



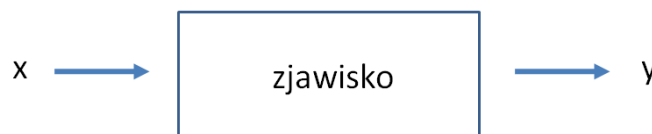
## ZASTOSOWANIE TECHNIK DATA MINING W BADANIACH NAUKOWYCH

Grzegorz Harańczyk, StatSoft Polska Sp. z o.o.

Zakres zastosowań analizy danych w różnych dziedzinach badań naukowych stale się poszerza. Wynika to w głównej mierze z coraz powszechniejszego przekonania, że przy rozwiązywaniu różnego rodzaju zagadnień poznawczych i praktycznych trzeba opierać się na empirycznych danych, opisujących badane zjawiska i procesy. Z drugiej strony dostępnych obecnie jest coraz więcej informacji, prawie wszystko jest mierzone, a pomiary archiwizowane. Bardzo szybko rośnie wolumen danych poddawanych analizie, stawiane są coraz to nowe problemy oraz rosną oczekiwania wobec uzyskiwanych wyników i modeli. W związku z tym musi poszerzać się również zakres stosowanych technik analizy danych.

### Wprowadzenie

Badania empiryczne polegają na przeprowadzaniu eksperymentów lub po prostu obserwowaniu pewnych zjawisk i wnioskowaniu na podstawie poczynionych obserwacji. Celem prowadzenia takich badań jest zrozumienie badanego zjawiska i przewidywanie jego przebiegu (i podobnych mu) w przyszłości. Przebieg badanego zjawiska opisywany jest za pomocą wielu mierzonych wartości, a sam jego wynik może zależeć od wielu czynników, te również są obserwowane podczas badań. Reasumując, schemat obiektu badań wygląda następująco:



Rys. 1. Schemat obserwowanego zjawiska.

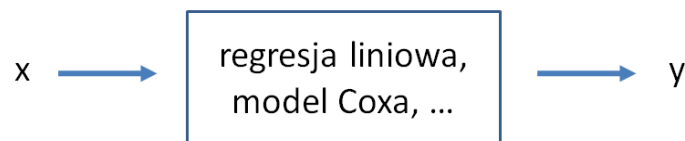
gdzie  $x$  to wektor wielkości wejściowych, natomiast  $y$  wielkość wyjściowa (wektor wielkości wyjściowych).

Najczęściej badania te prowadzone są po to, aby znaleźć odpowiedź na konkretnie postawione pytanie. Sama postać modelu lub badana hipoteza sformułowane są przed rozpoczęciem badań. Często jednak analiza wykonywana jest również wtórnie na zebranych już

danych - dane zbierane są dla innych badań lub są po prostu rejestrem pewnych zdarzeń, czy też są wielokrotnie wykorzystywane w poszukiwaniu najlepszego modelu. Należy jednak rozróżnić te dwie sytuacje, mając na uwadze ewentualne konsekwencje (por. tzw. *data snooping* [1, 9]).

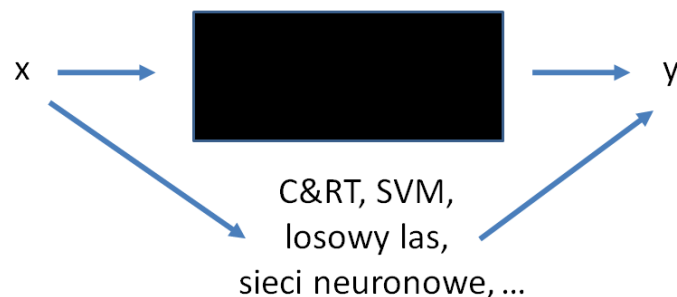
## Modelowanie statystyczne – dwa podejścia – wzorce a mechanizm

