Współczesna Gospodarka

Contemporary Economy
Electronic Scientific Journal
www.wspolczesnagospodarka.pl

Vol. 6 Issue 4 (2015) 123-134
ISSN 2082-677X

PRZESTRZENNO – CZASOWA ANALIZA NASILENIA WYPADKÓW DROGOWYCH W POLSCE

Kinga Kądziołka

Streszczenie

W artykule dokonano identyfikacji skupień powiatów charakteryzujących się ponadprzeciętną liczbą i natężeniem wypadków drogowych oraz podjęto próbę prognozowania dziennej liczby wypadków drogowych w Polsce z wykorzystaniem lasów losowych i sztucznych sieci neuronowych. Porównano błędy sieci, w której uwzględniono wszystkie analizowane zmienne objaśniające z siecią, w której liczbę zmiennych wejściowych zredukowano w oparciu o wykres ważności zmiennych, uzyskany dla lasu losowego. Najmniejszym przeciętnym absolutnym błędem procentowym na zbiorze treningowym i testowym charakteryzował się las losowy.

Słowa kluczowe: wypadki drogowe, dane dzienne, sieć neuronowa, las losowy

Wstęp

Wypadek drogowy to „zdarzenie drogowe, które pociągnęło za sobą ofiary w ludziach, w tym także u sprawcy tego zdarzenia, bez względu na sposób zakończenia sprawy[1]". Wypadki drogowe stanowią problem społeczny generujący koszty związane zarówno z leczeniem ofiar, naprawą pojazdów, odszkodowaniami, koszty postępowań sądowych, ale także koszty niematerialne związane z cierpieniem ofiar wypadków i ich rodzin. W Polsce roczne straty z tytułu kosztów wypadków drogowych szacowane są na około 2% PKB[2].

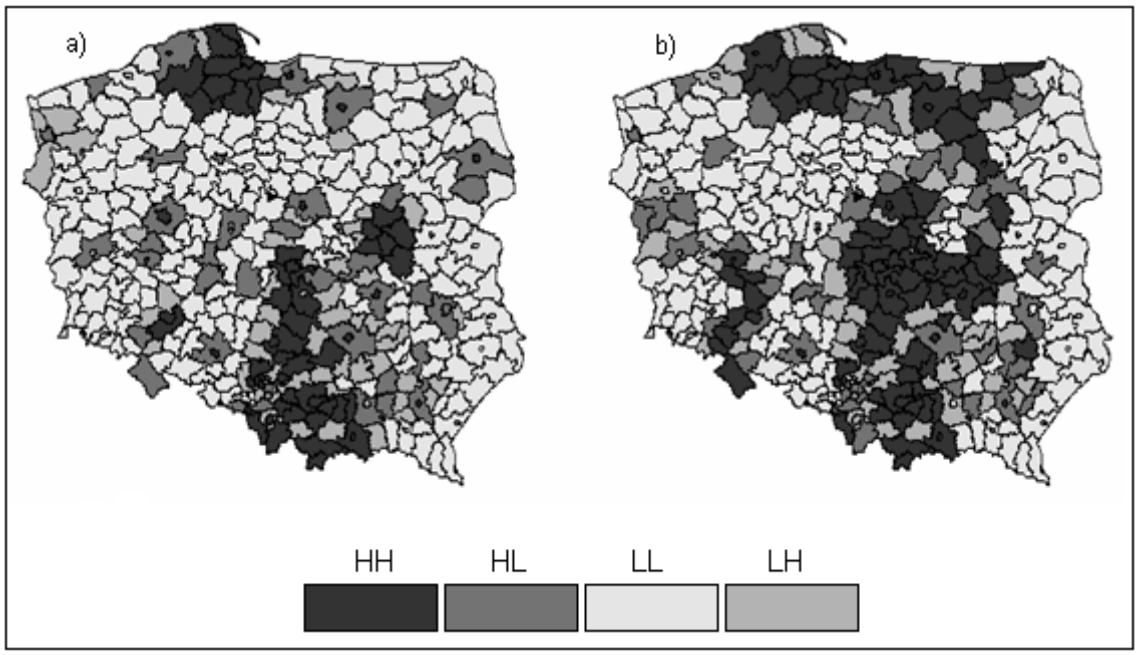
Aby porównywać zagrożenie wypadkami na obszarach różniących się powierzchnią i liczbą mieszkańców, można wykorzystywać takie wskaźniki jak gęstość wypadków[3] lub liczba wypadków na 100 tys. mieszkańców (wskaźnik ten nazwano w tym artykule natężeniem wypadków). Rysunek 1 przedstawia przestrzenne zróżnicowanie powiatów pod względem liczby i natężenia wypadków drogowych w 2014 r. Obszary podzielono na cztery grupy we-

[1] cyt. *Stan bezpieczeństwa ruchu drogowego oraz działania realizowane w tym zakresie w 2014 r.*, www.krbrd.gov.pl, 20.10.2015, Krajowa Rada Bezpieczeństwa Ruchu Drogowego, s. 2.

[2] por. J. Barcik., P., Czech, *The influence of road infrastructure on safety of road traffic – part 2*, Transport No. 69, Zeszyty Naukowe Politechniki Śląskiej, 2010, s. 16.

[3] Gęstość wypadków to liczba wypadków na 100 km dróg, por. *Stan bezpieczeństwa ruchu drogowego...*, s. 2.

dług przynależności do ćwiartek moranowskiego wykresu rozproszenia, co umożliwiło identyfikację skupień obszarów charakteryzujących się ponadprzeciętną liczbą/natężeniem zdarzeń. Moranowski wykres rozproszenia stanowi graficzną prezentację globalnej statystyki Morana. Na osi X znajduje się standaryzowana wartość analizowanej zmiennej a na osi Y standaryzowana zmienna opóźniona przestrzennie[4]. Obszary położone w I ćwiartce (ozn. HH, ang. *high – high*) charakteryzują się wysokimi wartościami badanej zmiennej oraz otoczone są sąsiadami, w których ta zmienna również przyjmuje wartości wysokie. Obszary położone w II ćwiartce (ozn.HL, ang. *high-low*) charakteryzują się wysokimi wartościami badanej zmiennej i są otoczone sąsiadami o niskich wartościach tej zmiennej. Obszary położone w III ćwiartce (ozn. LL, ang. *low-low*) charakteryzują się niskimi wartościami badanej zmiennej oraz otoczone są sąsiadami, w których ta zmienna również przyjmuje niskie wartości. Obszary położone w IV ćwiartce (ozn. LH, ang. *low – high*) charakteryzują się niskimi wartościami badanej zmiennej i otoczone są sąsiadami o wysokich wartościach tej zmiennej.



