www.biswbis.pb.edu.pl

CIVIL AND ENVIRONMENTAL ENGINEERING 2(2011) ISSN: 2081-3279 BUDOWNICTWO I INŻYNIERIA ŚRODOWISKA

ZAGADNIENIE WYBORU LOKALIZACJI Z WYKORZYSTANIEM METODYK DATA MINING

Marcin GAJZLER*

Instytut Konstrukcji Budowlanych, Politechnika Poznańska, ul. Piotrowo 5, 61-138 Poznań

Streszczenie: Zaprezentowano problem wyboru miejsca lokalizacji inwestycji – budynku mieszkalnego. Potencjalne miejsca lokalizacji zostały zawężone do miasta Poznania i najbliżej leżących miejscowości. Z analizą przypadku związane jest przedsiębiorstwo, które dotychczas działając w branży budowlanej – budownictwo przemysłowe, w sytuacji braku zleceń rozważa różne strategie funkcjonowania i rozwoju. Jedną z takich strategii jest rozpoczęcie działalności w sektorze przedsiębiorstw deweloperskich. Na podstawie różnych danych związanych z rynkiem nieruchomości przedsiębiorstwo rozważa między innymi różne lokalizacje chcąc uzyskać możliwe najlepszy wynik w zakresie ceny i okresu sprzedaży planowanej inwestycji. W takiej sytuacji decyzyjnej i w celu opracowania prognozy zaproponowano wykorzystanie metod data mining. W rezultacie sformułowano wnioski co do przydatności tych metod w analizowanym problemie, a także możliwości innych zastosowań..

Słowa kluczowe: zagadnienie lokalizacyjne, data mining, sztuczna inteligencja.

1. Wstęp

Zagadnienie lokalizacyjne w budownictwie jest znanym problemem i na przestrzeni wielu lat w zagadnieniu tym wykorzystywano szeroki wachlarz narzędzi (Warszawski, 1973). W przypadku prostych problemów o ograniczonej liczbie czynników wpływających na wybór lokalizacji można posłużyć się choćby metodami programowania liniowego. Są to najczęściej problemy jednokryterialne, które możemy zaliczyć do grupy problemów optymalizacyjnych. Przykładem takiego zagadnienia może być rozwiązywany z zastosowaniem algorytmu transportowego problem lokalizacji wytwórni poligonowej mieszanki betonowej. W tym przypadku załóżmy, że jedynym kryterium jest znany koszt transportu mieszanki betonowej jako pochodna pokonywanych odległości pomiędzy dostawcą i odbiorcą. Oprócz tego znane są zapotrzebowania określonych odbiorców, jak również możliwości produkcyjne samych wytwórni. Wymaga się, aby lokalizacja wytwórni uwzględniała minimalizację kosztów transportu. Tak zdefiniowany problem lokalizacyjny stanowi „podręcznikowe" zadanie, które rozwiązywane jest przy wykorzystaniu wspomnianego algorytmu transportowego (Kukuła, 1993).

Bardziej złożone przypadki związane z zagadnieniem lokalizacyjnym były i są rozwiązywane przy zastosowaniu metod analizy wielokryterialnej (Szwabowski i Deszcz, 2001). Przykładem tego są obecnie bardzo często opracowywane analizy związane z lokalizacją inwestycji o strategicznym znaczeniu lub w znacznym stopniu oddziaływujące na otoczenie i środowisko (Małopolskie Biuro Konsultingowo-Marketingowe, 2009). Często wykorzystywaną w takich analizach jest metoda AHP, metody Electre czy Promethee (Dziadosz, 2008; Dziadosz i in., 2010; Skorupka i Duchaczek, 2010).

W zakresie praktycznego zastosowania w budownictwie wymienionych wcześniej metod, zagadnienie lokalizacyjne sprowadzało się najczęściej do dwóch grup problemów. Pierwsza z grup związana była z etapem planowania inwestycji i dotyczyła lokalizacji planowanego obiektu, np. zakłady produkcyjnego, oczyszczalni ścieków czy spalarni śmieci. Obiekty te często charakteryzowała pewna uciążliwość dla otoczenia i środowiska, a ich lokalizacja wynikała z istniejącej, bądź również planowanej infrastruktury. Innym wariantem obiektu często występującym w takiej analizie był budynek mieszkalny wielolokalowy lub obiekt handlowo-usługowy jako obiekty, które w zadanym czasie miały generować przychód właścicielowi. Ich lokalizacja musiała reprezentować pewną szeroko rozumianą „atrakcyjność" tak, aby pokładane szanse w lokalizacji tego typu obiektów pozwoliły uzyskać wymierny rezultat dla inwestora czy właściciela.

Druga grupa problemów, w której wykorzystywano zagadnienie lokalizacyjne dotyczyła etapu realizacji.

* Autor odpowiedzialny za korespondencję. E-mail: marcin.gajzler@put.poznan.pl

Grupa ta związana była z organizacją przedsięwzięcia najczęściej w zakresie placu budowy. Problemy te analizowano w aspekcie lokalizacji elementów zaplecza budowy czy magazynów centralnych lub wytwórni mieszanki betonowej, z których zaopatrywane były poszczególne budowy (Koźniewski i Orłowski, 2001).

Obecnie zagadnienia lokalizacji w budownictwie dotyczą również problemów szczegółowych, np. dotyczących umiejscowienia elementu w konstrukcji, czy nawet określenia położenia poruszającego się obiektu w czasie (Oxley i Poskitt, 1996; Torrent i Caldas, 2009). Przy czym zagadnienia te nie sprowadzają się do wyboru lokalizacji, gdyż ta jest już przewidziana, lecz dążą do zdefiniowania miejsca lokalizacji i określonego wykorzystania tej informacji. W przypadku pierwszej przykładowej sytuacji zagadnienie tak rozumianej lokalizacji rozwijane jest w ramach systemów BIM, natomiast w przypadku drugiej zagadnienie to występuje w systemach zdalnej kontroli i monitoringu (przykładowo w oparciu o technologie GPS). Nie są to jednak, jak wspomniano, typowe i pierwotne elementy związane z zagadnieniem lokalizacyjnym.

