mgr inż. BARTOSZ BĘBEN
dr inż. ANDRZEJ TOMCZYK
Instytut Energoelektryki, Politechnika Wrocławska

Wykorzystanie metod eksploracji danych do analizy parametrów środowiskowych

Realizacja sieci sensorycznych dla potrzeb pomiarów czynników stanowiących zagrożenie w środowisku wymaga zaprojektowania i użytkowania dużych baz danych. Postęp w dziedzinie informatyki umożliwia gromadzenie dużych zbiorów danych o objętości przekraczającej niekiedy rząd terabajtów. Obecnie wyzwaniem staje się nie tyle efektywne przechowywanie dużej ilości danych, ale przede wszystkim ich analiza. Z tego względu powstało zapotrzebowanie na nowe metody i narzędzia informatyczne wspomagające odkrywanie wiedzy z danych. W artykule opisano zasady stosowania metod eksploracji danych (ang. Data Mining), określono sposoby podejścia do analizy powyższą metodą, przedstawiono główne zadania analizy eksploracji danych oraz załączono przykład implementacji analizy wybranych parametrów środowiskowych.

1. WSTĘP

Odkrywanie wiedzy z danych za pomocą wyszukiwania przydatnych informacji z dużych zbiorów danych lub baz danych jest nazywane eksploracją danych. Jest to dziedzina z pogranicza: informatyki, statystyki, zarządzania danymi oraz sztucznej inteligencji. Każda z wyżej wymienionych dziedzin zachowuje swój odrębny charakter, ale jednocześnie wspólnie wyznacza nowe problemy i nowe zasady rozwiązań. Zagadnienie analizy metodą eksploracji danych [2] można podzielić na trzy etapy:

- podejście typu „biała skrzynka”,
- przygotowanie danych wejściowych do ich analizy,
- traktowanie eksploracji danych jako zaplanowanego procesu analizy danych.

Technologia „białej skrzynki” uwzględnia oprócz danych wejściowych i wyjściowych również znajomość algorytmicznych struktur przetwarzania, tak aby dobrze poznać proces przetwarzania danych. Przygotowanie poprawności danych wejściowych jest ważnym i żmudnym procesem ich eksploracji, który według autorytetów w tej dziedzinie zajmuje około 60% czasu analizy, tj. uzupełnienie brakujących danych, poprawa błędnych danych, standaryzacja danych. Analiza metodami eksploracji nie może polegać tylko na zakupie programów typu „Data Mining”, wprowadzeniu danych wejściowych i czekaniu na wynik, ale na właściwie zaplanowanym procesie analizy danych.

Mając to na uwadze, można zaakceptować poniższą definicję eksploracji danych [1], jako:
„Eksploracja danych jest procesem odkrywania nowych, ważnych współzależności, wzorców, trendów dzięki przeszukiwaniu dużych baz danych za pomocą technik rozpoznawania wzorców i metod statystycznych oraz matematycznych”

2. OGÓLNIE STOSOWANE ZADANIA ANALIZY METODĄ EKSPLORACJI DANYCH

Zadania eksploracji danych [1], można sprowadzić do następujących zagadnień:

- opis,
- szacowanie (estymacja),
- przewidywanie (predykcja),
- klasyfikacja,
- grupowanie,
- odkrywanie reguł.

Opis analizowanych danych polega na znalezieniu wzorców i trendów badanych danych. Stosowanie formy graficznej badanych danych, nawet w formie wykresów w funkcji czasu, pozwala nam już inaczej spojrzeć na analizowane dane.

Zadanie szacowania (estymacji) danych jest podobne do ich klasyfikacji z wyjątkiem charakteru zmiennej celu, który jest numeryczny, a nie jakościowy. Prosta regresji liniowej wyznacza linię prostą, która najlepiej przybliża związek dwóch zmiennych według Metody Najmniejszych Kwadratów.

Przewidywanie (predykcja) jest podobne do klasyfikacji i szacowania (estymacji) z tym, że zajmuje się przyszłością a nie przeszłością.

W zadaniu klasyfikacji jest jakościowa zmienna celu, np. grupa wiekowa: młody, średni, starszy, która pozwala podzielić zbiór danych na trzy kategorie.

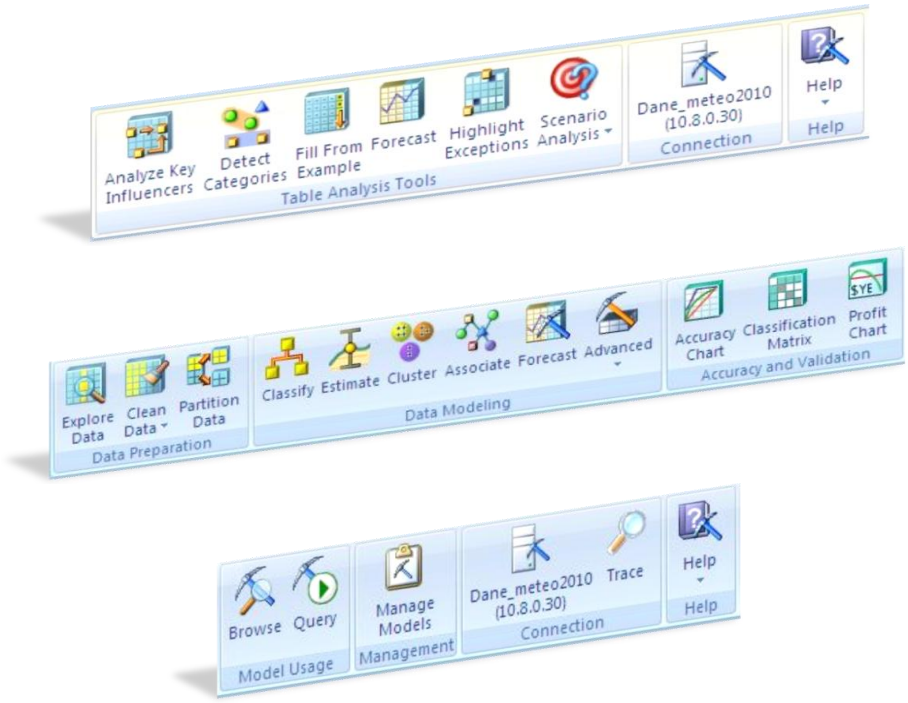
Najczęściej stosowanymi metodami podczas szacowania, przewidywania i klasyfikacji są algorytmy: K-najbliższych sąsiadów, drzew decyzyjnych i sieci neuronowych.