Rysunek 1. Podział powiatów wg moranowskiego wykresu rozproszenia: a) – liczba wypadków drogowych w 2014 r., b) – natężenie wypadków drogowych w 2014 r.

Źródło: opracowanie własne na podstawie danych GUS (Bank Danych Lokalnych).

Największym natężeniem wypadków drogowych w 2014 r. charakteryzowały się powiaty: m. Łódź (254), m. Rzeszów (220), powiat radomszczański (208), m. Częstochowa (201). Największa liczba wypadków drogowych miała miejsce w Łodzi (1791). Kolejne dwa miejsca przypadły na miasta: Kraków (1147) i Warszawę (1111).

K. Jamroz i W. Kustra (2010) wykorzystując dane przekrojowe badali (stosując uogólnione modele regresji liniowej) wpływ wybranych czynników (m.in.: typ drogi, średnioroczne dobowe natężenie ruchu, liczba pasów, procent długości terenu zabudowanego, udział pojazdów ciężarowych) na gęstość ofiar śmiertelnych na drogach krajowych w Polsce.

[4] Operator opóźnienia przestrzennego jest średnią ważoną z wartości zmiennej w regionach sąsiednich, zgodnie z zadeklarowaną macierzą wag, por. K. Kopczewska, *Ekonometria i statystyka przestrzenna z wykorzystaniem programu R CRAN*, Warszawa 2011, s. 70. Tutaj wykorzystana została binarna macierz wag standaryzowana wierszami zdefiniowana na podstawie kryterium wspólnej granicy.

Podejmowane były również analizy szeregów czasowych o częstotliwości rocznej i miesięcznej dotyczących wypadków drogowych w Polsce oraz porównywano wybrane charakterystyki bezpieczeństwa w Polsce w ramach poszczególnych województw a także na tle wybranych krajów. Badania takie prowadzili m. in. M. Dębowska – Mróz (2010), K. Chudy - Laskowska, T. Pisula (2014) oraz A. Wójcik (2015).

W tym artykule głównym przedmiotem analiz będą szeregi czasowe obrazujące dzienną liczbę wypadków drogowych w Polsce. Są to dane o złożonej sezonowości. Można zaobserwować większą liczbę zdarzeń w miesiącach letnich, która wynika ze wzmożonego ruchu drogowego, co zwiększa prawdopodobieństwo wypadku (rysunek 2). Ponadto przeciętnie najwięcej wypadków ma miejsce w piątki, soboty i poniedziałki a najmniej w niedziele (tablica 1). Większa liczba wypadków w piątki może mieć związek ze wzmożonym ruchem (np. wyjazdy wypoczynkowe). W niedziele też występuje wzmożony ruch, gdyż wiele osób wraca do domu po południu lub wieczorem. Przyczyną mniejszej liczby wypadków w niedziele może być skumulowanie tego ruchu, co może przyczyniać się do spowolnienia jazdy. Wpływ na liczbę wypadków ma też między innymi okres przedświąteczny i związane z tym wyjazdy na Boże Narodzenie, Wielkanoc, Wszystkich Świętych. Istotna jest też godzina. W Polsce najwięcej wypadków zdarza się w godzinach popołudniowych i wieczornych (między godziną 16 a 20). Najmniej wypadków ma miejsce w nocy i wczesnym rankiem – od północy do szóstej[5].

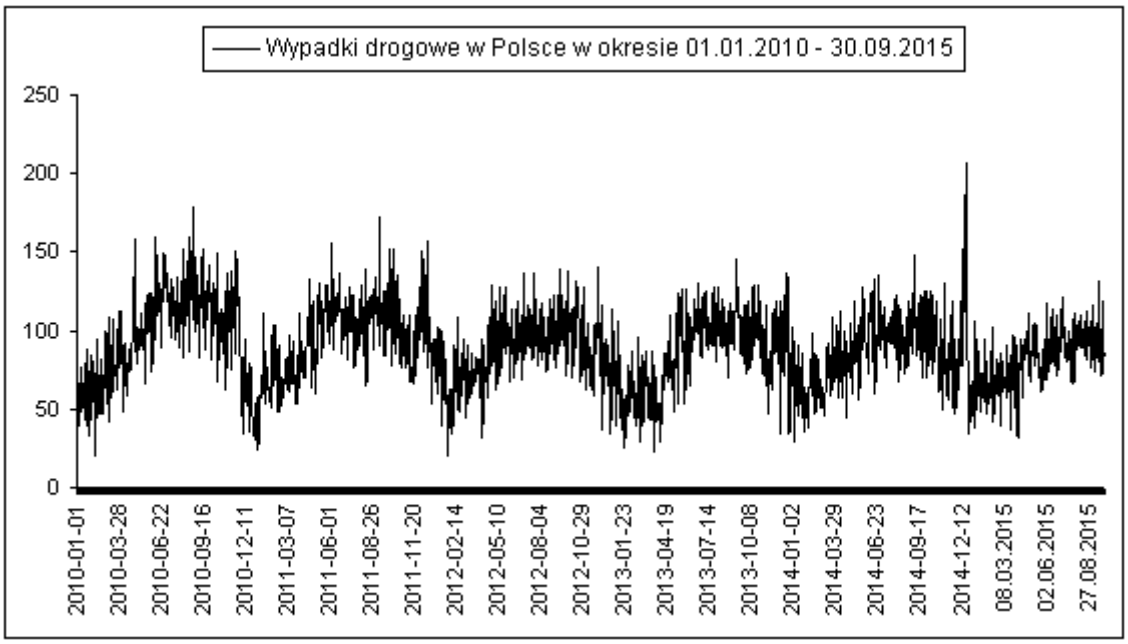
Tablica 1. Przeciętna liczba wypadków drogowych w Polsce

Dzień tygodnia	**Przeciętna liczba wypadków drogowych w Polsce w okresie 01.01.2010-30.09.2015**
Poniedziałek	90
Wtorek	87
Środa	87
Czwartek	87
Piątek	98
Sobota	91
Niedziela	79

Źródło: opracowanie własne na podstawie danych www.policja.pl.

