STUDIA INFORMATICA 2010
Volume 31 Number 2B (90)

Monika CHUCHRO, Adam PIÓRKOWSKI
Akademia Górniczo-Hutnicza, Katedra Geoinformatyki i Informatyki Stosowanej

WYKORZYSTANIE METOD I NARZĘDZI EKSPLORACJI DANYCH DO ANALIZY ZMIENNOŚCI NATĘŻENIA DOPŁYWU DO KOMUNALNYCH OCZYSZCZALNI ŚCIEKÓW[1]

Streszczenie. Eksploracja danych to obecnie najdynamiczniej rozwijające się zagadnienie związane z przetwarzaniem danych. Celem eksploracji (data mining, drążenie danych) jest uzyskiwanie nowej użytecznej wiedzy z dużych kolekcji danych przy użyciu specjalistycznych narzędzi statystycznych. W artykule dokonano porównania metod eksploracji w środowiskach Weka, R, Statistica i Microsoft SQL Server dla istniejącej bazy danych ilości dopływu ścieków do oczyszczalni. W tym przypadku dane wejściowe to szeregi czasowe o rozdzielczości dobowej, obejmujące długie okresy obserwacji (nawet kilkanaście lat).

Słowa kluczowe: eksploracja danych, drążenie danych, analiza szeregów czasowych, oczyszczalnie ścieków

METHODS AND TOOLS FOR DATA MINING OF INTENSITY VARIABILITY INLET TO MUNICIPAL WASTEWATER TREATMENT PLANT

Summary. Data mining is currently the fastest growing problem of processing data. The purpose of exploration (data mining, drilling of the data) is useful to obtain new knowledge from large data packets using specialized statistical tools. This article makes a comparison of methods for exploration in environments Weka, R, Statistica and Microsoft SQL Server database to an existing quantity of water inflow to the treatment plant. Inputs such application is a time series with daily resolution of the long periods of observation (even several years).

Keywords: data mining, time series analysis, sewage treatment, wastewater treatment

[1] Praca finansowana w ramach badań statutowych KGIS nr 11.11.140.561.

1. Wstęp

Dynamiczny rozwój technologii informatycznych pozwala na coraz większy obszar zastosowań rozwiązań cyfrowych. Powszechne stają się metody opisu i rejestracji procesów przemysłowych i handlowych, co nieraz generuje duże ilości danych. Zbiory takie pozwalają na dotychczas trudne w realizacji analizy. Wynikiem tych analiz mogą być nowe wnioski i spostrzeżenia, które pozwalają ulepszyć rozwiązania lub dostrzec niebezpieczeństwo czy też negatywne tendencje dotyczące środowiska. Przetwarzanie dużej ilości danych właśnie pod kątem wychwytywania nowych zależności nosi nazwę eksploracji danych [1] lub drążenia danych [2]. Wśród zastosowań przemysłowych dość istotną dziedziną są zagadnienia związane z gospodarowaniem wodą w systemach miejskich [3, 4].

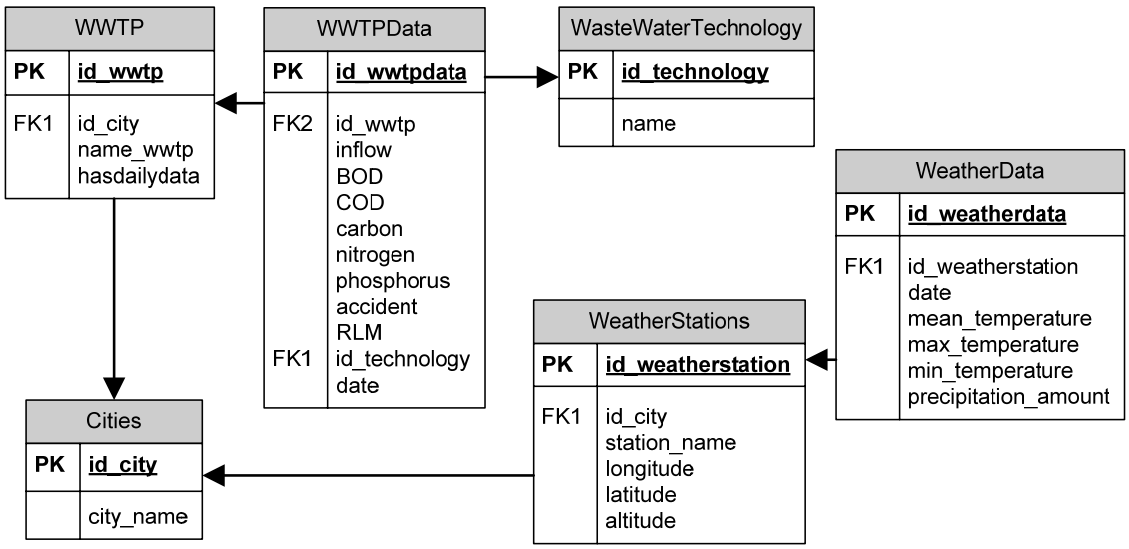
Duża część danych gromadzonych w bazach danych jako dziedzinę posiada czas. Do takich danych należą np. ceny walut, dane dotyczące przyjęć i wypisów pacjentów w szpitalach, występowanie plam na Słońcu. Wszystkie te informacje posiadają ważną wspólną cechę - są to realizacje procesów stochastycznych, zwane szeregami czasowymi [5]. Analiza szeregów czasowych opiera się na poszukiwaniu wzorców występujących w danych, określaniu struktury danych oraz predykcji przyszłych wartości [6]. Celem badawczym jest ocena dostępnego na rynku oprogramowania eksplorującego dane do analiz środowiskowych szeregów czasowych zamieszczonych w bazie danych.