2. Analiza strategii rozwoju przedsiębiorstwa

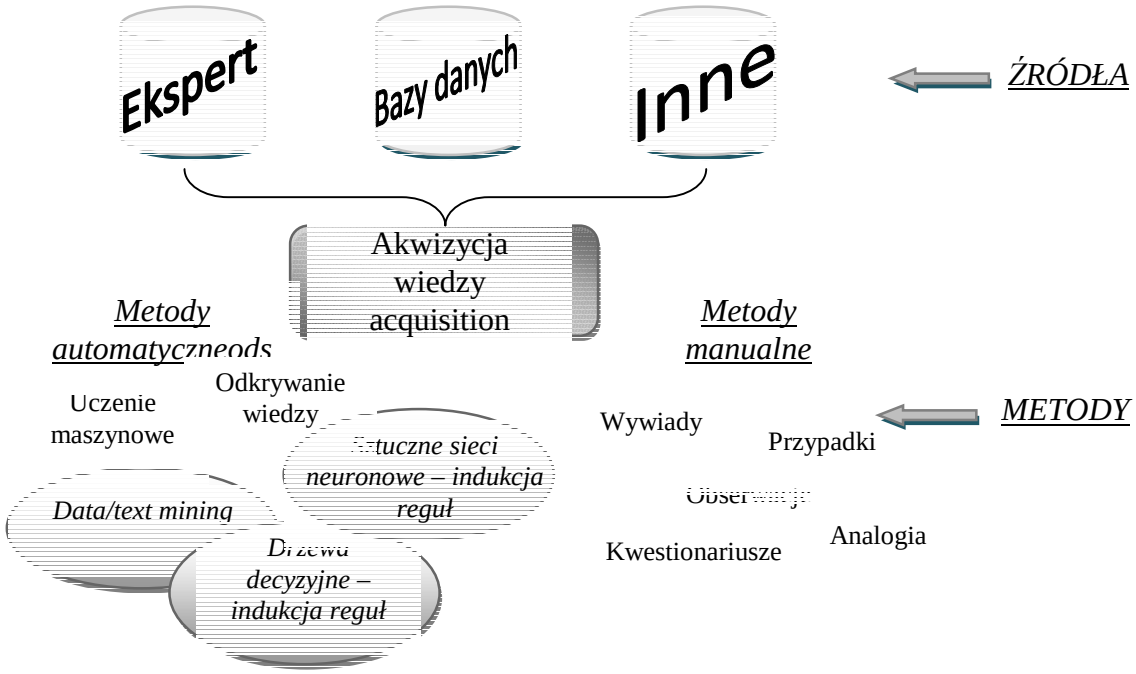
Analizowany przypadek zagadnienia lokalizacyjnego związany jest z działalnością przedsiębiorstwa budowlanego prowadzącego swoją zasadniczą działalność w zakresie wykonawstwa konstrukcji stalowych oraz montażu lekkiej obudowy w sposób nieprzerwany od 1992 roku. Przedsiębiorstwo to można zakwalifikować do grupy małych przedsiębiorstw. W toku swojej zasadniczej działalności począwszy od 2003 roku, przedsiębiorstwo sukcesywnie rozbudowywało swój park maszynowy i wzbogacało ofertę w zakresie budownictwa. Do ważniejszych inwestycji realizowanych przez przedsiębiorstwo w ostatnich latach można zaliczyć: wytwórnie tektur falistych, zakład produkcji papieru, przetwórnie pasz i wylęgarnia drobiu. W ramach tych inwestycji wykonywane były przeważnie obiekty halowe w technologii stalowo-żelbetowej.

Rok 2008 przyniósł obniżenie całej produkcji przemysłowej, co pociągnęło za sobą obniżenie ilości powstających obiektów budowlanych. Taka sytuacja wymusiła podjęcie pewnych działań mających na celu niwelację pustej przestrzeni w działalności zasadniczej przedsiębiorstwa, a powstałej w wyniku załamania rynku. Ze strony opisywanego przedsiębiorstwa podjęto działania mające na celu opracowanie nowych strategii działalności. Jedną z nich było skierowanie się ku działalności w sektorze budownictwa mieszkaniowego w roli przedsiębiorstwa deweloperskiego, z nastawieniem na wykorzystanie własnych mocy produkcyjnych w stopniu zależnym od sytuacji na rynku. Realizacja takiej strategii wymagała od przedsiębiorstwa przeprowadzenia wielu analiz. Dotyczyły one przede wszystkim sposobu finansowania inwestycji. Rozważano różny udział środków obcych w stosunku do własnych możliwości. Środki obce stanowił kredyt bankowy, a także ewentualne formy przedsprzedaży lokali mieszkalnych z tym, że pozyskanie tych środków uwarunkowane było szeregiem czynników. Jednym z nich była atrakcyjność nieruchomości w zakresie lokalizacji, standardu budynku, a także ceny sprzedaży. W związku z tym przedsiębiorstwo podjęło się przeprowadzenia analiz w zakresie lokalizacji obiektu, tak aby jego atrakcyjność znalazła odbicie w sprzedaży. Typowe podejście, które w rzeczywistości zastosowano, sprowadzało się do rozważenia możliwości kilku wybranych lokalizacji, dla których znana jest cena zakupu działki budowlanej oraz szereg czynników takich jak: infrastruktura komunikacyjna, handlowa, społeczna, uciążliwości, a których wpływ ma znaczenie na ogólną atrakcyjność sprzedaży. Alternatywne podejście, dzięki któremu możliwym jest poznanie różnego rodzaju powiązań i zależności pomiędzy analizowanymi czynnikami jest reprezentowane przez metodyki data mining.