Leo Breiman (twórca między innymi algorytmu losowego lasu) w swoim artykule *Statistical Modeling - The Two Cultures* [3] rozróżnił dwa podejścia do analizy danych. Z jednej strony możemy założyć, że dane generowane są przez pewien mechanizm, który obserwujemy z pewną dokładnością i chcemy ten mechanizm poznać. Z drugiej zaś strony może nas nie interesować sam mechanizm, a jedynie wynik przez niego generowany. W klasycznym podejściu zakładamy, że znamy model opisujący badany proces generujący dane. Zakładamy na przykład, że mechanizm generujący dane jest pewnej postaci (regresja liniowa, regresja logistyczna, model proporcjonalnego hazardu Coxa). Naszym celem jest jedynie oszacowanie jego parametrów lub przetestowanie zaplanowanej hipotezy. Schemat takiego podejścia uwzględniającego mechanizm zjawiska został zaprezentowany na rys. 2. Podejście to sprawdza się przede wszystkim podczas badania prostych zjawisk o nie bardzo skomplikowanym przebiegu.



Rys. 2. Podejście do analizy danych uwzględniające mechanizm zjawiska.

W drugim podejściu traktujemy badany mechanizm jako nieznaną, godzimy się z tym i szukamy jedynie związków pomiędzy wielkościami wejściowymi a wyjściowymi. Nie zakłada się tu często żadnej postaci modelu czy związku pomiędzy badanymi wielkościami, analiza oparta jest jedynie na danych. Wykorzystywane do takich celów metody analizy danych często określa się mianem nieparametrycznych.



Rys. 3. Podejście do analizy danych oparte jedynie na danych, nieuwzględniające mechanizmu zjawiska.



Porównując oba podejścia na schematach (rys. 2, 3) można zrozumieć Breimana, który dzieli analityków na pracujących/modelujących „wewnątrz skrzynki” albo „na zewnątrz” (*“pretty clear cut—are you modeling the inside of the box or not?”*). Podobny podział prezentuje Hand w pracy [5].

## Nowe metody

Pojawiają się pytania: czy drugie podejście, nieuwzględniające mechanizmu badanego zjawiska nie jest zbyt pójściem na skróty? Dlaczego w ogóle się pojawia i czy może ono działać? Postaramy się częściowo odpowiedzieć na te pytania.

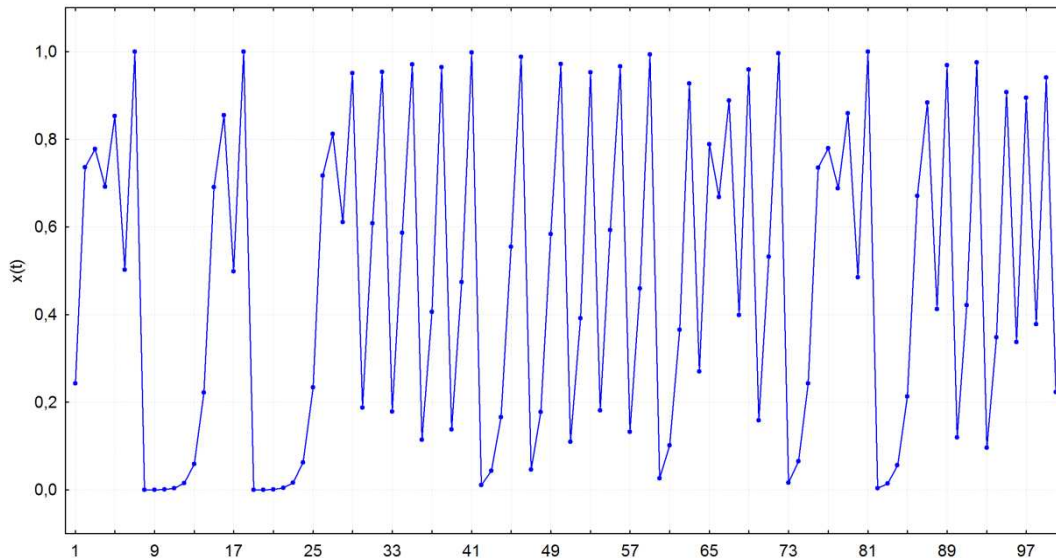
Z jednej strony stale i szybko rośnie ilość dostępnych danych (digitalizuje się wszystko – zbiory biblioteczne, wyniki obrazowania medycznego i inne), z drugiej strony przed analizą danych stawiane są coraz ambitniejsze cele. Badane są coraz bardziej skomplikowane struktury i zależności, w związku z czym, nie można od razu podać, czy choćby wstępnie zdefiniować modelu opisującego badane zjawisko. Dlatego coraz częściej wykorzystywane jest podejście oparte na danych, a nie mechanizmie zjawiska. Do rozwiązywania tego typu problemów – poszukiwania związków pomiędzy wielkościami wejściowymi i wyjściowymi – zostało stworzonych wiele nowych algorytmów, często określanymi mianem technik data mining. Nazwa wywodzi się z głównej cechy tych algorytmów – wielokrotnego iteracyjnego przeszukiwania zbioru danych w poszukiwaniu najlepszego modelu i ukrytych wzorców (kopania w danych). Data mining to jednak nie tylko nowe algorytmy analizy danych. Nowe techniki i algorytmy pociągają za sobą także nowe podejście i metodykę – ocena modelu nie na podstawie istotności statystycznej, ale na podstawie poprawności na zbiorze testowym. Opis metod i podejścia przedstawiony jest na przykład w pracach [2,4,6,7,10].

