STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

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

ANALIZA DANYCH ŚRODOWISKOWYCH METODAMI EKSPLORACJI DANYCH

Streszczenie. Eksploracja danych jest coraz popularniejszą metodą analizy danych środowiskowych. Głównym problemem podczas analizy takich danych jest silny wpływ składowej losowej oraz skomplikowane relacje między zmiennymi objaśniającymi. Celem podjętych badań jest wyznaczenie składowych szeregu czasowego dotyczącego parametrów produkcji biogazu w oczyszczalni ścieków oraz jego predykcja.

Słowa kluczowe: eksploracja danych, data mining, sieci neuronowe, regresja wieloraka, normalizacja, analiza szeregów czasowych

ENVIRONMENTAL DATA ANALYSIS USING DATA MINING METHODS

Summary. Data mining became popular method of environmental data analysis. Strong influence of irregular variable and complicated relationship between data are main problems during data modeling. This examination purpose is to identifying patterns and predict future values of time series of biogas production from wastewater treatment plant.

Keywords: data mining, artificial neural network, multiple regression, normalization, time series analysis.

1. Wstęp

Ilość danych znajdujących się w różnych bazach danych z każdym rokiem podwaja się. W 2010 roku, według raportu Digital Universe Study firmy IDC, miało powstać 1,2 mln petabajtów informacji [1]. Duża część tych danych opisuje zjawiska środowiskowe. Informacje dotyczące pogody, przyrostu pni drzew, poziomu hałasu, a także wielu innych zjawisk można opisać w postaci ciągu danych, zapisywanych z określonym krokiem czasowym. Zbiór takich

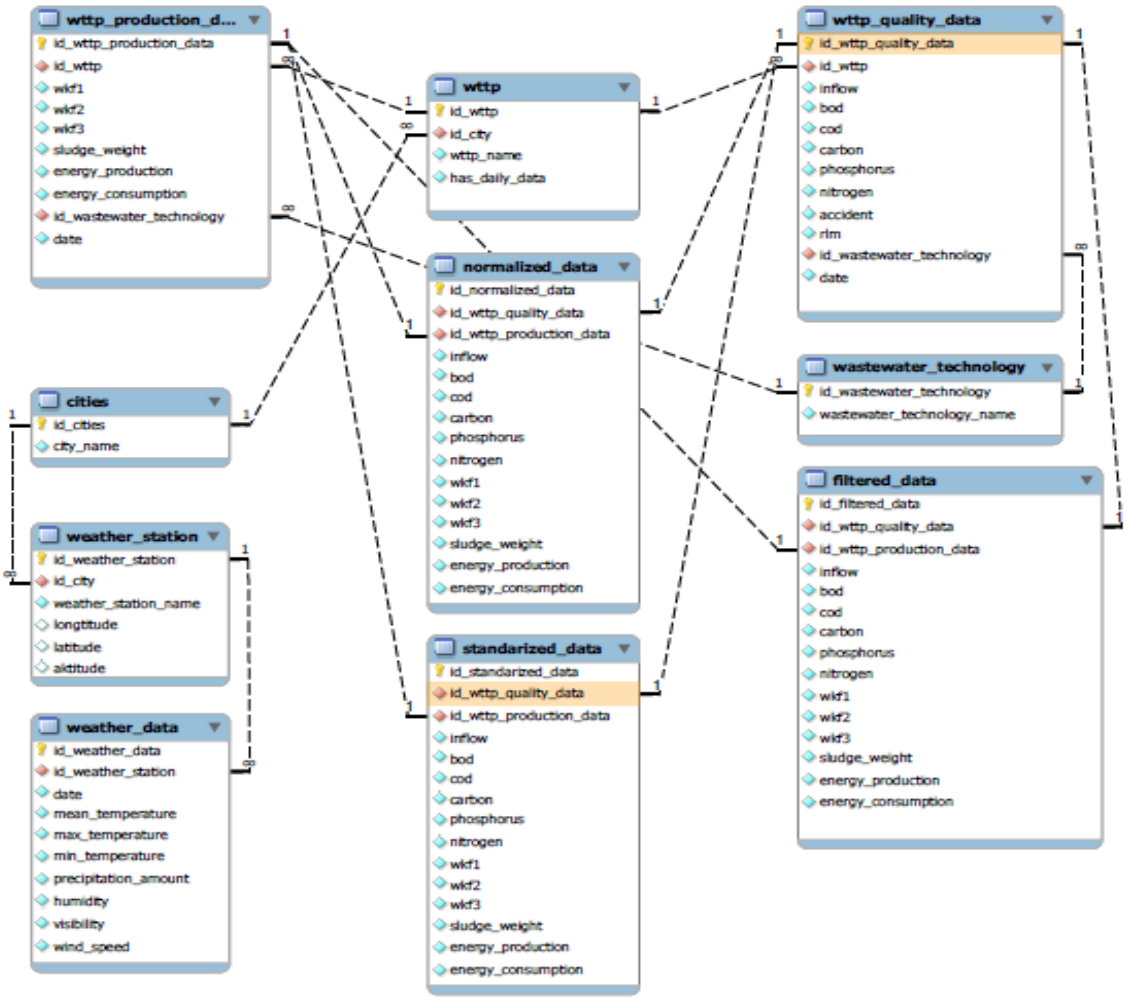
danych określany jest jako szereg czasowy [2]. Cechami charakterystycznymi dla środowiskowych szeregów czasowych jest silne zaszumienie danych, występowanie składowej nieregularnej oraz wielu zmiennych objaśniających. Dane takie są trudne do analizy oraz interpretacji ze względu na współistnienie wielu zmiennych, mających wpływ na analizowane zjawisko, nieposiadających wyraźnej cykliczności oraz trendu, często skorelowanych ze sobą i skointegrowanych [2].