3. Data mining

Metody data mining są określane w języku polskim jako eksploracja, drążenie danych. Są to stosunkowo młode metody, gdyż pierwsze zastosowania miały miejsce w latach dziewięćdziesiątych ubiegłego stulecia. Odmianą data mining jest metoda text mining, a więc analogiczne podejście zmierzające do eksploracji dokumentów tekstowych (Gajzler 2010). W przypuszczeniach można spodziewać się kolejnych wariantów analizy eksploracyjnej rodzaju foto/picture mining, co by było kolejnym elementem ewolucji data mining. Wszystkie wspomniane metody, w tym również te, o których rozwoju można jedynie przypuszczać, związane i rozpatrywane są w aspekcie pozyskiwania wiedzy. Można je zaliczyć do nowoczesnych i automatycznych metod akwizycji wiedzy (rys. 1). Dysponując wiedzą można wyciągać określone wnioski, natomiast w dalszej perspektywie rozwijać szereg interesujących narzędzi wspomagających działania człowieka. Niewątpliwym przykładem są wszelkiego rodzaju systemy wspomagające podejmowanie decyzji, w których wiedza oraz dane stanowią podstawy zasobów takiego systemu, natomiast proces akwizycji wiedzy stanowi często wąskie gardło budowy wspomnianych systemów.

Opisywana metodyka data mining zaliczana jest do grupy metod eksploracyjnej analizy danych, przy czym celem poszukiwań w tej metodzie nie są dane lecz wiedza. Dane zwykło się interpretować jako „surowiec" informacyjny, natomiast wiedzę jako zbiór posiadanych i powiązanych ze sobą informacji. Wynika z tego proste rozumienie idei data mining – od danych poprzez ich eksplorację do wiedzy. Definicji data mining jest wiele. Większość z nich oscyluje wokół definicji brzmiącej następująco – „data mining to proces badania i analizy danych metodami automatycznymi lub półautomatycznymi w celu odkrycia znaczących reguł i wzorców" (Berry i Linoff, 1997). W nurcie data mining można zauważyć występowanie i przenikanie się trzech dziedzin: baz danych, metod statystycznych i uczenia



Rys. 1. Metody akwizycji wiedzy (opr. własne)

maszynowego. Idea data mining polega więc na stosowaniu w dużym stopniu zautomatyzowanych i inteligentnych metod, które pozwalają na relatywnie szybkie odkrycia wiedzy. Literatura nie podaje dokładnej i sprecyzowanej klasyfikacji metod data mining. Wśród tych metod znajdują się jednak metody dobrze znane, które stosowane są niezależnie i jako takie nie są typowo utożsamiane z data mining. Spośród nich można wymienić: sztuczne sieci neuronowe, drzewa decyzyjne, algorytmy genetyczne, systemy regułowe, zbiory przybliżone, funkcje dyskryminacyjne czy metody klasteryzacyjne. Metody te stosowane są w różnych problemach. Zestawienie przykładowych metod w powiązaniu z klasami problemów występujących i analizowanych w data mining przedstawiono w tablicy 1 (Tadeusiewicz, 2006).

Na czym więc polega odmienność i sedno techniki data mining? Jest to przede wszystkim opisywane podejście, w którym nawet z przypadkowych czy „wyeksploatowanych" danych udaje się pozyskać „nową" wiedzę np. w postaci wykrycia zależności czy pewnych anomalii. Podstawową różnicą pomiędzy podejściem klasycznym (statystycznym), a metodyką data mining jest to, że w pierwszym przypadku dokonujemy najczęściej weryfikacji pewnych modeli opartych na wiedzy teoretycznej, natomiast w data mining często pozbawieni jesteśmy wiedzy teoretycznej o zjawisku i w związku z tym dokonujemy intuicyjnej eksploracji licząc na nowe odkrycie. Podejście data mining jest podejściem usystematyzowanym metodycznie i można je opisać następująco:

– utworzenie/pozyskanie zbioru danych,
– wstępna analiza i przetworzenie danych,
– wykonanie właściwych obliczeń,
– weryfikacja poprawności uzyskanych wyników,
– interpretacja wyników i wykorzystanie ich w procesie decyzyjnym.

Tab. 1. Przegląd podstawowych problemów i odpowiadających metod w data mining (Tadeusiewicz 2006)

Rodzaj problemu	Właściwe metody
Klasyfikacja wzorcowa	– perceptronowe sztuczne sieci neuronowe – drzewa decyzyjne – systemy regułowe – zbiory przybliżone – funkcje dyskryminacyjne
Klasyfikacja bezwzorcowa	– samouczące się sztuczne sieci neuronowe – algorytmy genetyczne – metody taksonomiczne – metody graficzne
Szeregi czasowe	– perceptronowe oraz radialne sztuczne sieci neuronowe – metody analizy sygnałów – metody badania sekwencji
Opis zależności w zjawisku	– perceptronowe oraz radialne sztuczne sieci neuronowe – statystyczne metody pomiaru zależności – metody analizy współwystępowania – zbiory przybliżone
Problemy wyboru	– algorytmy genetyczne – zbiory przybliżone – rekurencyjne sztuczne sieci neuronowe

Na każdym z tych etapów dysponujemy odpowiednimi metodami i narzędziami, które wspomagają naszą analizę i pomagają wyciągnąć wnioski. Sam proces pod względem technicznym realizowany jest przez wyspecjalizowane środowiska komputerowe. Te najbardziej znane to Statistica Data Miner oraz Oracle Data Miner.

