Ekonomia. Rynek, gospodarka, społeczeństwo
45(2016), s. 53−81
DOI: 10.17451/eko/45/2016/186
ISBN: 0137-3056
www.ekonomia.wne.uw.edu.pl

Determinanty przestępczości w Polsce. Analiza zależności z wykorzystaniem drzew regresyjnych

Kinga Kądziołka*

Streszczenie

Celem artykułu była identyfikacja zależności między przestępczością a wybranymi charakterystykami powiatów w 2014 roku z wykorzystaniem drzew regresyjnych. Do wygenerowania drzewa wykorzystana została nieobciążona metoda rekurencyjnego podziału. W trakcie kolejnych podziałów przestrzeni zmiennych istotne okazały się następujące czynniki objaśniające natężenie przestępstw stwierdzonych ogółem: wskaźnik urbanizacji, odsetek gospodarstw jednoosobowych, natężenie przestępstw stwierdzonych w powiatach sąsiednich, współczynnik rozwodów oraz udzielone noclegi w przeliczeniu na 1000 ludności. Do identyfikacji zależności między wybranymi charakterystykami obszarów a przestępczością wykorzystano również las losowy zbudowany z wielu drzew regresyjnych. Uzyskane dla lasów losowych rankingi ważności predyktorów ujawniły szczególnie silny związek między przestępczością a urbanizacją.

Słowa kluczowe: determinanty przestępczości, drzewo regresyjne, las losowy, dane przekrojowe
Kody JEL: C1, K42, R1
DOI: 10.17451/eko/45/2016/186

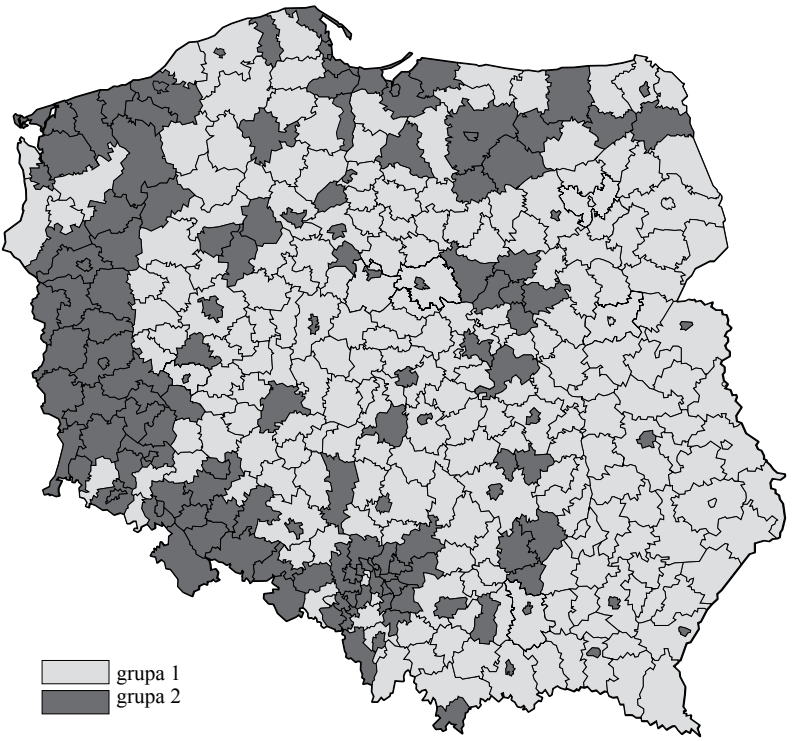
* Prokuratura Okręgowa w Katowicach, e-mail: kinga_kadziolka@onet.pl.

1. Wprowadzenie

Przestępczość jest zjawiskiem, które nie występuje z takim samym nasileniem na całym obszarze Polski. Ponadprzeciętnym natężeniem przestępstw charakteryzują się głównie obszary miejskie, a także powiaty zlokalizowane w pobliżu zachodniej granicy. Natężenie przestępstw (współczynnik przestępczości) to liczba przestępstw stwierdzonych[1] dla danego roku obliczeniowego, przypadająca na pewną umownie przyjmowaną stałą liczbę ludności zamieszkałej na danym terenie (Bułat *et al*. 2007, 71). Najczęściej przyjmuje się liczbę przestępstw przypadającą na 100 tys. lub 10 tys. ludności. W prowadzonych tu analizach natężenie przestępstw wyznaczano jako liczbę przestępstw stwierdzonych, przypadającą na 100 tys. ludności. Na mapie (Rycina 1) przedstawiono przestrzenne zróżnicowanie powiatów pod względem natężenia przestępstw stwierdzonych ogółem w 2014 roku. Obszary zostały podzielone na dwie grupy:

- grupa 1: obiekty, dla których: $x_i \leq \bar{x}$
- grupa 2: obiekty, dla których: $x_i > \bar{x}$

gdzie x_i oznacza wartość natężenia przestępstw w i – tym powiecie, $\bar{x}$ – przeciętna wartość natężenia przestępstw w powiatach w 2014 roku.



Rycina 1. Przestrzenne zróżnicowanie natężenia przestępstw w 2014 roku

Źródło: opracowanie własne na podstawie danych GUS.

[1] Przestępczość stwierdzona to ogół czynów, których charakter jako przestępstw został potwierdzony w wyniku postępowania przygotowawczego (Hołyst 2001, 52).

W 2014 roku największym natężeniem przestępstw stwierdzonych charakteryzowały się powiaty: m. Sopot (6334), m. Legnica (6012), m. Katowice (5734), m. Wałbrzych (5564), m. Nowy Sącz (5002), m. Wrocław (4453). Najmniejsze natężenie przestępstw stwierdzonych charakteryzowało powiat krośnieński (475). W Warszawie natężenie przestępstw stwierdzonych ogółem w 2014 roku wynosiło 2889 i plasowało stolicę na 54 miejscu pod względem zagrożenia przestępczością stwierdzoną (licząc od obszaru charakteryzującego się największym natężeniem przestępstw do obszaru o najmniejszym zagrożeniu przestępczością stwierdzoną). Większość powiatów charakteryzujących się ponadprzeciętnym natężeniem przestępstw stwierdzonych zlokalizowana jest na terenie województw graniczących z Niemcami. Skupienie obszarów wysokiej przestępczości można też zaobserwować m.in. na terenie województwa śląskiego oraz wokół Warszawy.

Celem artykułu jest identyfikacja zależności między natężeniem przestępstw stwierdzonych w powiatach w 2014 roku a wybranymi charakterystykami obszarów z wykorzystaniem drzew regresyjnych. Drzewa regresyjne dostarczają wyników łatwych w interpretacji i wizualizacji. Ponadto ta metoda, w przeciwieństwie do metod ekonometrycznych (wykorzystywanych w pracach polskich autorów do identyfikacji czynników wpływających na przestępczość), nie zakłada znajomości postaci analitycznej związku między zmienną objaśnianą a zmiennymi objaśniającymi. Nie jest także wymagane dokonywanie transformacji danych wejściowych ani redukcji początkowego zestawu potencjalnych zmiennych objaśniających, gdyż ich dobór następuje automatycznie na podstawie przyjętego wcześniej kryterium (Gatnar 2001, 8). Źródłem informacji będą ogólnodostępne dane publikowane na stronach internetowych Głównego Urzędu Statystycznego (Bank Danych Lokalnych) oraz Centralnego Zarządu Służby Więziennej.

Możliwość zastosowania drzew regresyjnych do identyfikacji zależności między przestępczością a wybranymi charakterystykami obszarów zaprezentowana została w pracy Arbii i Tabasso (2013). Autorzy analizowali zależność między natężeniem zabójstw a wybranymi charakterystykami społeczno-ekonomicznymi na południowym obszarze USA. Wykorzystali do tego cztery rodzaje modeli drzew regresyjnych. Pierwszy model zawierał jako zmienne objaśniające tylko społeczno-ekonomiczne charakterystyki obszarów. W modelu drugim dodatkowo uwzględniono współrzędne geograficzne środków ciężkości obszarów. W modelu trzecim wykorzystano jako zmienne objaśniające czynniki społeczno-ekonomiczne oraz opóźnioną przestrzennie zmienną objaśnianą. W modelu tym współrzędne geograficzne zostały pominięte. W ostatnim z modeli wykorzystano jako zmienne objaśniające społeczno-ekonomiczne charakterystyki obszarów, współrzędne geograficzne oraz opóźnioną przestrzennie zmienną objaśnianą. W każdym z modeli zestaw zmiennych o charakterze społeczno-ekonomicznym był ten sam. Różnice polegały jedynie na uwzględnionych zmiennych o charakterze przestrzennym. Uwzględnienie wśród zmiennych objaśniających przestrzennych charakterystyk

obszarów pozwoliło zredukować autokorelację przestrzenną reszt w porównaniu z modelem, który wykorzystywał tylko zmienne o charakterze społeczno-ekonomicznym jako predyktory[2].

Modele drzew regresyjnych były również stosowane do identyfikacji zależności między natężeniem przestępstw a wybranymi czynnikami w Polsce. Kądziołka (2016) do analizy czynników przestępczości na poziomie powiatów województwa śląskiego w 2014 roku wykorzystała drzewo regresyjne z uwzględnioną opóźnioną przestrzenie zmienną objaśnianą. Natomiast w pracy Kądziołki (2015a) porównano pod względem współczynnika pseudo-R^2 najlepszy z uzyskanych modeli opóźnienia przestrzennego z modelem drzewa regresyjnego, w którym wśród zmiennych objaśniających uwzględniono taki sam (jak w przypadku modeli opóźnienia przestrzennego) początkowy zestaw charakterystyk obszarów oraz współrzędne geograficzne środków ciężkości powiatów i opóźnioną przestrzennie zmienną objaśnianą. Model drzewa regresyjnego charakteryzował się nieco lepszym dopasowaniem pod względem współczynnika pseudo-R^2 niż najlepszy z uzyskanych modeli opóźnienia przestrzennego.

Oprócz pojedynczego drzewa regresyjnego wykorzystany zostanie również las losowy zbudowany z wielu drzew regresyjnych. Zastosowanie lasu losowego pozwoli na redukcję skokowego charakteru prognoz natężenia przestępstw uzyskiwanych w przypadku pojedynczego drzewa oraz wygenerowanie rankingu ważności predyktorów pod względem ich wpływu na zmienną objaśnianą. W pracach dotyczących identyfikacji czynników wpływających na przestępczość w Polsce (w przeciwieństwie do zagranicznych prac) metoda lasu losowego nie była jeszcze stosowana.

Kolejna część artykułu zawiera przegląd wybranych prac, w których podejmowane były próby identyfikacji zależności między przestępczością a czynnikami uznawanymi w literaturze za wpływające na przestępczość. W następnej części dokonano charakterystyki wykorzystanych danych i metod. Następnie zaprezentowano wyniki i wnioski z przeprowadzonych analiz oraz podsumowano rezultaty.

