PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107

Taksonomia 17 2010

Małgorzata Misztal
Uniwersytet Łódzki

O ZASTOSOWANIU METODY REKURENCYJNEGO PODZIAŁU W ANALIZIE PRZEŻYCIA

Streszczenie: Analiza przeżycia obejmuje metody analizy danych, w których badaną zmienną jest czas do pojawienia się danego zdarzenia (czas przeżycia). Do najpopularniejszych metod analizy czasów przeżycia należą krzywe przeżycia Kaplana-Meiera oraz model proporcjonalnego hazardu Coksa. Alternatywą dla tych metod może być metoda rekurencyjnego podziału.

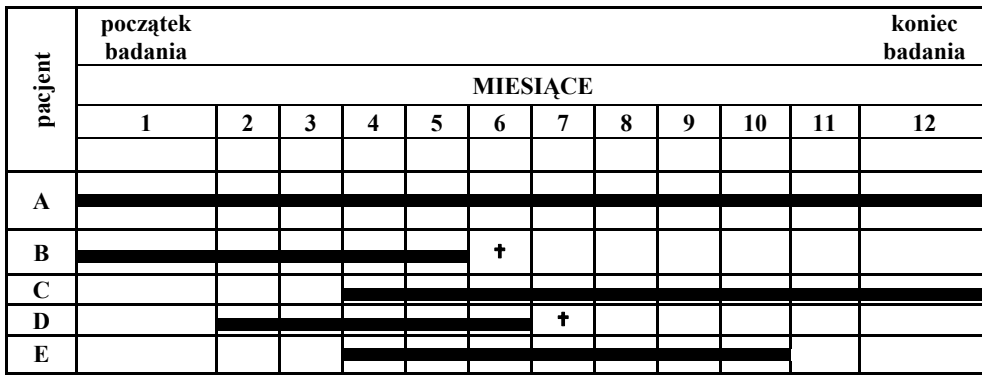
W artykule przedstawiono wyniki zastosowania pojedynczych i zagregowanych modeli drzew przeżycia (*survival trees*) do analizy czasów przeżycia pacjentów ze schorzeniami kardiologicznymi oraz przybliżono możliwości wykonywania niezbędnych obliczeń z wykorzystaniem środowiska R.

1. Wstęp

Analiza przeżycia obejmuje metody analizy danych, w których badaną zmienną jest czas do pojawienia się danego zdarzenia. *Zdarzeniem* jest dowolna zmiana stanu badanej cechy pierwotnej – np. w diagnostyce medycznej zdarzeniem może być zgon pacjenta, w kontroli jakości – awaria urządzenia, a w analizie ryzyka kredytowego – zaprzestanie spłacania zaciągniętego kredytu. Okres pomiędzy stanem początkowym a chwilą wystąpienia zdarzenia to tzw. czas przeżycia.

W analizie przeżycia mamy do czynienia z *obserwacjami uciętymi* (*cenzurowanymi*), czyli takimi obserwacjami, dla których mamy pewne informacje o czasie przeżycia, ale czas ten nie jest dokładnie znany. Na rysunku 1 przedstawiono przykłady obserwacji uciętych w badaniach medycznych.

Pacjent A był włączony do badania w dniu jego rozpoczęcia i żyje w dniu zakończenia badań – jest to obserwacja ucięta (wiemy, że przeżył 12 miesięcy, dokładnego czasu przeżycia nie znamy). Pacjent B został włączony do badania w chwili jego rozpoczęcia i zmarł w 5 miesiącu obserwacji – znamy jego czas przeżycia: 5 miesięcy. Pacjent C był włączony do badania od 4 miesiąca trwania badania i żyje w dniu zakończenia badań (obserwacja ucięta). Pacjent D został włączony do badania w 2 miesiącu, a zmarł w 6 miesiącu (znamy czas przeżycia – 5 miesięcy). Pacjent E wreszcie był włączony do badania od 4 miesiąca, tracimy z nim kontakt w 10 miesiącu badania (obserwacja ucięta, wiemy, że przeżył co najmniej 7 miesięcy).



Rys. 1. Przykłady obserwacji uciętych

Źródło: opracowanie własne.