4. Analiza przypadku

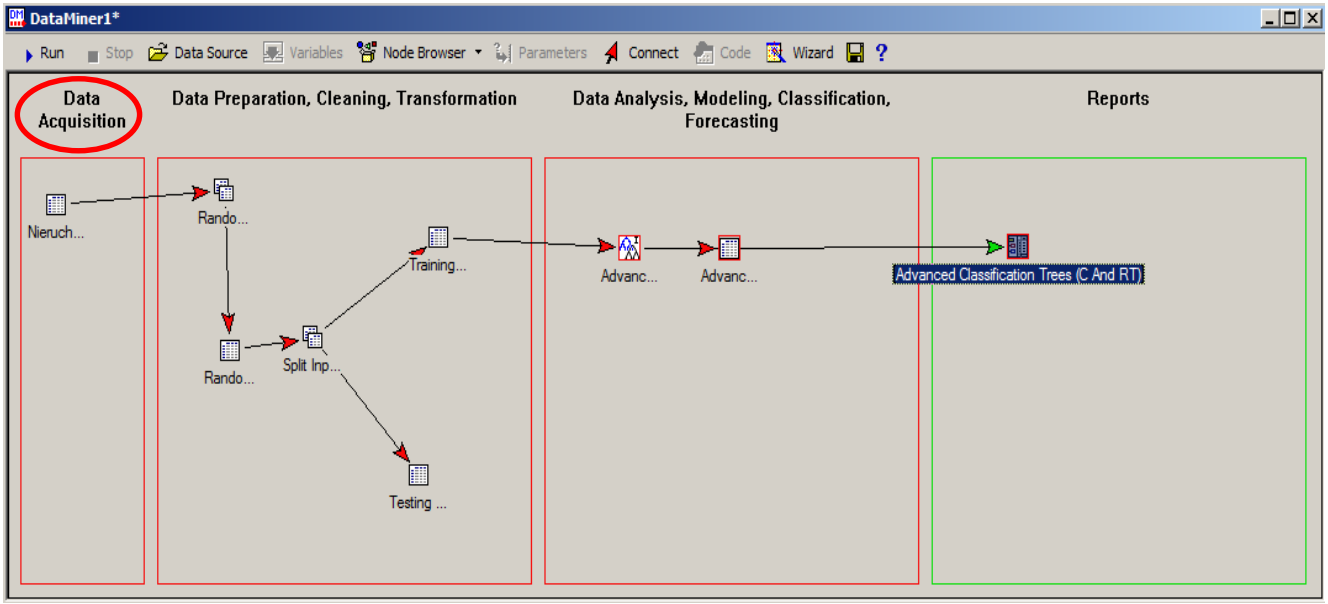
Analizowany tutaj problem lokalizacji inwestycji został rozwiązany w rzeczywistości w oparciu o analizę wielokryterialną. Przyjęto podejście polegające na rozważaniu kilku dostępnych lokalizacji (z punktu widzenia samej możliwości pozyskania działki budowlanej) z uwzględnieniem ceny działki, warunków prowadzenia inwestycji, a także samej technologii budowy i standardu wykończenia budynku. W rezultacie porównywano ze sobą 12 wariantów, z których wyłoniono jeden jako propozycję lokalizacji inwestycji, a także technologii wykonania i standardu wykończenia budynku. Należy przyznać, że metoda ta okazała się skuteczna i z punktu widzenia przedsiębiorstwa przyniosła określone rozwiązanie.

Z drugiej strony warto spojrzeć na inne możliwości prowadzenia takich analiz. Proponowana metodyka data mining (rys. 2) pozwala przecież z często przypadkowych danych uzyskać wiedzę, która może służyć w rozwiązywaniu problemów decyzyjnych. Dysponując podstawowymi informacjami o rynku nieruchomości w Poznaniu dokonano próby pozyskania wiedzy na ten temat i wykorzystania jej w problemie lokalizacji inwestycji budowlanej. W analizie zagadnienia wykorzystano pakiet oprogramowania Statistica Data Miner firmy Statsoft. Pakiet ten zawiera szereg narzędzi skupiających się wokół metodyk data mining, a także bardzo szeroki wachlarz narzędzi obróbki statystycznej.

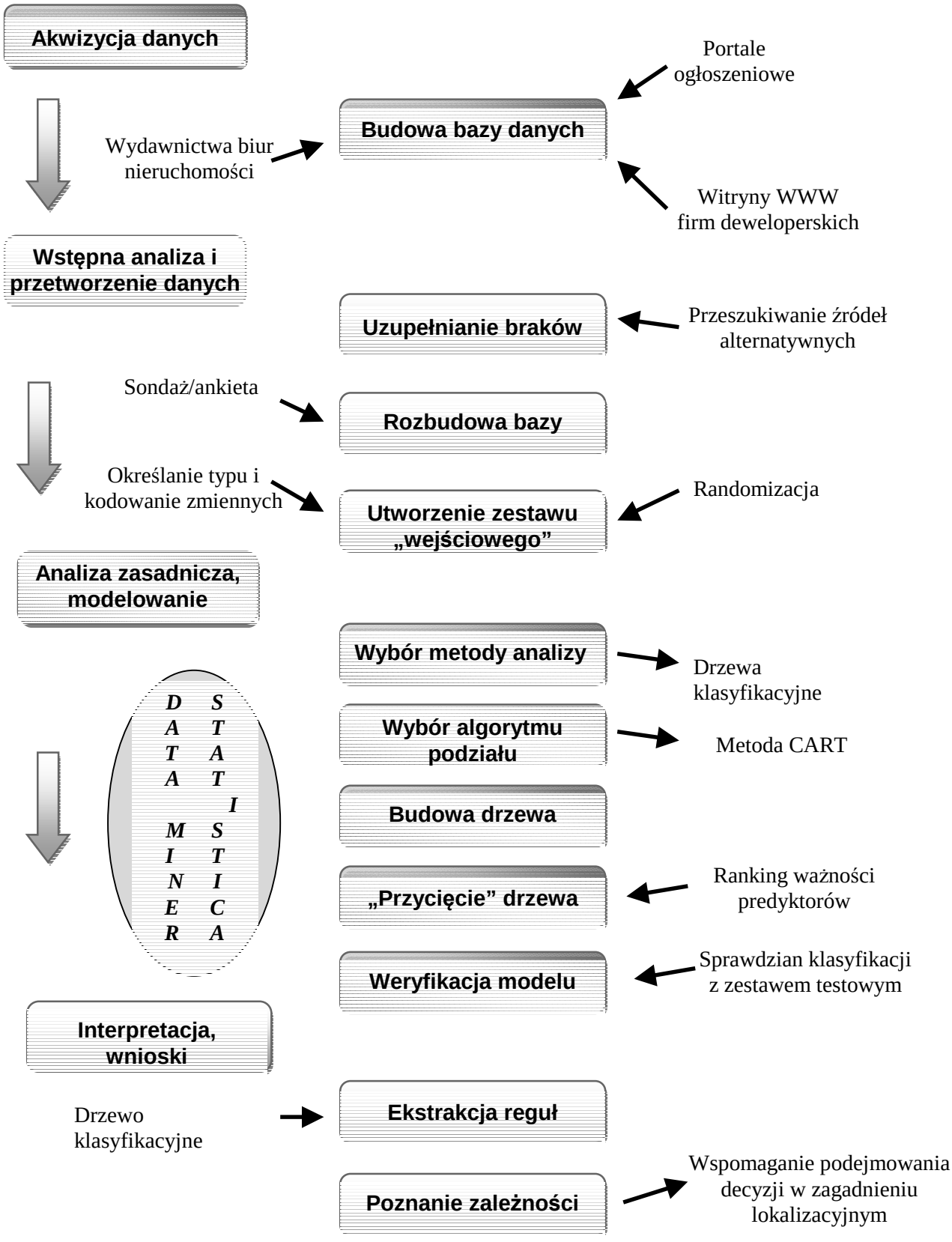
Jak wskazano w punkcie 3 metoda data mining polega na usystematyzowanym podejściu. Pierwszy z tych etapów dotyczy pozyskania danych. Pozyskiwanie danych rozpoczęto od przeszukiwania dostępnych baz danych ofert nieruchomości lokalowych na rynku pierwotnym. W tym celu wykorzystano zasoby internetowych portali ogłoszeniowych, witryny internetowe firm deweloperskich, a także wydawnictwa papierowe biur obrotu nieruchomościami. Wskazane źródła zawierały oferty sprzedaży nieruchomości lokalowych na terenie miasta Poznania oraz gmin ościennych. Przy tym ilość parametrów opisująca daną ofertę była różna. Na potrzeby analizy przyjęto zestaw parametrów opisujący daną nieruchomość w postaci najczęściej występujących. Były to:

- *położenie inwestycji* (dzielnica), gdzie przyjęto następujące kody: G – Grunwald, J – Jeżyce, N – Nowe Miasto, S – Stare Miasto, W – Wilda, P – tereny podmiejskie,
- *rodzaj zabudowy* zakodowany: W – wielorodzinna, S – szeregowa, J – jednorodzinna,
- *technologia budowy*: T – tradycyjna, M – monolityczna, S – szkieletowa,
- *odległość od linii komunikacji miejskiej* (tramwajowej lub autobusowej) wyrażona w metrach,
- *odległość od elementów infrastruktury społecznej* (zdrowie, oświata, rekreacja) wyrażona w metrach,
- *odległość od średnich lub dużych centrów handlowych* wyrażona w metrach,
- *wielkość lokalu* wyrażona w metrach kwadratowych powierzchni użytkowej,
- *liczba pomieszczeń*,
- *cena ofertowa* brutto za metr kwadratowy powierzchni użytkowej.

Zebranie danych w formie bazy jest realizacją pierwszego etapu metodyki data mining odpowiadającemu akwizycji danych (rys. 3).



Rys. 3. Widok podstawowych etapów procesu data mining w Statistica Data Miner



Rys. 2. Algorytm procesu Data Mining dla analizowanego problemu

Drugim etapem jest wstępna analiza i przetworzenie danych. W tym zakresie działania ukierunkowane są na sprawdzenie kompletności zbioru danych, a więc w miarę możliwości należy uzupełnić braki w zestawie danych. W tym celu możliwe są następujące podejścia: przeszukanie źródeł alternatywnych, szacowanie brakujących informacji, usunięcie przypadku zawierającego braki (niestety powoduje to zubożenie próbki, którą dysponujemy w analizie), usunięcie zmiennej odpowiadającej brakującym wartościom (podobnie jak wcześniej automatycznie zubażamy analizę oraz przyszły model). W analizowanym przypadku wszelkie braki występujące w bazie danych zostały uzupełnione przez indywidualne oszacowanie brakujących danych. Braki najczęściej związane były z umiejscowieniem nieruchomości względem komunikacji, infrastruktury społecznej oraz centrów handlowych. W oparciu o znajomość lokalizacji nieruchomości szacowano te odległości wykorzystując ogólnodostępne plany miejskie. Proces ten został zrealizowany jednocześnie z niewielką rozbudową bazy danych. Do stworzonej na podstawie opisanej wyżej akwizycji danych bazy wprowadzono zależną „atrakcyjność". Jest ona wynikiem określenia przez zainteresowane gremium ogólno pojętej atrakcyjności i potencjalnego zainteresowania ofertą. Zależna została skwantyfikowana na trzy poziomy: *wysoka* (H), *średnia* (M), *niska* (L). W celu przypisania zmiennej do konkretnego przypadku posłużono się badaniem ankietowym, gdzie zainteresowani przypisywali wybrany poziom atrakcyjności. Wyniki badania uśredniano przypisując wcześniej określonemu poziomowi atrakcyjności współczynniki liczbowe. Łatwo się domyśleć, że zależna atrakcyjność jest pochodną wszystkich czynników, a nie tylko tych uwzględniających czynniki bezpośrednio związane z lokalizacją. Pomimo takiej świadomości w dalszym ciągu będziemy zmierzali do wyciągnięcia wniosków dotyczących właśnie atrakcyjności samej lokalizacji.

Kolejny aspekt związany z etapem przygotowania i przetworzenia danych dotyczy wartości nietypowych. O ile jest nam znany przedział wartości oraz ogólny trend wskazujący na ich zachowanie się – problem nietypowych danych może być łatwo i szybko wychwycony. Jednak duża część przypadków, co do których nie posiadamy wiedzy w zakresie dopuszczalnych wartości, nie daje nam takiej możliwości. Inny aspekt to przyjmowanie przez zmienne nietypowych wartości, co związane jest z rzeczywistą obserwacją zjawiska i zachodzącymi tam zdarzeniami, często o charakterze anomalii. W tym przypadku pewnego rodzaju „standaryzacja" byłaby zjawiskiem niekorzystnym, gdyż zachodząca anomalia nie zostałaby rozpoznana. W przypadku problemu lokalizacyjnego występowały przypadki „pseudo" anomalii, a związane były z dużą zmiennością cen nieruchomości i tym samym różną atrakcyjnością w ramach jednej i tej samej dzielnicy. Aby pozbyć się tego zjawiska nietypowości należałoby rozszerzyć liczbę wartości przyjmowanych przez zmienną „dzielnica". Z racji tego, że zmienna „dzielnica" była zmienną skategoryzowaną, a wartości opisujące ją miały postać opisową zaniechano dalszego rozszerzania przedziału wartości. Powstały problem rozwiązałby się w przypadku, gdyby zmienna „dzielnica" była zmienną ciągłą i opisywana była wartościami ilościowymi, np. przy wykorzystaniu koordynatów GPS, co by pozwoliło na precyzyjne wskazanie nieruchomości i dokładniej odzwierciedliłoby prawidłowości związane z ceną i atrakcyjnością obecnie rozmywające się w ramach jednej dzielnicy.

