

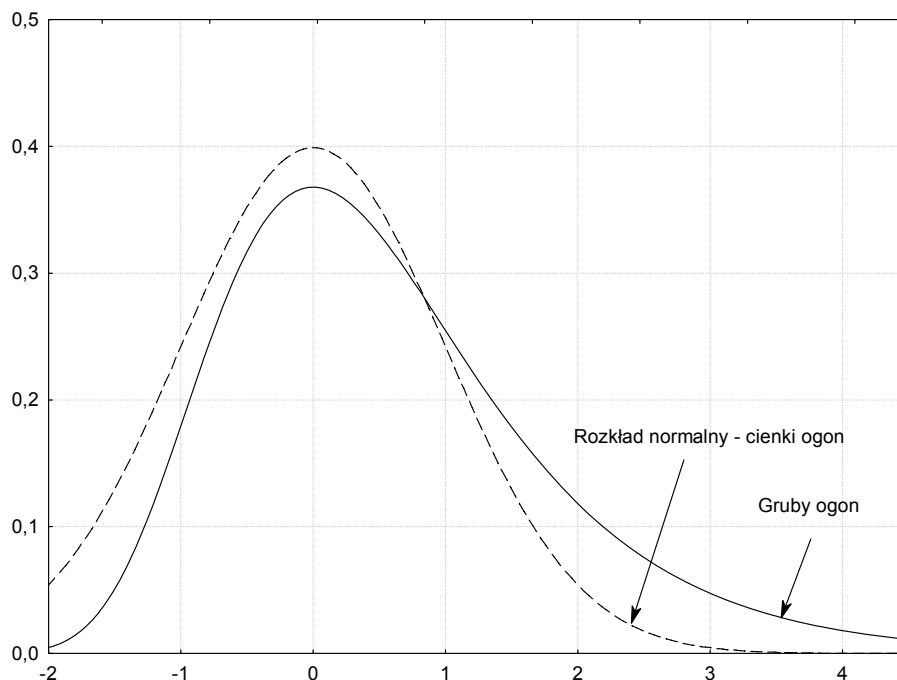


DOPASOWANIE ROZKŁADU I EKSPERYMENT SYMULACYJNY NA PRZYKŁADZIE DANYCH O WYPADKACH DROGOWYCH

Krzysztof Suwada, StatSoft Polska Sp. z o.o.

Często zdarza się, że nie satysfakcjonuje nas sama eksploracja danych, ujawnienie zawartych w nich zależności czy zbudowanie modelu. Dysponując już zbudowanym modelem, chcemy zobaczyć, w jaki sposób badane zjawisko może wyglądać np. w przyszłym roku. Dokonując symulacji, możemy modyfikować parametry modelu, uwzględniając w ten sposób np. nowe uregulowania prawne czy inne wydarzenia mogące mieć wpływ na interesujące nas zjawisko. Przykładem tego typu działań może być sytuacja, która miała miejsce na rynku bankowym, kiedy ze względu na kryzys banki zaostriżyły politykę kredytową.

Takie podejście jest szczególnie cenne przy modelowaniu danych ubezpieczeniowych, w których występują specyficzne typy rozkładów prawdopodobieństwa, tzw. rozkłady o grubych ogonach. W pewnym uproszczeniu można powiedzieć, że rozkład z grubymi ogonami to taki, w którym prawdopodobieństwo otrzymania wartości odległej od średniej jest stosunkowo duże (tzn. większe niż w przypadku rozkładu normalnego).



Rys. 1. Cienkie i grube ogony.



Przeprowadzając analizy, chcielibyśmy uwzględnić nie tylko 99,9% klientów, którzy generują niewysokie szkody, ale i sytuacje ekstremalnie wysokich odszkodowań, których wysokość może sięgać setek tysięcy złotych.

Do tego celu konieczne jest narzędzie, które nie tylko umożliwi zbudowanie modelu, ale i przeprowadzenie eksperymentów symulacyjnych. Takim narzędziem w środowisku *STATISTICA* jest moduł *Rozkłady i symulacje*.

