STUDIA INFORMATICA 2008
Volume 29 Number 1 (75)

Marcin GORAWSKI, Jarosław ŻYCIŃSKI
Politechnika Śląska, Instytut Informatyki

STATYSTYCZNA I EKSPLORACYJNA ANALIZA CZASU ŻYCIA SYSTEMÓW POMP GŁĘBINOWYCH ESP

Streszczenie. Systemy elektrycznych pomp głębinowych (ESP) wykorzystują jedną z metod sztucznego podnoszenia ropy naftowej, udoskonalającą proces produkcji w rezerwuarze. W artykule przeprowadzona jest statystyczna i eksploracyjna analiza czasu życia systemów ESP. W tym celu zastosowano estymator Kaplan-Meier oraz różne algorytmy eksploracji danych.

Słowa kluczowe: elektryczne pompy głębinowe (ESP), estymator Kaplan-Meier, algorytmy eksploracji danych, analiza czasu życia

STATISTICAL AND DATA MINING METHODS FOR SURVIVAL ANALYSIS OF ELECTRICAL SUBMERSIBLE PUMP SYSTEMS

Summary. Electrical submersible pump (ESP) systems are one of the more commonly used artificial lift methods that improve oil production from the well. This review of the literature describes survival analysis of ESP systems using statistical and data mining methodologies. Statistical analysis is based on the Kaplan-Meier estimator, while data mining utilizes a few traditional data mining algorithms.

Keywords: electrical submersible pumps (ESP), Kaplan-Meier estimator, data mining algorithms, survival analysis

1. Wstęp

Dokładny czas życia wielu istniejących i używanych produktów czy systemów nie jest znany w chwili ich powstania. Oszacowanie tego czasu jest bardzo często trudnym bądź niewykonalnym zadaniem. Jest to poważny problem, szczególnie wtedy, gdy system powinien pracować ciągle i bez przerwy. Reprezentatywnym przykładem takiego systemu jest szyb naftowy, który jest także motywujący nasze badania opisane w niniejszym artykule.

Ropa naftowa jest jednym z najważniejszych surowców naturalnych, na którą zapotrzebowanie ciągle wzrasta. Dlatego proces wydobycia ropy naftowej jest ciągle udoskonalany. Jednym z systemów stosowanych do kontrolowania szybkości wydobycia, a także zwiększającym ilość wydobytej cieczy z pokładów (rezerwuaru) jest system ESP (ang. *Electrical Submersible Pumps*) – system pomp głębinowych. Pompy pracują w bardzo trudnych warunkach fizycznych, takich jak temperatura czy ciśnienie i są narażone na uszkodzenia. Przerwy w produkcji ropy naftowej powodują ogromne straty dla przedsiębiorstw wydobywczych, co w konsekwencji przekłada się na wzrost jej ceny rynkowej i wpływać może na gospodarki wielu krajów.

Wiele szybów naftowych zlokalizowanych jest w odległych i trudno dostępnych miejscach. Dlatego ciągłe nadzorowanie pracy systemów ESP jest możliwe tylko przez zdalne metody. Warto jednak interweniować w pracę tego systemu, w momencie, gdy istnieje ryzyko jego awarii.

Niniejszy artykuł przedstawia próbę przeprowadzenia efektywnej analizy czasu życia systemów ESP. Wyróżniono dwa podejścia do problemu.

W pierwszej kolejności będzie utworzone narzędzie, które na bazie wykorzystania metody statystycznej pozwoli specjaliście nadzorującemu systemy ESP dokonać analizy czasu życia systemów ESP podobnych do pracującego systemu. Na tej podstawie specjalista oceni, jakie jest statystycznie zagrożenie awarii systemu ESP.