Podejście to działa w praktyce, a nawet wydaje się, że w pewnych sytuacjach jest koniecznością. Szczególnie popularne i wykorzystywane jest w zastosowaniach komercyjnych. Coraz częściej wykorzystywane jest również z powodzeniem w badaniach naukowych. Musi być ono jednak umiejętnie stosowane. Dla przykładu, cytując Breimana, modelowanie polegające jedynie na badaniu związków pomiędzy wejściem a wyjściem takie, że uzyskany model traktujemy jako model mechanizmu, prowadzi często do wątpliwych naukowych wniosków oraz nietrafnych teorii.

Jedno podejście oczywiście nie wyklucza drugiego, wręcz mogą się one uzupełniać na różnych etapach procesu poznawczego. Często w pierwszym etapie wykorzystanie podejścia opartego jedynie na danych daje nam pewną wiedzę i pozwala wykorzystać ją na przykład na etapie planowania nowych badań, których celem będzie zweryfikowanie postawionych hipotez. Poniżej zostanie zaprezentowane krótkie omówienie dwóch zagadnień będących motywacją do używania technik data mining w analizie danych. Zaprezentowano dwa charakterystyczne zastosowania tego typu technik – modele nieparametryczne oraz problemy klasyfikacyjne, w których liczba predyktorów znacznie przewyższa liczbę badanych obiektów (tzw. problemy  $p \gg N$ ).