2. Badania determinant przestępczości ze szczególnym uwzględnieniem Polski

W ramach poszczególnych nurtów kryminologii powstało wiele teorii wyjaśniających istotę, etiologię i uwarunkowania przestępczości. Szczególne znaczenie w wyjaśnianiu przyczyn przestępczości przypisywane jest czynnikom o charak-

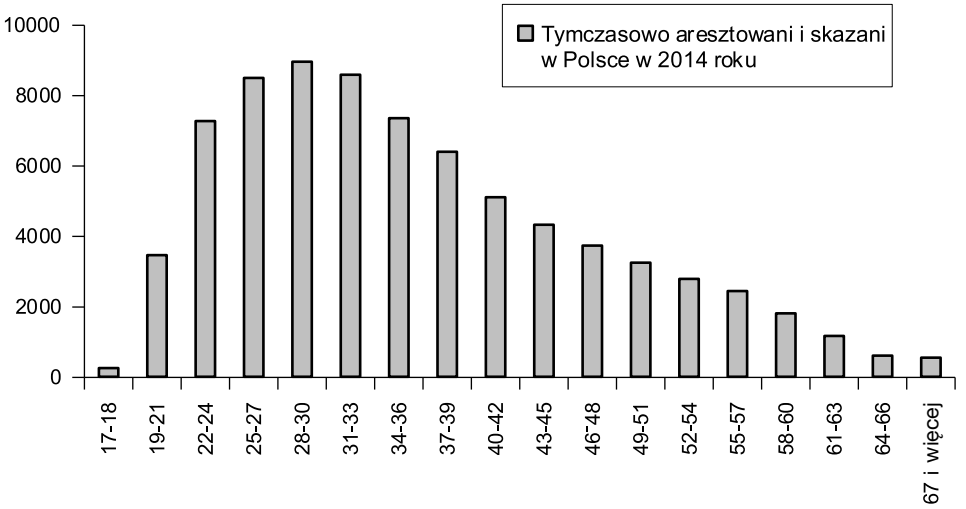
[2] Należy mieć jednak na uwadze, że wpływ na wartość statystyk przestrzennych i istotność autokorelacji przestrzennej mają m.in. sposób zdefiniowania macierzy wag przestrzennych oraz metoda weryfikacji istotności autokorelacji przestrzennej (Kądziołka 2014c).

terze społeczno-ekonomicznym, takim jak: bezrobocie, ubóstwo, zróżnicowanie dochodów, rozwój gospodarczy, wykształcenie, rozbicie rodziny, zmiany składu mieszkańców, zróżnicowanie rasowe, m.in.: Becker (1968), Ehrlich (1973), Groves i Sampson (1989), Besci (1999), Bogacka (2009), Kiersztyn (2008), Szczepaniec (2012), Florczak (2013), Kądziołka (2015a). Należy jednak mieć na uwadze, że brakuje jednej, „uniwersalnej" teorii, która w pełni tłumaczyłaby zachowania przestępcze, a założenia dotyczące wpływu poszczególnych czynników na przestępczość mogą się różnić w ramach poszczególnych teorii. Przykładowo, zgodnie z ekonomiczną teorią przestępczości Beckera (1968), w myśl której przestępstwo jest wynikiem racjonalnej kalkulacji zysków i strat z jego popełnienia, wzrost bezrobocia powinien przyczynić się do wzrostu przestępczości z uwagi na zmniejszenie kosztów straconych możliwości w przypadku osób bezrobotnych. Z kolei według teorii działań rutynowych Cohena i Felsona (1979) wzrost bezrobocia powinien przyczynić się do spadku przestępczości, gdyż osoby bezrobotne będą więcej czasu przebywać w domach, przez co maleje prawdopodobieństwo stania się ofiarą przestępstwa oraz wzrasta poziom ochrony mienia. Pierwszy z przedstawionych efektów oddziaływania bezrobocia na przestępczość nazywany jest w literaturze efektem motywacji (*motivation effect*), a drugi efektem możliwości (*opportunity effect*). Zwracana jest uwaga, że oba przedstawione mechanizmy (efekt motywacji i możliwości) mogą działać jednocześnie, znosząc się nawzajem (Kiersztyn 2008; Meyer i Sridharan 2005). Z powodu występowania obu tych efektów uzyskiwane zależności między przestępczością a pewnymi charakterystykami mogą nie mieć stałego, uniwersalnego charakteru.

Wpływ sytuacji społeczno-ekonomicznej na przestępczość akcentuje również teoria dezorganizacji społecznej, rozwinięta przez Shawa i McKaya (1942). Opiera się ona na założeniu, że istnienie patologii społecznych zależy nie od właściwości pojedynczych jednostek, ale od właściwości społeczno-kulturowych obszarów, na których te jednostki funkcjonują (Bogacka 2012, 74). Badając przestępczość młodzieży w zdegradowanych dzielnicach Chicago, autorzy ci doszli do wniosku, że jest ona pochodną trzech czynników: deprywacji ekonomicznej, częstych zmian składu mieszkańców oraz różnic etnicznych i kulturowych między członkami społeczności. Przy czym za podstawową zmienną uznali warunki ekonomiczne (Kiersztyn 2008, 64). Groves i Sampson (1989) rozwinęli teorię dezorganizacji społecznej, uwzględniając następujące źródła dezorganizacji: status socjoekonomiczny, mobilność mieszkańców, zróżnicowanie rasowe, rozpad rodziny, urbanizacja. W literaturze zwraca się uwagę, że osoby o niskim statusie socjoekonomicznym wykazują niewielką chęć uczestniczenia w lokalnych inicjatywach (Kiersztyn 2008, 64). Z kolei częste zmiany składu mieszkańców oraz zróżnicowanie etniczne i kulturowe społeczności utrudniają wykształcenie się więzi między nimi.

Wśród czynników o charakterze demograficznym, uznawanych w literaturze za wpływające na przestępczość, wskazuje się m.in. płeć i wiek potencjalnego

sprawcy. Ze statystyk policyjnych oraz więziennych wyłania się obraz typowego sprawcy, którym jest mężczyzna. Przykładowo, w Polsce w 2014 roku kobiety stanowiły 3,58% ogółu skazanych i tymczasowo aresztowanych. Ponadto sprawcami przestępstw w dużej mierze są osoby młode. W 2014 roku ponad 66% ogółu tymczasowo aresztowanych i skazanych w Polsce stanowiły osoby w wieku do 39 lat[3] (Rycina 2).



Rycina 2. Tymczasowo aresztowani i skazani w 2014 roku według wieku

Źródło: opracowanie na podstawie danych Centralnego Zarządu Służby Więziennej.

W literaturze zwraca się również uwagę na wpływ efektywności pracy organów ścigania oraz surowości kary na przestępczość. Według teoretycznego modelu przestępczości Beckera (1968) wzrost prawdopodobieństwa schwytania i ukarania sprawcy (jak również wzrost surowości kary) powinien przyczynić się do spadku przestępczości.

Ehrlich (1973) zapoczątkował badania czynników wpływających na przestępczość z wykorzystaniem metod ekonometrycznych, prezentując ekonometryczny model przestępczości oszacowany na podstawie danych przekrojowych z lat 40., 50. i 60. dla Stanów Zjednoczonych. W modelu tym zmienną objaśnianą była liczba przestępstw stwierdzonych przypadających na jednego mieszkańca, a zmiennymi objaśniającymi były m.in.: liczba mężczyzn przypadających na 100 kobiet, mediana zarobków w danej populacji, wydatki na policję, odsetek mężczyzn w wieku 14−24 lat, stopa bezrobocia wśród mężczyzn w wieku 14−24 oraz 35−39 lat, przeciętna liczba lat edukacji osób powyżej 25 lat. W zagranicznych pracach problematyka identyfikacji zależności między przestępczością a wybranymi charakterystykami obszarów jest często podejmowana. W tym celu wykorzystywane

[3] Obliczono na podstawie danych Centralnego Zarządu Służby Więziennej.

są zaawansowane metody statystyczne, ekonometryczne i metody *data mining*. Analizy prowadzone są z wykorzystaniem różnych typów danych, co pozwala uwzględnić zmiany zjawiska w czasie i w przestrzeni, m.in. Groves i Sampson (1989), Besci (1999), Entorf i Spengler (2000), Gorr, Olligschlaeger i Thompson (2003), Cracolici i Uberti (2008), Han (2009), Falcone i Lombardo (2011), Lauridsen, Zeren i Ari (2013), Cherain i Dawson (2015).

Jak dotąd powstało mało prac dotyczących zagadnienia identyfikacji czynników wpływających na przestępczość w Polsce, w których do badania zależności zastosowano zaawansowane metody statystyczne czy ekonometryczne. W Tabeli 1 przedstawiono wybrane prace, w których autorzy podejmowali próby identyfikacji zależności między przestępczością a wybranymi charakterystykami obszarów na różnych poziomach agregacji danych, jak dane ogólnopolskie, województwa, podregiony, powiaty oraz w ograniczeniu do konkretnego obszaru kraju. Pewien wyjątek stanowi tu praca Sypion-Dutkowskiej (2014), w której autorka analizowała dane „punktowe" dotyczące wybranych rodzajów przestępstw z wykorzystaniem systemów GIS. W odróżnieniu od pozostałych wymienionych prac w pracy Sypion-Dutkowskiej główna uwaga została skoncentrowana na środowiskowo-przestrzennych determinantach przestępczości. Autorka badała wpływ sposobów zagospodarowania i użytkowania przestrzeni na natężenie przestępstw „pospolitych"[4] w Szczecinie.

[4] Do tak nazwanej kategorii przestępstw autorka zaliczała następujące czyny: bójki i pobicia, kradzież rzeczy cudzej – inne, kradzież mieszkaniowa, kradzież rozbójnicza, kradzież samochodu, kradzież w placówkach handlowych, kradzież z samochodu, kradzież z włamaniem do mieszkania, kradzież z włamaniem do sklepu, kradzież z włamaniem do samochodu kradzież z włamaniem do innych obiektów, kradzież z włamaniem do piwnicy lub strychu, krótkotrwałe użycie pojazdu, przestępstwo narkotykowe, rozbój, uszkodzenie mienia, wymuszenie rozbójnicze (Sypion-Dutkowska 2014, 16).