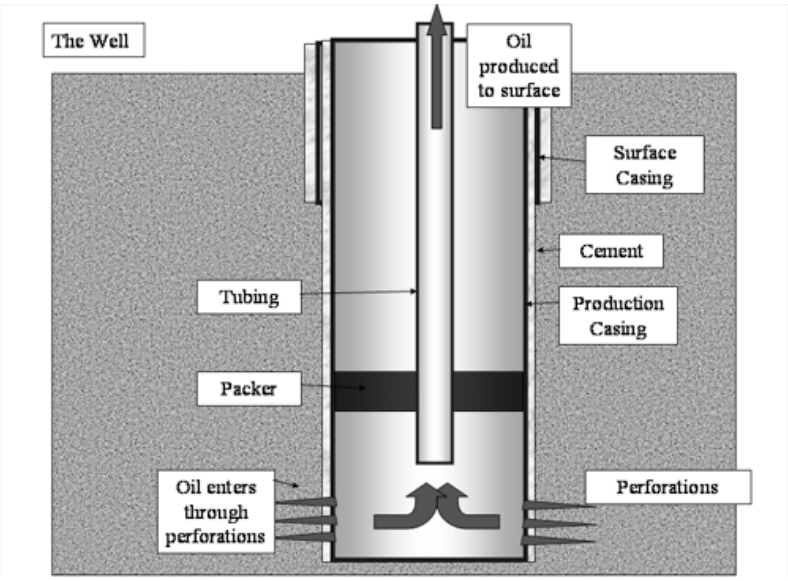
W drugiej części artykułu będzie przedstawione sprawdzenie przydatności wybranych algorytmów eksploracji danych do prognozowania czasu życia systemów ESP. Zostanie również podjęta próba analizy i odpowiedzi na pytanie, jakie czynniki wpływają na czas życia systemu ESP. Taka wiedza pozwala specjaliście skoncentrować się na takim projektowaniu systemu ESP, aby osiągnąć jak największe korzyści. Podczas gdy na właściwości fizyczne lub chemiczne zachodzące w szybie człowiek ma ograniczony wpływ, niektóre cechy, takie jak moc silnika napędzającego pompę czy rodzaj materiałów użytych do budowy różnych elementów, są wybierane przez człowieka. Dlatego metody eksploracyjne mogą znaleźć zastosowanie w przemyśle wydobycia ropy naftowej.

2. Wprowadzenie do problematyki artykułu

Ropa naftowa występuje na całym globie ziemskim. Znajduje się zarówno na lądzie, jak i pod dnem oceanów czy mórz. Także głębokość jej występowania oraz warstwy otaczające ją różnią się. W związku z powyższym techniki wydobycia ropy naftowej są różnorodne.

W celu wydobycia ropy naftowej należy przeprowadzić kilka czynności. Po pierwsze, trzeba zbudować szyb naftowy (rys. 1). Następnym etapem jest uruchomienie produkcji

w szybie naftowym (ang. *completion*). Za pomocą specjalnych „pistoletów" (ang. *perforating guns*) tworzy się małe dziury w rurze okładzinowej (ang. *casing*), zwane perforacjami (ang. *perforations*). Przez nie wypływa ropa dzięki ciśnieniu. Transportowana jest ona w szybie przez odpowiednią rurę, zwaną tubą (ang. *tubing)*. Jednakże ciśnienie, pod jakim wydobywa się płyn, może być zbyt małe lub może zmaleć podczas procesu wydobycia. W takich sytuacjach często stosuje się „sztuczne podnoszenie" (ang. *artificial lift*).



Rys. 1. Schemat budowy szybu naftowego (źródło: Wikipedia)
Fig. 1. Schematic of an oil well (source: Wikipedia)

Każdy proces, który dodaje energię do przepływu wydobywanej ropy naftowej w celu zainicjowania lub poprawienia produkcji, nazywa się sztucznym podnoszeniem.

Sztuczne podnoszenie jest używane wtedy, gdy płyn nie ma wystarczającego ciśnienia, by sam dotarł na powierzchnię ziemi. Jedną ze stosowanych metod jest zastosowanie elektrycznych pomp głębinowych (ang. *Electrical Submersible Pump*).

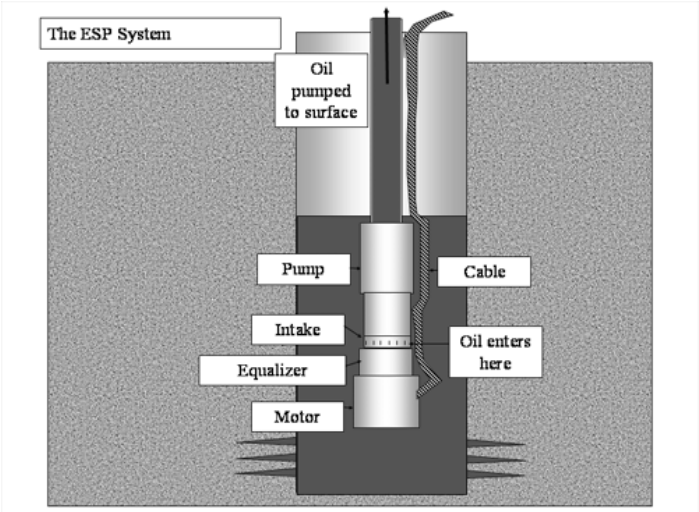
2.1. Elektryczne pompy głębinowe

Elektryczne pompy głębinowe oznacza się w skrócie ESP (rys. 2). Są one najszybciej rozwijającą się metodą sztucznego podnoszenia ropy naftowej. Przez pojęcie systemu ESP rozumie się system złożony z kilku części: pompy odśrodkowej, silnika, kabla z zasilaniem elektrycznym, urządzeń ochraniających oraz sterujących całym systemem.

Zasada działania systemu ESP jest następująca:

- Poprzez kabel doprowadzany jest prąd zasilający silnik.
- Silnik napędza zanurzoną w płynie pompę odśrodkową.
- Pompa dostarcza energię płynowi, który jest wypychany na powierzchnię.