Do podstawowych pojęć w analizie przeżycia należy także funkcja przeżycia i funkcja hazardu. *Funkcja przeżycia* podaje prawdopodobieństwo, że dany obiekt przeżyje dłużej niż pewien przyjęty czas t (innymi słowy – oblicza prawdopodobieństwo, że obiekt dożył aż do czasu t). *Funkcja hazardu* (zwana też *funkcją intensywności* lub *natężenia zdarzeń* – por. [Balicki 2006]) podaje wartość pojawienia się zdarzenia na jednostkę czasu (innymi słowy – jest to warunkowa funkcja gęstości wystąpienia zdarzenia w chwili t, pod warunkiem że do tej chwili obiekt takiego zdarzenia nie doznał).

Typowym zagadnieniem w analizie przeżycia jest poszukiwanie zmiennych istotnie wpływających na czas przeżycia. Występowanie obserwacji uciętych wyklucza zastosowanie tradycyjnych metod statystycznych, np. analizy regresji. Do najczęściej wykorzystywanych metod w analizie przeżycia należą: estymacja funkcji przeżycia metodą Kaplana-Meiera oraz model proporcjonalnego hazardu Coksa. Jako alternatywę do tych metod można zaproponować metodę rekurencyjnego podziału, której graficzną prezentacją są drzewa przeżycia (*survival trees*).

W pracy przedstawiono przykład zastosowania wymienionych metod (ze szczególnym uwzględnieniem metody rekurencyjnego podziału) w diagnostyce medycznej oraz przybliżono możliwości wykonywania niezbędnych obliczeń z wykorzystaniem środowiska R.

2. Materiał i metody

Analizie poddano dane rzeczywiste dotyczące 118 pacjentów chorych na HCM (kardiomiopatię przerostową) opisanych zestawem 35 zmiennych charakteryzujących stan pacjenta. Interesującym nas zdarzeniem (punktem końcowym) jest zgon pacjenta. Szukamy zmiennych istotnie wpływających na wystąpienie punktu koń-

cowego oraz czas przeżycia pacjenta. Do analizy wielowymiarowej wykorzystano 9 zmiennych, które po przeprowadzeniu analizy jednowymiarowej okazały się mieć istotny wpływ na wystąpienie punktu końcowego.

W rozważanym przykładzie mamy do czynienia z cenzurowaniem losowym (por. [Balicki 2006]). Pacjenci wchodzą do badania w różnym okresie, a obserwacja każdego pacjenta kończy się po 48 miesiącach. Obserwacje ucięte dotyczą tych wszystkich pacjentów, u których nie wystąpił zgon (są to obserwacje prawostronnie cenzurowane).

Do analizy czasu przeżycia oraz identyfikacji zmiennych istotnie wpływających na wystąpienie punktu końcowego i czas przeżycia wykorzystano metodę Kaplana-Meiera, regresję Coksa oraz drzewa przeżycia.

Metoda Kaplana-Meiera szacuje parametry funkcji przeżycia na podstawie danych zawierających czas przeżycia oraz informację o wystąpieniu lub nie punktu końcowego. Graficzną prezentacją tej metody jest krzywa przeżycia Kaplana-Meiera.

Analiza Kaplana-Meiera pozwala porównywać czas przeżycia w różnie zdefiniowanych grupach (np. według płci), ale nie daje możliwości modelowania wystąpienia zdarzenia w zależności od określonych mierzalnych czynników. Do tego celu można wykorzystać model proporcjonalnego hazardu Coksa postaci [Stanisz 2005, s. 355]:

$$h(t : x_1, x_2, ..., x_n) = h_0(t) \exp(a_1 x_1 + a_2 x_2 + ... + a_n x_n), \quad (1)$$

gdzie $h(t : x_1, x_2, ..., x_n)$ oznacza wynikowy hazard przy danych *n* zmiennych towarzyszących i odpowiednim czasie przeżycia. Wielkość $h_0(t)$ nosi nazwę hazardu odniesienia lub zerowej linii hazardu i oznacza wielkość hazardu, gdy wszystkie zmienne niezależne przyjmują wartość 0. Model proporcjonalnego hazardu Coksa zakłada, że iloraz hazardu dla dwóch zmiennych niezależnych *x* i *x'* jest niezależny od czasu (założenie proporcjonalności) oraz że istnieje log-liniowa zależność między zmiennymi niezależnymi a funkcją hazardu.

Użytecznym narzędziem do modelowania związków między czasem przeżycia a zestawem zmiennych objaśniających może być również metoda rekurencyjnego podziału (por. [Cappelli, Zhang 2007]). Wynika to z jej nieparametrycznego charakteru, braku wymagań co do rozkładu zmiennych i zależności między nimi.