2. Dane wejściowe

W rozważaniach dotyczących eksploracji danych jako przykład wykorzystano szeregi czasowe o rozdzielczości dobowej oraz miesięcznej, obejmujące lata od 2000 do 2007 włącznie. Dane pochodzą z komunalnych oczyszczalni ścieków zlokalizowanych w Baranowie Sandomierskim, Krakowie, Sandomierzu, Tarnobrzegu oraz Warszawie. W bazie danych umieszczono parametry mające wpływ na jakość oczyszczania ścieków. Jako podstawowe parametry pracy oczyszczalni przyjęto parametry w ściekach surowych: natężenie dopływu ścieków, pięciodniowe biochemiczne zapotrzebowanie na tlen (BZT_5), chemiczne zapotrzebowanie na tlen (ChZT), ogólny węgiel organiczny (OWO), zawartość azotu ogólnego (Nog), zawartość fosforu ogólnego (Pog). W bazie danych umieszczono także dane pogodowe, dotyczące analogicznego okresu. Dane pogodowe są potrzebne do oceny wpływu różnego rodzaju czynników, w tym pogody, na zmienność występującą w szeregach czasowych. Skutkiem silnego wpływu danych pogodowych na wartości parametrów oznaczanych w ściekach surowych może być obniżenie jakości predykcji analizowanych danych.

W skład danych pogodowych wchodzą: temperatura średniodobowa, temperatura maksymalna, temperatura minimalna, wysokość opadów, nasłonecznienie.

Dane w postaci tabel umieszczone są w relacyjnej bazie danych, utworzonej w języku SQL. Baza danych składa się z 6 tabel tworzących schemat (rys. 1), a zarządzanie nią odbywa się za pomocą wolnodostępnego systemu zarządzania bazą danych - MySQL.



Rys. 1. Struktura bazy danych
Fig. 1. Schema of database

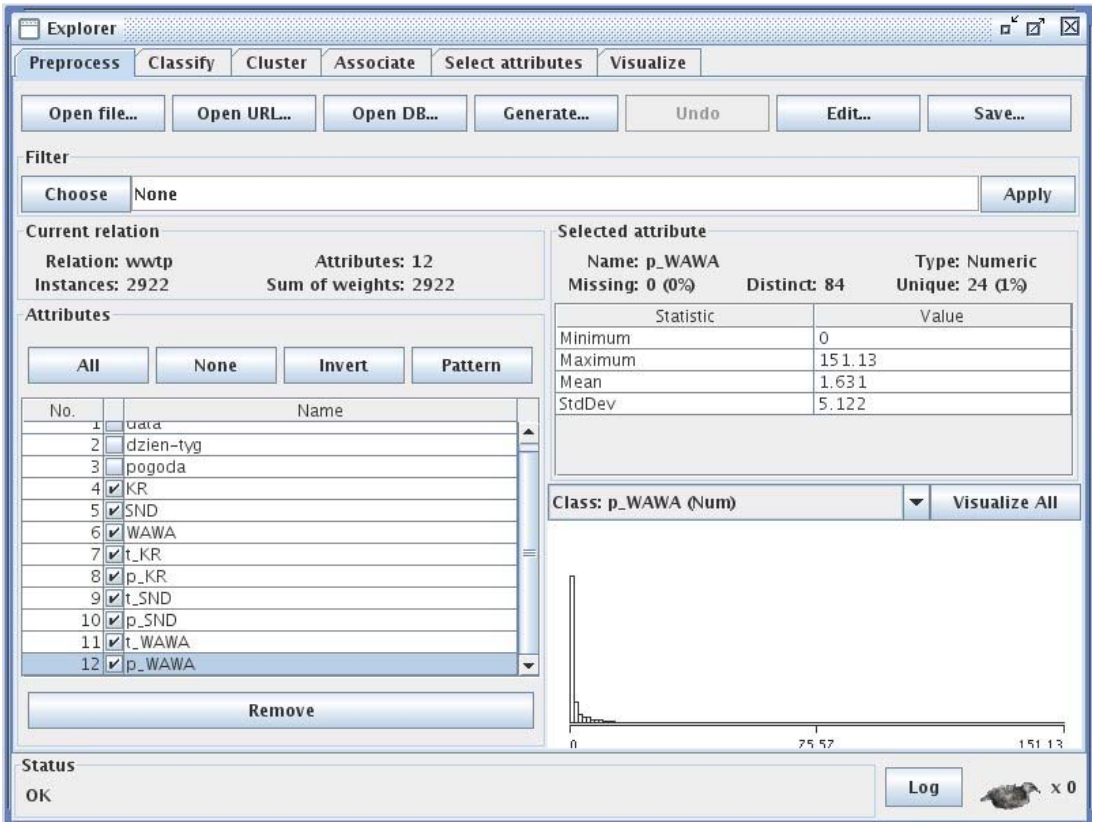
3. Narzędzia eksplorujące dane

Wybranie odpowiedniego narzędzia do analizy szeregów czasowych umieszczonych w bazie danych jest kluczowym zagadnieniem mającym wpływ na jakość analiz. Obecnie dostępnych jest wiele programów pozwalających na analizy *data mining* szeregów czasowych. Do testowania wybrano programy i środowiska dostępne na zasadzie *open source* oraz komercyjne:

- Weka,
- Statistica,
- R,
- Clementine,
- Rattle,
- RapidMiner Community Edition,
- Matlab,
- MS SQL Server 2008 (Analysis Services).

