
Dane osoby	Wiek
	Płeć
Dane pojazdu	Rodzaj pojazdu
	Marka pojazdu
	Długość historii ubezpieczenia pojazdu
	Czy wśród posiadaczy pojazdu występują osoby prawne?
	Czy pojazd ma więcej niż jednego posiadacza?
Historia osoby	Długość historii ubezpieczenia podmiotu
	Częstość szkód OC p.p.m. w ostatnich 5 latach
	Częstość szkód AC w ostatnich 5 latach
Dane geograficzne	Województwo i powiat
	Czy miasto powyżej 500 tys. mieszkańców?
	Czy miasto na prawach powiatu?
	Czy miasto wojewódzkie?

Źródło: opracowanie własne

4.2. Analiza opisowa

Pierwszym krokiem przeprowadzonej analizy empirycznej była analiza opisowa, która miała na celu eksplorację danych w ujęciu przestrzennym, szczególnie odpowiedź na następujące pytania:

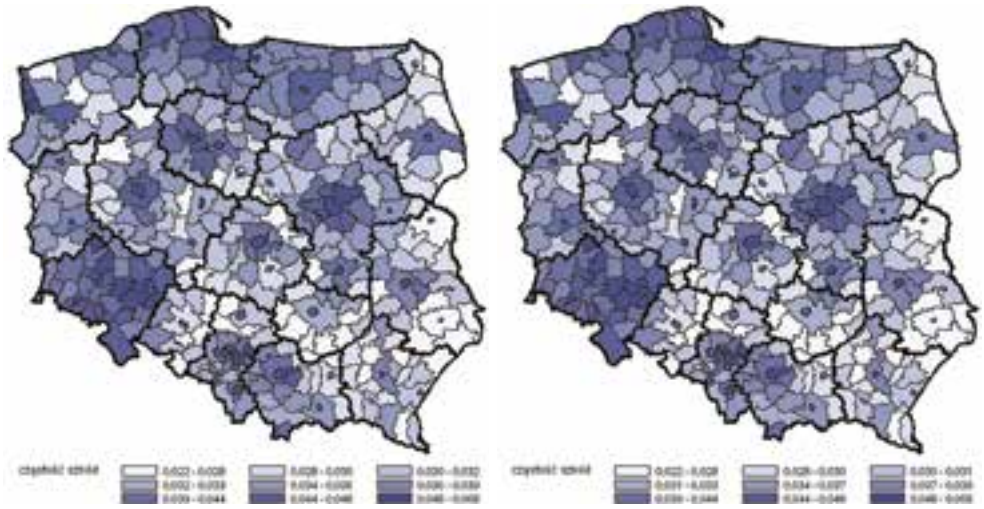
- Czy występuje przestrzenne zróżnicowanie częstości szkód?
- Czy występuje autokorelacja przestrzenna?
- Jak analizować wymiar przestrzenny, jeżeli pojazd ma wielu posiadaczy?

Ostatnie pytanie ma istotne znaczenie w sytuacji, gdy posiadaczami pojazdu są członkowie rodziny (np. ojciec i syn) i nie jest dostępna informacja o tym, kto jest głównym użytkownikiem pojazdu i gdzie ten pojazd będzie użytkowany.

Na kolejnych rysunkach została przedstawiona empiryczna częstość szkód w podziale na powiaty. Odrębnie zostały przeanalizowane ubezpieczenia OC p.p.m.