Tabela 1. Badania determinant przestępczości w Polsce na danych zagregowanych

Autor	Dane/okres	Analizowane kategorie przestępstw	Wykorzystane metody
Bobrowska i Piasecka (2002)	Dane przekrojowe – województwa w latach 1990–1999	Przestępczość stwierdzona ogółem	Analiza współczynników korelacji liniowej
Sztaudynger i Sztaudynger (2003)	Szeregi czasowe – roczne dane ogólnopolskie za okres 1978–2002	Przestępczość stwierdzona ogółem	Klasyczna metoda najmniejszych kwadratów
Frieske (2007)	Dane przekrojowe – województwa 1998 r.	Przestępczość stwierdzona ogółem, kradzieże	Analiza współczynników korelacji kolejnościowej Spearmana
Kiersztyn (2008)	Województwa (dawne 49 obszarów) – roczne dane przekrojowe i panelowe za okres 1991–1998	Przestępczość stwierdzona ogółem, przestępczość przeciwko mieniu, przestępczość przeciwko życiu i zdrowiu	Analiza współczynników korelacji liniowej, klasyczna metoda najmniejszych kwadratów, ekonometryczne modele dla danych panelowych
Bogacka (2009)	Dane przekrojowe – województwa, dane z lat 2002–2007	Przestępczość stwierdzona ogółem	Analiza współczynników korelacji liniowej, klasyczna metoda najmniejszych kwadratów
Lauridsen (2010)	Dane panelowe – podregiony w latach 2003–2005	Przestępczość stwierdzona ogółem	Klasyczny model dla danych panelowych, przestrzenne modele panelowe
Mordwa (2011)	Miasto Łódź – dane przekrojowe wg sektorów policyjnych, dane średnioroczne za lata 2006–2009	Kradzieże	Klasyczna metoda najmniejszych kwadratów, modele ekonometrii przestrzennej
Bieniek, Cichocki i Szczepaniec (2012)	Dane przekrojowe – powiaty 2008 r.	Przestępczość stwierdzona ogółem	Klasyczna metoda najmniejszych kwadratów
Bogacka (2012)	Powiaty województw graniczących z Niemcami – dane przekrojowe (średnioroczne) z lat 2006–2010	Przestępczość stwierdzona ogółem	Klasyczna metoda najmniejszych kwadratów
Florczak (2013)	Roczne dane ogólnopolskie za okres 1970–2008	Przestępstwa przeciwko mieniu, przestępstwa z użyciem przemocy, przestępstwa z art. 178 kk, inne	Równania regresji, analiza mnożnikowa
Kądziołka (2013)	Dane przekrojowe – powiaty w 2010 r.	Przestępczość stwierdzona ogółem	Analiza współczynników korelacji liniowej

Autor	Dane/okres	Analizowane kategorie przestępstw	Wykorzystane metody
Kądziołka (2014a)	Dane przekrojowe – podregiony 2006 r.	Przestępczość stwierdzona ogółem, przestępczość przeciwko mieniu, przestępczość przeciwko życiu i zdrowiu, przestępczość przeciwko rodzinie i opiece	Klasyczna metoda najmniejszych kwadratów
Kądziołka (2014b)	Dane panelowe – województwa, dane roczne za okres 2005–2012	Przestępczość przeciwko mieniu	Model panelowy z efektami ustalonymi
Sypion-Dutkowska (2014)	Szczecin – różne strefy odległości, dane roczne za okres 2006–2010	Przestępczość pospolita	Analizy geoinformacyjne
Kądziołka (2015a)	Szeregi czasowe o różnej częstotliwości (dane ogólnopolskie oraz dotyczące woj. śląskiego), dane przekrojowe – województwa, podregiony, powiaty – wybrane okresy z lat 1970–2012	Przestępczość stwierdzona ogółem, przestępczość przeciwko mieniu, przestępczość przeciwko życiu i zdrowiu, przestępczość przeciwko rodzinie i opiece, przestępczość gospodarcza, przestępczość drogowa	Klasyczna metoda najmniejszych kwadratów, model opóźnienia przestrzennego, drzewa regresyjne, systemy neuronowo – rozmyte, wielowymiarowa analiza porównawcza
Kądziołka (2015b)	Dane przekrojowe – podregiony 2012 r.	Przestępczość stwierdzona ogółem	Klasyczna metoda najmniejszych kwadratów, metoda Warda
Kądziołka (2015c)	Dane przekrojowe – podregiony 2012 r.	Przestępczość przeciwko mieniu	Uogólniona metoda najmniejszych kwadratów z korektą heteroskedastyczności, miernik syntetyczny
Kądziołka (2015d)	Województwa – dane przekrojowe, szeregi czasowe, dane panelowe za lata 2002–2012	Przestępczość stwierdzona ogółem, przestępczość przeciwko mieniu	Analiza współczynników korelacji liniowej, klasyczna metoda najmniejszych kwadratów, modele panelowe z efektami ustalonymi
Kądziołka (2016)	Województwo śląskie – dane przekrojowe (powiaty w 2014 r.), szeregi czasowe o różnej częstotliwości za okres 2009–2014	Przestępczość stwierdzona ogółem, kradzieże, kradzieże z włamaniem, bójki i pobicia	Ekonometryczny model z trendem i sezonowością, drzewo regresyjne

Źródło: opracowanie własne.

Wyniki prowadzonych w Polsce analiz zależności między wybranymi charakterystykami obszarów a przestępczością nie dają jednoznacznych odpowiedzi na pytanie dotyczące kierunku zależności między analizowanymi zmiennymi. Wpływ na uzyskiwane wyniki ma m.in. poziom agregacji oraz typ danych. Przykładowo, dla danych przekrojowych na poziomie województw w latach 2005−2012 wyznaczone współczynniki korelacji liniowej między stopą ubóstwa a natężeniem przestępstw stwierdzonych ogółem były ujemne. Natomiast w przypadku danych panelowych dotyczących województw w latach 2005−2012 współczynnik przy zmiennej określającej stopę ubóstwa był dodatni i istotnie różnił się od zera na przyjętym poziomie istotności 5%, co sugerowało, że wraz ze wzrostem stopy ubóstwa może wzrastać natężenie przestępstw. Natomiast ujemne współczynniki korelacji liniowej między stopą ubóstwa a natężeniem przestępstw (uzyskane w przypadku danych przekrojowych) mogą wynikać z tego, że na obszarach, gdzie więcej osób żyje w biedzie, występuje mniejsza liczba potencjalnych obiektów ataku (np. wartościowych rzeczy do kradzieży) sprawcy niż na obszarach charakteryzujących się mniejszym zagrożeniem ubóstwem (Kądziołka 2015d).

Analizy prowadzone na danych przekrojowych z wykorzystaniem regresji wielokrotnej wskazywały, że wybrane (arbitralnie przez autorów) zestawy zmiennych objaśniających w większym stopniu wyjaśniały zmienność natężenia stwierdzonych przestępstw przeciwko mieniu niż innych analizowanych kategorii przestępstw (Kiersztyn 2008; Kądziołka 2014a). Z uwagi na to, że dane przekrojowe dotyczące obszarów to dane przestrzenne, istotnym elementem jest ocena autokorelacji przestrzennej reszt modeli uzyskanych klasyczną metodą najmniejszych kwadratów. W przypadku modeli objaśniających natężenie wybranych kategorii przestępstw, zaprezentowanych w pracach Bogackiej (2012) i Kądziołki (2014a; 2015b; 2015c), autokorelacja przestrzenna reszt nie występowała i nie było potrzeby stosowania modeli ekonometrii przestrzennej (z wyjątkiem modelu objaśniającego natężenie przestępstw przeciwko rodzinie i opiece w pracy Kądziołki 2014a). W pracach Bogackiej (2009), Bieńka, Cichockiego i Szczepaniec (2012), Kądziołki (2014b) aspekt ten został pominięty. W pracy Kądziołki (2015a) do identyfikacji zależności między natężeniem przestępstw przeciwko mieniu a wybranymi charakterystykami powiatów w 2012 roku wykorzystano modele ekonometrii przestrzennej. Opóźniona przestrzennie zmienna zależna, będąca średnią ważoną (zgodnie z zadeklarowaną macierzą wag) natężenia stwierdzonych przestępstw przeciwko mieniu w lokalizacjach sąsiednich, okazała się istotnym czynnikiem wpływającym na natężenie przestępstw przeciwko mieniu w danej lokalizacji.

Jednym z problemów pojawiających się podczas prób identyfikacji zależności między natężeniem przestępstw a wybranymi czynnikami jest dobór zmiennych objaśniających do modelu. Brakuje wskazań literaturowych, który zestaw zmiennych objaśniających jest najlepszy dla danej kategorii przestępstw. W prowadzonych w Polsce badaniach wpływu wybranych czynników na przestępczość zbiory

zmiennych objaśniających były zazwyczaj dobierane w sposób arbitralny. Niekiedy wybrane zmienne były silnie skorelowane ze sobą, co mogło mieć wpływ na uzyskiwane oszacowania parametrów modeli ekonometrycznych. W pracy Kądziołki (2015a) analizowano dopasowanie do danych empirycznych modeli opóźnienia przestrzennego w przypadku stosowania różnych metod redukcji początkowego zestawu zmiennych objaśniających natężenie stwierdzonych przestępstw przeciwko mieniu w powiatach w 2012 roku. Porównano wyniki uzyskane w przypadku sekwencyjnej eliminacji kolejnych nieistotnych zmiennych objaśniających, redukcji liczby zmiennych z wykorzystaniem metody Warda oraz metody głównych składowych. Najlepszym dopasowaniem do danych empirycznych charakteryzował się model, w którym dokonano sekwencyjnej eliminacji kolejnych nieistotnych statystycznie zmiennych objaśniających. Nieco gorszym dopasowaniem charakteryzował się model, w którym redukcji początkowego zestawu zmiennych dokonano z wykorzystaniem metody Warda. Jednakże w przypadku redukcji zbioru zmiennych z wykorzystaniem hierarchicznych metod grupowania uzyskany wynik zależy m.in. od zastosowanej metody podziału dendrogramu, sposobu zdefiniowania miary niepodobieństwa zmiennych czy sposobu wyboru reprezentantów uzyskanych grup zmiennych. Z kolei w przypadku wykorzystywania metody głównych składowych do redukcji liczby zmiennych objaśniających istnieją różne metody wyboru liczby składowych (np. kryterium Kaisera, kryterium osypiska Cattella, kryterium wyjaśnionej wariancji), co z kolei (przy wykorzystaniu składowych głównych jako zmiennych objaśniających w modelu) ma wpływ na uzyskiwane rezultaty.

W przytaczanych w tym artykule pracach (z wyjątkiem pracy Sypion-Dutkowskiej 2014, która rozważała przestępczość rejestrowaną[5]) analizowana była przestępczość stwierdzona, nie zaś rzeczywista przestępczość, której rozmiar nie jest znany. Wpływ na rozmiar przestępczości stwierdzonej, ujętej w statystykach policyjnych, mają m.in. zmiany prawa. Obowiązujący Kodeks karny podlega ciągłym modyfikacjom, np. poprzez podnoszenie granicznej kwoty, poniżej której kradzież traktowana jest jak wykroczenie, a nie przestępstwo, czy kwalifikowanie jako przestępstw czynów, które dotychczas nimi nie były (np. stalking), i dlatego do porównań nasilenia przestępczości w różnych okresach należy podchodzić ostrożnie, gdyż nie zawsze mniejszej liczbie przestępstw ujętych w statystykach policyjnych odpowiada rzeczywisty spadek przestępczości.