Weka (Waikato Environment for Knowledge Analysis) jest to program rozpowszechniany jako oprogramowanie typu *open source* na licencji GNU (General Public License) [7].

Program można zainstalować pod kontrolą każdego systemu operacyjnego. W projekcie wykorzystano wersję Weka 3.7.0.



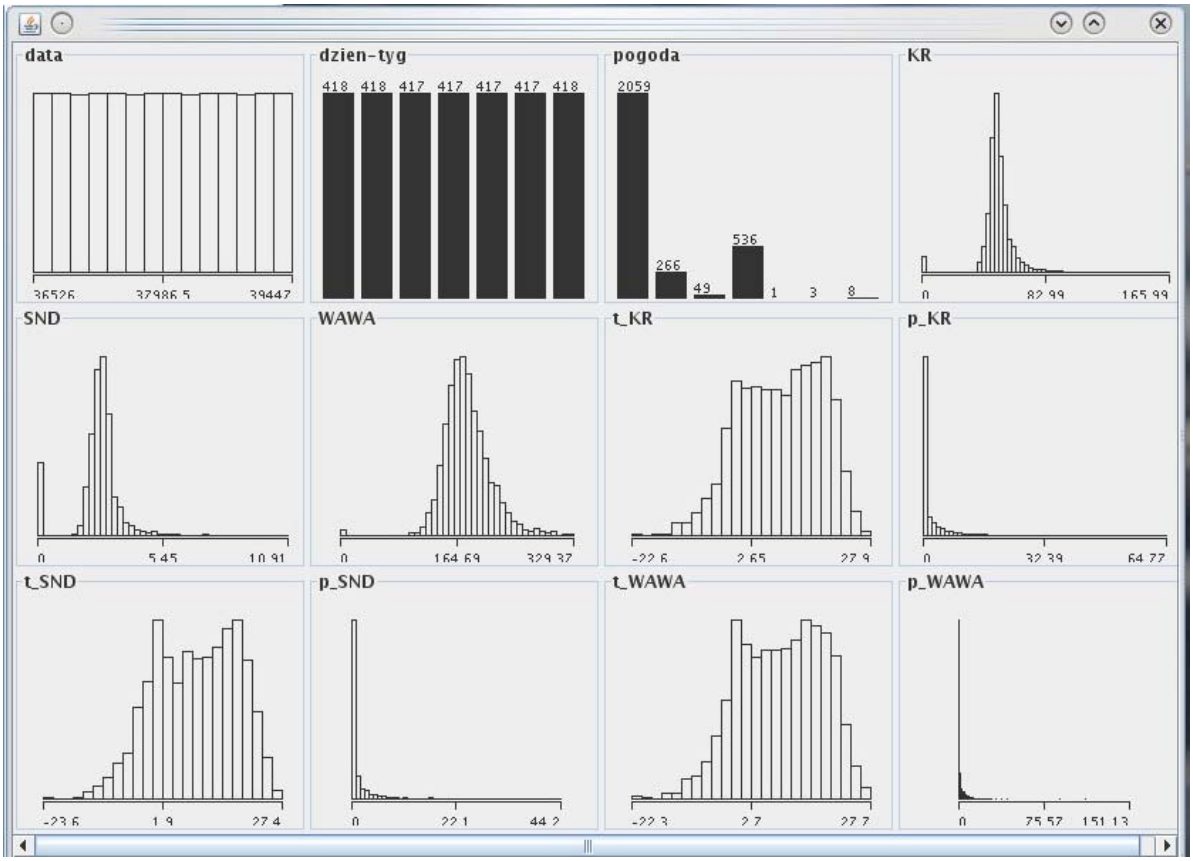
Rys. 2. Analiza danych w programie Weka
Fig. 2. Data analysis In Weka application

Po zainstalowania programu i uruchomieniu pojawia się okno z menu. Dla większości użytkowników najistotniejsza jest zakładka Applications, która zawiera panel umożliwiający analizę danych – Explorer (rys. 2) oraz panel Experimenter - pozwalający na testowanie hipotez, a także panel KnowledgeFlow, który w zasadzie zawiera funkcjonalność „Explorera" i jest oparty na technologii *drag-and-drop*. Dodatkowo w zakładce Applications znajduje się panel Simple CLI – uruchamia środowisko typu „wiersz poleceń" (dla zaawansowanych użytkowników). Wersja Weki 3.7.0 została także wyposażona w osobne zakładki: Program, Tools, Visualization, Help ułatwiające użytkownikom obsługę programu.

Statistica jest komercyjnym oprogramowaniem firmy StatSoft [8]. Program jest dostępny wyłącznie dla platformy Windows. Do analiz wybrano wersję Statistica 8.0 w języku polskim, wraz z dodatkiem uaktualniającym z 27 listopada 2008. Program posiada zakładkę zawierającą pełen pakiet analiz zakresu *data mining*. Podczas uruchamiania programu otwiera się projekt data mining, który pozwala na wykonanie analiz oraz śledzenie wszelkich zmian w tworzonym projekcie (rys. 4).