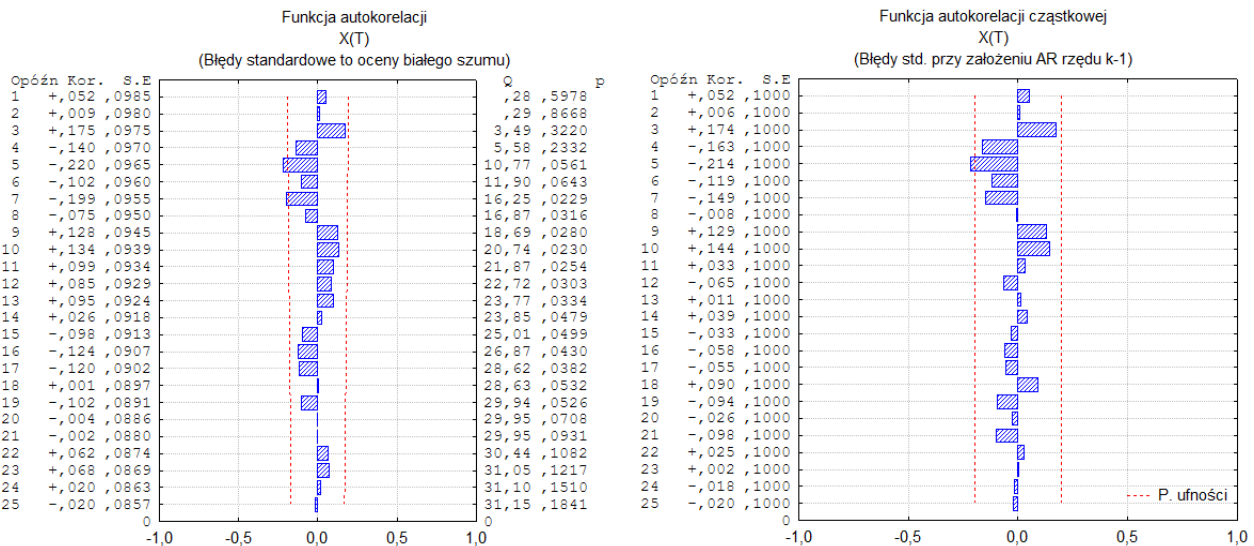
## Przykład – prognozowanie metodami nieparametrycznymi

Założmy, że pewna wielkość  $x(t)$  mierzona była w równych odstępach czasu. Poniżej znajduje się wykres liniowy tego szeregu czasowego.



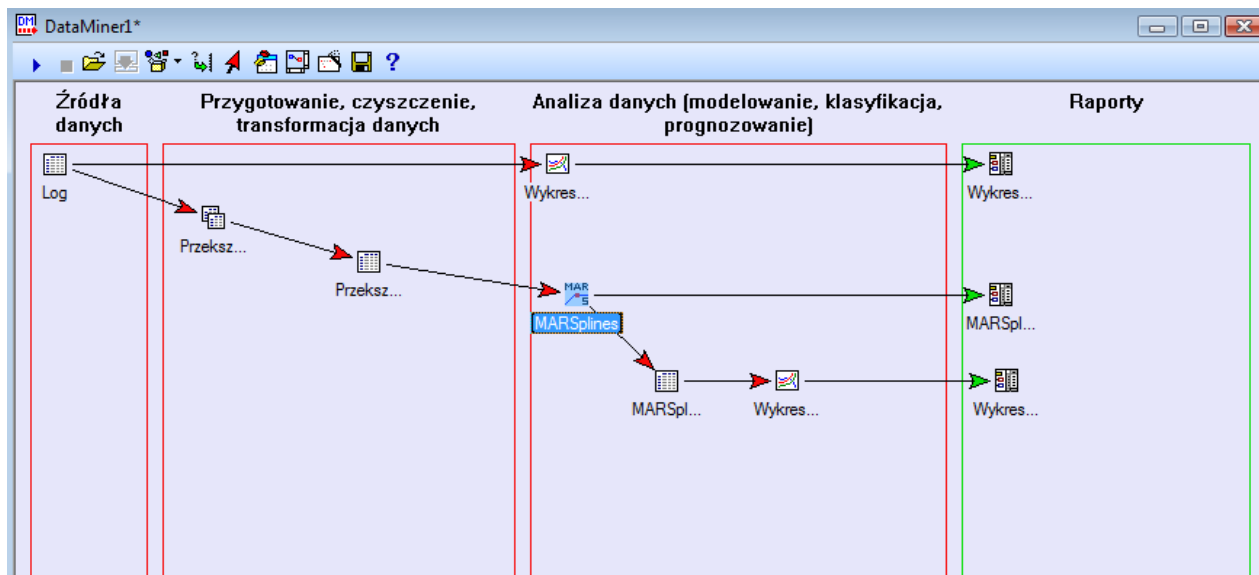
Rys. 4. Przebieg wartości dla zmiennej  $x(t)$ .