[5] Przestępczość rejestrowana to liczba zdarzeń rejestrowanych i wstępnie kwalifikowanych jako przestępstwa przez organy ścigania (Sypion-Dutkowska 2014, 15).

3. Charakterystyka wykorzystanych danych i metod

Analizowano dane przekrojowe zagregowane na poziomie powiatów dla 2014 roku[6]. Jako potencjalne zmienne objaśniające natężenie przestępstw stwierdzonych w powiatach uwzględniono następujące czynniki: **stb_dl** – stopa bezrobocia długoterminowego[7]; **pom_sp** – udział osób w gospodarstwach domowych korzystających z pomocy społecznej w ludności ogółem; **gimn20_39** – udział osób w wieku 20–39 lat mających wykształcenie co najwyżej gimnazjalne wśród ogółu osób w tym wieku; **zar** – przeciętne miesięczne wynagrodzenie brutto; **urb** – wskaźnik urbanizacji; **gzal** – gęstość zaludnienia (ludność na km^2); **gosp_1os** – odsetek gospodarstw jednoosobowych; **kobiety** – kobiety na 100 mężczyzn; **rozw** – rozwody na 1000 ludności; **migr** – migracje brutto na 1000 ludności[8]; **nocl** – udzielone noclegi na 1000 ludności; **wws** – wskaźnik wykrywalności sprawców[9];**wsp1** – długość geograficzna środka ciężkości powiatu; **wsp2** – szerokość geograficzna środka ciężkości powiatu; **op_npog** − natężenie przestępstw stwierdzonych w sąsiednich lokalizacjach (powiatach). Wartości tej zmiennej były średnią ważoną z wartości natężenia przestępstw stwierdzonych w powiatach sąsiednich, zgodnie z zadeklarowaną macierzą wag[10].

Przy wyborze potencjalnych zmiennych objaśniających kierowano się dostępnością danych, które pochodzą ze strony Głównego Urzędu Statystycznego (Bank Danych Lokalnych) oraz wskazaniami wybranych teorii kryminologicznych. Uwzględnienie wśród zmiennych objaśniających czynników określających bezrobocie i ubóstwo podyktowane było m.in. ekonomiczną teorią przestępczości Beckera (1968). Istniejące teorie kryminologiczne nie precyzują jednak, który rodzaj bezrobocia najsilniej oddziałuje na przestępczość. Tutaj uwzględniono bezrobocie długoterminowe, gdyż jest ono szczególnie groźnym zjawiskiem na rynku pracy i pociąga za sobą szereg negatywnych skutków (Kądziołka 2015d). Wśród czynników o charakterze demograficznym uwzględniono współczynnik feminizacji oraz odsetek młodych osób mających niskie wykształcenie. Oprócz struktury wieku

[6] Informacje dotyczące odsetka gospodarstw jednoosobowych oraz osób w wieku 20–39 lat mających wykształcenie co najwyżej gimnazjalne pochodzą z danych Narodowego Spisu Powszechnego 2011. Pozostałe charakterystyki obszarów obejmują dane dotyczące powiatów w 2014 roku.

[7] Tj. procentowy udział bezrobotnych zarejestrowanych dłużej niż rok wśród aktywnych zawodowo.

[8] Współczynnik migracji brutto definiowany jest jako suma liczby imigrantów i emigrantów (Mielecka-Kubień 2013, 24).

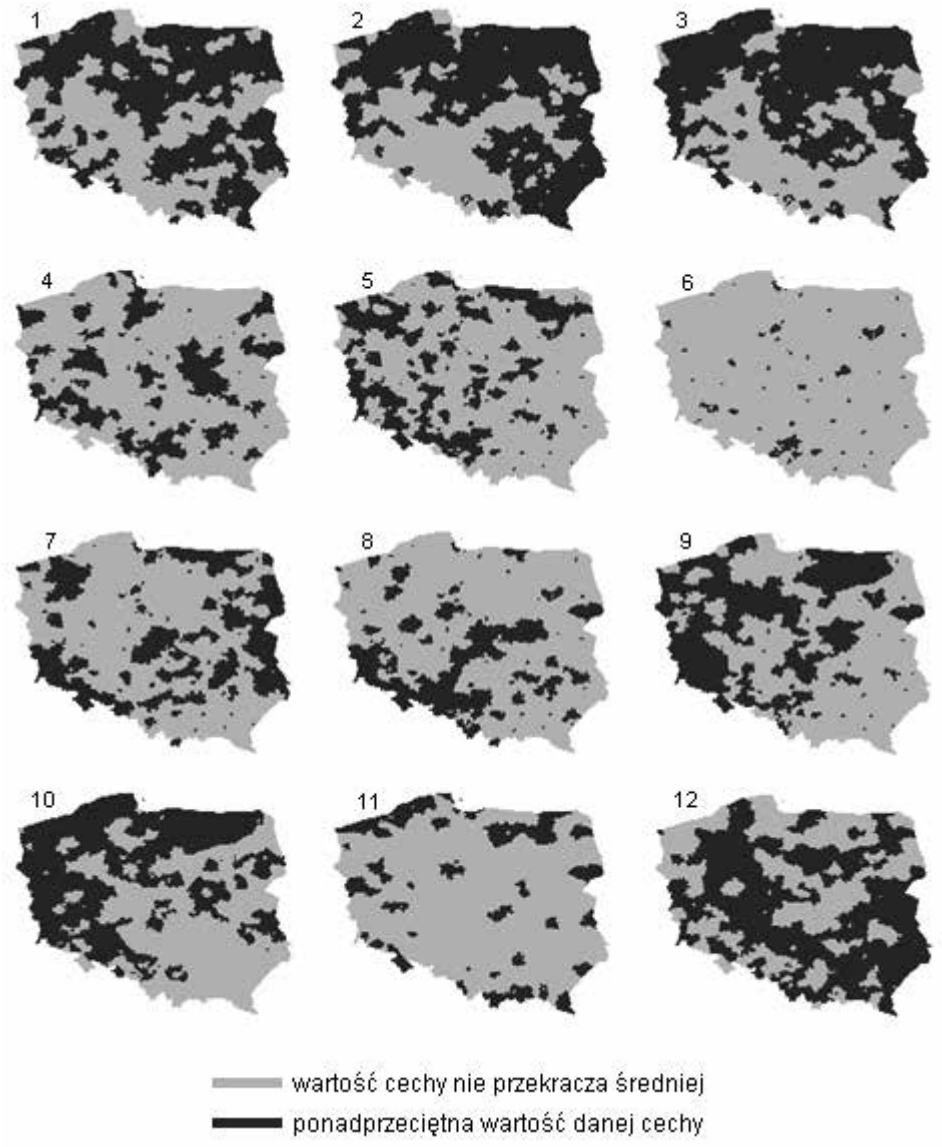
[9] Wskaźnik wykrywalności sprawców przestępstw wyraża się jako stosunek procentowy liczby przestępstw wykrytych do liczby przestępstw stwierdzonych. Z kolei przestępstwo wykryte to przestępstwo stwierdzone, którego popełnienie zarzucono co najmniej jednej osobie i w zakończonym postępowaniu przygotowawczym przyjęto, że popełniła ona przestępstwo (Bułat *et al*. 2007, 71−72). W przypadku tej zmiennej dla powiatu olsztyńskiego oraz m. Olsztyn występowały brakujące wartości dla danych z lat 2013 i 2014. Aby nie pomijać tej cechy, przyjęto dla tych powiatów wartość wskaźnika wykrywalności sprawców z 2012 roku.

[10] Wykorzystana została standaryzowana wierszami binarna macierz wag określona zgodnie z kryterium wspólnej granicy (Kopczewska 2011, 55−56).

uwzględniono również poziom wykształcenia, gdyż osoby podejrzane o popełnienie przestępstwa (a także skazani za przestępstwa) często legitymują się niskim wykształceniem oraz brakiem kwalifikacji zawodowych. W Polsce większość skazanych, zwłaszcza na kary pozbawienia wolności, ma wykształcenie podstawowe lub zawodowe, które uzyskuje często dzięki pobytowi w więzieniu i nauce w szkołach przywięziennych (Woźniakowska 2006, 7). Ponadto poziom wykształcenia wpływa również na możliwość znalezienia pracy, jej atrakcyjność oraz wysokość wynagrodzenia, co z kolei może wpływać na podejmowanie działań niezgodnych z prawem. Wybór zmiennej określającej przeciętne miesięczne wynagrodzenie podyktowany był wynikami wcześniejszych badań (na danych przekrojowych dla podregionów), wskazującymi, że zmienna ta była istotnym czynnikiem objaśniającym natężenie wybranych kategorii przestępstw (Kądziołka 2014a). Ponadto w literaturze wskazuje się, że z większym przeciętnym wynagrodzeniem może być związane większe jego zróżnicowanie, a osoby osiągające niewspółmiernie niskie zarobki w porównaniu z najzamożniejszą częścią obywateli mogą być bardziej skłonne do podejmowania działań niezgodnych z prawem (Sztaudynger i Sztaudynger 2003, 129). Wybór takich cech, jak: wskaźnik urbanizacji, gęstość zaludnienia czy odsetek gospodarstw jednoosobowych, był podyktowany wskazaniami teorii sposobności przestępczych. Analizując dane przekrojowe na różnych poziomach agregacji (województwa, podregiony, powiaty), stwierdzono istotną dodatnią korelację między natężeniem przestępstw a wskaźnikiem urbanizacji (Kądziołka 2015a). Szczególnie silna zależność występowała między wskaźnikiem urbanizacji a natężeniem przestępstw przeciwko mieniu, które są charakterystyczne dla obszarów miejskich, dających potencjalnemu sprawcy większą anonimowość niż obszary wiejskie. Z większą gęstością zaludnienia związana jest większa „dostępność" potencjalnych ofiar przestępstwa niż na obszarach charakteryzujących się mniejszą gęstością zaludnienia. Z kolei w przypadku gospodarstw jednoosobowych mniejszy jest poziom ochrony mienia niż w przypadku gospodarstw wieloosobowych. Wśród czynników związanych z efektywnością pracy organów ścigania wykorzystano wskaźnik wykrywalności sprawców. Natomiast takie czynniki jak rozbicie rodziny czy zmiany składu/mobilność mieszkańców mogą stanowić źródła dezorganizacji społecznej, na co wskazywali Groves i Sampson (1989). W związku z tym wśród potencjalnych zmiennych objaśniających uwzględniono współczynnik rozwodów oraz współczynnik migracji brutto. Z kolei wykorzystanie zmiennej określającej udzielone noclegi miało na celu uwzględnienie charakteru analizowanych obszarów (wyróżnienie miejscowości „turystycznych"), gdyż z większym natężeniem przyjeżdżających turystów czy kuracjuszy może być związana większa podaż okazji przestępczych. Na Rycinie 3 przedstawiono przestrzenne zróżnicowanie powiatów pod względem analizowanych zmiennych objaśniających (pominięto na mapach zmienne wsp1, wsp2 i op_npog). Przyjęto następujące oznaczenia: (1) – przestrzenne zróżnicowanie

powiatów według zmiennej stb_dl; (2) − przestrzenne zróżnicowanie powiatów według zmiennej pom_sp; (3) − przestrzenne zróżnicowanie powiatów według zmiennej gimn20_39; (4) − przestrzenne zróżnicowanie powiatów według zmiennej zar; (5) − przestrzenne zróżnicowanie powiatów według zmiennej urb; (6) − przestrzenne zróżnicowanie powiatów według zmiennej gzal; (7) − przestrzenne zróżnicowanie powiatówwedług zmiennej gosp_1os; (8) − przestrzenne zróżnicowanie powiatów według zmiennej kobiety; (9) − przestrzenne zróżnicowanie powiatów według zmiennej rozw; (10) − przestrzenne zróżnicowanie powiatów według zmiennej migr; (11) − przestrzenne zróżnicowanie powiatów według zmiennej nocl; (12) − przestrzenne zróżnicowanie powiatów według zmiennej wws.