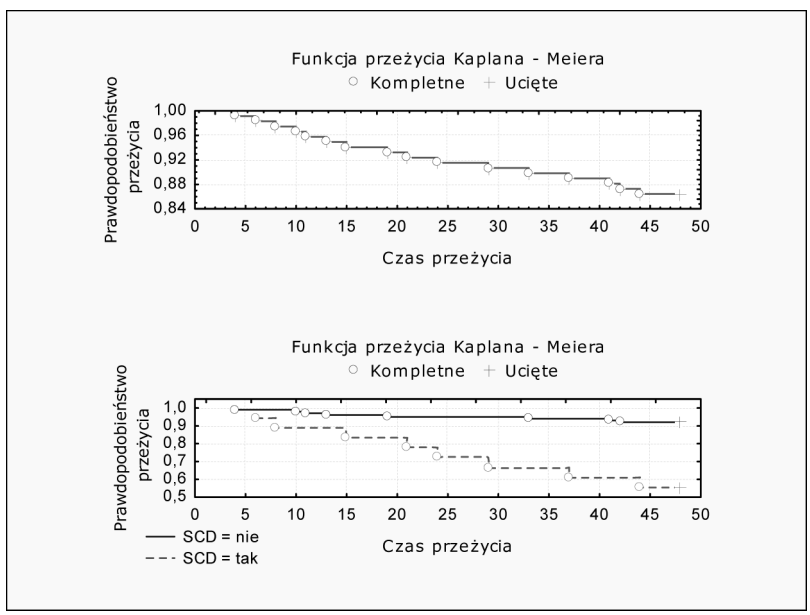
Najogólniej dokonywany jest tu podział zbioru obiektów na jednorodne podzbiory, a następnie w każdym z węzłów końcowych szacowane są funkcje przeżycia, np. metodą Kaplana-Meiera. Mamy zatem możliwość wyodrębnienia podgrup pacjentów charakteryzujących się podobnym rozkładem czasu przeżycia. Graficzną prezentacją metody będą drzewa przeżycia.

Wszystkie obliczenia wykonano z wykorzystaniem pakietu STATISTICA oraz środowiska R (pakiety: `survival`, `rpart`, `ipred`, `party`, `RandomSurvivalForest`).

3. Wyniki

Na rysunku 2 przedstawiono krzywą przeżycia Kaplana-Meiera dla danych ogółem oraz krzywe przeżycia w grupach wyodrębnionych według zmiennej SCD (nagła śmierć sercowa w rodzinie).

Jak widać, prawdopodobieństwo przeżycia do 48 miesiąca jest dość wysokie – ok. 86%. Ciekawsze są wyniki uzyskane w podgrupach według rodzinnego obciążenia wystąpieniem SCD. Uzyskane krzywe przeżycia różnią się istotnie (p = 0,00004 – test log-rank). Brak obciążenia SCD w sposób istotny przedłuża czas przeżycia w porównaniu z osobami, u których w rodzinie miały miejsce przypadki SCD (w tej grupie prawdopodobieństwo przeżycia do 48 miesiąca wynosi zaledwie 55%).



Rys. 2. Krzywe przeżycia Kaplana-Meiera dla danych ogółem i według wystąpienia SCD

Źródło: opracowanie własne.

Analogicznie można analizować czasy przeżycia w podgrupach wyodrębnionych ze względu na wartości innych zmiennych.

Oszacowania parametrów modelu Coksa dokonano z wykorzystaniem środowiska R (pakiet `survival`).

```
> require(survival)
> HCM.cox=coxph(Surv(Czas,PK)~.,data=HCM, control=coxph.control
(iter.max=100))
> HCM.cox.zph=cox.zph(HCM.cox)   # założenie proporcjonalności.
```

Założenie proporcjonalności dla analizowanego modelu zostało spełnione[1]. Cały uzyskany model jest istotny statystycznie ($p = 0{,}0004$).

Oszacowania parametrów modelu przedstawia tab. 1. 6 zmiennych okazało się mieć istotny statystycznie wpływ na wystąpienie punktu końcowego. Uzyskane wyniki są zgodne z interpretacją medyczną.

Podany w ostatniej kolumnie współczynnik ryzyka definiowany jest jako stosunek wartości ryzyka dla jednego pacjenta do wartości ryzyka dla innego pacjenta [Stanisz 2005].