Podjęta zostanie próba zastosowania sieci neuronowych i lasów losowych do prognozowania dziennej liczby wypadków drogowych w Polsce. Metody te, w przeciwieństwie do metod ekonometrycznych, nie zakładają znajomości rozkładów cech ani postaci analitycznej związku między nimi. W przypadku lasów losowych nie jest także wymagane dokonywanie specyfikacji predyktorów, jakie należy uwzględnić w modelu, gdyż dobór zmiennych objaśniających następuje automatycznie (w oparciu o przyjęte wcześniej kryterium). Ponadto wśród zmiennych objaśniających w przypadku lasów losowych mogą znajdować się zarówno zmienne ilościowe jak też jakościowe i nie ma potrzeby dokonywania ich przekształceń[6]. Modele budowane będą na podstawie tzw. zbioru treningowego, stanowiącego podzbiór całego zbioru danych, natomiast oceniane będą na podstawie działania na danych, które nie zostały wykorzystane do budowy modelu (tzw. zbiór testowy). Prezentowane obliczenia wykonano z wykorzystaniem darmowego programu R.

[5] por. B. Hołyst, *Kryminologia*, Wydawnictwo Prawnicze LexisNexis, Warszawa 2007, s. 695.
[6] por. E. Gatnar, *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, 2001, s. 8.



Rysunek 2. Wypadki drogowe w Polsce w okresie 01.01.2010 – 30.09.2015

Źródło: opracowanie własne na podstawie danych www.policja.pl.

1. Prognozowanie dziennej liczby wypadków drogowych z wykorzystaniem lasów losowych

Dane dzienne dotyczące wypadków drogowych były przedmiotem analizy w pracy K. Kądziołki (2015), w której podjęto próbę modelowania złożonej sezonowości z wykorzystaniem modeli ekonometrycznych ze zmiennymi 0-1 oraz drzew regresyjnych[7]. Wadą drzew regresyjnych jest to, że uzyskiwane prognozy mają skokowy charakter oraz niestabilność, co oznacza, że nawet niewielka zmiana zbioru uczącego może doprowadzić do całkiem innej sekwencji podziału[8]. W ograniczeniu zjawiska niestabilności oraz braku ciągłości prognoz może być pomocna metoda lasów losowych, polegająca na wykorzystaniu wielu różnych drzew regresyjnych.

Drzewo regresyjne to graf spójny, acykliczny, który stanowi graficzną prezentację modelu postaci[9]: $y = f(\mathbf{x}_i) = \sum_{k=1}^{K} \alpha_k \mathbf{I}(\mathbf{x}_i \in R_k)$, gdzie y – zmienna zależna, R_k – segment przestrzeni zmiennych objaśniających, α_k - parametry modelu ($k=1,...,K$), **I** – funkcja wskaźnikowa określona następująco: **I**(q) = 1, gdy warunek q jest prawdziwy oraz **I**(q) = 0 w przeciwnym przypadku. Parametry α_k wyznaczane są następująco: $\alpha_k = \frac{1}{N(k)} \sum_{\mathbf{x}_i \in R_k} y_i$, gdzie *N(k)* – liczba elementów znajdujących się w segmencie R_k, y_i - wartości przyjmowane przez zmienną zależną w

[7] por. K. Kądziołka, *Determinanty przestępczości w Polsce. Aspekt ekonomiczno – społeczny w ujęciu modelowania ekonometrycznego*, niepublikowana rozprawa doktorska, Uniwersytet Ekonomiczny w Katowicach, 2015.
[8] por. K. Fijorek i in., *Prognozowanie cen energii elektrycznej na rynku dnia następnego metodami data mining*, „Rynek energii" 12/2010.
[9] por. E. Gatnar, *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji, Warszawa 2008*, s. 37- 44.

segmencie R_k. Szczegółowe informacje dotyczące metod budowy drzew regresyjnych i klasyfikacyjnych można znaleźć m. in. w pracach E. Gatnara (2001, 2008).

W podstawowej wersji algorytm lasu losowego działa według następującego schematu[10]:

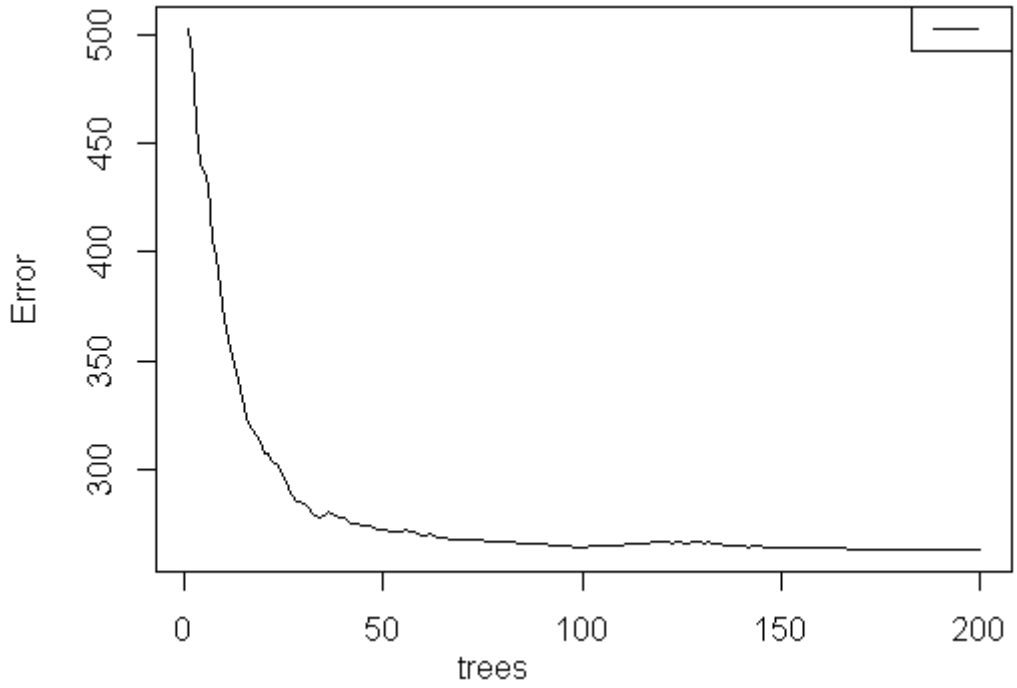
1) Ustal liczbę modeli bazowych (tutaj drzew regresyjnych) M oraz liczbę zmiennych K
2) Dla każdego j=1,...,M wykonaj następujące kroki:
    a. Wylosuj próbę uczącą U_j ze zbioru treningowego
    b. Zbuduj maksymalne drzewo D_m na podstawie próby U_m, losując w każdym węźle drzewa K zmiennych, spośród których najlepsza dobierana jest do modelu
3) Dokonaj predykcji modelu zagregowanego stosując uśrednianie wyników predykcji wszystkich M modeli

W wykorzystanym modelu lasu losowego zmienną objaśnianą była liczba wypadków drogowych w chwili *t*. Analizowano dane dzienne za okres 01.01.2010-30.09.2015. Uwzględniono 38 zmiennych objaśniających:

wyp_1 – liczba wypadków drogowych w chwili *t*-1
wyp_2 – liczba wypadków drogowych w chwili *t*-2
...
wyp_31 – liczba wypadków drogowych w chwili *t*-31
dzien – zmienna określająca dzień tygodnia przyjmująca wartości ze zbioru: {pon, wt, sr, czw, pt, sob, niedz}
swieto – zmienna przyjmująca wartość „tak" jeśli dany dzień (*t*) jest dniem świątecznym oraz „nie" w przeciwnym przypadku
swieto_1 – zmienna przyjmująca wartość „tak" jeśli w chwili *t*-1 był dzień świąteczny oraz „nie" w przeciwnym przypadku
nast_swieto - zmienna przyjmująca wartość „tak" jeśli w chwili *t*+1 będzie dzień świąteczny oraz „nie" w przeciwnym przypadku
mc – zmienna określająca miesiąc, która przyjmuje wartości ze zbioru {sty, luty,...,gru}
dzien_w_mies – zmienna określająca, który jest dzień miesiąca w chwili *t*
t – numer dnia (*t*=32,...,2039; uwzględniając opóźnienia zmiennej objaśnianej pominięto pierwsze 31 obserwacji)

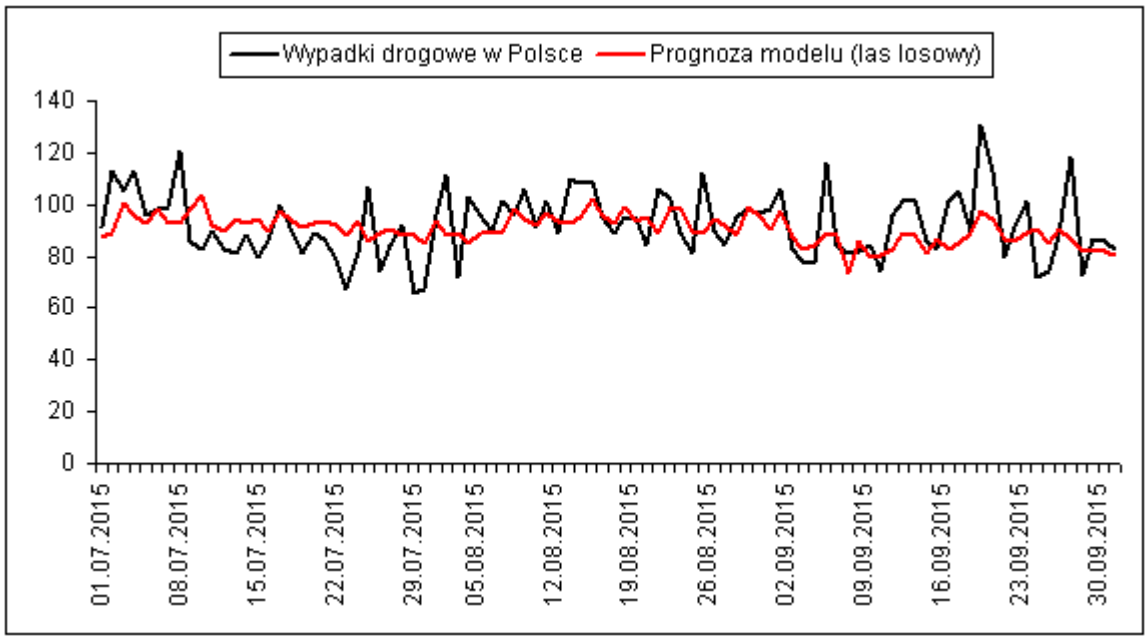
Wykorzystano las losowy zbudowany z 200 drzew regresyjnych (tj. M=200). Na każdym etapie konstrukcji drzew wybierano w sposób losowy 13 zmiennych (tj. K=13) spośród 38 zmiennych objaśniających. Zbiór treningowy stanowiły dane z okresu 01.02.2010 – 30.06.2015, natomiast pozostałe obserwacje za okres 01.07.2015 – 30.09.2015 stanowiły zbiór testowy. Dla uzyskanego modelu przeciętny absolutny błąd procentowy na zbiorze treningowym wynosił 6,62% a na zbiorze testowym 10,53%. W przypadku wykorzystania pojedynczego drzewa regresyjnego przeciętny absolutny błąd procentowy na zbiorze treningowym wynosił 16,92% a na zbiorze testowym 13,65%. W porównaniu z pojedynczym drzewem, błędy lasu losowego były mniejsze i nie występował problem skokowych wartości prognoz. Rysunek 3 przedstawia wartość błędu lasu losowego w zależności od liczby drzew regresyjnych. Rysunek 4 przedstawia dane rzeczywiste dotyczące dziennej liczby wypadków drogowych w Polsce dla zbioru testowego oraz wartości teoretyczne uzyskane z wykorzystaniem lasu losowego.