Podejrzmy do problemu klasycznie i zbadajmy na początek autokorelacje dla tego szeregu. Na rysunku poniżej widać brak istotnych autokorelacji. Czy ten szereg można przewidywać na podstawie wcześniejszych wartości?



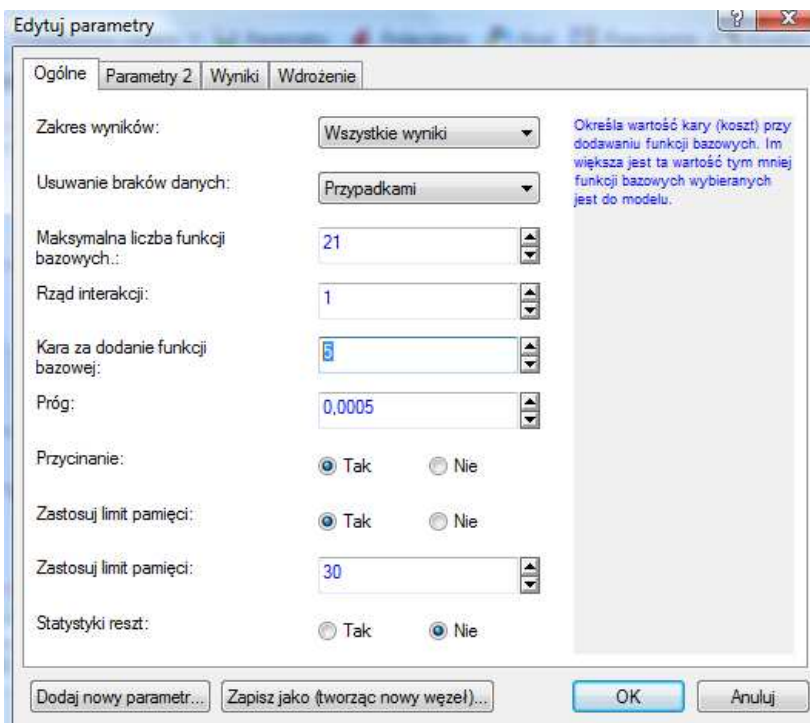
Rys. 5. Autokorelacje i autokorelacje cząstkowe dla zmiennej  $x(t)$ .

Brak autokorelacji nie oznacza, że nie można zbudować modelu, obserwowane związki mogą być nieliniowe.



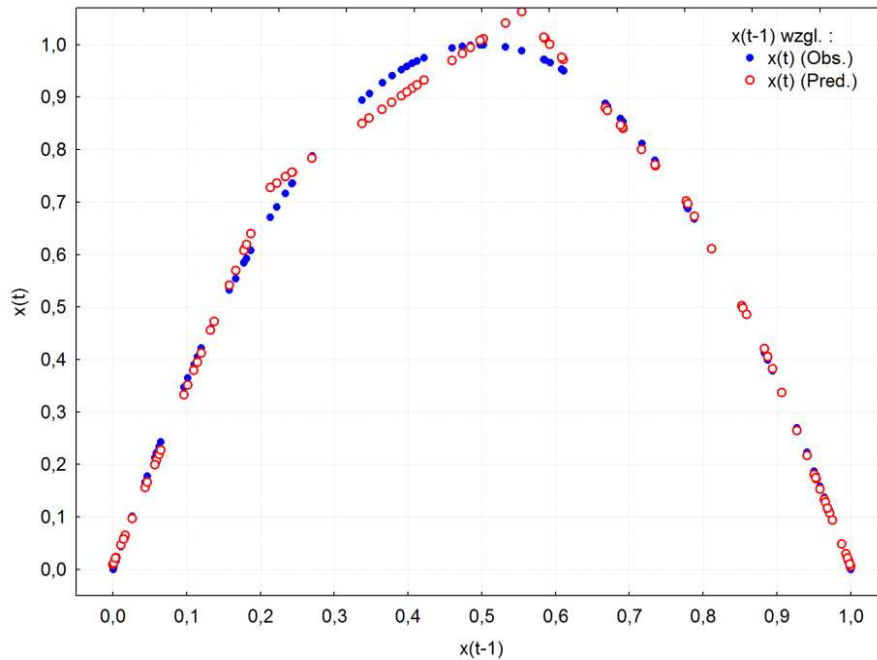
Rys. 6. Projekt analizy dla danych *Log.sta*.