R jest darmowym środowiskiem na licencji GNU służącym do analizy statystycznej. Środowisko można zainstalować w każdym systemie operacyjnym. Aktualna wersja oprogra-

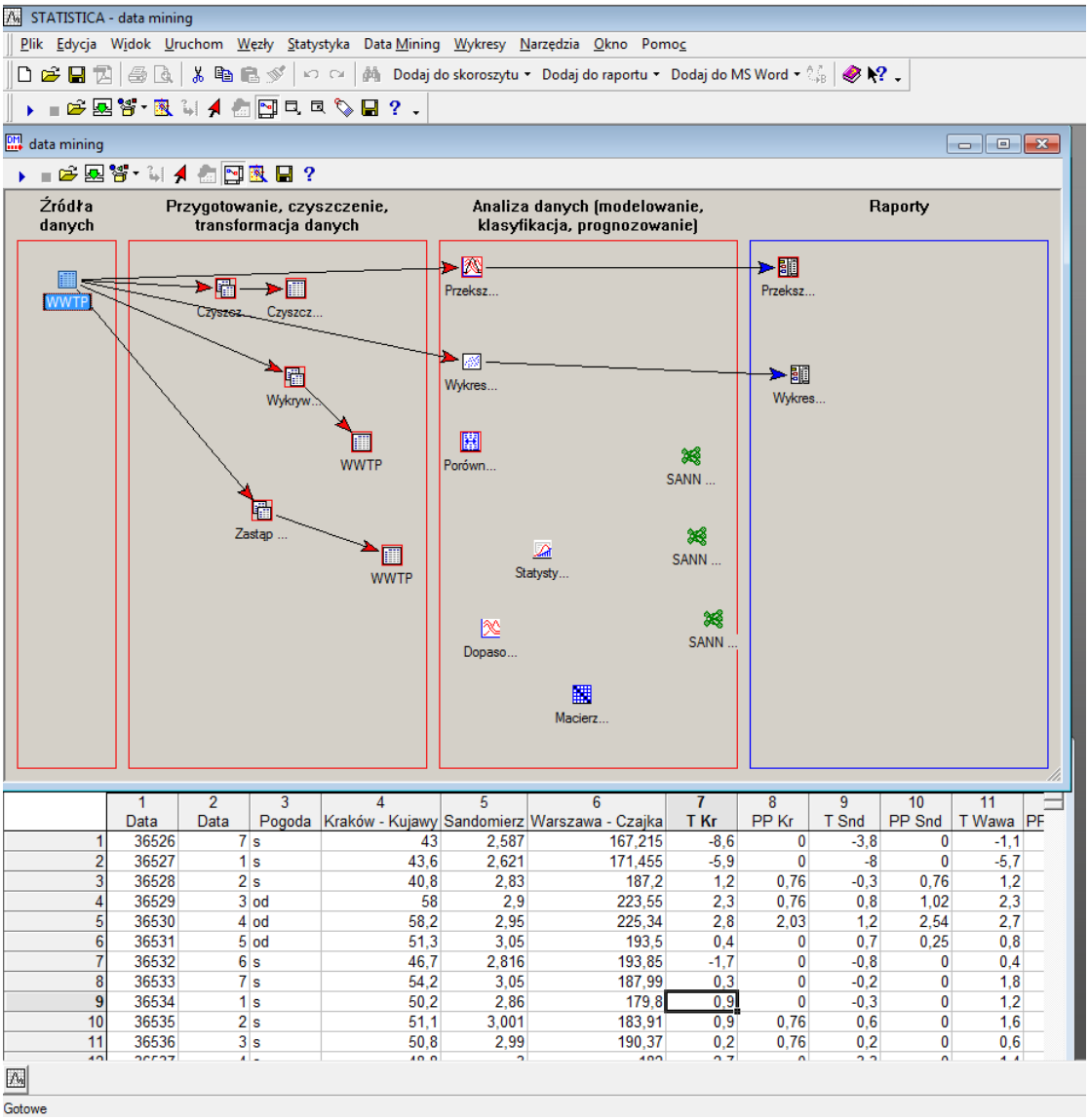
mowania została opublikowana 2009-12-14 (wersja 2.10.1). Podobnie jak Weka, środowisko jest dostępne wyłącznie w języku angielskim. Zaletą R jest możliwość implementacji w komercyjnych programach, w tym w Statistica 8.0 [9]. Obsługa środowiska następuje w trybie tekstowym, poprzez wpisywanie komend w konsoli. Komendy, wraz z dokumentacją, zostały zebrane w pakiety, dostępne na stronie projektu R [10].



Rys. 3. Analiza danych w programie Weka
Fig. 3. Data analysis In Weka application

Program firm IBM SPSS- PASW Modeler 13 (Clementine) jest programem z graficznym interfejsem do analiz data mining. Clementine zawiera algorytmy umożliwiające automatyczną analizę danych. Program składa się z czterech modułów: asocjacji, klasyfikacji, segmentacji oraz publikowania rozwiązań. IBM SPSS Modeler 13 działa jedynie w środowisku Windows, jest dostępny w kilku wersjach językowych, także w języku polskim [11].