Na zaprezentowanych mapach można przykładowo zauważyć, że powiaty charakteryzujące się wysoką stopą bezrobocia długoterminowego oraz wysokim odsetkiem osób korzystających z pomocy społecznej zlokalizowane są w większości na obszarach województw: zachodniopomorskiego, warmińsko-mazurskiego, kujawsko-pomorskiego, świętokrzyskiego, lubelskiego i podkarpackiego. Dla większości powiatów województw zachodniopomorskiego i warmińsko-mazurskiego charakterystyczne są też: wysoki odsetek osób młodych legitymujących się niskim wykształceniem, wysokie współczynniki rozwodów oraz duże zmiany składu mieszkańców. Wysokie współczynniki rozwodów i współczynniki migracji brutto charakterystyczne są również dla powiatów zlokalizowanych w pobliżu granicy z Niemcami, gdzie występują niespotykane w innych częściach kraju uwarunkowania dla zagranicznych migracji zarobkowych, co może pociągać za sobą negatywne zjawisko, jakim jest tzw. problem eurosierot (Arendt i Kryńska 2011, 70). Z kolei skupienia obszarów charakteryzujących się ponadprzeciętnym miesięcznym wynagrodzeniem zlokalizowane są wokół większych miast, jak Warszawa, Poznań, Wrocław, Katowice, Kraków.



Rycina 3. Przestrzenne zróżnicowanie powiatów według wybranych cech
Źródło: opracowanie własne na podstawie danych GUS.

Do identyfikacji zależności między wybranymi charakterystykami powiatów a natężeniem przestępstw stwierdzonych ogółem na tych obszarach wykorzystane zostanie drzewo regresyjne.

Drzewo regresyjne jest to graf spójny, acykliczny, który stanowi graficzną prezentację modelu postaci (Gatnar 2008, 37–44):

$$y = f(\mathrm{x}_i) = \sum_{k=1}^{K} \alpha_k \mathrm{I}(\mathrm{x}_i \in R_k) \tag{1}$$

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:

$$\alpha_k = \frac{1}{N(k)} \sum_{\mathrm{x}_i \in R_k} y_i \tag{2}$$

gdzie $N(k)$ – liczba elementów znajdujących się w segmencie R_k; y_i − wartości przyjmowane przez zmienną zależną w segmencie R_k.

Wadą drzew regresyjnych jest skokowy charakter zależności między wartościami empirycznymi i teoretycznymi. W związku z tym w prowadzonych badaniach wykorzystano również metodę lasu losowego (*random forest*) celem zredukowania braku ciągłości prognoz. Algorytm *random forest* działa według następującego schematu (Gatnar 2008, 158):

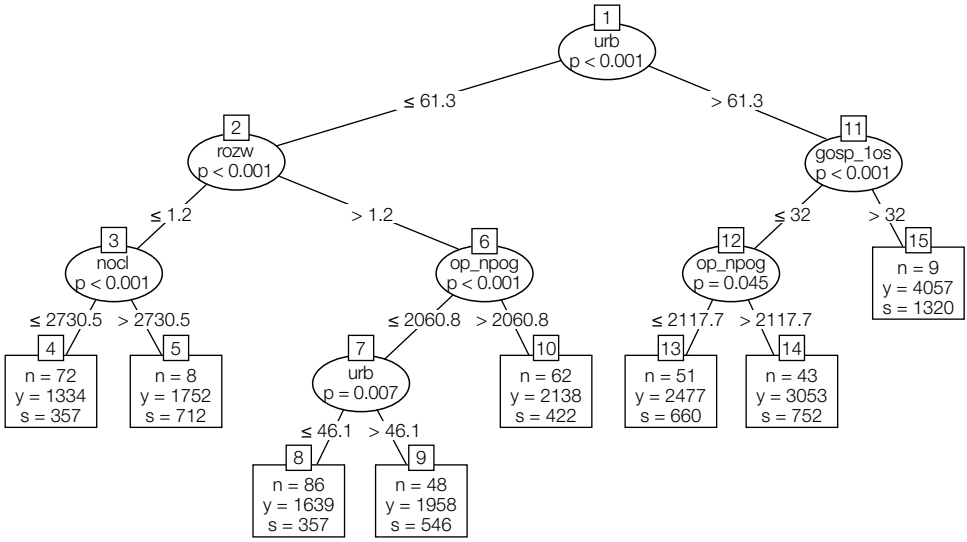
1. Ustal liczbę modeli bazowych (tutaj drzew regresyjnych) M oraz liczbę zmiennych K wybieranych losowo spośród zmiennych objaśniających na każdym etapie budowy drzew.
2. Dla każdego j=1,..., M wykonaj następujące kroki:
   a. Wylosuj próbę uczącą Uj ze zbioru treningowego.
   b. Zbuduj maksymalne drzewo Dm na podstawie próby Um, 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.

Do wygenerowania drzewa regresyjnego wykorzystano funkcję *ctree* pakietu *party* programu R. Funkcja ta do budowy drzew wykorzystuje nieobciążoną metodę rekurencyjnego podziału (*unbiased recursive partitioning*) zaproponowaną przez Hothorna, Hornika i Zeileisa (2006). W metodzie tej podstawą wyboru zmiennych objaśniających stanowiących podstawę podziału jest warunkowy rozkład statystyki mierzącej siłę związku między zmienną objaśnianą i zmiennymi objaśniającymi (Rozmus 2009, 138). W literaturze zwraca się uwagę, że ta metoda generowania drzew umożliwia bardziej obiektywny wybór zmiennych stanowiących podstawę podziału niż algorytm wyczerpującego przeszukiwania, i sugeruje

się jej wykorzystanie, gdy badacza interesuje ustalenie, które zmienne w istotny sposób wpływają na zmienną objaśnianą (Rozmus 2009, 145).

4. Wyniki i wnioski

Na Rycinie 4 przedstawiono uzyskane drzewo regresyjne. Węzły końcowe (liście) zawierają informacje o liczbie elementów w danym segmencie (n), teoretyczną wartość zmiennej objaśnianej (y), będącą średnią wartości natężenia przestępstw dla powiatów z poszczególnych grup, odchylenie standardowe wartości empirycznych zmiennej objaśnianej w danym segmencie (s) oraz p-value (p). W małych kwadratach znajdują się numery węzłów.



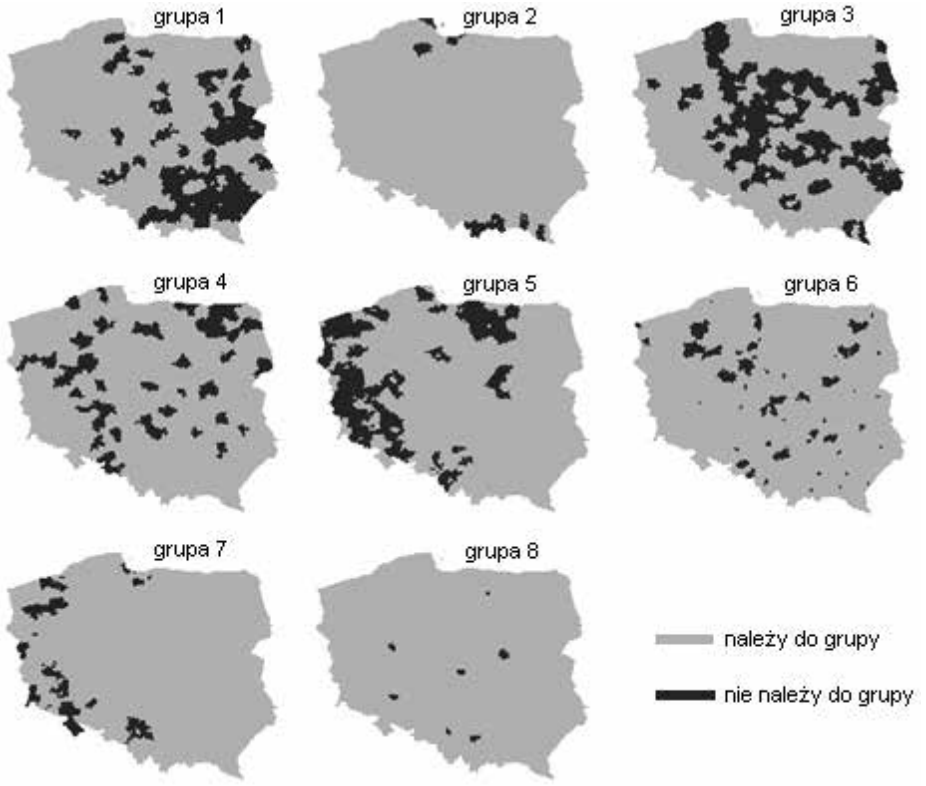
Rycina 4. Drzewo regresyjne objaśniające natężenie przestępstw stwierdzonych ogółem

Źródło: opracowanie własne na podstawie danych GUS.