Grupowanie danych zajmuje się łączeniem rekordów i obserwacji w klasy podobnych obiektów. Grupowanie różni się od klasyfikacji tym, że nie posiada zmiennej celu. Algorytm grupowania próbuje podzielić cały zbiór danych na stosunkowo zgodne grupy, gdzie podobieństwo rekordów w grupie jest maksymalizowane, a podobieństwo spoza grupy jest minimalizowane.

Odkrywanie reguł polega na szukaniu, które atrybuty są „powiązane ze sobą”, i jest często nazywana analizą koszykową na zasadzie analogii z „koszykiem zakupów”. Zadanie odkrywania reguł polega na odkrywaniu relacji pomiędzy dwoma lub wieloma atrybutami i przybiera zazwyczaj charakter reguł asocjacyjnych typu : „jeśli poprzednik to następnik”.

3. ZASTOSOWANE NARZĘDZIE ANALIZY DANYCH – MICROSOFT EXCEL DATA MINING ADD INS [3]

Microsoft SQL Server 200, w skład którego wchodzi Analysis Services, posiada możliwość eksploracji danych. Wbudowane algorytmy mogą być użyte do wyszukiwania zależności w dużych zbiorach danych. W celu skorzystania z możliwości data miningowych, SQL Servera w programie Microsoft Excel 2007, koniecznym jest zainstalowanie darmowego dodatku Microsoft SQL Server 2008 Data Mining Add-ins for Microsoft Office 2007. Poprawna instalacja powoduje powstanie dwóch nowych zakładek: *Analyze* oraz *Data Mining*. W skład każdej z nich wchodzi kilka kreatorów pozwalających na eksplorację danych za pomocą kilku metod. Na podstawie konkretnych danych system, w sposób automatyczny, dobiera algorytm i wymagane wartości współczynników. Możliwe jest tworzenie oraz porównywanie modeli udostępnionych poprzez Server Analysis Services.



Rys. 1. Funkcje dostępne w pakiecie

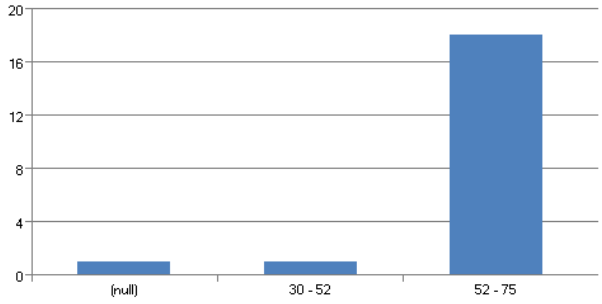
W celu korzystania z zakładki analizy koniecznym jest skonfigurowanie połączenia z SQL Server Analisys Services, które pozwala na użycie algorytów data miningowych dostępnych w SQL Server do analizy danych w arkuszu Excel. Następnie Dane w arkuszu należy sformatować do postaci tabelarycznej.

Analizie dla przykładu poddano następujące parametry środowiskowe (mierzone na dachu budynku), tj.:
– prędkość wiatru [m/s],
– temperatura powietrza [^{o}C],
– wilgotność względna powietrza [% RH],
– natężenie oświetlenia [W/m^{2}],
– od 2010.01.01 godz. 00:00 do 2010.12.31 godz. 23:59 m z rozdzielczością 30 sekund.

3.1. Przygotowanie danych

Pierwszym krokiem jest pobranie danych przy użyciu źródła danych ODBC, które pozwala na jednolity dostęp do baz danych zarówno lokalnych, jak i zdalnych różnego typu i złożoności. Konieczna jest następnie ich wstępna obróbka przy użyciu dostępnych metod, tzn:
a) Poznaj dane – *Explore data*

Funkcja Explore Data pozwala na graficzne pokazanie różnorodności danych.



Rys. 2. Przykładowy wynik działania funkcji Explore data

Oś Y – to liczba wystąpień, oś X – to zakres wartości w danej grupie.

b) Wypełnij na podstawie przykładu – *Fill from example*

Pozwala na wypełnianie luk w danych na podstawie wzorca.

Tablica 1

Przykładowy wynik działania funkcji Fill from example

Data	Wilgotność [%RH]
2010-06-27 05:19	75
2010-06-27 05:20	74,7
2010-06-27 05:20	74,7
2010-06-27 05:21	74,9
2010-06-27 05:21	74,3
2010-06-27 05:22	74,3
2010-06-27 05:22	74,6
2010-06-27 05:23	74,6
2010-06-27 05:23	74,8
2010-06-27 05:24	
2010-06-27 05:24	74,5
2010-06-27 05:24	74,5
2010-06-27 05:25	74,5
2010-06-27 05:25	74,6
2010-06-27 05:26	30
2010-06-27 05:26	74,5
2010-06-27 05:27	74,3
2010-06-27 05:27	74,3
2010-06-27 05:28	74,4
2010-06-27 05:28	74,3

Data	Wilgotność [%RH]
2010-06-27 05:19	75
2010-06-27 05:20	74,7
2010-06-27 05:20	74,7
2010-06-27 05:21	74,9
2010-06-27 05:21	74,3
2010-06-27 05:22	74,3
2010-06-27 05:22	74,6
2010-06-27 05:23	74,6
2010-06-27 05:23	74,8
2010-06-27 05:24	74,5
2010-06-27 05:24	74,5
2010-06-27 05:24	74,5
2010-06-27 05:25	74,5
2010-06-27 05:25	74,6
2010-06-27 05:26	30
2010-06-27 05:26	74,5
2010-06-27 05:27	74,3
2010-06-27 05:27	74,3
2010-06-27 05:28	74,4
2010-06-27 05:28	74,3

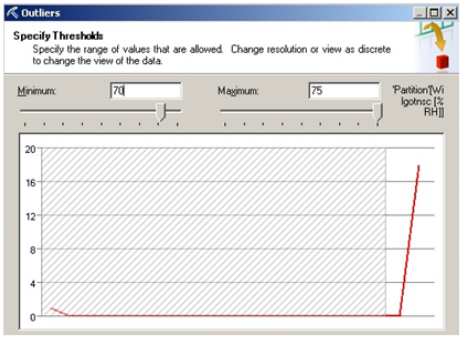
c) Wyczyść dane – *Clean data*

Narzędzie to pozwala na usunięcie danych wychodzących poza ustalony zakres. Funkcja jest przydatna wówczas, gdy np. dane pomiarowe obarczone są pojedynczymi błędnymi wartościami, spowodowanymi np. zakłóceniami itp.