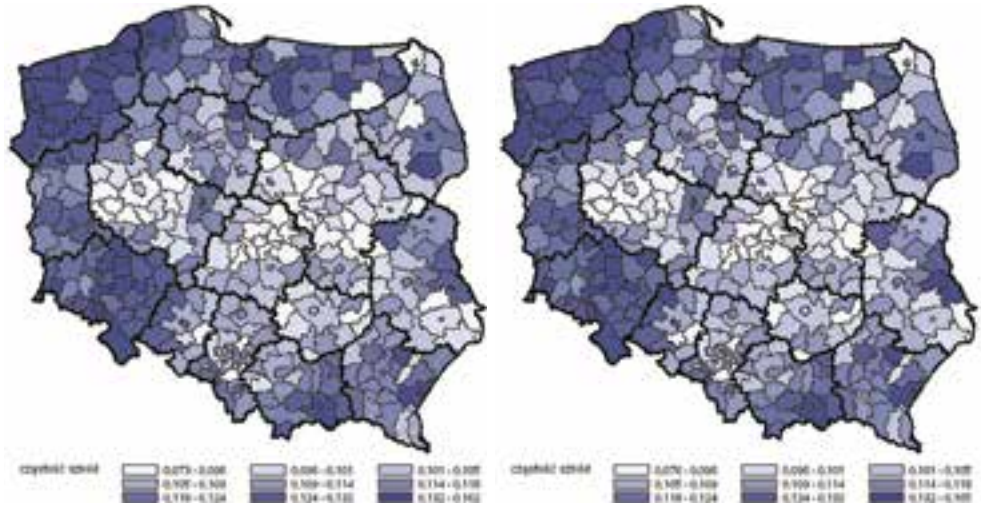
[21] Ekspozycja na ryzyko wyrażona jest w latach i określa, przez jaką część badanego roku kalendarzowego umowa była aktywna.

i AC. Umowy i szkody zostały przypisane do powiatu na dwa sposoby – według adresu najstarszego i najmłodszego posiadacza pojazdu. Na mapę w podziale na powiaty zostały naniesione granice województw.



Rysunek 1. Częstość szkód w ubezpieczeniach OC p.p.m. według powiatu zamieszkania najstarszego (po lewej) oraz najmłodszego (po prawej) posiadacza pojazdu

Źródło: opracowanie własne



Rysunek 2. Częstość szkód w ubezpieczeniach AC według powiatu zamieszkania najstarszego (po lewej) oraz najmłodszego (po prawej) posiadacza pojazdu

Źródło: opracowanie własne

Z analizy przedstawionych kartogramów wynika, że w obu rodzajach ubezpieczeń częstość szkód jest istotnie zróżnicowana przestrzennie. W przypadku OC można zaobserwować wyraźne występowanie wyższej częstości szkód wokół dużych ośrodków miejskich (np. miast wojewódzkich), natomiast w przypadku ubezpieczeń AC widoczna jest wysoka częstość szkód w województwach zachodnich. Warto również zwrócić uwagę na fakt, że w niektórych przypadkach powiaty z niższą i wyższą częstością szkód są oddzielone granicą województwa i różnią się od siebie mimo swojej geograficznej bliskości. Sugeruje to, że w analizach przestrzennych warto rozważać zarówno poziom powiatu, jak i poziom województwa.

Zaobserwowana koncentracja powiatów o wysokiej częstości szkód sugeruje występowanie autokorelacji przestrzennej. W celu weryfikacji tej tezy obliczone zostały współczynniki autokorelacji przestrzennej *I* Morana oraz *C* Geary'ego[22]. Każda z miar została wyznaczona dla binarnej macierzy sąsiedztwa oraz macierzy sąsiedztwa opartej na centroidach. Dodatkowo obliczenia wykonano odrębnie dla powiatu zamieszkania najstarszego oraz najmłodszego posiadacza. Wyniki dla badanego zbioru danych zostały przedstawione w tabeli nr 2 oraz uzupełnione wartościami dla roku 2014, pochodzącymi z pracy K. Gali. Pierwsza liczba dla 2015 r. opisuje autokorelację częstości szkód według powiatu zamieszkania najstarszego, natomiast druga – najmłodszego posiadacza pojazdu.