Informacje ogólne

Moduł *Rozkłady i symulacje* jest dostępny z poziomu menu *Statystyka/Podstawowe/Więcej rozkładów*. Składa się on z dwóch części:

- ◆ *Dopasuj rozkład* – umożliwia sprawdzenie, który rozkład najlepiej pasuje do danych, oraz estymację jego parametrów przy zachowaniu korelacji między poszczególnymi zmiennymi.
- ◆ *Uruchom symulację* – umożliwia wygenerowanie nowych danych na podstawie informacji o rozkładzie, zachowując strukturę korelacji między zmiennymi.

Do dyspozycji mamy szereg rozkładów ciągłych, w tym mieszankę rozkładów normalnych, rozkład Johnsona i uogólniony rozkład wartości ekstremalnej oraz dyskretnych. Jako miarę dobroci dopasowania możemy wybrać jeden z trzech testów:

- ◆ Kołmogorowa-Smirnowa,
- ◆ Chi-kwadrat,
- ◆ Andersona-Darlinga (tylko dla zmiennych ciągłych).

Moduł dostarcza także szeregu innych narzędzi służących do oceny dobroci dopasowania (np. histogram z dopasowaniem, dystrybuanta empiryczna) pozwalających wybrać najlepiej dopasowany rozkład.

Ważną cechą modułu jest to, że estymuje on nie tylko parametry dla poszczególnych zmiennych, ale uwzględnia korelacje między nimi. Dzięki zachowaniu struktury korelacyjnej, procedury generujące dane zawarte w module mogą z powodzeniem dostarczyć materiału do analiz typu *what-if*.

Modelowanie

Pochodzenie danych i ich charakterystyka

Wysokości szkód w wypadkach komunikacyjnych wykazują cechy charakterystyczne dla rozkładów o ciężkich ogonach. Zdecydowana większość roszczeń nie jest zbyt wysoka, są to odszkodowania przyznawane w przypadkach stłuczek i niezbyt groźnych wypadków, w których nie zachodzi potrzeba kasacji pojazdu – są to kwoty od kilkuset do kilku tysięcy złotych. Zdarzają się jednak sytuacje, gdy odszkodowanie musi pokryć zobowiązania nie tylko dotyczące naprawy samochodu, ale i kosztów leczenia ofiar wypadków, lub gdy



uszkodzony został samochód luksusowy wart kilkaset tysięcy złotych. Dane dotyczące wypadków drogowych pochodzą ze strony internetowej Policji. Zawierają one informacje o liczbie wypadków oraz liczbie osób zabitych i rannych.

Do dyspozycji mamy informacje miesięczne. Jak łatwo zauważyć, liczba wypadków w poszczególnych miesiącach nie jest stała i ten fakt należy uwzględnić w modelowaniu. Także zmienne dotyczące liczby wypadków oraz osób rannych i zabitych nie są niezależne.

	1	2	3	4	5			
	Rok	Miesiąc	Wypadki	Zabici	Ranni			
1	2007	Styczeń	4127	503	5029			
2	2007	Luty	2854	307	3631			
3	2007	Marzec	3579	366	4495			
4	2007	Kwiecień	3784	396	4829			
5	2007	Maj	4244	402	5540			
6	2007	Czerwiec	4288	467	5742			
7	2007	Lipiec	4537	541	6047			
8	2007	Sierpień	4395	509	5782			
9	2007	Wrzesień	4517	529	5796			
10	2007	Październik	4603	527	5816			
11	2007	Listopad	4365	492	5349			
12	2007	Grudzień	4243	544	5168			
13	2008	Styczeń	2528	252	3150			
14	2008	Luty	2599	237	3346			
15	2008	Marzec	2940	318	3729			
16	2008	Kwiecień	3522	366	4564			
17	2008	Maj	4049	417	5167			
18	2008	Czerwiec	4361	394	5547			
19	2008	Lipiec	4576	514	6070			
20	2008	Sierpień	4233	512	5566			
21	2008	Wrzesień	4527	538	5631			
22	2008	Październik	4588	568	5656			
23	2008	Listopad	4151	512	4919			
24	2008	Grudzień	4802	615	5778			

Rys. 2. Liczba wypadków.