Modelowanie badanej zależności przeprowadzimy za pomocą metody *MARSplines*. Można ją wykonać, wybierając z menu **Data mining** pozycję **MARSplines (Multivariate Adaptive Regression Splines)** lub, pracując w przestrzeni roboczej programu **STATISTICA Data Miner**, wybrać węzeł **MARSplines**. W oknie definiowania analizy wybieramy zmienne: jako **Ilościowe zależne** zmienną  $x(t)$ , natomiast jako **Ilościowe predktry** wybieramy  $x(t-1)$ . Na karcie **Ogólne** węzła **MARSplines** można zmienić wartość **Kara za dodanie funkcji bazowej** z 2 na 5.

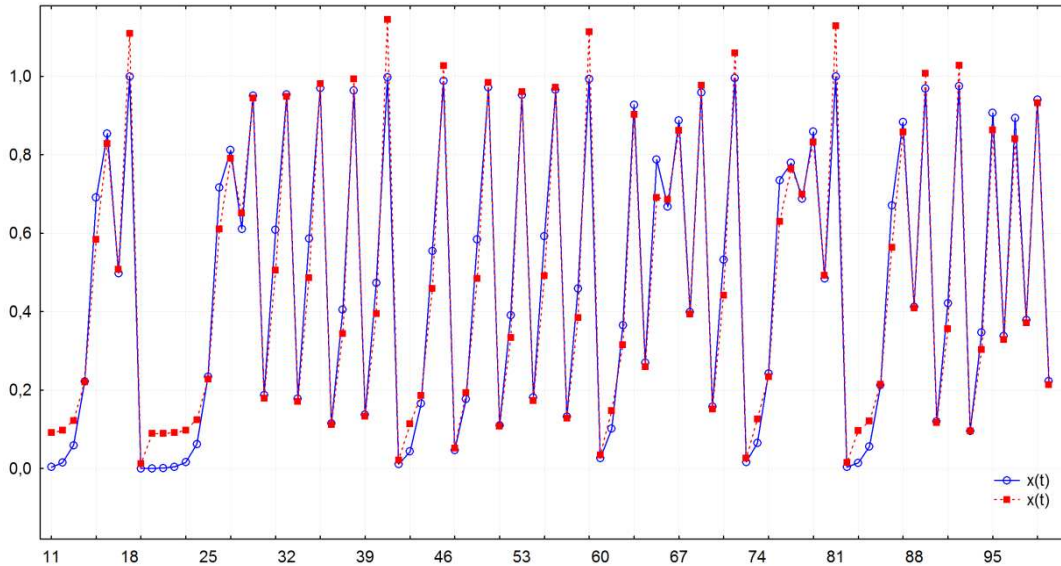


Rys. 7. Okno definiowania analizy *MARSplines*.

Zależność rzeczywista i wyznaczona przez model pomiędzy  $x(t)$  oraz  $x(t-1)$  została zaprezentowana na wykresie poniżej.



Rys. 8. Zależność pomiędzy wartością  $x(t-1)$  a  $x(t)$ .