Uzyskano podział obszarów na 8 grup (Rycina 5) scharakteryzowanych przez warunki:

grupa 1: powiaty, dla których (urb≤61,3%) i (rozw≤1,2) i (nocl≤2730,5), grupa 2: powiaty, dla których (urb≤61,3%) i (rozw≤1,2) i (nocl>2730,5), grupa 3: powiaty, dla których (urb≤61,3%) i (rozw>1,2) i (op_npog≤2060,8) i (urb≤46,1%), grupa 4: powiaty, dla których (urb≤61,3%) i (rozw>1,2) i (op_npog≤2060,8) i (urb>46,1%), grupa 5: powiaty, dla których (urb≤61,3%) i (rozw>1,2) i (op_npog>2060,8), grupa 6: powiaty, dla których (urb>61,3%) i (gosp_1os≤32%)

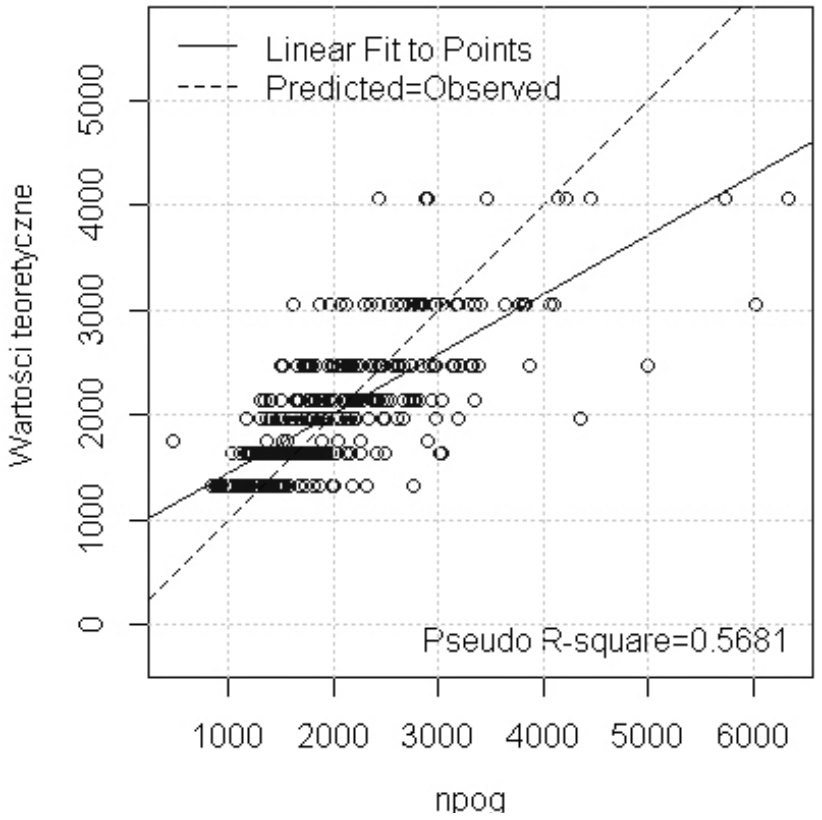
i (op_npog≤2117,7), grupa 7: powiaty, dla których (urb>61,3%) i (gosp_1os≤32%) i (op_npog>2117,7), grupa 8: powiaty, dla których (urb>61,3%) i (gosp_1os>32%).



Rycina 5. Podział powiatów na 8 grup według drzewa regresyjnego
Źródło: opracowanie własne na podstawie danych GUS.

Model drzewa regresyjnego jest prosty w interpretacji, gdyż sekwencja podziałów (od korzenia drzewa do liścia) generuje reguły postaci „JEŻELI…TO…". Każdej z grup odpowiada reguła decyzyjna. Przykładowo (grupa 8) sekwencja wierzchołków 1–11–15 generuje regułę: JEŻELI [(*urb* > 61.3) i (*gosp1_ os* > 32)] TO (y = 4057), co oznacza, że przeciętne natężenie przestępstw stwierdzonych ogółem w powiatach, w których więcej niż 61,3% osób mieszka w miastach oraz gospodarstwa jednoosobowe stanowią ponad 32% ogółu gospodarstw, wynosi 4057. Do grupy 8 należy dziewięć miast na prawach powiatu: Warszawa, Poznań, Wrocław, Kraków, Katowice, Chorzów, Sopot, Olsztyn i Łódź. Z kolei przeciętnie najmniejsze natężenie przestępstw stwierdzonych ogółem (wynoszące 1334) charakterystyczne jest dla powiatów należących do grupy 1. Są to powiaty, w których co najwyżej 63% osób mieszka w miastach, współczynnik rozwodów nie przekracza 1,2, a liczba udzielonych noclegów w przeliczeniu na 1000 ludności nie przekracza 2730,5.

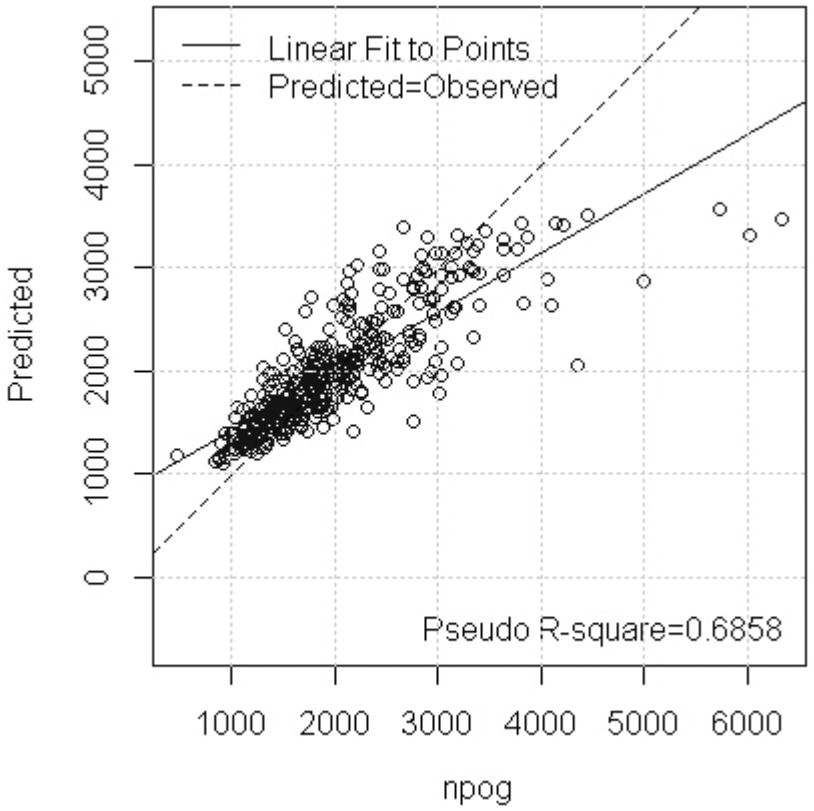
W trakcie kolejnych podziałów dokonywanych podczas generowania drzewa najbardziej istotne (dla poszczególnych podzbiorów danych) okazały się takie charakterystyki, jak: wskaźnik urbanizacji, odsetek gospodarstw jednoosobowych, natężenie przestępstw w sąsiednich powiatach, współczynnik rozwodów i udzielone noclegi na 1000 ludności. Dla zaprezentowanego na Rycinie 4 modelu współczynnik pseudo-R^2=0,5681. Wygenerowano również drzewo regresyjne, w którym wśród zmiennych objaśniających pominięto współrzędne geograficzne oraz opóźnioną przestrzenie zmienną objaśnianą. Dla uzyskanego w ten sposób modelu współczynnik pseudo-R^2=0,5601. Nieznacznie lepszy pod względem współczynnika pseudo-R^2 okazał się model, w którym wśród zmiennych objaśniających uwzględniono opóźnioną przestrzennie zmienną objaśnianą. Wykorzystanie drzew regresyjnych pozwoliło zidentyfikować grupy powiatów podobnych pod względem wybranych charakterystyk (zmiennych objaśniających) i określić przeciętny poziom natężenia przestępstw w ramach poszczególnych grup, jednakże zależność między wartościami empirycznymi i teoretycznymi miała skokowy charakter (Rycina 6). Ponadto przeciętny absolutny procentowy błąd prognozy był wysoki, wynosił bowiem 19,26%.



Rycina 6. Wartości empiryczne i teoretyczne (drzewo regresyjne)

Źródło: opracowanie własne na podstawie danych GUS.

W celu ograniczenia braku ciągłości prognoz wykorzystano las losowy zbudowany z 50 drzew regresyjnych (tj. M=50). Na każdym etapie konstrukcji drzew wybierano w sposób losowy 5 zmiennych (tj. K=5) spośród 15 zmiennych objaśniających[11]. Do wygenerowania modelu wykorzystano pakiet *rattle* programu R. Na Rycinie 7 przedstawiono zależność między wartościami empirycznymi i teoretycznymi w przypadku lasu losowego. Dla uzyskanego modelu współczynnik pseudo-R^2=0,6858. Przeciętny absolutny procentowy błąd prognozy w przypadku lasu losowego wynosił 15,57%.



Rycina 7. Wartości empiryczne i teoretyczne (las losowy)
Źródło: opracowanie własne na podstawie danych GUS.

Wykorzystanie lasu losowego zbudowanego z wielu drzew regresyjnych pozwoliło zredukować problem braku ciągłości prognoz. Ponadto model lasu losowego charakteryzował się lepszym dopasowaniem danych teoretycznych do empirycznych oraz mniejszym przeciętnym absolutnym procentowym błędem

[11] Przyjęto parametr K=5, gdyż w literaturze zalecane jest dla problemów regresyjnych losowanie K=N/3 zmiennych, gdzie N oznacza liczbę wszystkich zmiennych objaśniających (Liaw i Wiener 2002, 20).

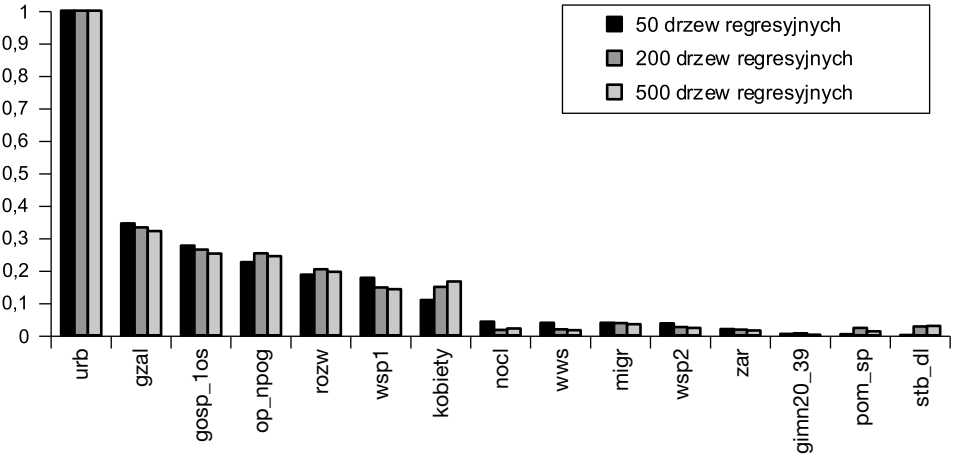
prognozy. Porównano również rezultaty uzyskane w przypadku lasów losowych zbudowanych z 200 i 500 drzew regresyjnych. Uzyskane modele były porównywalne pod względem współczynnika pseudo-R^2 oraz przeciętnego absolutnego procentowego błędu prognozy (ozn. MAPE) z modelem zbudowanym z 50 drzew (Tabela 2). Zatem zwiększanie liczby drzew regresyjnych nie prowadziło do istotnego polepszenia jakości lasów losowych w sensie przyjętych miar.