Tablica 2

Przykładowy wynik działania funkcji *Clean data* – usunięcie wiersza spoza zakresu

Data	Wilgotność [%RH]
2010-06-27 05:19	75
2010-06-27 05:20	74,7
2010-06-27 05:20	74,7
2010-06-27 05:21	74,9
2010-06-27 05:21	74,3
2010-06-27 05:22	74,3
2010-06-27 05:22	74,6
2010-06-27 05:23	74,6
2010-06-27 05:23	74,8
2010-06-27 05:24	74,5
2010-06-27 05:24	74,5
2010-06-27 05:24	74,5
2010-06-27 05:25	74,5
2010-06-27 05:25	74,6
2010-06-27 05:26	30
2010-06-27 05:26	74,5
2010-06-27 05:27	74,3
2010-06-27 05:27	74,3
2010-06-27 05:28	74,4
2010-06-27 05:28	74,3



3.2. Analiza danych

a) Wstępna analiza

Tablica 3

Podstawowa statystyka danych dotyczących roku 2010 w zakresie: wilgotności, temperatury, natężenia oświetlenia i prędkości wiatru

	Wilgotność [%RH]	Temperatura [°C]	Natężenie oświetlenia *W/m²*	Prędkość wiatru [m/s]
Średnia	77.59	8.84	114.46	2.63
Błąd standardowy	0.01637	0.9956	0.1914	0.0019
Mediana	81.4	9.6	3.1	2.1
Odchylenie standardowe	16.72	10.17	195.57	1.96
Wariancja próbki	2797.36	1034.54	382487.54	38.47
Zakres	73.4	54.3	1228.9	22.0
Minimum	24.8	-18.3	0	0
Maksimum	98.2	36.0	1228.9	22.0

b) Korelacja

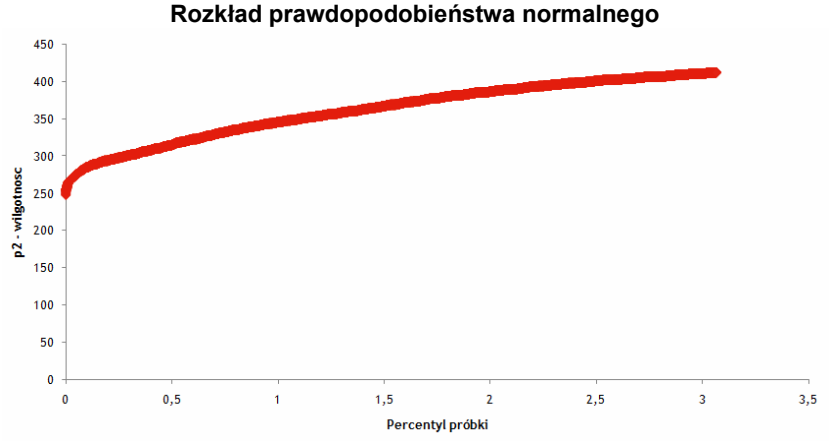
Tablica 4

Współczynniki korelacji wzajemnej w zakresie: wilgotności, temperatury, natężenia oświetlenia i prędkości wiatru

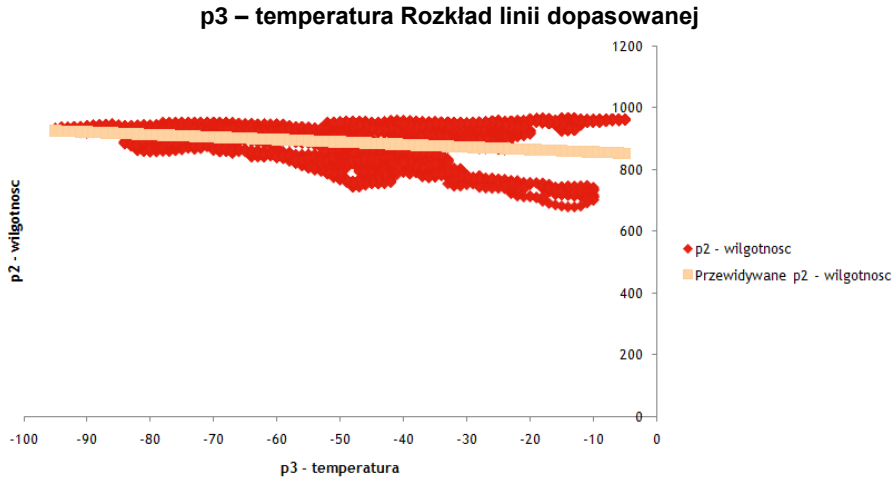
	Wilgotność	Temperatura	Natężenie oświetlenia	Prędkość wiatru
Wilgotność	1			
Temperatura	-0.49	1		
Natężenie oświetlenia	-0.57	0.43	1	
Prędkość wiatru	-0.06	0.038	0.05	1

Obliczenia współczynników korelacji potwierdzają znaną prawdę, że wraz ze wzrostem temperatury maleje wilgotność, wraz ze wzrostem natężenia oświetlenia maleje wilgotność oraz wraz ze wzrostem natężenia oświetlenia rośnie temperatura.

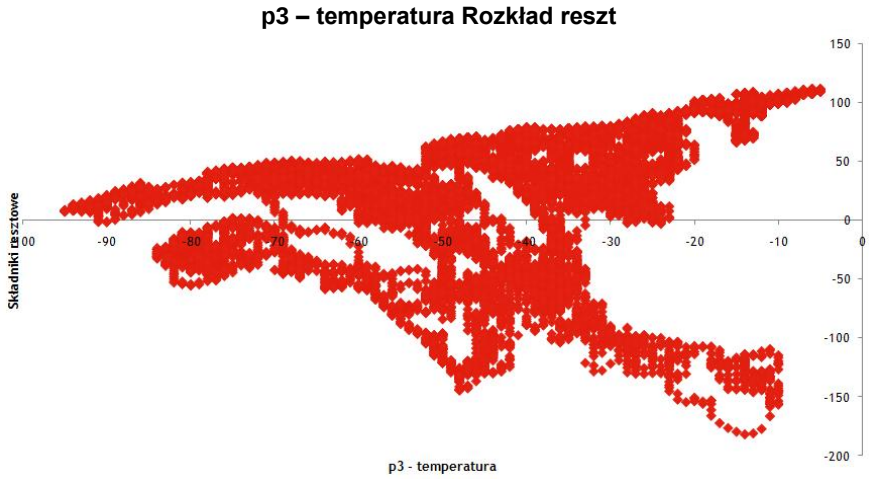
c) Regresja temperatury i wilgotności



Rys. 3. Rozkład normalny wilgotności dla określonej średniej i odchylenia standardowego wg wzoru: Rozkład Normalny (x, średnia, Odchylenie standardowe)