Ważnym etapem analizy danych środowiskowych jest ich dobre przygotowanie do eksploracji danych. Ten etap pracy polega na wykryciu potencjalnych zmiennych, mogących mieć wpływ na analizowane zjawisko, czyszczenie danych oraz wstępne transformacje danych. W przypadku danych środowiskowych często konieczne jest zastosowanie transformacji Boxa-Coxa w celu normalizacji danych. Bardzo często korzysta się także z filtrów górnoprzepustowych oraz różnych metod uzupełniania brakujących danych i usuwania wartości błędnych [3].

Następnym etapem poszukiwania prawidłowości i wzorców w danych jest zastosowanie metod eksploracji danych. W zależności od założeń badawczych można zastosować tutaj metody ilościowe lub jakościowe. Do najbardziej popularnych metod eksploracji danych zaliczyć można: sieci neuronowe, drzewa klasyfikacyjne i regresyjne, MARS, algorytmy genetyczne oraz regresję wieloraką [4].

2. Dane wejściowe

Dane środowiskowe wykorzystane do analiz pochodzą z oczyszczalni ścieków typu PUB2, według klasyfikacji zawartej w Krajowym Programie Oczyszczania Ścieków Komunalnych. Jest to oczyszczalnia biologiczna z podwyższonym usuwaniem związków azotu i fosforu, spełniająca standardy odprowadzanych ścieków dla aglomeracji powyżej 15 000 RLM [5]. Zmienną zależną wybraną do analiz jest suma produkcji wytworzonego biogazu przez trzy wydzielone komory fermentacyjne (WKF). Dane mają rozdzielczość dobową i obejmują okres od 01.01.2005 do 31.12.2007 roku. Informacje dotyczące pracy oczyszczalni ścieków, produkcji biogazu oraz warunków pogodowych zostały umieszczone w relacyjnej bazie danych. Baza danych składa się z 10 tabel tworzących schemat (rys. 1), a zarządzanie nią odbywa się za pomocą wolno dostępnego systemu zarządzania bazą danych – MySQL.



Rys. 1. Schemat bazy danych
Fig. 1. Database schema

3. Podstawowe funkcje programu Statisica 9.0

Do analizy danych spośród obszernej puli programów komercyjnych oraz *open source* wybrano program Statistica 9.0 firmy StatSoft [7]. Jest to program do statystycznej analizy danych, pozwalający na korzystanie z baz danych. Dodatkowo Statistica ma rozbudowane środowisko programistyczne, oparte na Visual Basic oraz pozwala na korzystanie z języka R. Dużą zaletą programu Statistica jest łatwy w obsłudze graficzny interfejs oraz specjalne okno startowe, pozwalające na modelowanie analiz *data mining*. Program ten posiada rozbudowaną bibliotekę w językach polskim oraz angielskim, zawierającą filmy, opisy zagadnień statystycznych i funkcji programu. Poszczególne funkcje programu rozmieszczone są w kilku panelach. Panel Data Mining zawiera liczne analizy takie, jak: drzewa klasyfikacyjne i regresyjne, losowy las, uogólnione modele addytywne, MARS, analiza skupień, analiza składowych niezależnych, analizy sekwencji i klasyfikacji, analizy koszykowe, automatyczne sieci