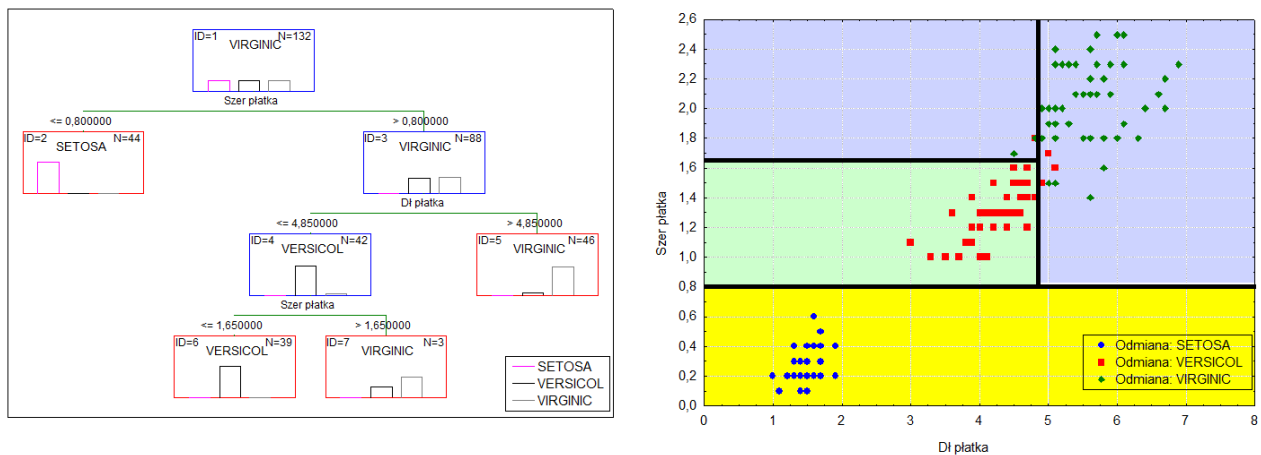
Rys. 9. Wartości obserwowane i prognoza metodą *MARSplines*.

Analizowane dane zostały wygenerowane za pomocą następującej formuły  $x_{t+1}=4x_t(1-x_t)$ . Szereg tego typu nie jest jednak wyłącznie tworem teoretycznym. Jest wiele procesów, które zachowują się według schematu tego typu. Jednym z nich jest, przedstawiony przez R. Maya [8], rozwój populacji. May badał zachowania populacji ryb: jeśli nie ma czynników hamujących (np. drapieżników), to populacja ryb rośnie, aż zapełni dostępne

środowisko. Jednak gdy liczba ryb jest tak duża, że zaczyna brakować pożywienia, rozpoczyna się zmniejszenie populacji. Związki nieliniowe trudniej się modeluje, bo trzeba więcej wiedzieć o samym procesie. Trudniej jest zgadnąć postać równania, dlatego warto korzystać z metod nieparametrycznych, w których nie trzeba znać modelu już na początku procesu modelowania. W tym przypadku oczywiście postać badanej zależności jest wręcz trywialna, jednak w rzeczywistych zastosowaniach, gdzie uwzględniamy więcej wielkości wejściowych, a nie tylko poprzedni stan populacji, pojawiają się znacznie bardziej skomplikowane zależności.

## Przykład – klasyfikacja ( $p \gg N$ )

Kolejnym typowym zadaniem analizy danych jest przewidywanie wartości zmiennej jakościowej, czyli problem klasyfikacyjny. Na podstawie wybranych cech mamy za zadanie przyporządkować badany obiekt do jednej z kilku klas (np. na podstawie wyników badań zdiagnozować pacjenta: chory - zdrowy). W tym wypadku również nierzadko nie wiemy wcześniej, na podstawie których cech i jakich poziomów ich wartości to przyporządkowanie będzie dokonywane. Naszym celem jest zbudowanie modelu klasyfikacyjnego. W tym wypadku również związek pomiędzy zmiennymi wejściowymi (cechami badanego obiektu) a zmienną wyjściową (klasą, do której faktycznie należał obiekt) może być nieliniowy i skomplikowany. Zatem również nie zawsze można podać jego postać. Jest jednak wiele metod, które same dobierają tak nieliniowe podziały przestrzeni cech, aby otrzymać jak najlepszą jakość modelu klasyfikacyjnego.



Rys. 10. Drzewo decyzyjne C&RT oraz wyznaczony przez nie podział przestrzeni cech.