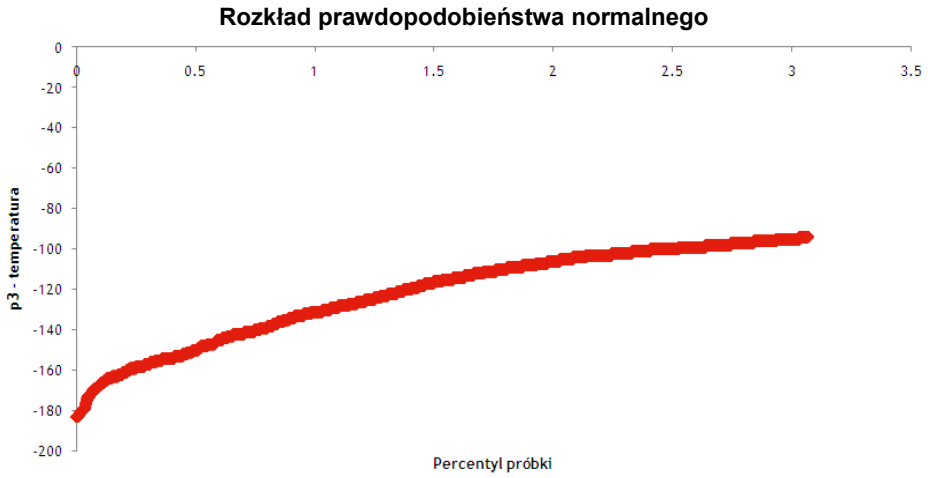


Rys. 4. Zależność wartości prognozowanej wilgotności w funkcji temperatury

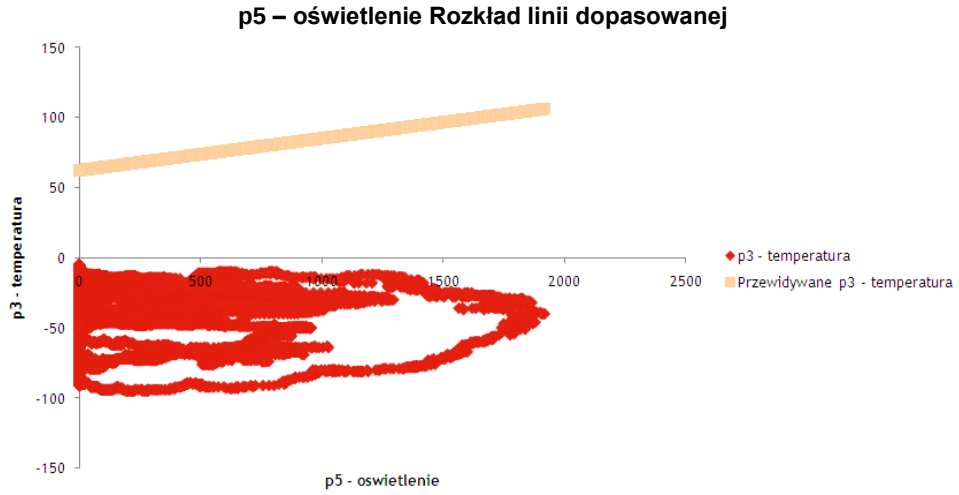


Rys. 5. Rozkład różnic wartości prognozowanych i rzeczywistych temperatury

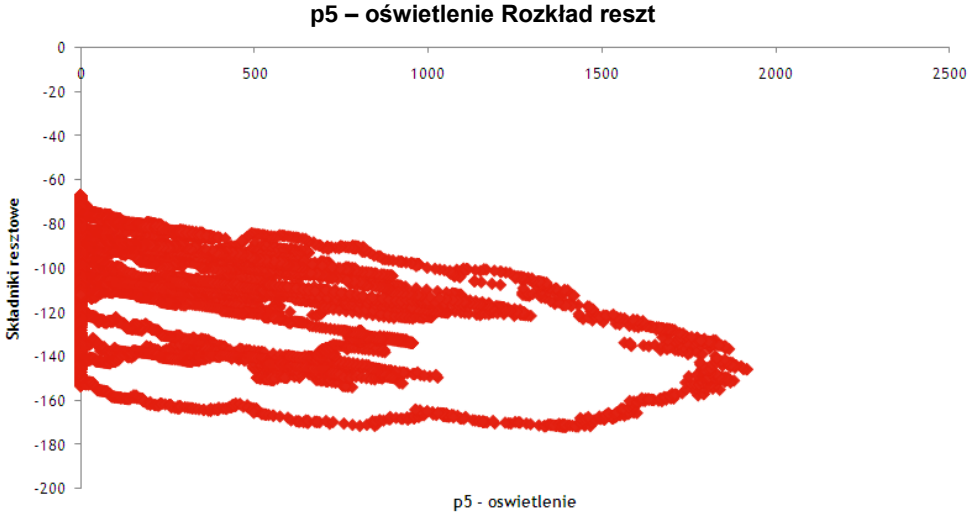
d) Regresja temperatury i oświetlenia



Rys. 6. Rozkład normalny temperatury dla określonej średniej i odchylenia standardowego wg wzoru: Rozkład Normalny (x, średnia, Odchylenie standardowe)

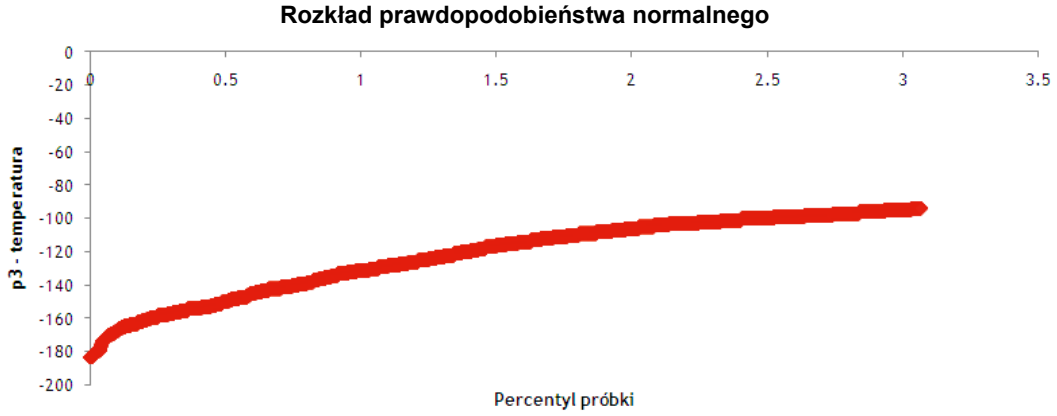


Rys. 7. Zależność wartości prognozowanej temperatury w funkcji natężenia oświetlenia