Tabela 2. Wartości współczynników autokorelacji przestrzennej

Rodzaj umowy	Rok	*I* Morana		*C* Geary'ego	
		binarna	centroidy	binarna	centroidy
AC	*2014*	*0,588*	*0,607*	*0,401*	*0,382*
	2015	0,584 / 0,580	0,595 / 0,591	0,417 / 0,420	0,405 / 0,407
OC	*2014*	*0,447*	*0,468*	*0,432*	*0,473*
	2015	0,460 / 0,460	0,470 / 0,471	0,486 / 0,485	0,502 / 0,500

Źródło: opracowanie własne oraz K. Gala, op.cit., s. 105

Uzyskane wartości współczynników autokorelacji przestrzennej wskazują na występowanie dodatniej autokorelacji przestrzennej i stabilność tej relacji w kolejnych latach. Potwierdza to zasadność stosowania metod statystyki przestrzennej, należy mieć jednak na uwadze, że w tej części analizy nie zostały uwzględnione inne zmienne taryfowe, które opisywałyby strukturę populacji w poszczególnych jednostkach geograficznych. Wyniki bardziej szczegółowej

[22] B. Suchecki (red.), op.cit., s. 112–114.

analizy z wykorzystaniem uogólnionych modeli liniowych i uwzględnieniem dodatkowych zmiennych objaśniających zostały przedstawione w dalszej części pracy.

Na zakończenie analizy opisowej zbadano, która definicja wymiaru geograficznego – według adresu najstarszego oraz najmłodszego posiadacza – daje najlepsze wyniki w objaśnianiu częstości szkód. W tym celu oszacowane zostały parametry dwóch modeli regresji Poissona, w których jedynymi zmiennymi objaśniającymi były województwo oraz zmienne na poziomie powiatu. W tabeli nr 3 zostały przedstawione wartości kryterium informacyjnego BIC dla obu modeli. Pogrubioną i podkreśloną czcionką zaznaczono niższą (lepszą) wartość tego kryterium.

Tabela 3. Wartości kryterium informacyjnego BIC dla modeli regresji Poissona

Rodzaj umowy	Najstarszy posiadacz	Najmłodszy posiadacz
OC	4 133 051	**4 132 863**
AC	**1 859 131**	1 859 147

Źródło: opracowanie własne

Uzyskane wyniki są bardzo podobne dla obu definicji wymiaru geograficznego. Wynika stąd, że na podstawie analizy wizualnej (kartogramy) oraz prostych analiz statystycznych nie można rozstrzygnąć, którego posiadacza należy uwzględnić w analizie ryzyka ubezpieczeniowego. Zagadnienie to zostanie poruszone w kolejnym punkcie niniejszego artykułu.

4.3. Estymacja parametrów modeli statystycznych

Poniżej przedstawiono wyniki modelowania statystycznego z wykorzystaniem uogólnionych modeli liniowych. Przeprowadzona analiza miała na celu weryfikację występowania efektów przestrzennych z uwzględnieniem innych zmiennych taryfowych. W tym celu z dostępnego zbioru danych wybrano losowo 1,25 mln umów ubezpieczenia OC p.p.m. Ograniczenie liczby obserwacji miało na celu przyspieszenie procedury estymacji, a także odzwierciedlenie rzeczywistości, w której zakład ubezpieczeń dysponuje wiedzą tylko o swoim portfelu ubezpieczeń, przez co niektóre jednostki geograficzne mogą być reprezentowane niezbyt licznie.

Otrzymany zbiór danych został podzielony na dwa rozłączne zbiory – zbiór uczący (1 mln obserwacji), służący do estymacji parametrów modeli, oraz zbiór walidacyjny (250 tys. obserwacji), służący do porównania dopasowania modeli

do danych, które nie brały udziału w procesie estymacji. Oszacowane zostały parametry czterech modeli:

1. **Modelu bazowego** – najlepszego modelu uzyskanego metodą postępującego wyboru zmiennych objaśniających na podstawie kryterium BIC, bez uwzględnienia zmiennych geograficznych.
2. **Modelu z danymi na poziomie powiatu** – najlepszego modelu uzyskanego metodą postępującego wyboru zmiennych objaśniających na podstawie kryterium BIC z uwzględnieniem zmiennych geograficznych.
3. **Modelu z losowymi efektami przestrzennymi –** modelu wielopoziomowego, w którym składnik systematyczny ma specyfikację identyczną z modelem 2, natomiast model dla efektów losowych nie uwzględnia autokorelacji.
4. **Modelu ze skorelowanymi efektami przestrzennymi** – modelu wielopoziomowego analogicznego do modelu 3, w którym została uwzględniona korelacja między składnikiem losowym dla różnych obszarów.