neuronowe i wiele innych. Program Statistica pozwala na tworzenie makr, zapamiętywanie modeli i wdrażanie ich na nowych zestawach danych. Utworzone projekty mogą być zapamiętane jako skrypty pmml, C/C++. Program zawiera panel do obsługi baz danych oraz do wysyłania i edytowania zapytań za pomocą interfejsu graficznego [7].

4. Przygotowanie danych do analizy

4.1. Usuwanie braków danych i błędnych informacji

Zgromadzone dane w bazach danych mogą mieć braki, błędne informacje lub źle zakodowane zmienne jakościowe. Może być to spowodowane błędem aparatury pomiarowej, błędem zapisu, awarią sprzętu lub brakiem dostępu do informacji. Istnieje wiele metod postępowania w takiej sytuacji. Wybór metody zależy od osoby analizującej dane oraz od samych danych. Jeśli zmienna objaśniająca ma wiele luk, a informacje jakie zawiera mogą być pominięte lub też zastąpione przez inną zmienną, to taką zmienną objaśniającą można usunąć lub zignorować [8]. W przypadku gdy braki danych występują w kilku zmiennych objaśniających, można zastosować:

- wartość globalną, która będzie wskazywać na brak danych,
- ręczne uzupełnianie braków danych, oparte na danych zewnętrznych,
- zastosowanie średniej, mediany lub średniej ruchomej,
- użycie najbardziej prawdopodobnej wartości [6] [3].

4.2. Standaryzacja danych, kodowanie zmiennych jakościowych

Jeśli analizowane dane wykazują rozkład normalny, można wykonać standaryzację danych, która umożliwia obliczenie prawdopodobieństwa występowania zmiennej *X*. Transformacja ta pozwala także na porównanie kilku zmiennych o różnej średniej i odchyleniu standardowym [2].

Przed przystąpieniem do transformacji danych należy sprawdzić czy dane wykazują rozkład normalny. Najprostszą metodą jest wizualizacja danych w postaci histogramu. Można także wykonać test Kołmogorowa-Smirnowa albo Shapiro-Wilka [8].

Standaryzacja jest transformacją wykorzystującą rozkład wartości w poszczególnych cechach, w wyniku której otrzymuje się średnią wartość oczekiwaną 0 i wariancję wynoszącą 1 (1). Transformację tę wykorzystuje się w przypadku danych o rozkładzie normalnym.

$$x_i' = \frac{x_i - \bar{x}}{s}, \tag{1}$$

gdzie: x_i' – wartość zestandaryzowana, x_i – wartość rzeczywista, $\bar{x}$ – wartość średnia z próby, s – odchylenie standardowe z próby.

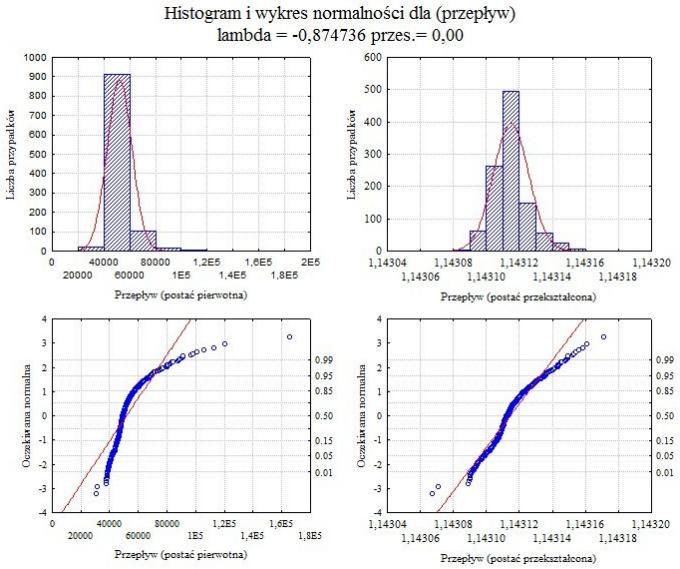
Dane jakościowe, zakodowane poszczególnymi słowami, warto zakodować w postaci binarnej lub w notacji, wykorzystującej wartości −1, 0, 1. Takie sposoby kodowania pozwalają na wykorzystanie zmiennych jakościowych w analizie regresji wielorakiej, sieciach neuronowych i drzewach regresyjnych [8].