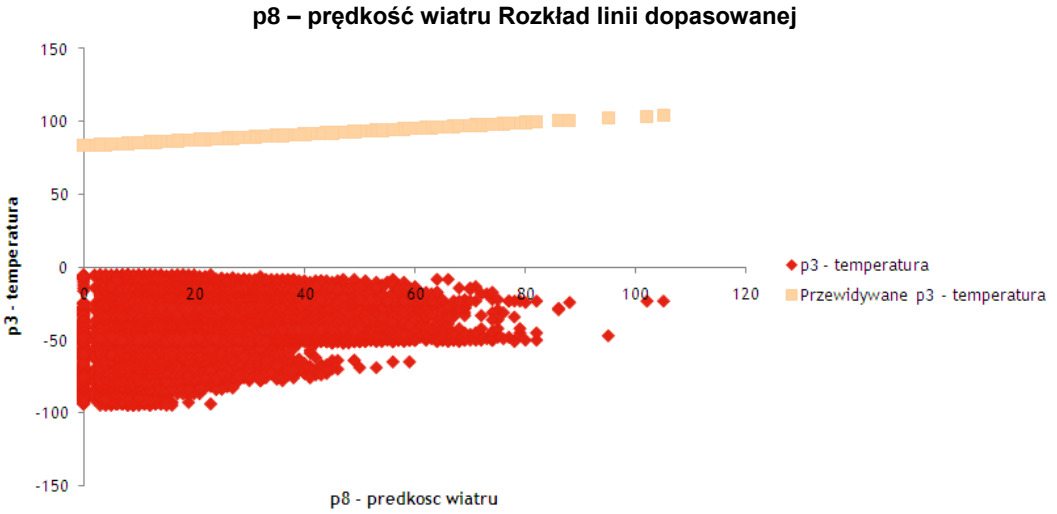


Rys .8. Rozkład różnic wartości prognozowanych i rzeczywistych natężenia oświetlenia

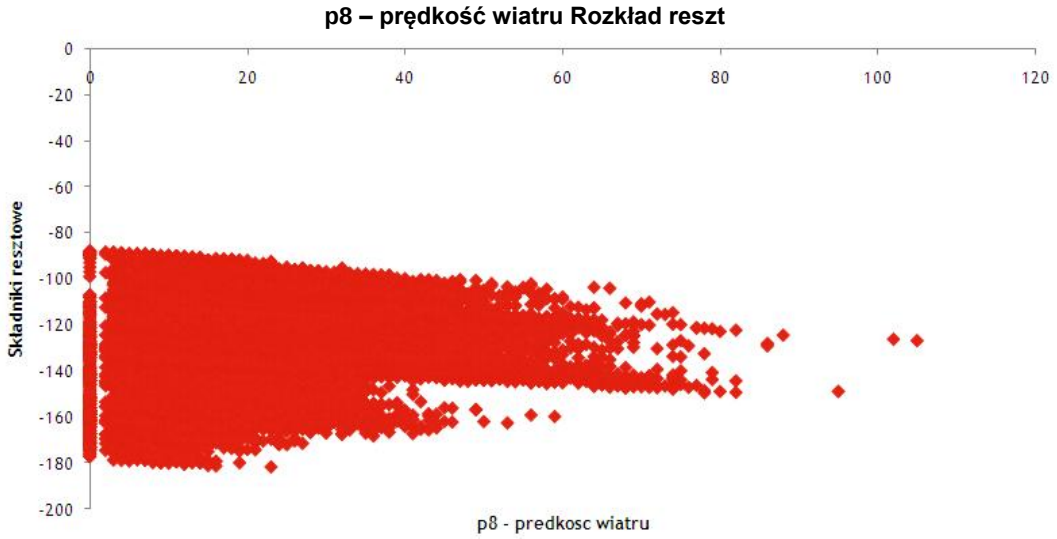
e) Regresja temperatury i prędkości wiatru



Rys. 9. Rozkład normalny temperatury dla określonej średniej i odchylenia standardowego wg wzoru: Rozkład Normalny (x, średnia, Odchylenie standardowe)