Informacje o wysokościach szkód zostały na potrzeby prezentacji wygenerowane w programie *STATISTICA* z pewnego rozkładu o ciężkich ogonach. Arkusz zawiera prawie 60 000 przypadków, czyli mniej więcej tyle, ile wypadków rocznie zdarza się na polskich drogach. Dodatkowo arkusz zawiera zmienną *Szkoda zgłoszona (w tys PLN)*, która ma oddawać sytuację, gdy poszkodowany z pewnym prawdopodobieństwem decyduje się nie zgłaszać szkody, jeżeli kwota jest niewielka – w tym przypadku 500 zł.



	1 Szkoda rzeczywista	2 Szkoda zgloszona (tys PLN)
1	6,49418658	6,49
2	3,34693565	3,35
3	0,989205807	0,99
4	1,55835104	1,56
5	1,20301648	1,2
6	2,50559203	2,51
7	3,63948674	3,64
8	2,34933036	2,35
9	1,25833748	1,26
10	3,52957365	3,53
11	2,41887646	2,42
12	2,74415721	2,74
13	2,88084786	2,88
14	2,34751496	2,35
15	4,23464719	4,23
16	1,75886944	1,76
17	1,92814619	1,93
18	6,58139932	6,58

Rys. 3. Wysokości szkód.

Opis modelu

W celu opisu wysokości szkód za dany okres czasu stosować będziemy zmienne losowe o rozkładach złożonych postaci:

$$Y = X_1 + X_2 + \dots + X_N,$$

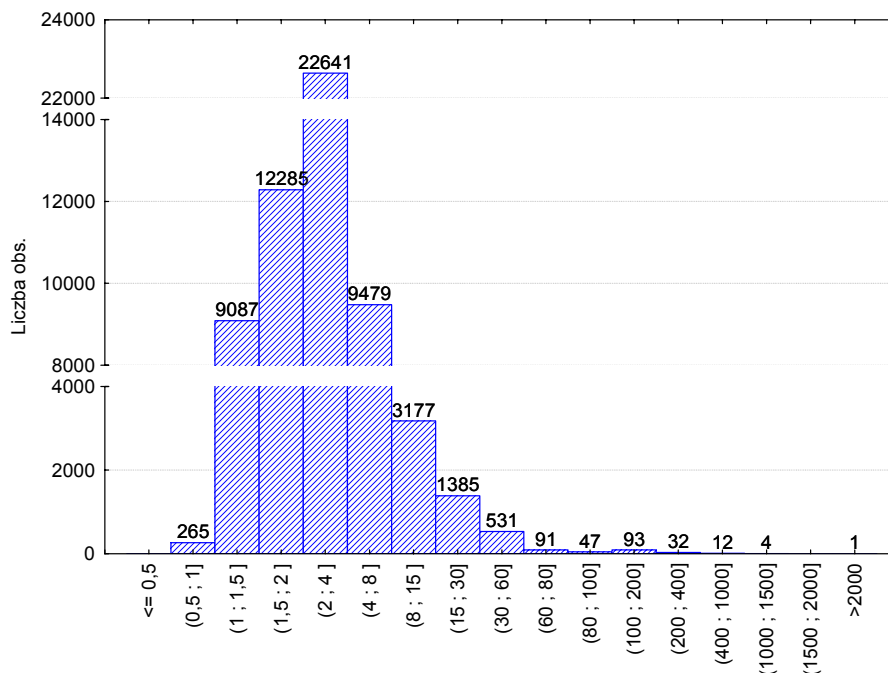
gdzie X_i są zmiennymi niezależnymi o zadanym rozkładzie, a N jest zmienną losową dyskretną niezależną od X_i . Do modelowania wysokości pojedynczej szkody dobrze nadają się np. mieszanki rozkładów gaussowskich i uogólniony rozkład wartości ekstremalnej, a do modelowania liczby szkód w danym miesiącu, np. rozkład Poissona.