4.3. Normalizacja danych

Jeśli dane nie wykazują rozkładu normalnego, co jest częstym zjawiskiem w przypadku danych środowiskowych, należy wykonać ich normalizację. Jednym ze sposobów nadania danym rozkładu normalnego jest ich zlogarytmowanie [2]. Drugą metodą jest wykonanie przekształcenia Boxa-Coxa, przedstawionego na rys. 2. Tak przekształcone dane mogą być wykorzystane do testów i analiz, które w założeniach wymagają rozkładu normalnego zmiennych. Przekształcenie Boxa-Coxa (2) można także użyć w celu poprawy jakości modelu. W przypadku zmiennych o wykazujących skośność lub rozkład odmienny od normalnego, zastosowanie przekształcenia w stosunku do jednej zmiennej lub obu często wpływa na uzyskanie większej dokładności modelu [8].

$$\widetilde{X} = \begin{cases} \dfrac{(x+\alpha)^{\lambda}}{\lambda}, & \lambda \neq 0 \\ \log(x+\alpha), & \lambda = 0 \end{cases} \tag{2}$$

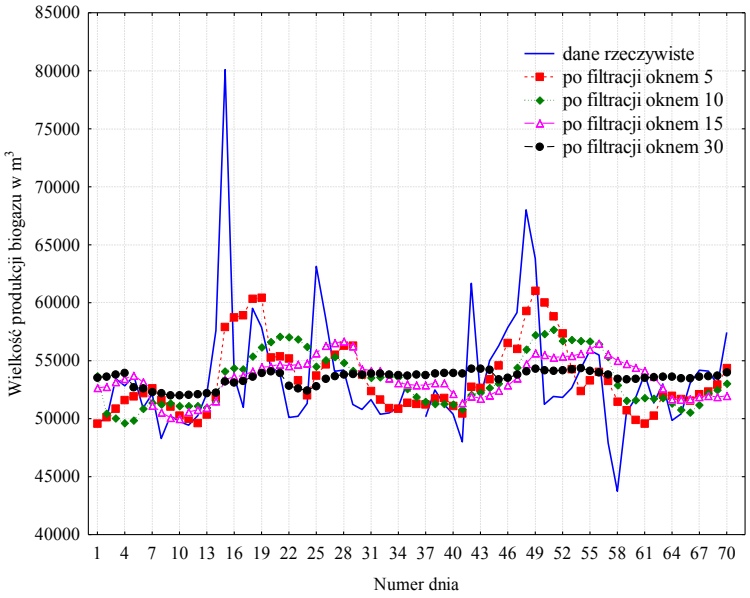
gdzie: $\widetilde{X}$ – zmienna przekształcona, λ – główny parametr przekształcenia, α – parametr przesunięcia zmiennej.



Rys. 2. Przekształcenie Boxa-Coxa danych wykazujących leptokurtyczność i wysoką skośność
Fig. 2. Leptokurtic and strong bias data of Box-Cox transform

4.4. Filtracja

Modelowanie zjawiska na podstawie silnie zaszumionych danych jest działaniem trudnym i często nieskutecznym. W przypadku analizy danych silnie zaszumionych należy wcześniej wykonać filtrację dolnoprzepustową, która osłabia wysokie częstotliwości i przepuszcza niskie. Rezultatem takiej filtracji jest wygładzony szereg. Dzięki temu można identyfikować zmiany procesu maskowanego przez proces szumowy. W zależności od wielkości okna filtra można osiągnąć różny stopień wygładzenia. Przykładem filtra dolnoprzepustowego jest średnia ruchoma [9]. W przypadku gdy analizie będzie poddany składnik krótkookresowy szeregu czasowego, można wykonać filtracje górnoprzepustową, której efekt będzie przeciwny do efektu filtru dolnoprzepustowego, widocznego na rys. 3 [8].