Rattle jest to program opierający się na środowisku R, a także na bibliotece graficznej GTK2. Program można zainstalować pod każdym systemem operacyjnym, a użytkowanie jest możliwe na licencji GNU. 15 stycznia 2010 roku udostępniono najnowszą wersję programu 2.5.15. Rattle pozwala na klasyfikację, asocjację, boosting oraz tworzenie modeli liniowych. Program swoim wyglądem przypomina projekt Weka. Tab-bar zawiera narzędzia eksploracji danych, takie jak: eksploracja, testy, transformacje, klasteryzacja, asocjacja, modelowanie, ewaluacja i zapisywanie [12,13].



	1 Data	2 Data	3 Pogoda	4 Kraków - Kujawy	5 Sandomierz	6 Warszawa - Czajka	7 T Kr	8 PP Kr	9 T Snd	10 PP Snd	11 T Wawa	PF
1	36526	7	s	43	2,587	167,215	-8,6	0	-3,8	0	-1,1	
2	36527	1	s	43,6	2,621	171,455	-5,9	0	-8	0	-5,7	
3	36528	2	s	40,8	2,83	187,2	1,2	0,76	-0,3	0,76	1,2	
4	36529	3	od	58	2,9	223,55	2,3	0,76	0,8	1,02	2,3	
5	36530	4	od	58,2	2,95	225,34	2,8	2,03	1,2	2,54	2,7	
6	36531	5	od	51,3	3,05	193,5	0,4	0	0,7	0,25	0,8	
7	36532	6	s	46,7	2,816	193,85	-1,7	0	-0,8	0	0,4	
8	36533	7	s	54,2	3,05	187,99	0,3	0	-0,2	0	1,8	
9	36534	1	s	50,2	2,86	179,8	0,9	0	-0,3	0	1,2	
10	36535	2	s	51,1	3,001	183,91	0,9	0,76	0,6	0	1,6	
11	36536	3	s	50,8	2,99	190,37	0,2	0,76	0,2	0	0,6	

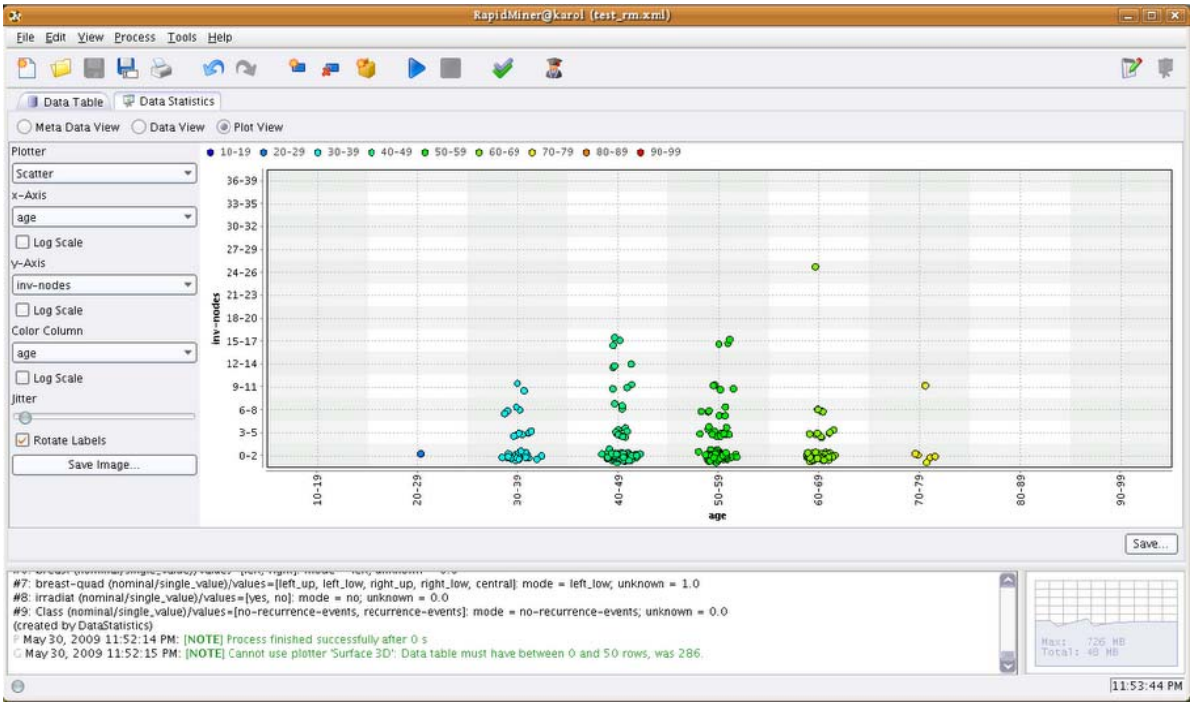
Rys. 4. Widok modułu Data Miner w programie Statistica 8.0
Fig. 4. A view of Data Miner in Statistica 8.0

RapidMiner Community Edition (YALE) wersja 5.0.001 wykorzystuje oprogramowanie Weka. Oprócz ok. 100 aplikacji Weki pozwala dodatkowo na używanie 400 operatorów. Program dostępny jest w dwóch wersjach licencji: OEM oraz GNU oraz został przystosowany do skomplikowanych, wielkich zbiorów danych. RapidMiner został napisany w Java, dlatego też może pracować pod kontrolą większości systemów operacyjnych [14]. Interfejs programu jest przyjazny dla użytkowników (rys. 5). Górne menu zawiera podstawowe funkcje programu, natomiast z prawej strony ekranu znajduje się okno zawierające zakładki: Operators i Repositories.