Tabela 1. Oszacowania parametrów modelu Coksa

Zmienna	*b* – ocena współczynnika	SE	*p*	exp(*b*) – współczynnik ryzyka (hazard względny)
SCD	2,5855	0,7568	**0,0006**	13,2701
Omd	1,4447	0,6860	**0,0352**	4,2408
LA	0,0636	0,0620	0,3051	1,0656
Grad	1,8855	0,7760	**0,0151**	6,5897
Em	–1,1212	0,3998	**0,0050**	0,3259
Gr_max	0,0740	0,0737	0,3154	1,0768
BNP	0,0013	0,0012	0,2872	1,0013
E.Vp	1,3385	0,6146	**0,0294**	3,8132
E.Em	–0,2965	0,1390	**0,0329**	0,7434

Źródło: obliczenia własne.

Najwyższą wartość współczynnika hazardu uzyskujemy dla zmiennej SCD – ryzyko zgonu jest ponad 13 razy wyższe u pacjenta, u którego w rodzinie miał miejsce epizod SCD niż u pacjenta nieobciążonego dziedzicznie.

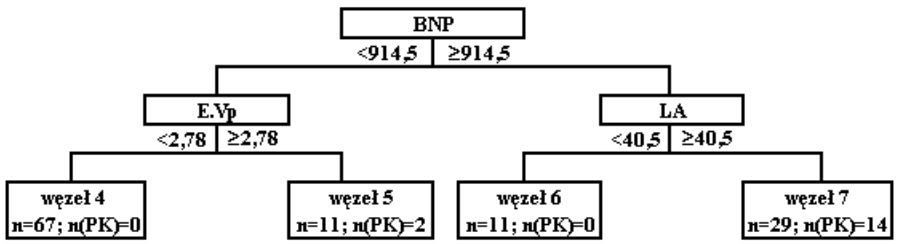
Drzewa przeżycia (*survival trees*) można utworzyć, korzystając m.in. z pakietów `rpart` i `party`.

Drzewo przeżycia uzyskane z wykorzystaniem pakietu `rpart` przedstawiono na rys. 3. Drzewo ma 4 węzły końcowe. W liściach podano liczbę obserwacji w danym węźle oraz liczbę pacjentów, u których wystąpił punkt końcowy (PK).

```
> require(rpart)
> HCM.rp=rpart(Surv(Czas,PK)~.,data=HCM, control=rpart.control
(minsplit=5, cp=0.05))
> print(HCM.rp)
```

Jak łatwo zauważyć, w węzłach 4 i 6 brak jest pacjentów, u których nastąpił zgon. Większość takich osób znalazła się w węźle 7, a 2 osoby w węźle 5. Krzywe przeżycia przedstawiono na rys. 4. Zwróćmy uwagę, że tylko ok. 52% pacjentów, którzy znaleźli się w węźle 7, przeżyło do 48 miesiąca.

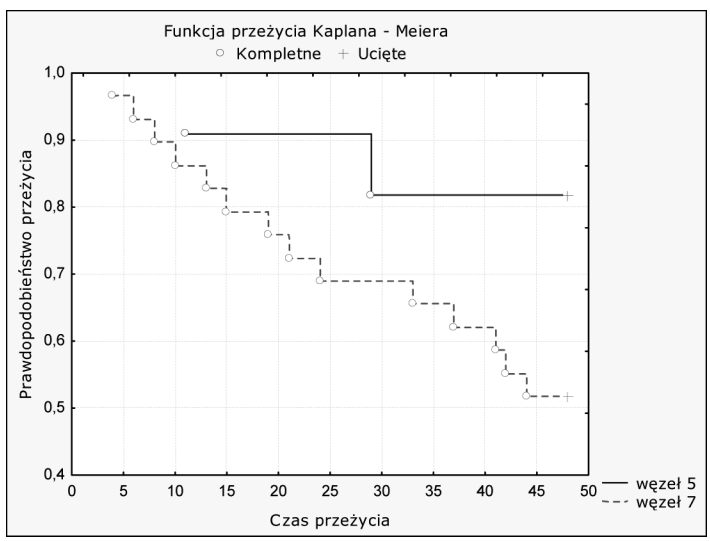
[1] Ze względu na ograniczoną objętość pracy brak prezentacji szczegółowych wyników.



Rys. 3. Drzewo przeżycia – rpart

Źródło: opracowanie własne.

W wyniku zastosowania pakietu party (por. [Hothorn, Hornik, Zeileis 2006]) dostajemy drzewo przedstawione na rys. 5.