[10] Por. E. Gatnar, *Podejście wielomodelowe...*, s. 158.



Rysunek 3. Wartość błędu lasu losowego w zależności od liczby drzew regresyjnych

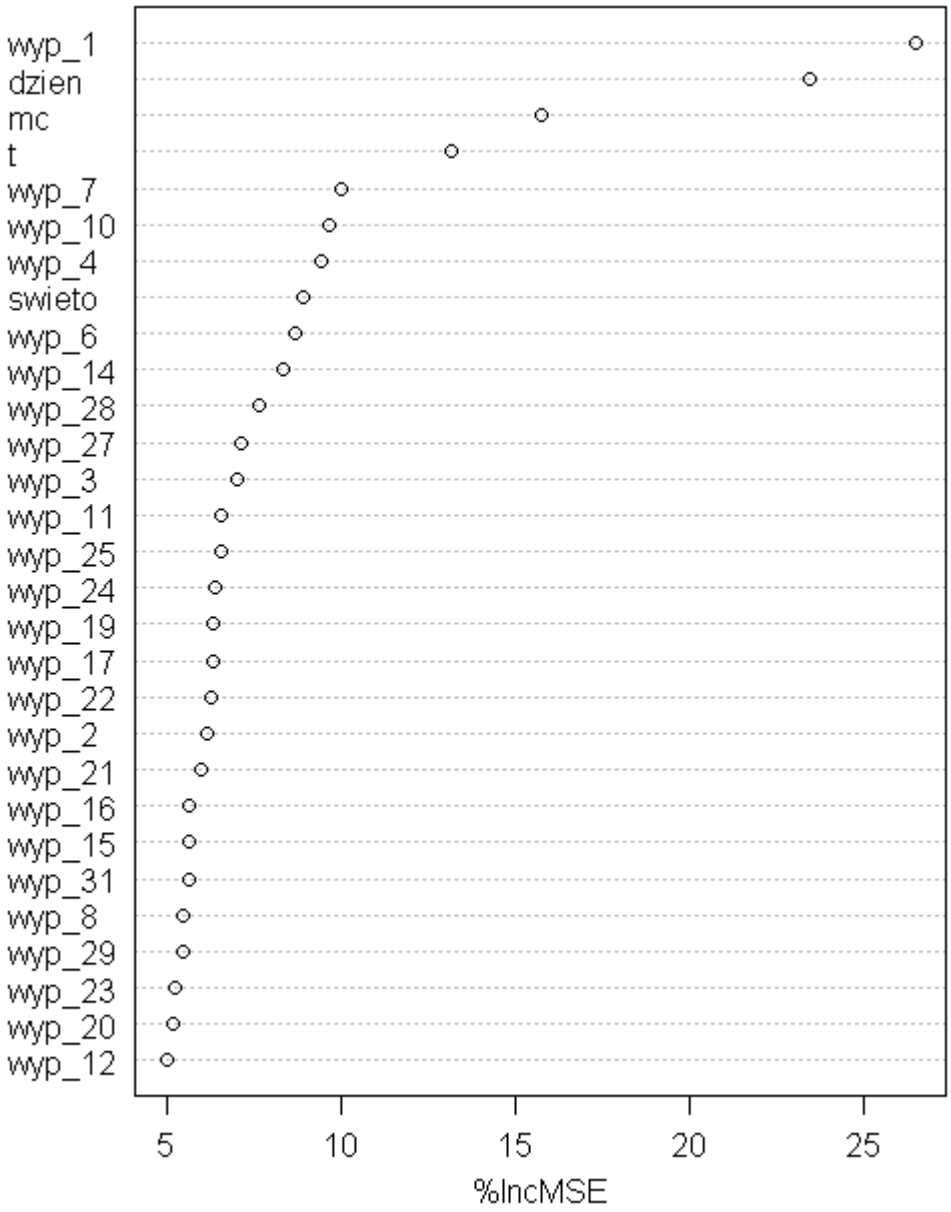
Źródło: opracowanie własne.



Rysunek 4. Wypadki drogowe w Polsce – dane empiryczne i teoretyczne

Źródło: opracowanie własne.

Algorytm RandomForest umożliwia wygenerowanie tzw. wykresu ważności zmiennych, który obrazuje stopień przyrostu błędu prognozy, gdy daną zmienną wyłączy się z modelu[11]. Rysunek 5 przedstawia wykres ważności zmiennych dla analizowanego problemu. Przedstawiono tylko te zmienne objaśniające, dla których %IncMSE>5.