W zakładce Operators znajdują się wszystkie potrzebne funkcje oraz algorytmy do analiz data mining podzielone na grupy:

- Process Control – pozwala na sterowanie oraz optymalizacje procesów data mining,

- Utillity – pozwala na tworzenie makr oraz zawiera różne możliwości zapisu danych, generowanie danych oraz wykonanie wcześniej zapisanych procesów,
- Repository Access – pozwala na magazynowanie oraz odczyt obiektów IO z bazy danych,
- Import,
- Eksport,
- Data Transformation – pozwala na przeprowadzenie analiz data mining od przygotowania danych, przez filtracje do agregacji,
- Modeling – zawiera narzędzia do klasyfikacji, regresji, asocjacji, grupowania, segmentacji oraz wielu innych analiz,
- Evaluation – pozwala na walidacje modeli, graficzną ewaluacje oraz ocenę istotności.



Rys. 5. Widok aplikacji RapidMiner Community Edition
Fig. 5. A view of RapidMiner Community Edition application

Do środowisk pozwalających na analizę statystyczną dużej ilości danych dołączył produkt SQL Server firmy Microsoft. Wydana w 2005 roku wersja zawierała serwis udostępniający kompletny zestaw funkcji eksploracji danych. Kolejna wersja, SQL Server 2008, została rozszerzona w dziedzinie możliwości integracji pakietu Analysis Services (gdyż tak się moduł eksplorujący dane nazywa) z pakietem MS Office 2007 Excell, przez co łatwo można definiować zapytania i kryteria drążenia danych. Producent dostarczył zaimplementowane takie metody analizy danych, jak:

- Analyze Key Influencers – służąca do wykrywania kluczowych czynników wpływających na wynik,

- Detect Categories – pomagająca w identyfikowaniu i segmentowaniu danych w oparciu o wspólne kategorie,
- Forecast – pozwalająca na przewidywanie przyszłych wartości w oparciu o trendy wprowadzonych danych,
- Prediction Calculator – umożliwiająca dokonanie analizy nowych przypadków,
- Shopping Basket Analysis – w zastosowaniach handlowych wyznacza koszyk towarów kupowanych często razem,
- Scenariusz: What If – metoda szacująca zmiany, jakie wywołuje modyfikacja jednej zmiennej,
- Scenariusz: Goal Seeking – poszukiwanie zmian danych, które pozwalają na osiągnięcie celu w danym kryterium.

Wśród dostępnych algorytmów w MS SQL Server 2008 są zaimplementowane mechanizmy sieci neuronowych.

4. Analiza danych

Metody data mining zawierają szeroki zakres wygodnych narzędzi dla całego procesu przekształcania danych w użyteczną wiedzę, począwszy od wprowadzenia danych lub pobrania ich z zewnętrznego źródła, aż do utworzenia wynikowego raportu [15]. Przekształcanie danych odbywa się w kilku etapach. Pierwszym etapem jest przygotowanie danych do dalszych analiz. W dalszych etapach dane są agregowane, klasteryzowane, filtrowane. Ostatnim etapem analiz jest tworzenie modeli, predykcja oraz ewaluacja [16].

Dane dotyczące podstawowych parametrów pracy oczyszczalni ścieków oraz informacje pogodowe zawarte w bazie danych zostały poddane przetwarzaniu w wybranych programach. Pierwszym etapem analizy szeregów czasowych jest tzw. czyszczenie danych. Polega ono na usunięciu braków danych, standaryzacji, znalezieniu wartości ekstremalnych, a w przypadku silnie skośnych danych konieczna jest ich normalizacja za pomocą przekształcenia Box - Cox [6, 17]. Dzięki znajomości rozkładu danych oraz podstawowych statystyk można rozpocząć drugi etap analiz, do którego należy modelowanie, klasyfikacja oraz prognozowanie [18]. Jako narzędzie do prognozowania szeregów czasowych wybrano automatyczne sieci neuronowe. Do analizy wybrano typy sieci uczonych z nauczycielem: MLP (percepton wielowarstwowy) i RBF (radialna funkcja bazowa) [19]. Liczba warstw sieci i neuronów w poszczególnych warstwach była ustalana automatycznie przez program. Obie wybrane sieci są jednokierunkowe oraz nie występuje w nich sprzężenie zwrotne. Sieci RBF prowadzą do wykrycia bardziej złożonych związków w danych, w tym celu wymagają jednak większej liczby warstw, przez co obliczenia są bardziej czasochłonne [12].

5. Wyniki analiz

Analizę szeregów czasowych oraz ich prognozowanie za pomocą sieci neuronowych wykonano w wybranych środowiskach, opisanych w niniejszym artykule. Rezultaty analiz są bardzo do siebie zbliżone ze względu na używanie tych samych algorytmów w każdym programie.