Rezultatem działań zawartych w etapie przygotowania i przetworzenia danych jest baza danych, która będzie następnie podstawą zasadniczej analizy data mining. Baza danych została zdefiniowana w środowisku Statistica Data Miner jako arkusz danych interpretowany dalej jako zbiór zmiennych wejściowych i odpowiadająca zmienna zależna (rys. 4). Część z danych ma charakter zmiennych ciągłych, natomiast pozostała część to zmienne

Data: Nieruchomości (10v by 100c)

	1 Dzielnica	2 Zabudowa	3 Technologia	4 Komunikacja	5 Infrastruktura społeczna	6 Handel/ usługi	7 Wielkość lokalu	8 Ilość pomieszczeń	9 Cena za metr	10 Atrakcyjność
1	J	W	T	600	1200	2500	28,6	1	6500	H
2	J	W	T	600	1200	2500	42,5	3	6390	H
3	G	W	T	350	800	1100	36,3	2	6711	M
4	G	W	T	1100	2000	1100	84,3	4	5798	M
5	G	W	T	1100	2000	1100	68,2	3	6100	M
6	S	W	Ż	450	1000	1300	95,1	4	5153	H
7	N	W	T	300	650	2600	56,4	3	5654	H
8	N	W	Ż	350	1400	3500	82,6	4	6740	M
9	N	W	Ż	350	1400	3500	119	5	6230	M
10	S	W	T	400	550	1200	39,6	3	6950	H
11	S	W	T	400	550	1200	47,1	3	6160	H
12	S	W	T	900	650	2000	37,8	2	6800	H
13	S	W	T	900	650	2000	67,4	4	5950	H
14	J	W	Ż	2300	1900	3500	38,5	2	6300	M
15	P	S	T	4100	6300	5600	112,8	5	4950	L
16	G	W	Ż	2600	2800	3200	64,5	4	5479	M
17	P	W	T	3100	4500	4000	56,4	3	4120	M
18	P	W	T	3100	4500	4000	71,8	5	4030	M
19	P	J	T	4900	5600	5000	128,6	5	3850	L

Rys. 4. Fragment bazy danych wraz z oceną atrakcyjności nieruchomości jako arkusz danych w Statistica Data Miner

skategoryzowane. Wspomniana już wprowadzona zależna atrakcyjność, ma również postać zmiennej skategoryzowanej o trzech klasach.

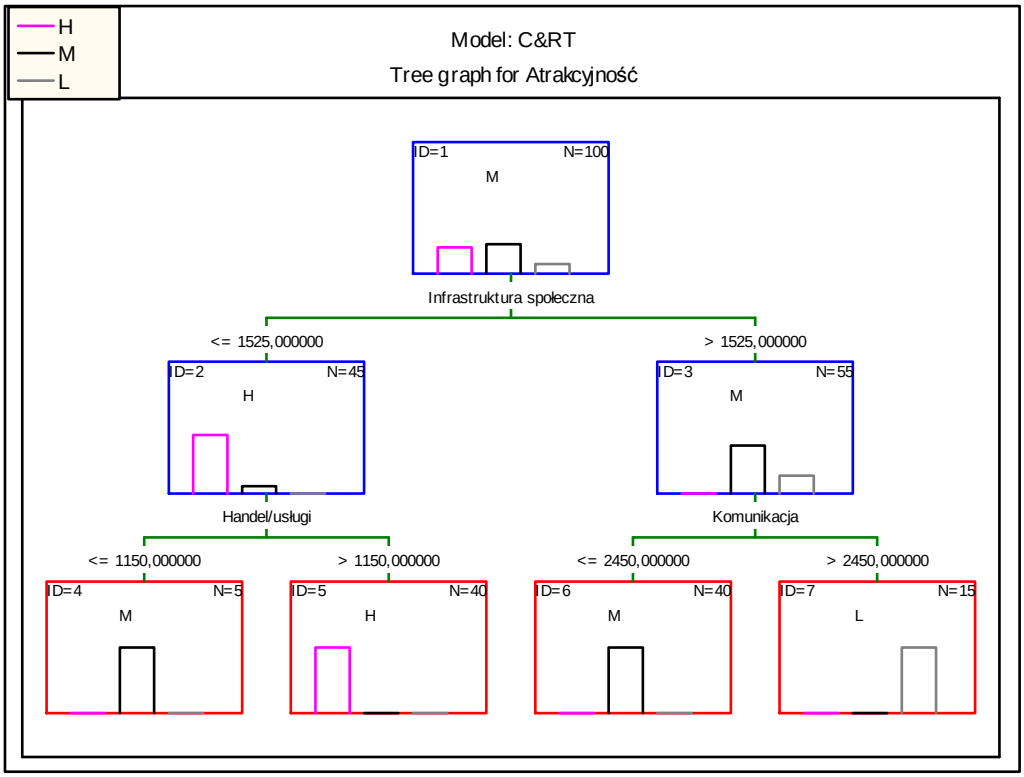
Zasadniczą analizę data mining przeprowadzono w środowisku Statistica Data Miner. Daje ono wiele możliwości zarówno w zakresie etapu przygotowania i wstępnej obróbki danych, a także wyboru samych metod analiz. W analizie lokalizacji inwestycji wykorzystano drzewa klasyfikacyjne. Są one stosunkowo dobrą i prostą wizualizacją zależności występujących pomiędzy zmienną zależną a predyktorami. Dodatkowo na podstawie struktury drzewa, korzenia, gałęzi i liści można w łatwy sposób generować reguły, które dalej można wykorzystać w budowie systemów regułowych na potrzeby wspomagania podejmowania decyzji. Ta możliwość wydaje się szczególnie cenna w związku z budową systemów ekspertowych, które podlegając różnym ewolucjom stanowić mogą wartościowe narzędzie w podejmowaniu decyzji.