Parametry modeli 3 i 4 zostały oszacowane za pomocą iteracyjnej procedury opisanej w punkcie 3. Do obliczeń przyjęto binarną macierz sąsiedztwa. Wyniki podsumowano w tabelach 4 i 5. W tabeli 4 uwzględniono tylko zmienne istotne statystycznie w co najmniej jednym modelu. Warto zwrócić uwagę, że płeć, która nie może być obecnie stosowana jako czynnik taryfikacyjny, okazała się nieistotna statystycznie z punktu widzenia częstości szkód OC p.p.m.

Tabela 4. Podsumowanie wyników estymacji

Zmienna	Model 1	Model 2	Model 3	Model 4
Wiek (najmłodszy)	<= 25 lat 36–45 lat	<= 25 lat 36–45 lat	<= 25 lat 36–45 lat	<= 25 lat 36–45 lat
Rodzaj pojazdu	Samochody pozostałe, przyczepy	Samochody pozostałe, przyczepy	Samochody pozostałe, przyczepy	Samochody pozostałe, przyczepy
Marka pojazdu	Inne, Fiat	Inne, Fiat	Inne, Fiat	Inne, Fiat
Historia pojazdu	Brak historii –	Brak historii –	Brak historii –	Brak historii –
Czy osoba prawna	+	+	+	+
Liczba posiadaczy	–	Brak wpływu	Brak wpływu	Brak wpływu
Częstość szkód OC (najstarszy)	Brak historii, >= 1	Brak historii, >= 1	Brak historii, >= 1	Brak historii, >= 1
Częstość szkód AC (najstarszy)	Brak historii, >= 1	Brak historii, >= 1	Brak historii, >= 1	Brak historii, >= 1
Historia posiadacza (najmłodszy)	Brak historii, 8–10 lat	Brak historii, 8–10 lat	Brak historii, 8–10 lat	Brak historii, 8–10 lat

Zmienna	Model 1	Model 2	Model 3	Model 4
Miasto na prawach powiatu (najstarszy)	Nie dotyczy	+	+	+
Miasto wojewódzkie (najstarszy)	Nie dotyczy	+	+	+
Województwo (najstarszy)	Nie dotyczy	Dolnośląskie, świętokrzyskie	Łódzkie, świętokrzyskie	Łódzkie, świętokrzyskie

Źródło: opracowanie własne

W przypadku zmiennych numerycznych oraz zmiennych dychotomicznych znak „+" oznacza wzrost oczekiwanej częstości szkód wraz ze wzrostem wartości zmiennej lub wystąpieniem wartości „1". W przypadku zmiennych nominalnych na pierwszym miejscu podano kategorię z najwyższą oczekiwaną częstością szkód, a na drugim miejscu – z najniższą.

Tabela 5. Oceny parametrów strukturalnych w zmodyfikowanym modelu Bühlmanna-Strauba

Parametr	Model 3	Model 4
σ^2	0,2982	0,2967
τ^2	0,0059	0,0058
ρ	–	0,5147

Źródło: opracowanie własne

Uzyskane wyniki wskazują, że autokorelacja składnika losowego pozostaje istotna nawet po uwzględnieniu wielu zmiennych taryfowych. Oznacza to, że zasadne jest poszukiwanie zmiennych opisujących jednostkę geograficzną (np. gęstość zaludnienia, liczba pojazdów na mieszkańca itp.), aby lepiej objaśnić badane zjawisko. Istnieje również potrzeba budowy modeli statystycznych, które będą w stanie uwzględnić zaobserwowane własności rozkładu przestrzennego częstości szkód.