Rys. 4. Krzywa przeżycia dla węzłów 5 i 7

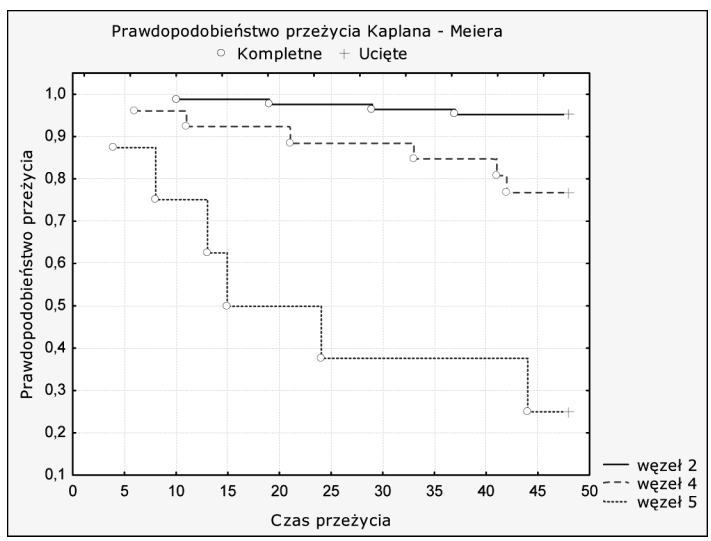
Źródło: opracowanie własne.

Uzyskane drzewo ma 3 liście. Krzywe przeżycia w każdym z liści przedstawiono na rys. 5. Jak widać, zdecydowanie najmniej przeżywających do 48 miesiąca (zaledwie 25%) obserwujemy w węźle 5 (osoby, dla których LA – wielkość przedsionka ≥ 44 mm i maksymalna grubość ściany Gr_max ≥25 mm).

Poprawę stabilności utworzonych modeli drzew można uzyskać, wykorzystując modele zagregowane. Problematykę tę podejmował np. Breiman [2002] – rozszerzenie metody *Random Forests* na potrzeby analizy przeżycia, oraz Hothorn i in.

[2004; 2005] – agregacja modeli za pomocą metody *bagging* (implementacja w pakiecie `ipred`).

Interesującą propozycję przedstawiono także w pracy [Ishwaran i in. 2008]. Autorzy zaproponowali sposób wykorzystania algorytmu *Random Forests* do analizy danych prawostronnie cenzurowanych. Procedura tworzy *n* pojedynczych drzew przeżycia, łączonych następnie w jeden model zagregowany, za pomocą którego szacowana jest skumulowana funkcja hazardu (CHF). Zaimplementowane zostały 4 metody wyboru zmiennych do podziału w węzłach (oparte głównie na testach typu log-rank). Do oceny jakości predykcji, na podstawie zbioru OOB, szacowana jest wartość współczynnika zgodności Harrella (*Harrell's concordance index*). Użytecznym dodatkiem jest możliwość oceny przydatności zmiennych objaśniających do prognozowania czasu przeżycia. Wszystkie obliczenia wykonać można z wykorzystaniem pakietu `randomSurvivalForest`:

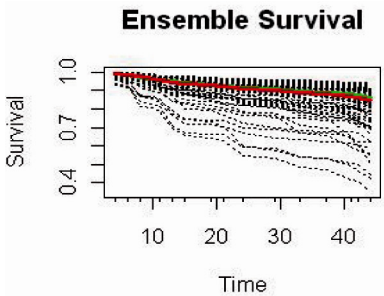


Rys. 5. Krzywe przeżycia w liściach drzewa.

Źródło: opracowanie własne.

```
> require(randomSurvivalForest)
> HCM.rsf=rsf(Survrsf(Czas, PK)~., data=HCM, ntree=100,
forest=TRUE)
> print(HCM.rsf)
> plot.ensemble(HCM.rsf)          # prezentacja graficzna wyników
> plot.error(HCM.rsf)             # prezentacja graficzna oceny błędu
oraz rankingu zmiennych
```

Agregacji poddano 100 drzew. Uzyskane wyniki podsumowano na rys. 6. Oszacowany z wykorzystaniem zbioru OOB błąd wynosi ok. 15% (por. rys. 7).




```
Sample size: 118
Number of deaths: 16
Number of trees: 100
Minimum terminal node size: 3
Average no. of terminal nodes: 2.74
No. of variables tried at each split: 3
Total no. of variables: 9
Splitting rule: logrank
Estimate of error rate: 14.73%
```