Tabela 2. Jakość dopasowania modeli lasu losowego

Liczba drzew regresyjnych	pseudo-R^2	MAPE
50	0,6858	15,57%
200	0,6867	15,52%
500	0,6871	15,52%

Źródło: opracowanie własne na podstawie danych GUS.

Rozważając pojedyncze drzewo regresyjne, należy mieć na uwadze, że dana zmienna jest najlepsza (w sensie przyjętego kryterium) wyłącznie na danym etapie budowy drzewa, tzn. kolejność pojawiania się zmiennych w modelu nie świadczy o sile ich wpływu na zmienną zależną (Jefmański i Kusterka-Jefmańska 2012,215). Jednakże dzięki wykorzystaniu lasów losowych możliwe staje się dokonanie globalnej oceny ważności poszczególnych predyktorów pod względem ich wpływu na zmienną zależną. Na Rycinie 8 pokazano wartość miary ważności poszczególnych predyktorów (wyrażoną w skali od 0 do 1), przy czym wartości najwyższe oznaczają największy wpływ danej zmiennej na zmienną zależną.



Rycina 8. Ranking ważności predyktorów dla lasów losowych

Źródło: opracowanie własne na podstawie danych GUS.

W przypadku rozważanych trzech modeli lasów losowych występowała istotna zgodność uporządkowania predyktorów pod względem ich ważności – współczynniki korelacji kolejnościowej Spearmana były dodatnie i istotne statystycznie na przyjętym poziomie istotności 1%. Przedstawiony na Rycinie 8 ranking pokazuje, że największy wpływ na natężenie przestępstw stwierdzonych w powiatach w 2014 roku miał wskaźnik urbanizacji. Uzyskany wynik nie jest zaskoczeniem, gdyż − jak już wspomniano − przestępczość w Polsce jest głównie problemem „miejskim". Relatywnie silną zależność zaobserwowano również między natężeniem przestępstw a gęstością zaludnienia, odsetkiem gospodarstw jednoosobowych czy natężeniem przestępstw w sąsiednich powiatach.

Istotnym zagadnieniem, które zostało niejako pominięte w niniejszej pracy, jest ocena odporności wykorzystanych metod. Odporność można rozpatrywać w różnych aspektach, m.in. jako niewrażliwość na występowanie obserwacji odstających w zbiorze danych, niewrażliwość na występowanie losowych zakłóceń wartości cech czy zmiennych nieistotnych (Trzęsiok 2015, 76). W tym artykule problem zasygnalizowany zostanie na przykładzie odporności na obserwacje odstające. Sprawdzone zostanie, czy analizowane metody prowadzą do uzyskania modeli, dla których wartości miar predykcji nie zmieniają się istotnie po usunięciu obserwacji odstających. W pierwszej kolejności zidentyfikowano obserwacje odstające. W tym celu wykorzystano kryterium opierające się na odległości Mahalanobisa, opisane m.in. w pracy Trzęsiok (2015). Na podstawie tej metody wyróżniono 24 obserwacje odstające tj.: powiat kołobrzeski, m. Sopot, m. Świętochłowice, m. Świnoujście, powiat legnicki, m. Jastrzębie-Zdrój, powiat tatrzański, powiat lubiński, m. Chorzów, m. Warszawa, powiat bełchatowski, m. Łódź, m. Siemianowice Śląskie, m. Katowice, powiat kamieński, m. Legnica, powiat kościański, m. Ostrołęka, powiat wrocławski, powiat lipnowski, m. Nowy Sącz, powiat leski, m. Gdańsk, powiat węgorzewski. Następnie wygenerowano model drzewa regresyjnego oraz lasu losowego (zbudowanego z 50 drzew) dla zbioru danych, z którego usunięto wskazane obserwacje odstające. Dla uzyskanego modelu drzewa regresyjnego pseudo-R^2=0,5874, MAPE=18,05%. W przypadku lasu losowego pseudo-R^2=0,721, MAPE=14,78%. Zatem wartości rozważanych miar dokładności predykcji nie zmieniły się istotnie w porównaniu z modelami generowanymi na całym zbiorze danych. Ranking ważności predyktorów uzyskany dla drzewa regresyjnego ponownie wskazywał, że największy wpływ na zmienną objaśnianą miał wskaźnik urbanizacji. Kolejne w rankingu były: współczynnik rozwodów, gęstość zaludnienia, natężenie przestępstw w sąsiednich powiatach, długość geograficzna środka ciężkości obszaru. Ostatnie miejsce w rankingu zajął wskaźnik wykrywalności sprawców, a drugie i trzecie od końca, odpowiednio, odsetek osób młodych z niskim wykształceniem i stopa bezrobocia długoterminowego. W przypadku pojedynczego drzewa na pierwszym etapie podziału (podobnie jak dla całego zbioru danych) został wybrany wskaźnik urbanizacji. Pozostały-

mi zmiennymi uwzględnionymi w trakcie dalszego podziału były: współczynnik rozwodów, natężenie przestępstw w sąsiednich powiatach, długość geograficzna środka ciężkości obszaru i przeciętne miesięczne wynagrodzenie brutto.

Problemem, jaki się pojawia przy próbach oceny odporności wybranych metod, jest m.in. wybór metody identyfikacji obserwacji odstających. Istnieje wiele metod identyfikacji takich danych i mogą one generować różne wyniki[12]. Pojawiają się też inne dylematy, np. kiedy usunięcie takich obserwacji jest uprawnione, lub czy dana obserwacja rzeczywiście jest nietypowa (czasami obserwacje oddalone obrazują poprawne, choć „nietypowe" i rzadkie zachowanie analizowanych zjawisk).

5. Podsumowanie

W artykule podjęto próbę identyfikacji zależności między natężeniem przestępstw a wybranymi charakterystykami powiatów w 2014 roku z wykorzystaniem drzewa regresyjnego. Uzyskano podział obszarów na osiem grup zróżnicowanych pod względem natężenia przestępstw stwierdzonych. Należy jednak mieć na uwadze, że zaprezentowane tu drzewo regresyjne stanowi jeden z wielu możliwych do uzyskania modeli. Postać końcowa drzewa zależy m.in. od zastosowanej metody podziału przestrzeni zmiennych czy sposobu przycinania drzew. Uzyskane modele można porównywać pod względem dopasowania do danych empirycznych, lecz w dalszym ciągu pozostaje problem wyboru zmiennych objaśniających, gdyż nawet niewielka zmiana początkowego zestawu potencjalnych zmiennych objaśniających może prowadzić do całkiem innej sekwencji podziału. Wykorzystanie lasu losowego zbudowanego z wielu drzew regresyjnych pozwoliło na zredukowanie skokowego charakteru prognoz natężenia przestępstw generowanych przez pojedyncze drzewo oraz umożliwiło dokonanie globalnej oceny wpływu poszczególnych predyktorów na zmienną objaśnianą. Z uzyskanych rankingów ważności zmiennych objaśniających wynikało, że przestępczość była silnie powiązana z takimi charakterystykami, jak: urbanizacja, gęstość zaludnienia, odsetek gospodarstw jednoosobowych lub natężenie przestępstw w powiatach sąsiednich.

Identyfikacja obszarów szczególnie zagrożonych przestępczością oraz charakterystyk tych obszarów jest istotna dla opracowywania strategii bezpieczeństwa na danym obszarze, prognozowania kosztów generowanych przez poszczególne kategorie przestępstw, odpowiedniego rozmieszczenia komisariatów, patroli policji czy szacowania kosztów związanych z funkcjonowaniem organów ścigania. Jednocześnie należy mieć na uwadze, że nie ma „prostego" przełożenia zależności uzyskanych dla danych zagregowanych na indywidualne zachowania jedno-

[12] Metody te opisane są m.in. w pracy Trzęsiok (2015).

stek. Rankingi ważności predyktorów wskazywały drugorzędną rolę czynników o charakterze społeczno-ekonomicznym, jak stopa bezrobocia długoterminowego, ubóstwo, odsetek młodych osób mających niskie wykształcenie. Tymczasem prowadzone w Polsce badania na danych indywidualnych dotyczących sprawców przestępstw wyraźnie wskazują na związek między sytuacją społeczno-ekonomiczną jednostki a podejmowaniem decyzji o popełnianiu przestępstw. Przykładowo w 2010 roku w Polsce najliczniejszą grupę podejrzanych o dokonanie czynu niezgodnego z prawem stanowiły osoby bezrobotne lub niepracujące i nieszukające pracy (Szymanowski 2012, 93).

Bibliografia

Arbia, Giuseppe i Myriam Tabasso. 2013. *Spatial econometric modeling of massive datasets: The contribution of data mining*. https://ideas.repec.org/p/wiw/wiwrsa/ersa13p1004.html (dostęp: 05.10.2015).

Arendt, Łukasz i Elżbieta Kryńska. 2011. *Rynek pracy i kierunki wzrostu aktywności zawodowej ludności na obszarze zachodnich województw Polski w kontekście prowadzonej polityki regionalnej*. http://polskazachodnia2020.pl/ekspertyzy.html (dostęp: 02.06.2013)

Becker, Gary S. 1968. „Crime and punishment: an economic approach". *Journal of Political Economy* 76 (2): 169−217.

Besci, Zsolt. 1999. „Economics and crime in the States". *Economic Review* 84 (1): 38−56, http://www.frbatlanta.org/filelegacydocs/zbecsi.pdf (dostęp: 28.12.2012).

Bieniek, Piotr, Stanisław Cichocki i Maria Szczepaniec. 2012. „Czynniki ekonomiczne a poziom przestępczości – badanie ekonometryczne". *Zeszyty Prawnicze* 12 (1): 147−172.

Bobrowska, Agnieszka i Aleksandra Piasecka. 2002. „Bezrobocie a przestępczość w Polsce – próba określenia związku przyczynowo – skutkowego tych zjawisk". W: *Demograficzne i społeczne aspekty rozwoju miast*, red. Janusz Słodczyk, 231−239. Opole: Wydawnictwo Uniwersytetu Opolskiego

Bogacka, Emilia. 2009. „Poziom i czynniki przestępczości w układzie regionalnym Polski". *Biuletyn Instytutu Geografii Społeczno-Ekonomicznej i Gospodarki Przestrzennej UAM Seria Rozwój Regionalny i Polityka Regionalna* 8: 33−43.

Bogacka, Emilia. 2012. *Struktura przestrzenna i czynniki przestępczości na obszarze nadgranicznym Polski z Niemcami*. Studia i Prace Geografii i Geologii 25. Poznań: Bogucki Wydawnictwo Naukowe.