Rys. 3. Filtracja danych natężenia dopływu ścieków do oczyszczalni filtrem dolnoprzepustowym
Fig. 3. Influent into wastewater treatment plant filtration with downpass filters

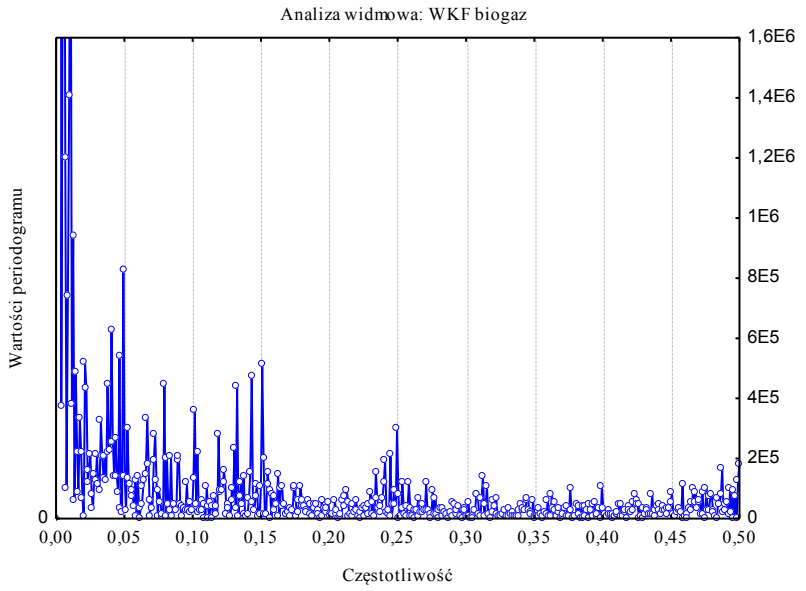
5. Analiza danych

Celem analiz jest ustalenie struktury szeregu czasowego sumy produkcji biogazu w wydzielonej komorze fermentacyjnej. W trakcie ustalania struktury szeregu czasowego wyznaczana jest tendencja rozwojowa, inaczej nazywana trendem, składowe systematyczne oraz składowa nieregularna. Oprócz ustalenia występowania poszczególnych składowych struktury w analizowanym szeregu czasowym lub ich braku, należy wyznaczyć rodzaj relacji pomiędzy tymi składowymi. Relacja pomiędzy składnikami szeregu czasowego może być addytywna, wtedy wartości składowych sumują się. W drugim rodzaju relacji wartość końcowa jest iloczynem składowych szeregu, czyli szereg ma charakter multiplikatywny. W trakcie analiz zostaną wykorzystane dane nieprzekształcone, a także poddane transformacjom, wy-

mienionym powyżej w artykule (wzory 1 i 2 oraz rys. 2 i 3). Jako zmienne objaśniające zostały wykorzystane dane pogodowe, parametry pracy oczyszczalni, wartości opóźnione sumy produkcji biogazu, zmienne jakościowe oznaczające dzień tygodnia, miesiąc, występowanie świąt oraz okres aktywności przedświątecznej.

Na podstawie wyników uzyskanych z analizy skupień, ARIMA oraz drzew regresyjnych zostały wybrane zmienne mogące mieć istotny statystycznie wpływ na kształtowanie się sumy produkcji biogazu w wydzielonych komorach fermentacyjnych (WKF). Dodatkowo, wykonano analizę widmową zmiennej objaśnianej. Wyniki analizy widmowej widoczne są na rys. 4. Na wykresie zauważalne są wzrosty koncentracji wartości szczytowych w pobliżu częstotliwości: 0.25, 0.15, 0.5. Analiza Fouriera wykazała występowanie składowej periodycznej o okresowościach 7, 14, 30 dni.

Eliminacja zmiennych w procesie regresji wielorakiej opierała się na usuwaniu zmiennych na podstawie wskaźnika istotności statystycznej *p*. Zmienne usuwane były pojedynczo, zaczynając od najwyższej wartości *p*, aż do uzyskania modelu zawierającego same istotne parametry. Następnie przeprowadzono analizę reszt, w celu sprawdzenia czy występuje jej autokorelacja. W przypadku występowania autokorelacji reszt lub autokorelacji cząstkowej wykonano modele reszt niezawierające wyrazu wolnego. Dla porównania wykonano trzy różne modele, zawierające kolejno nieprzekształcone dane, dane znormalizowane i dane poddane filtracji dolnoprzepustowej.