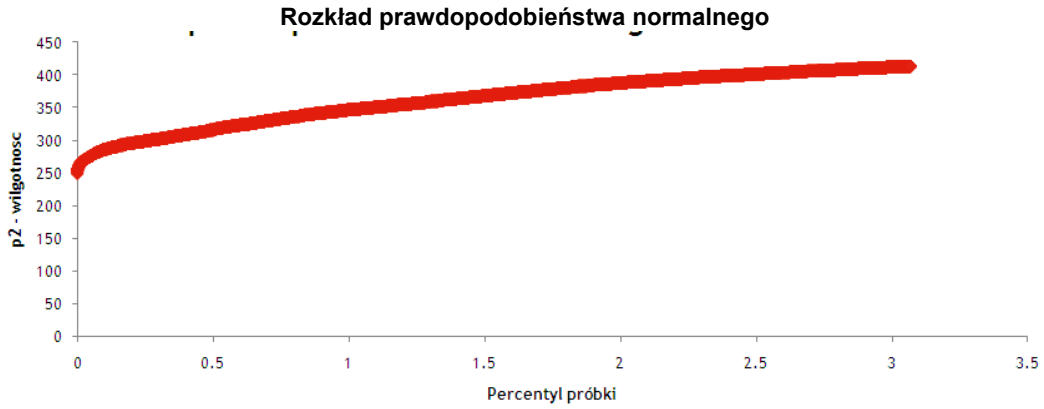


Rys. 10. Zależność wartości prognozowanej temperatury w funkcji prędkości wiatru

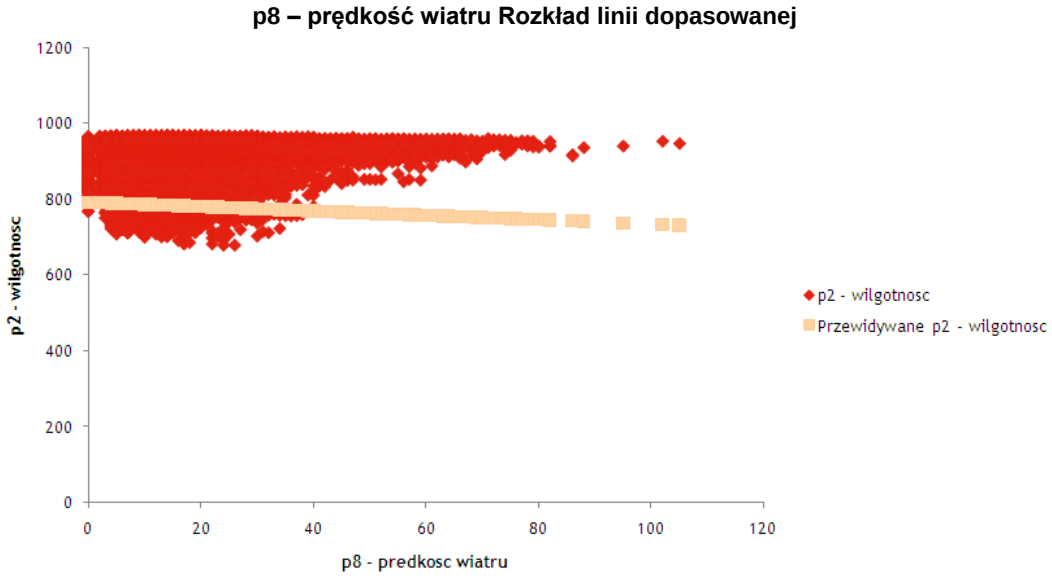


Rys. 11. Rozkład różnic wartości prognozowanych i rzeczywistych prędkości wiatru

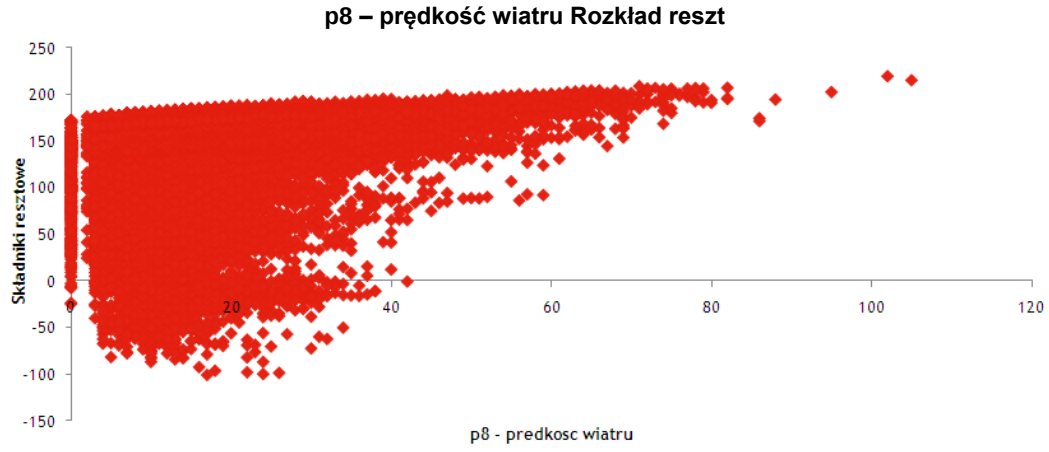
f) Regresja wilgotności i prędkości wiatru



Rys. 12. Rozkład normalny wilgotności dla określonej średniej i odchylenia standardowego wg wzoru: Rozkład Normalny (x, średnia, Odchylenie standardowe)