Wykorzystując drzewa klasyfikacyjne środowisko Data Miner daje do dyspozycji kilka algorytmów podziału drzew. Dwa podstawowe to CHAID (Chi-squared Automatic Interaction Detector) oraz CART (Classification and Regression Trees). Właściwości i różnice każdej z tych metod opisane są w dostępnych podręcznikach statystyki (Statsoft, 2006). We właściwej analizie wykorzystano drzewa klasyfikacyjne typu CART wykorzystując przy tym moduł drzew interakcyjnych, w których użytkownik Statistica Data Miner może w sposób manualny sterować budową drzewa. W budowie tego drzewa nie musimy uwzględniać wszystkich predyktorów, gdyż część z nich nie ma związku z lokalizacją. Otrzymamy wówczas już „przycięte" drzewo, które nie będzie charakteryzowało się nadmiernym, i z naszego punktu widzenia, niepotrzebnym rozrostem. Środowisko Data Miner w module drzew interakcyjnych pozwala również zbudować drzewo w sposób zautomatyzowany, natomiast chcąc uwzględnić tylko wybrane zmienne można posłużyć się metodą manualną i po kolei definiować zmienne i nadzorować podziały. Chcąc możliwie uprościć drzewo i uwzględnić jedynie wybrane zmienne posłużono się statystyką predyktorów określającą ich istotność dla badanej zależnej (rys. 5). Statystyka ta jest tworzona w środowisku Data Miner w sposób automatyczny i udostępniana użytkownikowi w postaci listy rankingowej zmiennych wejściowych, gdzie każdemu predyktorowi przypisywany jest współczynnik ważności z przedziału <0, 1> .

W analizowanym przypadku, ze wspomnianej statystyki wynika, że dla analizowanej próby najsilniej ze względu na „atrakcyjność" różnicują właśnie zmienne związane z lokalizacją, czyli w kolejności: odległość od infrastruktury społecznej, dzielnica (przy czym należy mieć świadomość o niejednoznaczności tej zmiennej), odległość od linii komunikacyjnych oraz centrów handlowych. W praktyce przy budowie drzewa z dostępnych predykatorów uwzględniono właśnie te związane bezpośrednio z lokalizacją, a więc nie uwzględniono takich zmiennych jak rodzaj zabudowy i technologia, wielkość lokalu, ilość pomieszczeń oraz cena. Oprócz wymienionych zdecydowano się nie uwzględniać również zmiennej „dzielnica" ze względu na komentowane wcześniej niejednoznaczności. Powstające drzewo będzie na pewno przez to uboższe, ale jednocześnie pozbawione będzie zbędnych (nie dotyczących problemu lokalizacji) i zafałszowanych (niejednoznacznych – dzielnica) gałęzi. W rezultacie otrzymano drzewo, z analizy którego można wyciągnąć wnioski dotyczące lokalizacji inwestycji i jej przypuszczalnej atrakcyjności dla zainteresowanych (rys. 6).

	Predictor Information for Node 1 (Nieruchomości) The order of predictors according to Statistic/df Model: C&RT				
	Split type	Improvement Statistic			
Infrastruktura społeczna	Automatic	0,307929			
Dzielnica	Automatic	0,235833			
Komunikacja	Automatic	0,196818			
Handel/usługi	Automatic	0,196818			
Zabudowa	Automatic	0,191471			
Wielkość lokalu	Automatic	0,140000			
Cena za metr	Automatic	0,121667			
Ilość pomieszczeń	Automatic	0,121667			
Technologia	Automatic	0,061667			

Rys. 5. Ranking istotności predykatorów dla zależnej „Atrakcyjność"



Rys. 6. Widok „przyciętego" drzewa klasyfikacyjnego dla zależnej „atrakcyjność"

Tak powstałe drzewo może służyć ekstrakcji reguł w oparciu o które można wnioskować o atrakcyjności planowanej lokalizacji. Przykładem takich zależności mogą być reguły:

– jeżeli lokalizacja inwestycji znajduje się w odległości mniejszej równej 1525 m od elementów infrastruktury społecznej oraz w odległości większej niż 1150 m od centrum handlowego to atrakcyjność takiej lokalizacji jest *wysoka*;
– jeżeli lokalizacja inwestycji znajduje się w odległości większej niż 1525 m od elementów infrastruktury społecznej oraz w odległości mniejszej równej 2450 m od linii komunikacyjnej to atrakcyjność takiej lokalizacji jest *średnia*;
– jeżeli lokalizacja inwestycji znajduje się w odległości mniejszej równej 1525 m od elementów infrastruktury społecznej oraz w odległości mniejszej równej 1150 m od centrum handlowego to atrakcyjność takiej lokalizacji jest *średnia*.

Na podstawie takich reguł można uogólnić, że atrakcyjna lokalizacja powinna znajdować się w bliskiej okolicy elementów infrastruktury społecznej, niedalekiej odległości od linii komunikacyjnej oraz powinna być oddalona od średnich i dużych centrów handlowych.

Nie mniej ciekawe wnioski można by wyciągnąć w zakresie czynników wpływających na cenę metra kwadratowego powierzchni użytkowej lokalu mieszkalnego. W takiej sytuacji w roli zmiennej zależnej należy uwzględnić cenę metra kwadratowego, natomiast pozostałe zmienne traktować jako predyktory. Dzięki temu możemy zbadać w jaki sposób predyktory związane z lokalizacją (dzielnica, odległości od linii komunikacji, infrastruktury społecznej i handlowej) oddziałują na cenę. Również poza pozyskaniem takiej wiedzy istnieje i sam aspekt techniczny związany z wykorzystaniem innych narzędzi analizy, np. sieci neuronowych czy regresji wielorakiej. Interesujące mogą się okazać „predyspozycje" określonych narzędzi do wykorzystania w określonym problemie.