Podstawowe informacje o parametrach pracy oczyszczalni uzyskano obliczając średnią arytmetyczną, odchylenie standardowe, skośność oraz współczynniki zmienności i korelacji. Najbardziej zróżnicowane pomiary zaobserwowano w oczyszczalni ścieków obsługującej miasto Sandomierz, odchylenie standardowe w przypadku natężenia dopływu ścieków do oczyszczalni dla tych danych wynosi 0.63, a współczynnik zmienności 22%. W pozostałych oczyszczalniach wartość współczynnika zmienności dla natężenia dopływu ścieków wynosi ok. 17%. Wartość tego współczynnika wskazuje na to, że znaczna ilość pomiarów miała wartość zbliżoną do średniej arytmetycznej (rys. 3), jednakże występują pomiary o wartości dużo wyższej i dużo niższej od wartości średniej.

Szeregi czasowe natężenia dopływu ścieków do oczyszczalni z Krakowa i Sandomierza wykazują silną prawostronną skośność, której wartość przekracza 3. W danych pochodzących z tych dwóch oczyszczalni zaobserwowano niewiele pomiarów bardzo niskiego natężenia dopływu. Wysokie natężenie dopływu było natomiast zjawiskiem dość powszechnym (rys. 3). Dane z Warszawy wykazują skośność, jednak jest ona dużo słabsza, niż w pozostałych badanych oczyszczalniach ścieków i wynosi 0.9. Różnice w skośności poszczególnych szeregów czasowych widoczne są na rys. 2. Skośność danych została usunięta przy użyciu przekształcenia Boxa-Coxa.

Tabela 1

Wyniki eksploracji danych dla wybranych sieci neuronowych

Nazwa sieci	Jakość			Aktywacja		algorytm uczenia
	uczenie	testowanie	walidacja	ukryte	wyjściowe	
MLP 14-2-1	0,591	0,664	0,585	liniowa	logistyczna	BFGS 22
RBF 14-21-1	0,597	0,615	0,594	Gaussa	liniowa	RBFT
RBF 21-21-1	0,603	0,652	0,619	Gaussa	liniowa	RBFT
MLP 7-4-1	0,638	0,670	0,629	Tanh	wykładnicza	BFGS 26

W trakcie poszukiwania najlepszego modelu predykcji środowiskowych szeregów czasowych wykonano kilkadziesiąt modeli sieci neuronowych dla każdej z analizowanych oczyszczalni ścieków. Jakość pierwszych modeli udowodniła przewagę sieci MLP nad RBF dla analizowanych danych. Sieci RBF, aby uzyskać jakość uczenia, testowania i walidacji, jakie osiągnęły sieci MLP, potrzebowały dużo większej liczby neuronów pośrednich. Jakość testowania i walidacji wahała się od 0,001 dla sieci RBF z kilkoma neuronami ukrytymi, do 0,6 dla sieci z kilkudziesięcioma neuronami ukrytymi. Dla sieci MLP wyniki te wahały się

w granicach od 0,2 do 0,6 jakości testowej i walidacji (tab. 1). Liczba danych wejściowych – liczba wartości opóźnionych była uzależniona od charakteru danych. Wykonano sieci dla 1, 7, 14, 21 i 28 liczby wartości opóźnionych. Celem było wykonanie sieci przewidującej wartości natężenia dopływów na dobę i dwie doby w przód. Rezultaty uzyskane przez pięć najlepszych sieci neuronowych widoczne są w tabeli 1.

6. Wnioski

Wykorzystanie wielu programów typu *open source* oraz komercyjnych pozwoliło na ocenę przydatności programów do analizy szeregów czasowych umieszczonych w bazie danych.

Wielką zaletą programów Statistica i Clementine jest możliwość zaimplementowania obsługi plików w różnych formatach. W tych programach można także korzystać z danych zawartych w bazach danych. Bardzo pomocny w analizach jest graficzny interfejs stworzony w taki sposób, że nawet początkujący analityk nie będzie miał z nim problemów. Bardzo dobrze funkcjonuje zestaw podręczników i pomocy. Wyniki analiz mogą być przedstawione w formie graficznej lub tabelarycznej. Puste miejsca braków danych mogą zostać niezapełnione, można także skorzystać z jednej z wielu metod uzupełniania, możliwych w trakcie procesu „czyszczenia" danych. Zaletą programu Statistica jest możliwość bezpośredniego uruchamiania skryptów R w programie.

Z programów na licencji GNU najbardziej funkcjonalnym narzędziem do analiz *data mining* jest RapidMiner Community Edition (YALE). Posiada łatwy w obsłudze graficzny interfejs, rozszerzoną użyteczność Weki, możliwość wykonywania komend z poziomu linii poleceń. Dodatkowo program posiada własny system wtyczek oraz dwa tryby działania: eksperta oraz początkującego. Obecnie z RapidMinera korzysta wiele znanych firm, w tym: Cisco Systems, Ford, Honda, HP, IBM, Philips.

Zaletą programu Rattle jest prosty interfejs, który początkowo wydaje się ubogi w funkcje. Jednakże zastosowanie R poszerza zakres funkcjonalności Rattle. Proces instalacji w systemie operacyjnym Linux jest mało skomplikowany, natomiast instalacja Rattle pod Windows przebiega w kilku etapach, podczas których instalowane są GTK, GGobi, R i XML package. Dodatkowo w celu rozszerzenia funkcjonalności projektu można zainstalować Emacs Speac S lub Tinn-R.

Weka jest dobrym rozwiązaniem dla większości użytkowników, pozwala na wykonanie większości popularnych analiz data mining. Zaletą programu jest dostępność wielu instruktaży oraz książek dotyczących zagadnień drążenia danych w programie Weka. Wadą programu jest konieczność wprowadzania plików w formie *csv* lub *arff*, po wcześniejszym usunięciu braków danych. Dane zawarte w bazach danych także nie mogą posiadać luk w rekordach.