Zdolności predykcyjne zbudowanych modeli na zbiorze walidacyjnym zostały ocenione na podstawie dwóch statystyk:

- Błąd średniokwadratowy (BŚK) – średni kwadrat różnicy między teoretyczną z modelu a wartością rzeczywistą dla danej obserwacji.
- *Lift* („podbicie") na poziomie *k%* – iloraz odsetka umów szkodowych (co najmniej jedna zaistniała szkoda) wśród *k%* obserwacji ze zbioru walidacyjnego z najwyższą wartością oczekiwaną wyznaczoną na podstawie modelu oraz odsetka umów szkodowych w całej populacji (trafność całkowicie losowej prognozy).

Wykorzystanie *liftu* jako kryterium oceny modelu zmiennej licznikowej podyktowane jest tym, że dla takich modeli nie można bezpośrednio zbudować macierzy trafności prognoz. Jest to konsekwencją tego, że liczba szkód może być dowolną nieujemną liczbą całkowitą, natomiast ze względu na dużą liczność umów bezszkodowych teoretyczna liczba szkód prognozowana przez model kształtuje się na niskim poziomie, przez co najlepszą prognozą byłby zawsze „brak szkód". Zakładając jednak, że zakład ubezpieczeń chce wykorzystać model do identyfikacji umów o podwyższonym ryzyku zajścia szkody, taka miara wydaje się odpowiednia.

Tabela 6. Podsumowanie dopasowania modeli do danych

Statystyka		Model 1	Model 2	Model 3	Model 4
BŚK	Zbiór uczący	0,0527	0,0528	**0,0526**	**0,0526**
	Zbiór walidacyjny	**0,0539**	**0,0539**	**0,0539**	**0,0539**
Lift	1%	2,4347	**2,6103**	2,5518	2,5518
	5%	1,9595	2,0344	**2,0438**	2,0391
	10%	1,7535	**1,8482**	1,8155	1,8167

Wytłuszczone zostały najlepsze wartości danej statystyki.
Źródło: opracowanie własne

Na podstawie uzyskanych wyników trudno jednoznacznie wskazać najlepszy model. Modele wielopoziomowe są lepiej dopasowane do zbioru uczącego, ale ich przewaga pod tym względem jest bardzo mała, natomiast wszystkie modele są tak samo dopasowane do zbioru walidacyjnego. Z kolei jeśli chodzi o *lift*, to na poziomie 1% i 10% najlepsze prognozy daje model 2, jednakże na poziomie 5% najlepsze prognozy daje model 3. Ostateczny wybór zależałby w tym przypadku od celu analizy, natomiast biorąc pod uwagę typową częstość szkód w ubezpieczeniach OC p.p.m. na poziomie ok. 0,03–0,05, *lift* na poziomie 5% wydaje się bardziej odpowiedni. Porównując modele z efektami losowymi, można zauważyć, że na poziomie 1% i 5% lepszy jest model bez autokorelacji, natomiast przewaga modelu 4 ujawnia się przy wartości statystyki *lift* na poziomie 10%.

Podsumowując, należy stwierdzić, że wszystkie zbudowane modele pozwalają na podniesienie skuteczności prognozowania wystąpienia szkód z tytułu umowy ubezpieczenia OC p.p.m. w stosunku do prognozy losowej. Wymiar geograficzny jest tutaj istotnym czynnikiem taryfikacyjnym – model nieuwzględniający tego elementu okazał się gorszy niż pozostałe rozważane modele. Wśród modeli uwzględniających przestrzenny wymiar danych w pewnych sytuacjach przewagę ma klasyczny uogólniony model liniowy ze zmiennymi opisującymi