Rys. 4. Analiza Fouriera produkcji biogazu w wydzielonych komorach fermentacyjnych
Fig. 4. Production of biogas in digestive tank Fourier analysis

Dobrym narzędziem do predykcji szeregów czasowych są sztuczne sieci neuronowe. Dużą zaletą stosowania sieci neuronowych jest ich zdolność do odwzorowania złożonych funkcji, a także nieliniowych relacji występujących w danych. Dodatkowo, sztuczne sieci neuro-

nowe są dość łatwe w użyciu. Projektowanie oraz uczenie sieci może się odbywać w sposób automatyczny. Do zaprojektowania sieci neuronowej wykorzystano panel „automatyczne poszukiwanie sieci". Funkcja ta pozwala na zbudowanie wstępnych modeli sieci neuronowych, które następnie można udoskonalać dzięki funkcji „projekt sieci użytkownika". W programie, w przypadku problemów regresyjnych istnieje wybór typu sieci: z perceptronem wielowarstwowym albo z radialną funkcją bazową (RBF) [7]. Perceptron wielowarstwowy (MLP) pozwala na osiągnięcie dobrych rezultatów predykcji zmiennej objaśnianej o jednolitym charakterze. W przypadku gdy w danych występują podgrupy różniące się średnią wartością, wariancją lub trendem, jakość predykcji dla sieci typu MLP znacznie spada. W takim przypadku lepsze rezultaty można osiągnąć stosując sieć neuronową typu RBF [10] [11]. Celem analizy było wykonanie modeli predykcyjnych, na podstawie informacji uzyskanych we wcześniejszych analizach. Ze względu na jednolity charakter danych, w których wartość średnia oraz wariancja nie zmieniają się istotnie statystycznie w czasie, wybrano sieć neuronową z perceptronem wielowarstwowym.

Perceptron wielowarstwowy (MLP) jest siecią jednokierunkową z nauczycielem [10]. MLP składa się z trzech typów neuronów: wejściowych, pośrednich oraz wyjściowych. Wszystkie neurony, wchodzące w skład perceptronu, dokonują agregacji danych wejściowych poprzez wyznaczenie sumy ważonych wejść [12, 13]. Funkcje aktywacji neuronów są zmienne: funkcja aktywacji neuronów wejściowych ma charakter sigmoidalny, neuronów ukrytych najczęściej nieliniowy lub logistyczny, a neuronów wyjściowych nieliniowy [11]. W początkowej fazie uczenia połączeniom między neuronami nadaje się niewielkie losowe wartości wag. W trakcie trwania uczenia wartości te są modyfikowane aż do uzyskania przez nie prawidłowej wartości, wynikającej z oczekiwanej funkcji decyzyjnej [11]. Sieć jest uczona na zbiorze testowym, następnie porównywana ze zbiorami testowym oraz walidacyjnym [13].

6. Wyniki

Szereg czasowy sumy produkcji biogazu na oczyszczalni ścieków ma składową sezonową o charakterze miesięcznym oraz składową cykliczną o charakterze tygodniowym. Sezonowość widoczna jest na wykresie ogólnym danych, a także w wynikach analizy Fouriera (rys. 4).

Analiza regresji wielorakiej udowodniła istotny wpływ niektórych zmiennych objaśniających na dane modelowane; są nimi:

- produkcja biogazu w WKF opóźnione o 1, 4 i 5 okresów, masa osadu nadmiernego i natężenie dopływu ścieków, opóźnione o 1 okres,

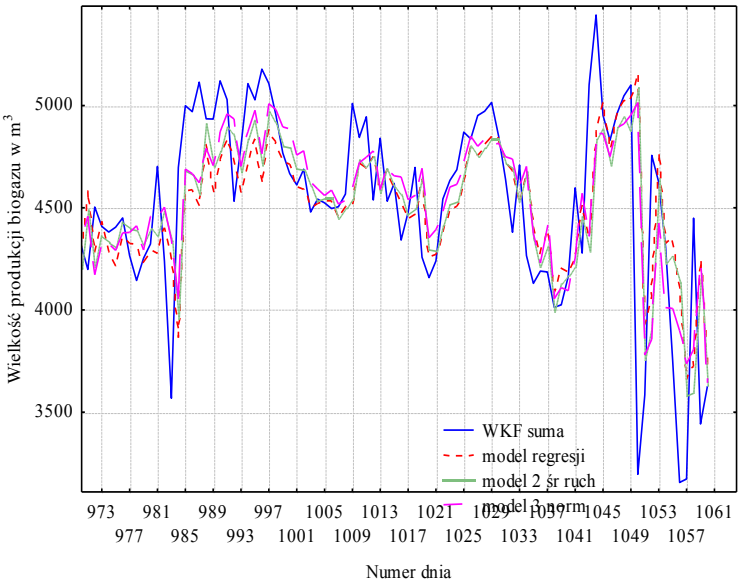
- dane pogodowe: temperatury średnia oraz minimalna, wysokości opadów opóźnione o 1 okres,
- dane w formie binarnej: styczeń, luty, listopad, niedziela, sobota.