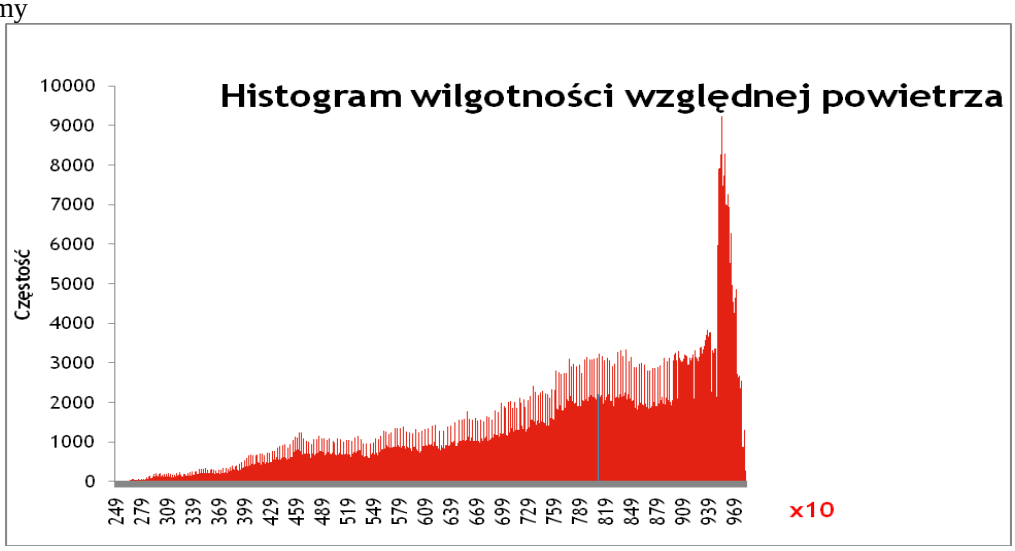


Rys. 13. Zależność wartości prognozowanej wilgotności w funkcji prędkości wiatru

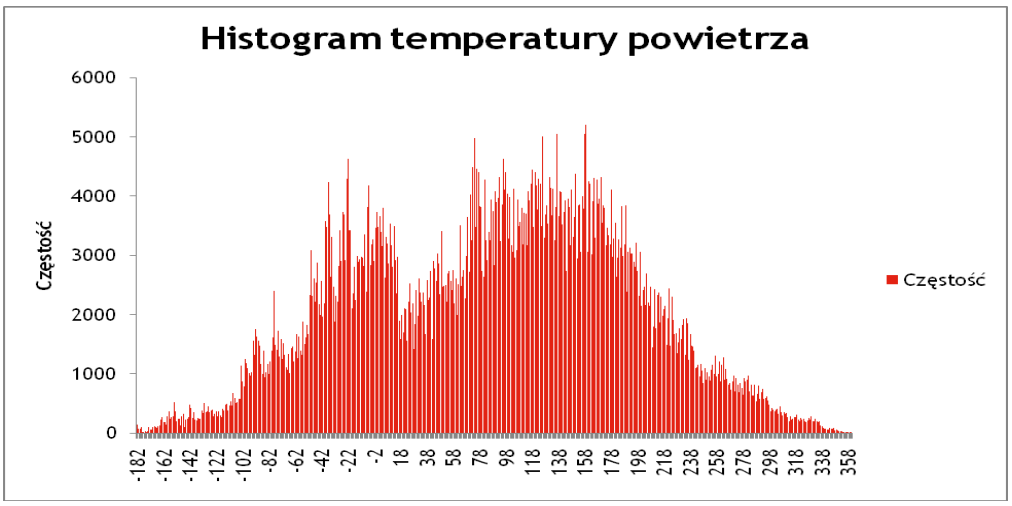


Rys. 14. Rozkład różnic wartości prognozowanych i rzeczywistych prędkości wiatru

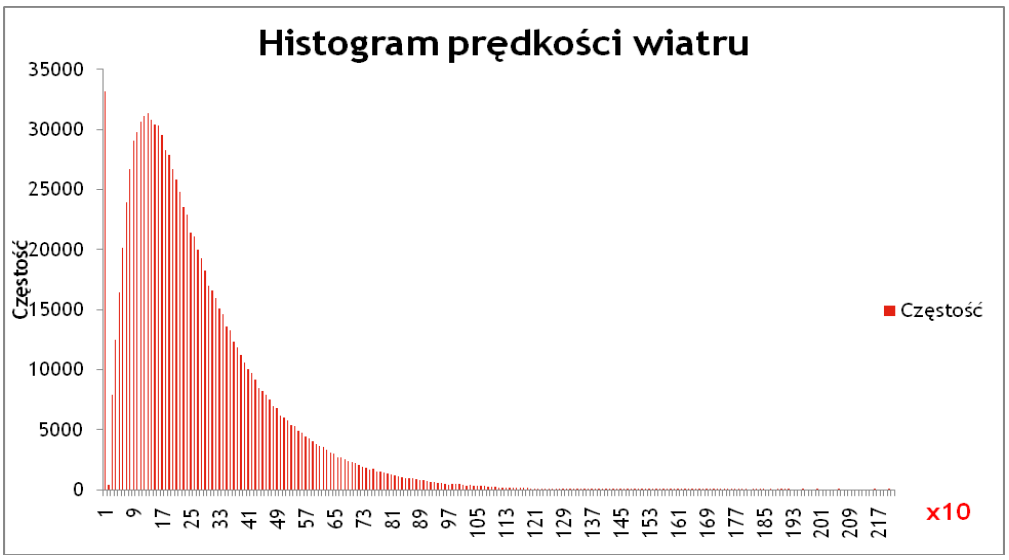
g) Histogramy



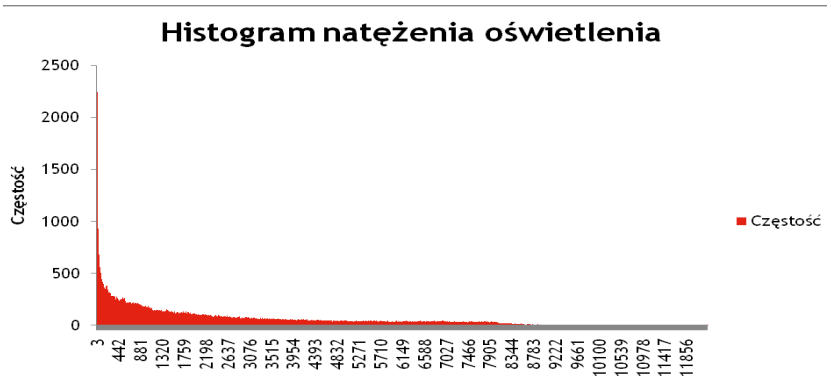
Rys. 15. Histogram wilgotności powietrza



Rys. 16. Histogram temperatury powietrza



Rys.17. Histogram prędkości wiatru



Rys. 18. Histogram natężenia oświetlenia

h) Wykrywanie wyjątków – Highlight exceptions

Procedura wykrywa na podstawie analizy trendów dane niepasujące do reszty. W przykładzie zamieszczonym w artykule została podświetlona liczba 14, ponieważ zaczynając od niej dane zmieniają monotoniczność z rosnącej na malejącą.

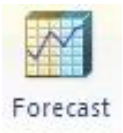
Tablica 5

Przykład wykrywania przypadków wartości, na które należy zwrócić uwagę

Data	Temperatura [°C]
2010-06-08 00:00	9
2010-06-08 01:00	9
2010-06-08 02:00	8
2010-06-08 03:00	7
2010-06-08 04:00	7
2010-06-08 05:00	7
2010-06-08 06:00	7
2010-06-08 07:00	8
2010-06-08 08:00	9
2010-06-08 09:00	10
2010-06-08 10:00	12
2010-06-08 11:00	15
2010-06-08 12:00	16
2010-06-08 13:00	17
2010-06-08 14:00	18
2010-06-08 15:00	14
2010-06-08 16:00	12
2010-06-08 17:00	12
2010-06-08 18:00	11
2010-06-08 19:00	11
2010-06-08 20:00	9
2010-06-08 21:00	9
2010-06-08 22:00	9
2010-06-08 23:00	9

Data	Temperatura [°C]
2010-06-08 00:00	9
2010-06-08 01:00	9
2010-06-08 02:00	8
2010-06-08 03:00	7
2010-06-08 04:00	7
2010-06-08 05:00	7
2010-06-08 06:00	7
2010-06-08 07:00	8
2010-06-08 08:00	9
2010-06-08 09:00	10
2010-06-08 10:00	12
2010-06-08 11:00	15
2010-06-08 12:00	16
2010-06-08 13:00	17
2010-06-08 14:00	18
2010-06-08 15:00	14
2010-06-08 16:00	12
2010-06-08 17:00	12
2010-06-08 18:00	11
2010-06-08 19:00	11
2010-06-08 20:00	9
2010-06-08 21:00	9
2010-06-08 22:00	9
2010-06-08 23:00	9

i) Przewidywanie – Forecast



Funkcja służy do przewidywania danych na podstawie istniejących serii. Na poniższym przykładzie został przedstawiony wynik przewidywania dwudniowej temperatury powietrza na podstawie pomiarów z trzech dni.

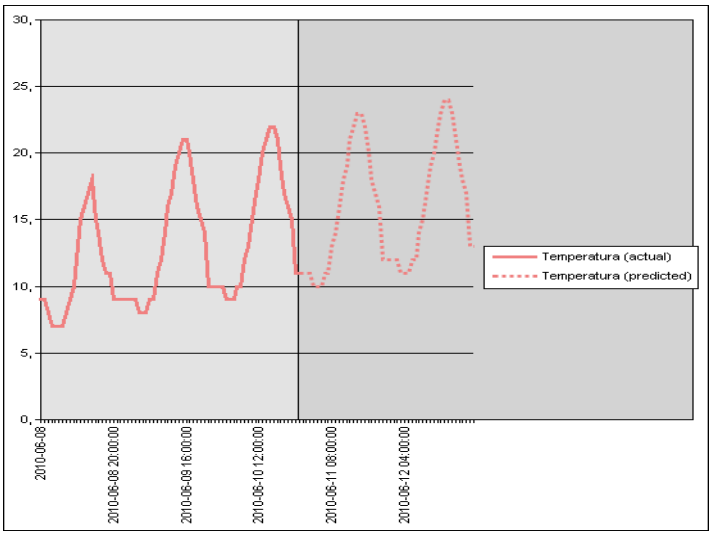
Tablica 6

Przykład przewidywania przebiegów temperatury

Data	Temperatura [°C]
2010-06-08 00:00	9
2010-06-08 01:00	9
2010-06-08 02:00	8
2010-06-08 03:00	7
2010-06-08 04:00	7
2010-06-08 05:00	7
2010-06-08 06:00	7
2010-06-08 07:00	8
2010-06-08 08:00	9
2010-06-08 09:00	10
2010-06-08 10:00	12
2010-06-08 11:00	15
2010-06-08 12:00	16
2010-06-08 13:00	17
2010-06-08 14:00	18
2010-06-08 15:00	15
2010-06-08 16:00	14
2010-06-08 17:00	12
2010-06-08 18:00	11
2010-06-08 19:00	11
2010-06-08 20:00	9
2010-06-08 21:00	9
2010-06-08 22:00	9
2010-06-08 23:00	9