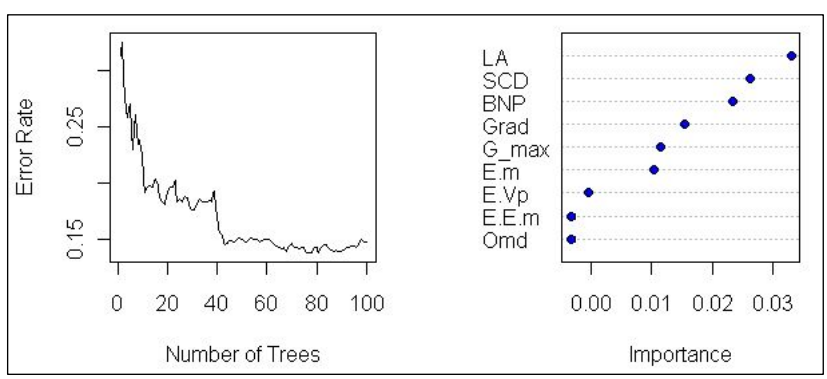
Rys. 6. Podsumowanie wyników dla modelu zagregowanego

Źródło: opracowanie własne.

Na rysunku 7 przedstawiono dodatkowo ranking ważności predyktorów. Wartości równe 0 lub ujemne oznaczają, że zmienna nie ma wpływu na kształtowanie się zmiennej zależnej. Im wyższa wartość, tym większa przydatność danej zmiennej w przewidywaniu czasu przeżycia – zatem najbardziej istotne zmienne to LA, SCD oraz BNP.

4. Uwagi końcowe

Metoda rekurencyjnego podziału może być stosowana jako uzupełnienie stosowanych zwykle metod – estymacji funkcji przeżycia metodą Kaplana-Meiera i modelu proporcjonalnego hazardu Coksa. Jej zaletą jest możliwość wyodrębnienia grup pacjentów podobnych pod względem czasu przeżycia i przejrzysta wizualizacja uzyskanych wyników w przypadku pojedynczych modeli.



Rys. 7. Błąd predykcji i ranking ważności predyktorów

Źródło: opracowanie własne.

Zwróćmy uwagę, że w prezentowanym przykładzie zmienne, dla których oceny parametrów w modelu Coksa były nieistotne statystycznie, okazały się przydatne do podziału pacjentów na bardziej jednorodne podgrupy.

Ocena przydatności drzew przeżycia, w sensie poprawy dokładności predykcji, w porównaniu z modelami tradycyjnie stosowanymi wymaga jednak przeprowadzenia dodatkowych analiz. Będzie to przedmiotem dalszych badań.

Literatura

Balicki A., *Analiza przeżycia i tablice wymieralności*, PWE, Warszawa 2006.

Breiman L., *How to Use Survival Forests*, http://stat-www.berkeley.edu/~breiman, 2002.

Cappelli C., Zhang H., *Survival Trees*, [w:] *Statistical Methods for Biostatistics and Related Fields*, W. Hardle, Y. Mori, P. Vieu (red.), Springer Berlin Heidelberg, 2007.

Hothorn T., Buhlmann P., Dudoit S., Molinaro A.M., van der Laan M.J., *Survival Ensembles*, U.C. Berkeley Division of Biostatistics Working Paper Series 2005 no 174, University of California, Berkeley; http://www.bepress.com/ucbbiostat/paper174.

Hothorn T., Hornik K., Zeileis A., *Unbiased recursive partitioning: a conditional inference framework*, „Journal of Computational and Graphical Statistics" 2006 vol. 15, no 3.

Hothorn T., Lausen B., Benner A., Radespiel-Troger M., *Bagging survival trees*, „Statistics in Medicine" 2004 no 23.

Ishwaran H., Kogalur U.B., Blackstone E.H., Lauer M.S., *Random survival forests*, „The Annals of Applied Statistics" 2008 vol. 2, no 3.

Stanisz A. (red.), *Biostatystyka*, Wydawnictwo UJ, Kraków 2005.

ON THE USE OF RECURSIVE PARTITIONING METHOD IN SURVIVAL ANALYSIS

Summary: Survival data deals with survival time – i.e. the time to the occurrence of an event of interest. In medical research the event of interest is usually the time to death of a patient after the diagnosis.

Tree-based models can be a very useful alternative to the most popular methods to analyze survival data: Kaplan-Meier survival curves and Cox proportional hazards regression.

The results of the application of Kaplan-Meier method, Cox model and single and aggregated survival trees to predict survival time for patients with HCM are presented in the paper. All the calculations were made using the R environment.