Rysunek 4. Wykres ważności zmiennych objaśniających

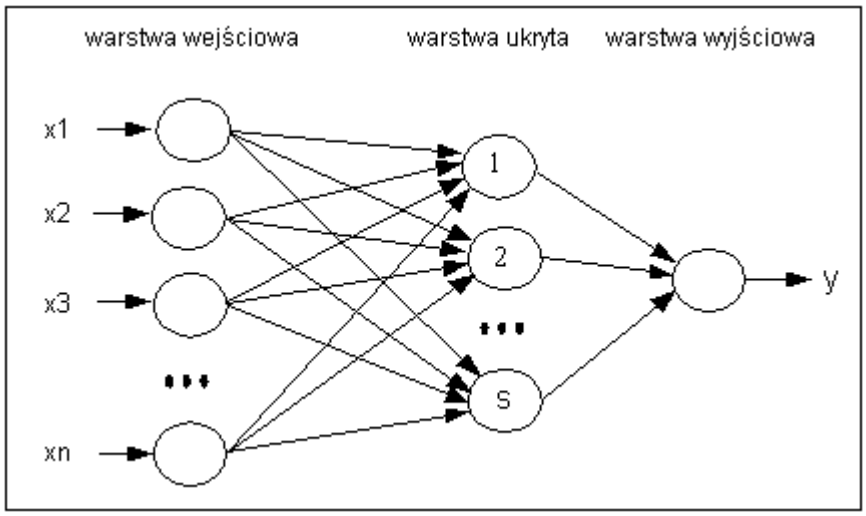
Źródło: opracowanie własne.

Uzyskany ranking zmiennych zostanie następnie wykorzystany do zredukowania liczby zmiennych wejściowych w przypadku modeli sieci neuronowych.

[11] por. K. Fijołek i in., *Prognozowanie cen energii…*

2. Prognozowanie dziennej liczby wypadków drogowych z wykorzystaniem sieci neuronowych

Do prognozowania liczby wypadków drogowych w Polsce wykorzystano perceptron wielowarstwowy z jedną warstwą ukrytą, sigmoidalnymi funkcjami aktywacji w warstwie ukrytej oraz liniowym wyjściem. W pracach zagranicznych autorów podejmujących tematykę modelowania dziennej liczby wypadków drogowych z wykorzystaniem sieci neuronowych wykorzystywano m. in. ten właśnie rodzaj sieci[12]. Celem umożliwienia porównania wyników uzyskanych za pomocą sieci neuronowych i lasu losowego wykorzystywano ten sam zbiór treningowy i testowy. Rysunek 5 przedstawia architekturę perceptronu z jedną warstwą ukrytą.



Rysunek 5. Architektura wielowarstwowego perceptronu z jedną warstwą ukrytą

Źródło: opracowanie własne.

Liczbę neuronów w warstwie ukrytej (ozn. s) ustalano jako średnią geometryczną liczby neuronów na wejściu i wyjściu sieci. Wyjście każdego elementu warstwy poprzedniej połączone było z wejściem każdego elementu warstwy następnej. Zadaniem elementów warstwy wejściowej jest wstępne przetworzenie sygnałów wejściowych (np. normalizacja lub skalowanie danych). Zasadnicze przetwarzanie neuronowe odbywa się w warstwach ukrytych oraz w warstwie wyjściowej[13]. Przed rozpoczęciem procesu uczenia sieci, zmienne wejściowe znormalizowano za pomocą następującej formuły: $x_{ij}^{*} = \frac{x_{ij} - \min_j \{x_{ij}\}}{\max_j \{x_{ij}\} - \min_j \{x_{ij}\}}$. Porównano przeciętne błędy procentowe dla dwóch rodzajów sieci. Pierwszą była sieć, w której uwzględniono na wejściu wszystkie zmienne objaśniające. Dla drugiej sieci wykorzystano

[12] m.in. prace: H.F. Bayata i in., *Modeling of monthly traffic accidents with the artificial neural network method*, International Journal of the Physical Sciences Vol. 6(2), 2011; Z. Quiang i in., *Traffic Accidents Forecasting Based on Neural Network and Principal Component Analysis*, Research Journal of Applied Sciences, Engineering and Technology 6(6), Maxwell Scientific Organization, 2013; K.S. Jadaan i in., *Prediction of Road Traffic Accidents in Jordan using Artificial Neural Network (ANN)*, Journal of Traffic and Logistics Engineering Vol. 2, No. 2, 2014; S. Sikka, *Prediction of Road Accidents in Delhi using Back Propagation Neural Network Model*, International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 5 No. 08, 2014.

[13] por. D. Witkowska, *Sztuczne sieci neuronowe i metody statystyczne. Wybrane zagadnienia finansowe,* Wydawnictwo C. H. Beck, Warszawa, 2002, s. 10.

następujące zmienne objaśniające: wyp_1, dzien, mc, t, wyp_7, wyp_10, wyp_4, swieto, wyp_6, wyp_14. W modelu sieci, w którym dokonano redukcji zmiennych wejściowych, uwzględniono arbitralnie tylko te zmienne, dla których %IncMSE>8 (było to 10 pierwszych zmiennych uzyskanych w rankingu ważności zmiennych). Dla kolejnych zmiennych (począwszy od zmiennej „wyp_28") zamieszczonych na wykresie ważności cech obserwowany jest łagodny przyrost błędu prognozy, gdy daną zmienną wyłączy się z modelu. W przypadku sieci neuronowych zmienne symboliczne zastąpiono przez zmienne 0-1, skutkiem czego sieć, w której uwzględniono wszystkie zmienne objaśniające miała 55 neuronów w warstwie wejściowej, 7 neuronów w warstwie ukrytej i 1 neuron w warstwie wyjściowej. Sieć, w której uwzględniono 10 pierwszych zmiennych z rankingu ważności zmiennych, po zamianie zmiennych symbolicznych na zmienne 0-1, miała na 27 neuronów w warstwie wejściowej, 5 neuronów w warstwie ukrytej i 1 neuron w warstwie wyjściowej. Tablica 2 przedstawia przeciętne absolutne błędy procentowe na zbiorze treningowym i testowym uzyskane przez sieci w przypadku, gdy liczba epok w procesie uczenia sieci wynosiła 1000. Wyznaczając błędy, wartości teoretyczne przekształcono do oryginalnego zakresu (przed normalizacją), aby umożliwić porównanie wyników z wynikami lasu losowego. Sieć, w której dokonano redukcji początkowego zbioru zmiennych objaśniających charakteryzowała się większym przeciętnym absolutnym błędem procentowym na zbiorze treningowym, jednakże na zbiorze testowym błąd ten był nieco mniejszy niż w przypadku sieci, w której uwzględniono wszystkie zmienne objaśniające.