Na dalszym etapie w programie *STATISTICA* do danych empirycznych zostanie dopasowany cały zestaw rozkładów ciągłych i dyskretnych. Na podstawie wskaźników dobroci dopasowania i wiedzy biznesowej dobrane zostaną rozkłady dla zmiennych X_i oraz N .

Estymacja parametrów

Rozkład wysokości pojedynczej szkody

Liczba danych, którymi dysponujemy, jest dosyć duża, co z jednej strony jest sytuacją pożądaną, a z drugiej może stać się źródłem problemów numerycznych. Dla bardzo dużych zbiorów danych warto rozważyć ograniczenie się do pewnej losowej próbki. W tym przykładzie wykorzystany zostanie pełny zbiór danych (prawie 60 000 przypadków) pochodzący z pewnego rozkładu ciągłego.



Rys. 4. Rozkład wysokości pojedynczej szkody.

Bardzo często zdarza się także, że ubezpieczeni nie zgłaszają małych szkód, ponieważ skutkowałoby to utratą zniżek na OC/AC, ale także tę sytuację będziemy chcieli uwzględnić w naszej symulacji. Jako próg „małej szkody” przyjmujemy kwotę 500 zł. Szkoda ta jest zgłaszana losowo z prawdopodobieństwem 0,2.

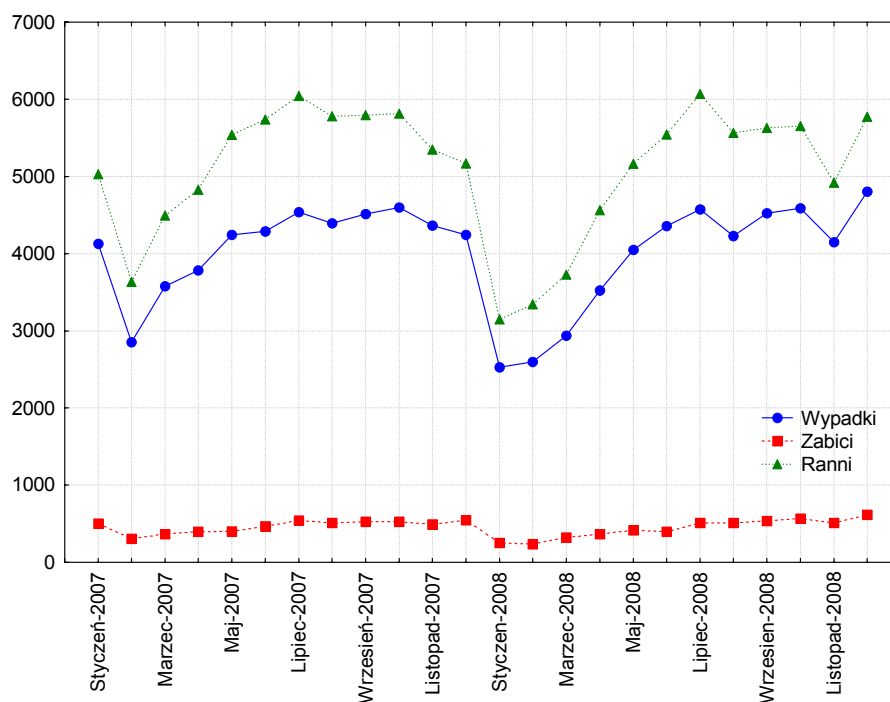
Jak widać na histogramie, sytuacja jest charakterystyczna dla rozkładów z ciężkimi ogonami. Raz na kilkadziesiąt tysięcy zdarza się bardzo wysoka szkoda rzędu 2 mln złotych.