Dane dotyczące produkcji biogazu, natężenia dopływu ścieków oraz masy osadu nadmiernego zostały zestandaryzowane, znormalizowane, a także przekształcone filtrem dolnoprzepustowym. W ten sposób uzyskano 3 modele regresji. W pierwszym modelu zostały uwzględnione dane zestandaryzowane, w drugim dane po zastosowaniu filtru dolnoprzepustowego, a w trzecim – znormalizowane. Rezultaty modelowania widoczne są w tabeli 1. Na rys. 5 przedstawiono dopasowanie wykonanych modeli do zmiennej prognozowanej. Rezultaty osiągnięte przy użyciu regresji wielorakiej pozwalają sądzić, że istnieje jeszcze co najmniej jedna zmienna, mająca istotny wpływ na dane modelowane, jednakże w przypadku zaszumionych danych środowiskowych wytłumaczenie 66% zmienności danych można uznać za zadowalające.

Tabela1

Wyniki regresji wielorakiej zstępującej

	Model regresji	**Model 2 śr. ruch.**	**Model 3 norm.**
R wielorakie	0,8066	0,8090	0,8133
Wielorakie R2	**0,6505**	**0,6545**	**0,6614**
Skorygowane R2	0,6468	0,6505	0,6562
p	0,0000	0,0000	0,0000
Błąd std. estymacji	309,65	307,83	305,17



Rys. 5. Dopasowanie modeli regresji do zmiennej prognozowanej
Fig. 5. Regression models matching to the original data

Na podobnej zasadzie, jak modele regresji wielorakiej zostały wykonane 3 grupy sieci neuronowych. Każda grupa sieci zawiera 5 perceptronów wielowarstwowych. Wyniki najlepszej sieci neuronowej z każdej grupy ukazane są w tabeli 2. Widoczne są tam wartości

współczynnika korelacji R Pearsona pomiędzy danymi wejściowych a utworzonymi modelami predykcyjnymi. Współczynnik korelacji R Pearsona – w przypadku sieci neuronowych, w programie Statistica – został oznaczony jako „jakość". Wartości współczynnika korelacji obliczane są na każdym etapie tworzenia sieci. W ten sposób każdą sieć charakteryzują trzy parametry „jakości". Najlepsze rezultaty, czyli najwyższą wartość współczynnika korelacji R Pearsona uzyskano dla danych znormalizowanych. Model ten objaśnia 74,5% zmienności danych. Wszystkie prezentowane w tabeli 2 modele mają dobrą moc predykcyjną. Średni bezwzględny błąd procentowy dla prognoz wygasłych nie przekroczył 10%.

Tabela 2

Wyniki sieci neuronowych

Rodzaj zmiennych	Nazwa sieci	Jakość (uczenie)	Jakość (testowanie)	Jakość (walidacja)
Standaryzowane	MLP 25-15-1	0,760	0,729	0,718
Średnia ruchoma	MLP 27-5-1	0,740	0,706	0,724
Normalizowane	MLP 27-6-1	0,763	0,702	0,745

7. Wnioski

Metody eksploracji danych są dobrym narzędziem do analizy zaszumionych środowiskowych szeregów czasowych. Najlepsze rezultaty modelowania i predykcji danych można uzyskać poprzez wykorzystanie kilku metod, które się uzupełniają. Zastosowanie analizy widmowej, analizy skupień, ARIMA oraz wyrównania wykładniczego ułatwia poznanie struktury wewnętrznej danych. Informacje takie pozwalają na wyodrębnienie trendu i składowych cyklicznych. Znajomość składowych szeregu czasowego i wielkości składowej losowej umożliwia wykonanie lepszych modeli predykcyjnych za pomocą regresji wielorakiej oraz sieci neuronowych. Dodatkowo, lepsze rezultaty modelowania można uzyskać poprzez wykorzystanie danych transformowanych: standaryzowanych, normalizowanych lub filtrowanych. Dobre rezultaty predykcji i możliwość kształtowania relacji można uzyskać dzięki regresji wielorakiej wstecznej. Wadą tej metody jest wymóg braku kointegracji zmiennych objaśniających. Ponadto, potrzebna jest znajomość struktury szeregu czasowego.