Średnie godzinowe temperatury z trzech dni

2010-06-09 00:00	9		2010-06-10 00:00	10	
2010-06-09 01:00	9		2010-06-10 01:00	10	
2010-06-09 02:00	9		2010-06-10 02:00	10	
2010-06-09 03:00	8		2010-06-10 03:00	9	
2010-06-09 04:00	8		2010-06-10 04:00	9	
2010-06-09 05:00	8		2010-06-10 05:00	9	
2010-06-09 06:00	9		2010-06-10 06:00	10	
2010-06-09 07:00	9		2010-06-10 07:00	10	
2010-06-09 08:00	11		2010-06-10 08:00	12	
2010-06-09 09:00	12		2010-06-10 09:00	13	
2010-06-09 10:00	14		2010-06-10 10:00	15	
2010-06-09 11:00	16		2010-06-10 11:00	17	
2010-06-09 12:00	17		2010-06-10 12:00	18	
2010-06-09 13:00	19		2010-06-10 13:00	20	
2010-06-09 14:00	20		2010-06-10 14:00	21	
2010-06-09 15:00	21		2010-06-10 15:00	22	
2010-06-09 16:00	21		2010-06-10 16:00	22	
2010-06-09 17:00	20		2010-06-10 17:00	21	
2010-06-09 18:00	18		2010-06-10 18:00	19	
2010-06-09 19:00	16		2010-06-10 19:00	17	
2010-06-09 20:00	15		2010-06-10 20:00	16	
2010-06-09 21:00	14		2010-06-10 21:00	15	
2010-06-09 22:00	10		2010-06-10 22:00	11	
2010-06-09 23:00	10		2010-06-10 23:00	11	



Rys. 21. Wykres byłej i przewidywanej temperatury

j) Wykrywanie wzoru – Fill from example

Oprócz przygotowania danych narzędzie to służy także do ich analizy. Pozwala na wyznaczenie odpowiedzi na podstawie wykrytych wzorów i niepełnych danych. W przykładzie wartości znajdujące się w pomarańczowych polach zostały wyznaczone automatycznie na podstawie pozostałych, pełnych wierszy.

Tablica 7

Przykład maszynowego przygotowania odpowiedzi

Data	Temperatura [°C]	Oświetlenie W/m²	Jest ładna pogoda?
2010-06-08 00:00	9	0	NIE
2010-06-08 01:00	9	0	NIE
2010-06-08 02:00	8	0	NIE
2010-06-08 03:00	7	0	NIE
2010-06-08 04:00	7	0	NIE
2010-06-08 05:00	7	5	NIE
2010-06-08 06:00	7	20	NIE
2010-06-08 07:00	8	30	NIE
2010-06-08 08:00	9	70	NIE
2010-06-08 09:00	10	90	NIE
2010-06-08 10:00	12	110	NIE
2010-06-08 11:00	15	130	NIE
2010-06-08 12:00	16	145	TAK
2010-06-08 13:00	17	120	TAK
2010-06-08 14:00	18	115	TAK
2010-06-08 15:00	15	112	TAK
2010-06-08 16:00	14	100	TAK
2010-06-08 17:00	12	90	NIE
2010-06-08 18:00	11	85	NIE
2010-06-08 19:00	11	70	NIE
2010-06-08 20:00	9	63	NIE
2010-06-08 21:00	9	30	NIE

PODSUMOWANIE

W artykule przedstawiono możliwości wykonania analizy danych wybranymi odpowiednimi metodami eksploracji. Zaprezentowano przydatność wymienionej metody do analizy i nakreślono metodę syntetyzacji dużej ilości danych.

Przeprowadzona analiza została wykonana dla danych pomiarowych uzyskanych z całego roku 2010. Dla sprawdzenia trendów i zakresów wartości zostanie powtórzona ona dla danych z roku 2011. Powstaje jednak zawsze pytanie, czy wyniki będą podobne ?

Literatura

1. *Daniel T. Larose*: Odkrywanie wiedzy z danych – Wprowadzenie do eksploracji danych. WNT, Warszawa 2006.
2. *Michael J.A. Berry, Gordon S. Linoff* : Data Mining Techniques. Wiley Publishing, Inc. Indiana 2004.
3. *Gordon S. Linoff*: Data Analysis Using SQL and Excel. Wiley Publishing, Inc. Indiana 2008.

Recenzent: prof. dr hab. inż. Bogdan Miedziński

THE USE OF DATA EXPLORATION METHODS FOR THE ANALYSIS OF ENVIRONMENTAL PARAMETERS

The development of sensor networks for measuring the indicators that bring some hazards for the environment requires to design and use huge data bases. The progress in information technology makes it possible to store big data bases whose capacity sometimes reaches the terabyte level. Nowadays, the challenge is not really to effectively store huge amounts of data but to analyze them. Therefore there is demand for new IT methods and tools supporting the discovery of knowledge in data. The article presents the principles how to use data mining methods, determines the ways to approach the analysis with these methods, and presents major tasks of the data mining analysis. Additionally, there is an example of implementing the analysis of selected environmental parameters.

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ДОБЫЧИ ДАННЫХ ДЛЯ АНАЛИЗА ПАРАМЕТРОВ ОКРУЖАЮЩЕЙ СРЕДЫ

Выполнение сенсорных сетей для измерений факторов, составляющих опасность в среде, требует проектирования и использования больших баз данных. Прогресс в области информатики позволяет накапливать большие фонды данных с объёмом, который иногда превышает порядок терабайтов. В настоящее время вызовом становится не столько эффективное хранение большого количества данных, сколько прежде всего их анализ. В связи с этим возникла потребность новых информатических методов и инструментов, поддерживающих открытие информации в данных. В статье описаны принципы использования методов добычи данных (анг. Data Mining), определены способы подхода к анализу вышеуказанным методом, представлены главные задания анализа добычи данных, а также приведен пример имплементации анализа выбранных параметров среды.