Bardzo popularną metodą są drzewa klasyfikacyjne. Wynikiem takiej analizy jest drzewo decyzyjne, na podstawie którego możemy przyporządkować obiekt do odpowiedniej klasy. Rozwinięciem tej metody są na przykład metody losowy las (ang. *random forest*) czy też wzmacniane drzewa (ang. *stochastic gradient boosting trees*). Są to zespoły prostych drzew, realizujące ideę głosowania modeli. Okazuje się, że zespół prostych drzew zazwyczaj daje zdecydowanie trafniejsze przewidywania niż pojedyncze, nawet bardzo złożone drzewo.



W przypadku losowego lasu dodatkową zaletą jest również korzystanie z losowego podzbioru predyktorów przy tworzeniu nowego podziału w drzewie. Podzbiory predyktorów są niezależne i wybierane spośród wszystkich dostępnych zmiennych, a sam ich wybór dla poszczególnych drzew odbywa się ze zwracaniem. Podejście to bardzo dobrze się sprawdza w problemach, gdy liczba analizowanych obiektów (przypadków) jest mniejsza od liczby badanych cech (zmiennych). Zagadnienia takie pojawiają się coraz częściej i dotyczą na przykład badań genetycznych. Powszechnym jest tam zgromadzenie danych o ekspresji tysięcy genów dla zbadanych kilkudziesięciu lub kilkuset pacjentów. Przykład takiej właśnie analizy opisuje Breiman w rozdziale 11.3, gdzie to zbadano ekspresję 4682 genów dla 81 osób. Do analizy z powodzeniem wykorzystano losowy las.

Tu warto pamiętać o tym, że nie oceniamy modelu poprzez dopasowanie do danych (szczególnie w takim przypadku byłoby to szkodliwe), ale na podstawie działania na nowych danych, w tym wypadku dokonujemy podziału danych na dwie próby: uczącą i testową. Model budujemy na podstawie próby uczącej. Oceniamy model na podstawie działania na próbie testowej. Zatem ocena modelu jest jak najbardziej praktyczna, symuluje dobroć modelu podczas stosowania dla nowych danych.

## Podsumowanie

Zaprezentowane dwa przykłady pokazują możliwości technik data mining w analizie danych. Wykorzystywanie niektórych z nich może dać lepsze rezultaty, a w pewnych sytuacjach te nowe techniki są po prostu niezbędne. Najślabszym punktem podejścia opartego na mechanizmie badanego zjawiska wydaje się być właśnie założenie, że dane generowane są zgodnie ze wskazanym modelem. Często problematyczne lub wręcz niemożliwe jest wybranie odpowiedniego modelu. Dodatkowo warto zwrócić uwagę na to, że dzięki wynikom uzyskanym dzięki drugiemu z opisywanych podejść realizujemy zarówno cel związany z trafnością przewidywania przebiegu badanego zjawiska, ale również czynimy pierwsze kroki w poznaniu mechanizmu nim sterującego.

## Literatura

1. Berk R., Brown L., Zhao L., *Statistical Inference After Model Selection*, J Quant Criminol, 2010, 26, 217-36.
2. Bishop C. M., *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
3. Breiman L., *Statistical Modeling - The Two Cultures*, Statistical Science, 2001, 16 (3), 199-231.
4. Hand D., Mannila H., Smyth P., *Eksploracja danych*, WNT, Warszawa, 2005.
5. Hand D., *Modern statistics: the myth and the magic*, J. R. Statist. Soc. A, 2009, 172, 287-306.
6. Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, Springer, New York, 2001.



7. Koronacki J., Ćwik J., *Statystyczne systemy uczące się*, WNT, Warszawa, 2005.
8. May R. M., *Simple mathematical models with very complicated dynamics*, Nature, 1976, 261 (5560), 459-67.
9. Sullivan R., Timmermann A., White H., *Data-snooping, technical trading rule performance, and the bootstrap*, The Journal of Finance, 1999, 5, 1647-91.
10. Vapnik V., *The Nature of Statistical Learning Theory*, Springer, New York, 1996.