5. Wnioski

Podsumowując należy wskazać, że w zakresie rozwiązywania problemu lokalizacji inwestycji przedstawiona metodyka może stanowić pewną alternatywę dla klasycznych metod, np. analizy wielokryterialnej. W analizowanym przypadku uzyskano wiedzę przybliżającą decydenta do wypracowania decyzji związanej z lokalizacją, która zapewni „atrakcyjność" planowanej inwestycji. Wnioski wypływające z analizy zagadnienia warto skonfrontować z innymi opracowaniami uwzględniającymi aspekt lokalizacji inwestycji realizowanych przez przedsiębiorstwa deweloperskie (Zima, 2007). Pamiętać należy przy tym, że każdy regionalny rynek nieruchomości ma swoją własną specyfikę, która może się przyczynić do istotnych różnic w formułowaniu stwierdzeń dotyczących lokalizacji.

Jednocześnie łatwo zauważyć w tym konkretnym zagadnieniu pewną ułomność proponowanej metody i przewagę „klasycznego" podejścia. Znacznie wygodniej analizować konkretne i przede wszystkim dostępne dla inwestora lokalizacje, z pełną znajomością ich

uwarunkowań – posługując się np. metodami analizy wielokryterialnej, niż prowadzić analizy ogólne. Te są natomiast przydatne przede wszystkim na etapie określania kryteriów wyboru oraz ich ważności. Do czego wobec tego mogą być przydatne analizy data mining? Odpowiedź wydaje się prosta – służą przede wszystkim poznawaniu nowych zależności, odkrywaniu powiązań, o których być może nie mieliśmy świadomości istnienia i związku. Biorąc pod uwagę fakt, że w zagadnieniach budownictwa w etapach planowania przedsięwzięcia oraz monitoringu realizacji występuje duża liczba danych – metody data mining mogą stanowić użyteczne narzędzie w zdobywaniu wiedzy.

Literatura

Berry M., Linoff G. (1997). Data mining techniques for marketing, sales and customer support. *Wiley*, New York.

Dziadosz A. (2008). Ocena i selekcja inwestycji budowlanych z wykorzystaniem analitycznego procesu hierarchicznego. *Czasopismo Techniczne*, 1-B/2008, 41-52.

Dziadosz A., Gajzler M., Szymański P. (2010). Problemy wyboru metody wspomagającej podejmowanie decyzji w budownictwie. *Czasopismo Techniczne*, 1-B/2010, 71-84.

Gajzler M. (2010). Text and data mining techniques in aspect of knowledge acquisition for decision support system in construction industry. Technological and Economic Development of Economy. Vilnius: *Technika*, Vol. 16, No. 2, 219-232.

Koźniewski E., Orłowski Z. (2001). Czynniki wpływające na lokalizację wytwórni mieszanki betonowej. *Inżynieria i Budownictwo*, 8/2001, 462-464.

Kukuła K. (red.) (1993). Badania operacyjne w przykładach i zadaniach. *PWN*, Warszawa.

Małopolskie Biuro Konsultingowo-Projektowe – ochrona środowiska s.c. (2009). Analiza wyboru lokalizacji Instalacji Termicznego Przekształcania Odpadów Komunalnych dla miasta Poznania wraz ze wstępną analizą wielokryterialną – opracowanie. WEB: *http://www.poznan.pl/mim/public/wos/attachments.html?co=show&instance=1039&parent=33042&lang=pl&id=69291*.

Oxley R., Poskitt J. (1996). Management Techniques Applied to the Construction Industry. *Wiley*, Blackwell.

Skorupka D., Duchaczek A. (2010). Zastosowanie metody AHP w optymalizacji procesów decyzyjnych związanych z realizacją przedsięwzięć logistycznych. *Zeszyty Naukowe WSOWL*, Rocznik XLII, No. 3, 54-62, Wrocław.

StatSoft (2006). Elektroniczny Podręcznik Statystyki PL, Kraków, WEB: *http://www.statsoft.pl/textbook/stathome.html*.

Szwabowski J., Deszcz J. (2001). Metody wielokryterialnej analizy porównawczej – podstawy teoretyczne i przykłady zastosowań w budownictwie. *Wydawnictwo Politechniki Śląskiej*, Gliwice.

Tadeusiewicz R. (2006). Data mining jako szansa na relatywnie tanie dokonywanie odkryć naukowych przez przekopywanie pozornie całkowicie wyeksploatowanych danych empirycznych. W: Materiały *Seminarium „Zastosowania statystyki i data mining w badaniach naukowych”, Statsoft Polska*, Kraków.

Torrent D.G., Caldas C.H. (2009). Methodology for Automating the Identification and Localization of Construction Components on Industrial Projects. *Journal of Computing in Civil Engineering*, Vol. 23, 3-13.

Warszawski A. (1973). Multi-Dimesional Location Problems. *Operational Research Quarterly*, Vol. 24, No. 2, 165-179.

Zima K. (2007). Analiza deweloperskich przedsięwzięć budowlanych z zastosowaniem logiki rozmytej. Praca doktorska. *Politechnika Krakowska*, Kraków.

THE ISSUE OF CHOOSING LOCATION WITH THE USE OF DATA MINING TECHNIQUES

Abstract: The paper presents the issue of choosing the location of the investment – a residential building. The potential locations were reduced to the city of Poznan and towns situated nearby. The case study is connected with an enterprise which was operating in the field of building trade – industrial building, and as a result of the lack of works – is forced to consider different operation and development strategies. One of these strategies is to start the activity in the development sector. Based on various data related to real estate market, the enterprise considers different locations wishing to achieve the best result when it comes to price and sales period of the planned investment. In such a decision situation and in order to develop a forecast the use of data mining techniques was suggested. It led to conclusions concerning the usefulness of these techniques in the analysed problem and also possibilities of other applications.