Po uruchomieniu modułu do dopasowywania rozkładów i przejrzaniu wyników estymacji do dalszego modelowania wybieramy rozkład uogólniony wartości ekstremalnej.

Rozkład liczby szkód

Dysponujemy danymi od 2007 r. i na tej podstawie chcemy zidentyfikować rozkład liczby szkód i estymować jego parametry. Ze względu na małą liczbę danych zakładamy, że rozkład liczby wypadków nie zmienia się z miesiąca na miesiąc. Na potrzeby tego przykładu wykorzystywana będzie tylko informacja o liczbie wypadków, ale dopasowany i symulowany będzie łączny rozkład razem z liczbą zabitych i rannych.

Jak widać na rys. 5, dane są silnie skorelowane, co potwierdza także macierz korelacji – należy o tym pamiętać w momencie uruchamiania symulacji.



Rys. 5. Liczba wypadków oraz zabitych i rannych w wypadkach drogowych w latach 2007 i 2008.

Tabela 1. Korelacje między zmiennymi.

	Wypadki	Zabici	Ranni
Wypadki	1,000000	0,925910	0,977618
Zabici	0,925910	1,000000	0,868995
Ranni	0,977618	0,868995	1,000000

Po uruchomieniu modułu do dopasowania rozkładów i przejrzeniu wyników, uwzględniając wiedzę biznesową do modelowania, wybieramy rozkład Poissona.

Symulacja

Generowanie danych

Na tym etapie dysponujemy już informacjami o tym, jakie są rozkłady poszczególnych zmiennych i jakie są między nimi zależności. Naszym celem jest określenie, jakie środki należy przeznaczyć na pokrycie zobowiązań ubezpieczyciela z tytułu wypłat z polis OC/AC.

Okres, za jaki chcemy otrzymać pewnego rodzaju prognozę, to jeden rok, musimy więc wylosować dwunastoelementową próbkę z rozkładu liczby szkód. Kolejnym krokiem jest wygenerowanie realizacji zmiennej losowej, która określi nam wysokość szkody.

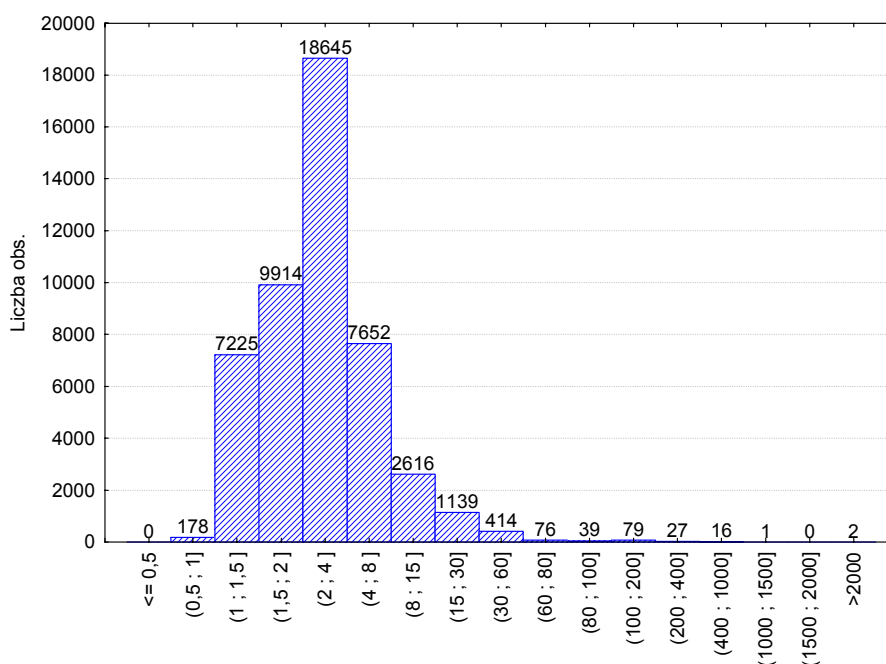


Przegląd wyników

Dla kalkulacji rezerw najważniejszym czynnikiem jest dobre oszacowanie rozkładu wysokości szkody. W naszym przypadku eksperyment symulacyjny daje następujące wyniki.