Tablica 2. Przeciętne absolutne procentowe błędy prognoz sieci

Architektura sieci	Przeciętny absolutny błąd procentowy na zbiorze treningowym	Przeciętny absolutny błąd procentowy na zbiorze testowym
55-7-1	9,57%	14,34%
25-5-1	13,31%	12,22%

Źródło: opracowanie własne.

3. Prowadzenie pojazdu w stanie nietrzeźwości a wypadki drogowe

Około 20% wypadków drogowych w Polsce powodowanych jest przez nietrzeźwych kierujących[14]. Analizując dane dzienne można zauważyć, że przeciętnie najwięcej osób zatrzymywanych podczas prowadzenia pojazdu w stanie nietrzeźwości jest w Polsce w niedziele. Jednocześnie w niedziele przeciętna liczba wypadków drogowych jest najmniejsza.

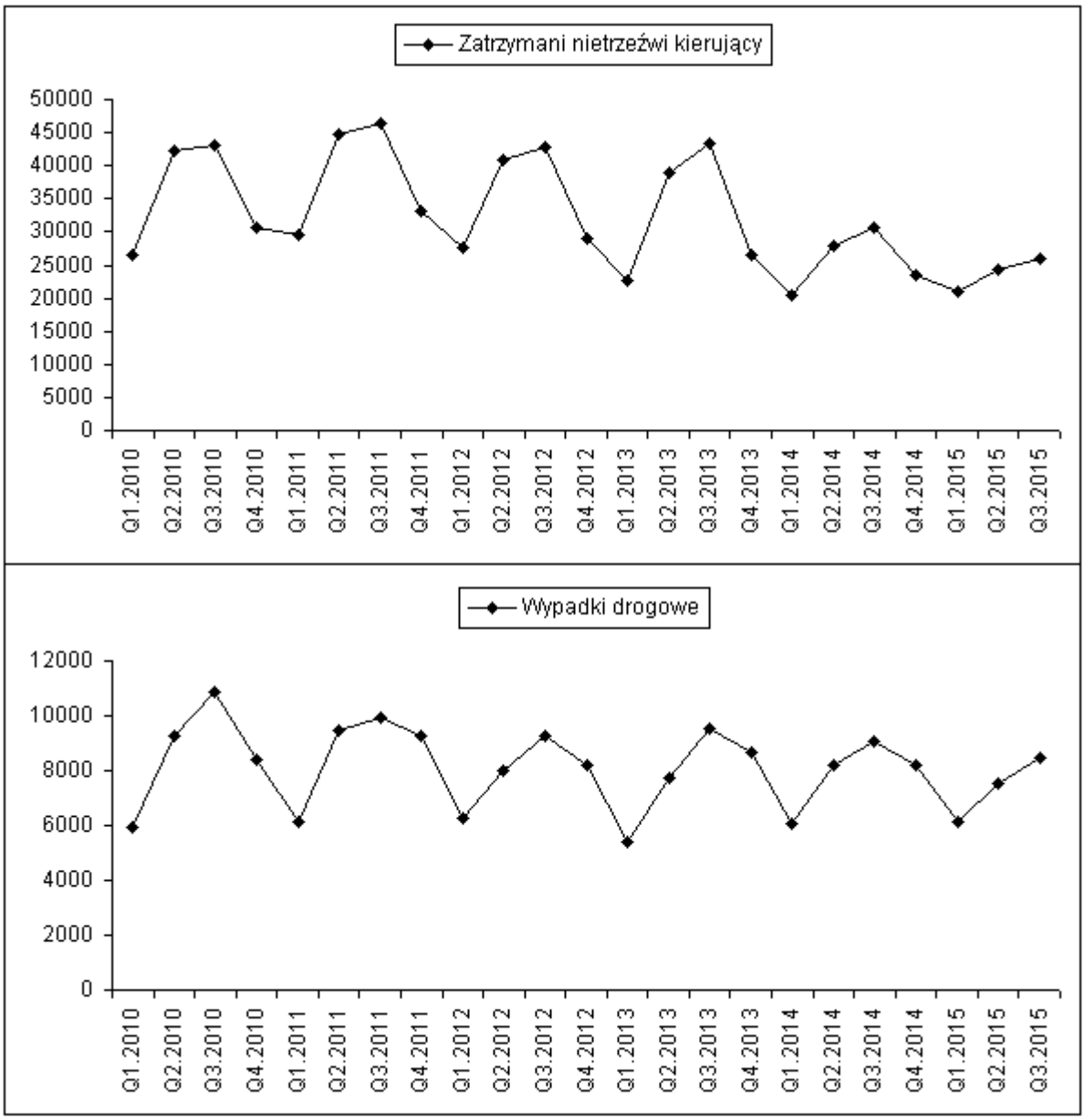
Tablica 3. Przeciętna liczba zatrzymanych nietrzeźwych kierujących

Dzień tygodnia	Przeciętna liczba zatrzymanych nietrzeźwych kierujących w okresie 01.01.2010-30.09.2015
Poniedziałek	357
Wtorek	293
Środa	298
Czwartek	311
Piątek	339
Sobota	425
Niedziela	452

Źródło: opracowanie własne na podstawie danych www.policja.pl.

[14] por. P. Bucoń, *Odpowiedzialność cywilna uczestników wypadku komunikacyjnego*, Oficyna 2008.

Agregując dane dzienne do danych kwartalnych (poprzez zsumowanie liczby zdarzeń w danym okresie) wyraźnie uwidacznia się sezonowość. Można zauważyć m. in. większą liczbę wypadków drogowych oraz zatrzymanych nietrzeźwych kierujących w trzecim kwartale niż w pozostałych (rysunek 6). Jednocześnie od drugiego kwartału 2014 r. można zaobserwować mniejszą liczbę zatrzymywanych nietrzeźwych kierujących niż w poprzednich okresach.



Rysunek 6. Wypadki drogowe i zatrzymani nietrzeźwi kierujący – dane kwartalne

Źródło: opracowanie własne.