Sieci neuronowe są dobrym narzędziem predykcyjnym zaszumionych szeregów czasowych. Pozwalają na modelowanie ilości danych wejściowych, a przez to można kształtować procesem jej uczenia. Metoda ta wymaga jednak praktyki, szczególnie w przypadku samodzielnego wykonywania modelu. Dodatkowo, należy pamiętać o zjawisku przeuczenia, które często osiągane jest w przypadku nieumiejętnego tworzenia sieci neuronowych.

Wykonane analizy oraz modele regresji i MLP pozwalają na dokładną analizę danych. W przypadku prognozy danych dla obu typów modeli średni bezwzględny błąd procentowy nie przekroczył 10%.

Praca finansowana w ramach grantu promotorskiego nr.18.18.140.942

BIBLIOGRAFIA

1. http://gigaom.files.wordpress.com/2010/05/2010-digital-universe-iview_5-4-10.pdf.
2. Chatfield C.: The analysis of time series. Chapman & Hall/CRC, Boca Raton 2004.
3. Larose D. T.: Metody i modele eksploracji danych. PWN, Warszawa 2008.
4. Chuchro M., Piórkowski A.: Wykorzystanie metod i narzędzi eksploracji danych do analizy zmienności natężenia dopływu do komunalnych oczyszczalni ścieków. Studia Informatica, Vol. 31, No. 2B (90), Wyd. Pol. Śląskiej, Gliwice 2010, s. 348÷358.
5. Opis programu Statistica 9.0 www.statsoft.pl.
6. Witten I. H., Frank E.: Data mining. Practical machine learning tools and techniques. Elsevier, San Francisco 2005.
7. Sikora M.: Data clearing and transformation- the first stage of data mining process. Studia Informatica, Vol. 25, No. 2 (58), Wyd. Pol. Śląskiej, Gliwice 2004, s. 127÷136.
8. Hand D., Mannila H., Smyth P.: Eksploracja danych. WNT, Warszawa 2005.
9. Gworek S., Urata A.: Wykorzystanie predyktorów typu neural network do prognozowania szeregów czasowych. Górnictwo i Geoinżynieria, nr 29, z. 4, 2005, s. 53÷62.
10. Tadeusiewicz R.: Wprowadzenie do praktyki stosowania sieci neuronowych. Czytelnia StatSoft (www.statsoft.pl).
11. Tadeusiewicz R.: Sieci neuronowe. Akademicka Oficyna Wydawnicza, Warszawa 1993.
12. Chrabałowska J., Halicka K.: Prognozowanie wielkości przychodów ze sprzedaży z wykorzystaniem modeli ARIMA oraz SSN. ZN Pol. Białostockiej. Ekonomia i Zarządzanie, z. 8, Białystok 2003.
13. Krajowy Program Oczyszczania Ścieków Komunalnych
14. http://www.kzgw.gov.pl/pl/Krajowy-program-oczyszczania-sciekow-komunalnych.html.

Recenzenci: Dr inż. Robert Brzeski
Prof. dr hab. inż. Alicja Wakulicz-Deja

Wpłynęło do Redakcji 16 stycznia 2011 r.

Abstract

Data mining is useful tool for environmental time series analysis. This kind of data could have complicated structure with strong influence of irregular component. Also many independent variables mould values of one dependent variable. These independent variables could be cross-correlated and cointegrated. Good preparation of data for analysis, helps to get statistically significant and proper models of analyzed data.

The paper focused on data preparation, preprocessing and prediction with multiple regression and artificial neural network. Preparation of data concentrated on creation of database (Fig.1), complement of gaps in time series. Also methods of data transformation were discussed as a preprocessing.

Main analysis concern on multiple regression and artificial neural network from three groups of data: standardized, normalized and after low-pass filtration. Results of these analyses are seen on Table 1 and Table 2 and in the Figure 5.

Adres

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