Tabela 2. Korelacje między zmiennymi.

	Wypadki	Zabici	Ranni
Wypadki	1,000000	0,840401	0,964127
Zabici	0,840401	1,000000	0,784827
Ranni	0,964127	0,784827	1,000000



Rys. 6. Wysymulowany rozkład wysokości pojedynczej szkody.

Widać, że wyniki są bardzo podobne do wyjściowego „rzeczywistego” rozkładu. Ujawniają się także grube ogony, co jest własnością bardzo pożądaną.

Co dalej

Wygenerowaliśmy dane według najprostszego z możliwych scenariuszy – przyszłość będzie bardzo podobna do przeszłości. Dużo bardziej interesująca jest zmiana parametrów symulacji tak, aby odzwierciedlały one różne mniej lub bardziej prawdopodobne scenariusze, np.:

- ◆ wzrost wartości samochodów – akcja rządowych dopłat do wymiany aut starszych niż 10 lat,
- ◆ spadek wartości samochodów,
- ◆ obniżenie lub podwyższenie wieku kierowców,



- ◆ zwiększenie liczby wypadków na drogach,
- ◆ trend wzrastający lub malejący liczby wypadków,
- ◆ drastyczne zwiększenie cen benzyny – mniejsza liczba samochodów na drogach.

Dzięki takim symulacjom możemy przygotować odpowiednie procedury na wypadek różnych scenariuszy, zanim one wystąpią. Symulacje pozwalają także sprawdzić, w jaki sposób wysokość środków zabezpieczonych na wypłatę odszkodowań powinna zmieniać się wraz ze zmianami na rynku, i pomaga odpowiedzieć na wiele pytań, wspierając wiedzę ekspercką nowoczesnym mechanizmem probabilistycznym.

Wiedza zdobyta w ten sposób jest bardzo cenna, co potwierdzają także nowoczesne regulacje prawne. Na przykład dla instytucji z sektora bankowego zarządzających w ten sposób ryzykiem wymagania kapitałowe są niższe i bardziej elastyczne, co może dać istotną przewagę nad konkurencją.

Podobne podejście można zastosować do bardzo wielu innych dziedzin, np.:

- ◆ przewidywanie rozregulowania maszyny w fabryce,
- ◆ szacowanie liczby pomyłek urzędników, które skutkują pozwami, i zabezpieczanie odpowiednich środków na odszkodowania,
- ◆ zabezpieczanie odpowiedniej liczby np. zapasowych komputerów, laptopów czy innych urządzeń lub środków finansowych na ich zakup,
- ◆ szacowanie szkód powodowanych przez drobne uszkodzenia magazynowanego sprzętu z uwzględnieniem zmian w asortymencie.

Podsumowanie

Moduł do dopasowywania rozkładów i symulacji jest bardzo użytecznym narzędziem w rękach sprawnego analityka. Pozwala przeprowadzać symulacje rozmaitych scenariuszy, według których można później testować rozmaite modele.

Moduł ten może stać się bardzo cenną częścią systemu do analizy i zarządzania ryzykiem, dzięki możliwości pełnej automatyzacji pracy (VBA), stając się narzędziem do generowania raportów *what-if* i *stress test*. W połączeniu z wiedzą biznesową i odpowiednią infrastrukturą informatyczną może stać się częścią zaawansowanego systemu, dając znaczącą przewagę nad konkurencją, dzięki zmniejszeniu wymaganych prawnie rezerw kapitałowych.

Literatura

1. Panjer, Harry H., *Operational Risk: Modeling Analytics*, Wiley & Sons, 2006.
2. Chernobai, Anna S., Rachev, Svetlozar T., Fabozii, Frank J., *Operational Risk: A guide to BASEL II Capital Requirements, Models, and Analysis*, Wiley & Sons, 2007.
3. www.policja.pl.