Podsumowanie

W pracy analizowano przestrzenno – czasowe zróżnicowanie nasilenia wypadków drogowych w Polsce. Wykorzystując metody statystyki przestrzennej zidentyfikowano skupienia powiatów charakteryzujących się ponadprzeciętną liczbą i natężeniem wypadków

drogowych. Podjęto próbę prognozowania dziennej liczby wypadków drogowych z wykorzystaniem wybranych metod data mining. Najmniejszym przeciętnym absolutnym błędem procentowym na zbiorze treningowym i testowym charakteryzował się model lasu losowego. W analizowanym przykładzie redukcja liczby zmiennych objaśniających wykorzystanych w przypadku sieci neuronowych doprowadziła do wzrostu przeciętnego absolutnego błędu procentowego na zbiorze treningowym, jednakże w porównaniu z siecią, w której uwzględniono wszystkie zmienne objaśniające, przeciętny absolutny błąd procentowy na zbiorze testowym nieco się zmniejszył. Należy mieć jednak na uwadze, że wpływ na wyniki uzyskiwane za pomocą sieci neuronowych ma między innymi sposób ustalenia parametrów początkowych modelu (jak np. architektura sieci, początkowe wartości wag czy liczba epok w procesie uczenia). Optymalizacja tych parametrów nie była tu przedmiotem analiz. Możliwe, że przy uwzględnieniu optymalizacji parametrów początkowych istniałaby sieć pozwalająca uzyskać lepsze rezultaty.

Podejmując próby prognozowania dziennej liczby wypadków drogowych należy mieć na uwadze, że istotnym czynnikiem wpływającym na ich nasilenie są warunki pogodowe. Z uwagi na to, że analizowano dane ogólnopolskie, czynniki te zostały tutaj pominięte. Natomiast analizując dane dzienne dotyczące wypadków na poziomie lokalnym (np. na terenie Katowic) wydaje się zasadne uwzględnienie wśród zmiennych objaśniających takich cech jak np. poziom widoczności (mgły) i opadów. Z kolei łącząc informacje o dziennej liczbie wypadków drogowych z danymi przestrzennymi można szczególną uwagę zwrócić na prognozowanie liczby zdarzeń na obszarach charakteryzujących się wysokim zagrożeniem wypadkami.

Literatura

1. Barcik J., Czech P., *The influence of road infrastructure on safety of road traffic – part 2*, Transport No. 69, Zeszyty Naukowe Politechniki Śląskiej, 2010
2. Bayata H. F. i in., *Modeling of monthly traffic accidents with the artificial neural network method*, International Journal of the Physical Sciences Vol. 6(2), 2011
3. Bucoń P., *Odpowiedzialność cywilna uczestników wypadku komunikacyjnego*, Oficyna 2008
4. Chudy – Laskowska K., Pisula T., *Prognoza liczy wypadków drogowych w Polsce*, „Logistyka" nr 6/2014
5. Dębowska – Mróz M., *Ocena bezpieczeństwa ruchu drogowego w Polsce*, „Logistyka" nr 6/2010
6. Fijołek K. i in., *Prognozowanie cen energii elektrycznej na rynku dnia następnego metodami data mining*, „Rynek energii" 12/2010
7. Gatnar E., *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, 2001
8. Gatnar E., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji,* Wydawnictwo Naukowe PWN, Warszawa 2008
9. Hołyst B., *Kryminologia*, Wydawnictwo Prawnicze LexisNexis, Warszawa 2007
10. Jadaan K. S. i in., *Prediction of Road Traffic Accidents in Jordan using Artificial Neural Network (ANN)*, Journal of Traffic and Logistics Engineering Vol. 2, No. 2, 2014
11. Jamroz K., Kustra W., *Analysis of factors influencing the density of fatalities on national roads in Poland*, Journal of KONBiN 1 (13) 2010
12. Kądziołka K., *Determinanty przestępczości w Polsce. Aspekt ekonomiczno – społeczny w ujęciu modelowania ekonometrycznego*, niepublikowana rozprawa doktorska, Uniwersytet Ekonomiczny w Katowicach, 2015

13. Kopczewska K., *Ekonometria i statystyka przestrzenna z wykorzystaniem programu R CRAN*, CeDeWu Sp. z o.o., Warszawa 2011
14. Quiang Z. i in., *Traffic Accidents Forecasting Based on Neural Network and Principal Component Analysis*, Research Journal of Applied Sciences, Engineering and Technology 6(6), Maxwell Scientific Organization, 2013
15. Sikka S., *Prediction of Road Accidents in Delhi using Back Propagation Neural Network Model*, International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 5 No. 08, 2014
16. Witkowska D., *Sztuczne sieci neuronowe i metody statystyczne. Wybrane zagadnienia finansowe,* Wydawnictwo C. H. Beck, Warszawa, 2002
17. Wójcik A., *Statystyczna analiza bezpieczeństwa w ruchu drogowym w układzie województw*, Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach nr 220, 2015
18. *Stan bezpieczeństwa ruchu drogowego oraz działania realizowane w tym zakresie w 2014r.*, www.krbrd.gov.pl, 20.10.2015
19. Strona internetowa Komendy Głównej Policji, www.policja.pl, 20.10.2015
20. Strona internetowa Głównego Urzędu Statystycznego (Bank Danych Lokalnych), www.stat.gov.pl, 20.10.2015

SPATIO – TEMPORAL ANALYSIS OF ROAD ACCIDENTS IN POLAND

Summary

This article attempts to model the daily number of road accidents in Poland using data mining methods such as random forests and artificial neural networks. There was compared the network, which takes into account all the analyzed explanatory variables to the network with reduced the number of input variables. The network with reduced set of input variables was characterized by a slightly lower average absolute percentage error on the test set than the network, which includes all explanatory variables. There was also identified clusters of poviats characterized by high risk of road accidents.

Keywords: road accident, neural network, random forest, Moran statistic

Kinga Kądziołka
Prokuratura Okręgowa w Katowicach
kinga_kadziolka@onet.pl