Bułat, Kamil, Paweł Czarniak, Anna Gorzelak, Krzysztof Grabowski, Magdalena Czub, Mikołaj Iwański, Przemysław Jakubek, Jan Jodłowski, Milena Małek, Sylwia Młodawska-Mąsior, Alicja Pieprz i Maria Stożek. 2007. *Kryminologia*. Warszawa: Oficyna a Wolters Kluwer Business.

Cherian, John i Mitchell Dawson. 2015. *RoboCop: Crime Classification and Prediction in San Francisco*. www.cs229.stanford.edu/proj2015/254_report.pdf (dostęp: 23.03.2016).

Cohen, Lawrence E. i Marcus Felson. 1979. „Social change and crime trends: a routine activity approach". *American Sociological Review* 44 (4): 588−608..

Cracolici, Maria F. i Teodora E. Uberti. 2008. „Geographical Distribution of Crime in Italian Provinces: A Spatial Econometric Analysis". *Social Science Research Network Electronic Paper Collection*. http://ssrn.com/abstract=1105082 (dostęp: 02.02.2014).

Ehrlich, Isaak. 1973. „Participation in illegitimate activities: a theoretical and empirical investigation". *The Journal of Political Economy* 81 (3): 521−565.

Entorf, Horst i Hannest Spengler. 2000. „Socioeconomic and demographic factors of crime in Germany. Evidence from panel data of the German states". *International Review of Law and Economics* 20 (1): 75−106.

Falcone, Marianna i Rosetta Lombardo. 2011. *Crime and Economic Performance. A Cluster Analysis of Panel Data on Italy's Nuts 3 Regions*. Working Paper no. 12–2011.Università della Calabria. www.ecostat.unical.it/RePEc/WorkingPapers/WP12_2011.pdf (dostęp: 05.03.2016)

Florczak, Waldemar. 2013. *Co wywołuje przestępczość i jak ją ograniczać? Wielowymiarowa analiza makroekonomiczna*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Frieske, Kazimierz. 2007. „Przestępczość w Polsce na przełomie stuleci. Stereotypy i realia". W: *Wymiary życia społecznego. Polska na przełomie XX i XXI wieku*, red. Mirosława Marody, 212−240. Warszawa: Wydawnictwo Naukowe SCHOLAR.

Gatnar, Eugeniusz. 2001. *Nieparametryczna metoda dyskryminacji i regresji*. Warszawa: Wydawnictwo Naukowe PWN.

Gatnar, Eugeniusz. 2008. *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji.* Warszawa: Wydawnictwo Naukowe PWN.

Gorr, Wilpen, Andreas Olligschlaeger i Yvonne Thompson. 2003. „Short-term forecasting of crime". *International Journal of Forecasting* 19: 579−594.

Groves, W. Byron i Robert J. Sampson. 1989. „Community structure and crime: testing social-disorganization theory". *The American Journal of Sociology* 94 (4): 774−802.

Han, Lu. 2009. *Economic Analyses of Crime in England and Wales*. University of Birmingham Research Archive e-theses repository. http://etheses.bham.ac.uk/584/ (dostęp: 05.03.2016)

Hołyst, Brunon. 2001. *Kryminologia*. Warszawa: Wydawnictwo Prawnicze LexisNexis.

Hothorn Torsten, Kurt Hornik i Achim Zeileis. 2006. „Unbiased recursive partitioning: A conditional inference framework". *Journal of Computational and Graphical Statistics*, 15 (3): 651−674.

Jefmański, Bartłomiej i Marta Kusterka-Jefmańska. 2012. „Determinanty satysfakcji klientów z usług jednostek administracji publicznej – na przykładzie urzędu miasta w Dzierżoniowie". W: *Orientacja na wyniki we współczesnej gospodarce*, red. Tadeusz Borys i Piotr Rogala, 208−212. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 265. Wrocław: Uniwersytet Ekonomiczny we Wrocławiu.

Kądziołka, Kinga. 2013. „Przestrzenne zróżnicowanie poziomu przestępczości w Polsce". W: *Problemy społeczno-ekonomiczne w uwarunkowaniach ryzy-*

ka i statystycznej nieokreśloności: metody i modele w rozwoju regionów, red. Włodzimierz Szkutnik, 101−114. Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Kądziołka, Kinga. 2014a. „Wpływ wybranych czynników na rozmiarprzestępczości stwierdzonej w Polsce". W: *Problemy społeczno-ekonomiczne w relacjach międzynarodowych. Analiza modelowa rozwoju regionów*, red. Włodzimierz Szkutnik, 46−63. Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Kądziołka, Kinga. 2014b. „Modele ekonomiczne w analizie zjawiska przestępczości". *Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach* 206 (14): 46−60.

Kądziołka, Kinga. 2014c. „Identyfikacja skupień obszarów wysokiej przestępczości z wykorzystaniem statystyki przestrzennej". W: *Rola informatyki w naukach ekonomicznych i społecznych. Innowacje i implikacje interdyscyplinarne*, *2/2014*, red. Zbigniew E. Zieliński, 110−121. Kielce: Wydawnictwo Wyższej Szkoły Handlowej.

Kądziołka, Kinga. 2015a. *Determinanty przestępczości w Polsce. Aspekt ekonomiczno-społeczny w ujęciu modelowania ekonometrycznego*. Niepublikowana rozprawa doktorska. Uniwersytet Ekonomiczny w Katowicach.

Kądziołka, Kinga. 2015b. „Analiza czynników wpływających na przestrzenne zróżnicowanie przestępczości w Polsce na poziomie podregionów". *Współczesna Gospodarka* 6 (3): 43−52. www.wspolczesnagospodarka.pl (dostęp: 05.10.2015).

Kądziołka, Kinga. 2015c. „Sytuacja społeczno-ekonomiczna mieszkańców a przestępczość w Polsce". W: *Rola informatyki w naukach ekonomicznych i społecznych. Innowacje i implikacje interdyscyplinarne, 1/2015*, red. Zbigniew E. Zieliński, 83−92. Kielce: Wydawnictwo Wyższej Szkoły Handlowej.

Kądziołka, Kinga. 2015d. „Bezrobocie, ubóstwo i przestępczość w Polsce. Analiza zależności na poziomie województw". *Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach* 242: 71−84.

Kądziołka, Kinga. 2016. „Przestrzenno-czasowa analiza zjawiska przestępczości w województwie śląskim". *Kwartalnik Prawo − Społeczeństwo – Ekonomia* 1: 81−95.

Kiersztyn, Anna. 2008. *Czy bieda czyni złodzieja? Związki między bezrobociem, ubóstwem a przestępczością*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Kopczewska, Katarzyna. 2011. *Ekonometria i statystyka przestrzenna z wykorzystaniem programu R CRAN*. Warszawa: CeDeWu.

Lauridsen, Jørgen. 2010. „Is Polish crime economically rational?". *The Journal of Regional Analysis & Policy* 40 (2): 125−131.

Lauridsen, T. Jørgen, Fatma Zeren i Ayşe Ari. 2013. „A spatial panel data analysis of crime rates in EU". *Discussion Papers on Business and Economics* no 2.

Liaw, Andy i Matthew Wiener. 2002. „Classification and regression by randomForest". *R News* 2 (3): 18−22. http://CRAN.R-project.org/doc/Rnews (dostęp31.07.2015)

Meyer, Jona i Sanjeev Sridharan. 2005. *Exploratory Spatial Data Approach to Identify the Context of Unemployment − Crime Linkages in Virginia, 1995−2000*. https://www.ncjrs.gov/pdffiles1/nij/grants/208937.pdf (dostęp: 05.08.2013).

Mielecka-Kubień, Zofia. 2013. „Migracje wojewódzkie na pobyt stały w województwie śląskim w 2010 roku w świetle praw migracji E.G. Ravensteina". W: *Perspektywy rozwoju górnego śląska. Analiza ekonometryczno-statystyczna*, red. Andrzej S. Barczak, 24−40. Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

Mordwa, Stanisław. 2011. „Kradzieże w przestrzeni Łodzi". *Acta Universitatis Lodziensis Folia Geographica Socio-Oeconomica* 11: 1−20.

Rozmus, Dorota. 2009. „Nieobciążona metoda rekurencyjnego podziału". W: *Zastosowania ekonometrii*, red. Andrzej S. Barczak, 137−146. Katowice: Akademia Ekonomiczna im. Karola Adamieckiego w Katowicach.

Shaw, Clifford i Henry D. McKay. 1942. *Juvenile Delinquency and Urban Areas*. Chicago: University of Chicago Press.

Sypion-Dutkowska, Natalia. 2014. *Uwarunkowania przestrzenne przestępczości w wielkim mieście w ujęciu GIS (na przykładzie Szczecina)*. Warszawa: Polska Akademia Nauk Komitet Przestrzennego Zagospodarowania Kraju.

Szczepaniec, Maria. 2012. *Teoria ekonomiczna w prawie karnym*. Warszawa: Wydawnictwo Uniwersytetu Kardynała Stefana Wyszyńskiego.

Sztaudynger, Jan J. i Marcin Sztaudynger. 2003. „Ekonometryczne modele przestępczości". *Ruch Prawniczy, Ekonomiczny i Socjologiczny* 3: 127−143.

Szymanowski, Teodor. 2012. *Recydywa w Polsce: zagadnienia prawa karnego, kryminologii i polityki karnej*. Warszawa: Wolters Kluwer Polska.

Trzęsiok, Joanna. 2015. „O odporności na obserwacje odstające wybranych nieparametrycznych metod regresji". *Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach* 227: 75−84.

Woźniakowska, Dagmara. 2006. *Skazani i byli skazani na rynku pracy – ocena problemu z punktu widzenia organizacji pozarządowych*. Fundacja Inicjatyw Społeczno-Ekonomicznych. http://www.fise.org.pl/files/1bezrobocie.org.pl/public/Raporty/DWozniakowska_raport_dot_wiezniow.pdf (dostęp: 25.01.2014).

Strona internetowa Centralnego Zarządu Służby Więziennej: http://sw.gov.pl/pl/o-sluzbie-wieziennej/statystyka/statystyka-roczna/(dostęp: 03.10.2015).

Strona internetowa Głównego Urzędu Statystycznego (Bank Danych Lokalnych): http://stat.gov.pl/bdl/app/strona.html? p_name=indeks (dostęp: 05.12.2015).

Determinants of crime rate in Poland. Analysis using regression trees

Abstract

The aim of this article is to identify relationships between crime rate and some socio – economic, demographic and environmental factors in the poviats of Poland. There were analysed cross – sectional data using regression tree. The following factors were found to significantly explain the intensity of crime rate: urbanisation, percentage of single-person households, provided accommodation per 1000 population, divorce's coefficient and the intensity of crime in the neighboring areas. Then the random forest was used to improve prediction's accuracy and generate rank of variable importance.

Keywords: crime rate, regression tree, random forest, determinants of crime
JEL Code: C1, K42, R1
DOI: 10.17451/eko/45/2016/186