region, natomiast w innych – modele wielopoziomowe. Wybór między tymi modelami byłby w praktyce kompromisem między złożonością modelu i jego zdolnością predykcyjną, uwzględniającym cel prowadzonej analizy (np. odsetek umów klasyfikowanych jako umowy o wysokim prawdopodobieństwie szkody). Model ze skorelowanymi efektami losowymi jest lepszy od standardowego modelu wielopoziomowego dopiero dla statystki *lift* na poziomie 10%. Biorąc jednak pod uwagę fakt, że model 3 jest (w przybliżeniu) zagnieżdżony w modelu 4, to można się spodziewać, że istnieją aspekty, w których model 4 daje pełniejszy obraz rzeczywistości niż model 3. Może być to kierunkiem dalszych badań.

5. Podsumowanie i kierunki dalszych badań

W niniejszym artykule przedstawiono zagadnienia dotyczące modelowania częstości szkód w ubezpieczeniach komunikacyjnych OC p.p.m. i AC. Zagadnienia te zostały zbadane zarówno od strony teoretycznej (budowa modeli uwzględniających autokorelację przestrzenną), jak i empirycznej (analiza danych pochodzących z bazy OI UFG).

Przeprowadzona analiza opisowa wykazała, że częstość szkód w ubezpieczeniach komunikacyjnych cechuje się przestrzennym zróżnicowanem i wyraźną autokorelacją przestrzenną.

Interesującym kierunkiem dalszych badań omawianego zagadnienia jest rozwój modeli taryfikacji *a priori* uwzględniających wymiar przestrzenny oraz badanie ich własności teoretycznych i zdolności predykcyjnych. Biorąc pod uwagę uzyskane wyniki analizy opisowej, szczególnie godne uwagi wydają się modele hierarchiczne (województwo – powiat). Istotną kwestią jest również wpływ przyjętych założeń (definicja wymiaru geograficznego, metoda pomiaru odległości, definicja sąsiedztwa) oraz dostępnych danych o jednostkach geograficznych (np. zmienne demograficzne lub ekonomiczne na poziomie województwa lub powiatu) na zdolność predykcyjną budowanych modeli statystycznych. Kolejnym krokiem może być budowa geograficznych zmiennych taryfowych, które w najlepszym stopniu pozwalają oceniać ryzyko ubezpieczeniowe.

Bibliografia

Anselin L., *Local Indicators of Spatial Association – LISA*, „Geographical Analysis" 1995, vol. 27, no. 2, s. 93–115.

Boskov M., Verrall R.J., *Premium rating by geographical area using spatial models*, „ASTIN Bulletin" 1994, vol. 24, iss. 1, s. 131–143.

Brouhns N., Denuit M., Masuy B., Verrall R., *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 2002, s. 1–26.

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

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

Gala K., *Taryfikacja a priori z uwzględnieniem efektów przestrzennych*, „Śląski Przegląd Statystyczny" 2017, nr 15(21), s. 99–113.

Lemaire J., Park S.C., Wang K.C., *The use of annual mileage as a rating variable*, „ASTIN Bulletin" 2016, vol. 46, iss. 1, s. 39–69.

Nelder J.A., Wedderburn R.W.M., *Generalized Linear Models*, „Journal of the Royal Statistical Society" 1972, Series A (General), vol. 135, s. 370–384.

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

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

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

Ustawa z dnia 22 maja 2003 r. o ubezpieczeniach obowiązkowych, Ubezpieczeniowym Funduszu Gwarancyjnym i Polskim Biurze Ubezpieczycieli Komunikacyjnych, tekst jedn.: Dz.U. 2018, poz. 473.

* * *

The *a priori* risk classification with spatial autocorrelation in automobile insurance

Abstract

The subject of this paper is to describe *a priori* risk classification models in automobile insurance which take into consideration the address of the insured. The extension of the generalized linear model with the multi-level factor to account for the correlation between the levels is presented. The analysis of empirical data from the Polish insurance indicates both significant spatial heterogeneity and spatial autocorrelation. Moreover, the inclusion of spatial variables in the risk classification models can improve their accuracy and effectiveness.

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