Biorąc pod uwagę funkcjonalność i zasób dostępnych analiz najlepszym rozwiązaniem jest pakiet R do zaawansowanych obliczeń statystycznych. Utrudnieniem dla początkujących użytkowników jest brak graficznego interfejsu, przez co wszystkie komendy są wpisywane w linii poleceń. Funkcjonalność pakietu R rozszerzają dodatkowe biblioteki dostępne na stronie projektu.

W dalszej części badań dokonane zostanie porównanie możliwości analizy danych za pomocą środowiska SQL Server Analysis Services. Eksploracja ta zostanie przeprowadzona na znacznie większym zbiorze danych o charakterze rozproszonym – konstruowana będzie hurtownia danych. Jako dodatkowy wątek poruszone zostanie wykorzystanie pakietów dla środowiska Matlab w zastosowaniu do eksploracji danych. Środowisko to bowiem jest bardzo popularne w ośrodkach badawczych.

BIBLIOGRAFIA

1. Morzy M.: Eksploracja danych - przegląd dostępnych metod i dziedzin zastosowań. VI edycja Hurtownie Danych i Business Intelligence, Centrum Promocji Informatyki, Warszawa, 11 kwietnia 2006.
2. Jagielski J., Skorupska I.: Metody pozyskiwania wiedzy z danych historycznych. Bazy danych: modele, technologie, narzędzia, pod red. S. Kozielskiego, B. Małysiak, P. Kasprowskiego i D. Mrozka, Wydawnictwa Komunikacji i Łączności, Warszawa 2005.
3. Rojek I.: Bazy danych i bazy wiedzy dla miejskiego systemu wodno-ściekowego. Bazy danych: nowe technologie, pod red. S. Kozielskiego, B. Małysiak, P. Kasprowskiego i D. Mrozka, Wydawnictwa Komunikacji i Łączności, Warszawa 2007.
4. Gorawski M., Kowalski D.: Klasteryzacja szeregów czasowych na przykładzie pomiarów zużycia mediów. Bazy danych: struktury, algorytmy, metody, pod red. S. Kozielskiego, B. Małysiak, P. Kasprowskiego i D. Mrozka, Wydawnictwa Komunikacji i Łączności, Warszawa 2006.
5. Box G., Jenkins J.: Analiza szeregów czasowych. Prognozowanie i sterowanie. Państwowe Wydawnictwo Naukowe, Warszawa 1983.
6. Larose D.T.: Metody i modele eksploracji danych. Wydawnictwo Naukowe PWN, Warszawa 2008.
7. Weka 3 Documentation, http://www.cs.waikato.ac.nz/ ~ml/weka/index_documentation.html.
8. Internetowy Podręcznik Statystyki, http://www.statsoft.pl/textbook/stathome.html
9. Komsta Ł.: Wprowadzenie do środowiska R. http://cran.r-project.org/doc/contrib/Komsta-Wprowadzenie.pdf
10. The R Project for Statistical Computing, http://www.r-project.org/.

11. Khabaza T., Shearer C.: Data Mining with Clementine. Integral Solution Limited, 20th Jan 1995.
12. Rattle: Gnome Cross Platform GUI for Data Mining using R, http://rattle.toga-ware.com/.
13. Data Mining Desktop Survival Guide, http://datamining.togaware.com/survivor/Data_Mining.html.
14. Rapid Miner, http://rapid-i.com/content/view/26/82/.
15. Chatfield C.: The Analysis of Time Series. An Introduction. Chapman &Hall/Crc, 2004.
16. Last M., Kandel A., Bunke H. (edts), Data Mining in Time Series Database, Series in Machine Perception Artificial Intelligence, Vol. 57, s. 67÷101.
17. Hand D., Mannila H., Smyth P.: Eksploracja danych. Wydawnictwa Naukowo-Techniczne, Warszawa 2005.
18. Witten I.H., Frank E.: Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, Sydney 2005.
19. Gworek S., Utrata A.: Wykorzystanie predyktorów typu neural network do prognozowania szeregów czasowych. Górnictwo i Geoinżynieria 2005, nr 29, z. 4, s. 53÷62.

Recenzenci: Prof. dr hab. inż. Andrzej Grzywak
Prof. dr hab. inż. Bolesław Pochopień

Wpłynęło do Redakcji 28 stycznia 2010 r.

Abstract

Data mining is currently the fastest growing domain of processing data. The purpose of data mining is useful to obtain new knowledge from large data packets using specialized statistical tools. This article makes a comparison of methods for exploration in environments Weka, R, Statistica and Microsoft SQL Server database to an existing quantity of water inflow to the treatment plant. Inputs such application is a time series with daily resolution of the long periods of observation (even several years).

Adres

Monika CHUCHRO: Akademia Górniczo-Hutnicza, Katedra Geoinformatyki i Informatyki Stosowanej, al. Mickiewicza 30, 30-059 Kraków, Polska, chuchro@geol.agh.edu.pl.
Adam PIÓRKOWSKI: Akademia Górniczo-Hutnicza, Katedra Geoinformatyki i Informatyki Stosowanej, al. Mickiewicza 30, 30-059 Kraków, Polska, pioro@agh.edu.pl.