System ESP może pracować w różnych warunkach, na różnych głębokościach i może być użyty do pompowania różnych rodzajów płynów. Dzięki zastosowaniu systemu ESP można kontrolować produkcję ropy naftowej, tj. ilość wydobytej ropy na powierzchnię oraz zwiększać jej wydobycie.



Rys. 2. Elektryczna pompa głębinowa (źródło: Wikipedia)
Fig. 2. Electrical submersible pump system (source: Wikipedia)

2.2. Nadzorowanie systemów ESP

Na całym świecie znajduje się ponad 100 000 systemów ESP [10]. Wiele z nich jest umieszczonych w szybach znajdujących się w trudno dostępnych miejscach na świecie, często w morzu. Dlatego dużo wygodniej jest nadzorować je zdalnie.

Warunki w szybie naftowym zmieniają się i bywają trudne do przewidzenia. Przykładowo, zmienia się ciśnienie i skład płynów, może też nastąpić zmiana temperatury. To wszystko i wiele innych czynników może doprowadzić do uszkodzenia i nieprawidłowego funkcjonowania systemu ESP. Wymiana urządzeń ESP jest kosztowna, w związku z powyższym warto jest zastosować system nadzorujący pracę systemów ESP. Takim systemem jest **espWatcher.**

2.3. System espWatcher

espWatcher jest systemem czasu rzeczywistego wspomagającym korporację *Schlumberger* w świadczeniu usług zdalnego nadzorowania pracy i kontrolowania systemów elektrycznych pomp głębinowych.

W odpowiednich miejscach szybu naftowego rozmieszczone są czujniki mierzące różne właściwości fizyczne. Dane z tych czujników są przesyłane drogą satelitarną do wyznaczonych serwerów, gdzie następnie są zapisywane i analizowane. W przypadku zakłóceń pracy

systemów ESP, odpowiedzialne osoby są alarmowane lub otrzymują ostrzeżenie, na przykład za pośrednictwem telefonu komórkowego.

espWatcher zawiera także informacje o tym, jak zbudowany jest szyb naftowy, dane dotyczące jego części składowych. Przy uwzględnieniu bieżącej sytuacji w szybie, na podstawie odczytów danych z sensorów, próbuje się obliczyć wskaźnik awarii pompy PFI (ang. *pump failure index*). Jest to możliwe tylko wtedy, gdy dysponuje się odpowiednią ilością danych. Tak więc tylko dla niektórych ESP można odczytać PFI.

3. Analiza statystyczna

W celu przeprowadzenia statystycznej analizy czasu życia systemów ESP została napisana w języku C# aplikacja o nazwie ESAP (ang. *ESP Survival Analysis Package*). Aplikacja ESAP używa ona relacyjnej bazy danych *firebird* jako repozytorium danych oraz statystycznej maszyny R w celu użycia nieparametrycznego estymatora Kaplan-Meier.

3.1. Nieparametryczny estymator Kaplan-Meier

W roku 1958 *Kaplan* i *Meier* zdefiniowali funkcję przeżycia, która bierze pod uwagę dane o czasie życia, które w pewnych momentach czasu przestają być analizowane, są „ocenzurowane" (ang. *censored*) [1]. Empiryczna funkcja przeżycia nie uwzględnia takich danych.

Inną spotykaną nazwą tego estymatora jest estymator limitu produktu (ang. *product limit estimator*).

Definicja 1

Niech: $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ oznacza posortowane dyskretne momenty czasu, w których następuje śmierć produktu. Nie bierze się tutaj pod uwagę momentów, w których przestało się analizować (ocenzurowało się) dany produkt. Dalej, niech n_j oznacza liczbę produktów żyjących (zagrożonych śmiercią, ang. *at risk*) bezpośrednio przed czasem $t_{(j)}$, włączając również te, które zginą w momencie czasu $t_{(j)}$, a d_j oznacza liczbę produktów umierających w czasie $t_{(j)}$. Jeżeli produkt jest ocenzurowany w czasie $t_{(j)}$ i w tym samym czasie następuje śmierć przynajmniej jednego z produktów, to zakłada się, że cenzurowanie następuje bezpośrednio po czasie $t_{(j)}$ i ocenzurowany produkt wlicza się do n_j. Nieparametryczny estymator Kaplan-Meier funkcji przeżycia definiujemy jako:

$$\hat{S}(t) = \prod_{j=1}^{i} \frac{n_j - d_j}{n_j} \qquad i = 1, \ldots, m. \tag{1}$$

Równanie (1) ma interpretację następującą:

Aby produkt przetrwał do czasu t, musi najpierw przeżyć do czasu $t_{(1)}$. Następnie musi przeżyć od $t_{(1)}$ do $t_{(2)}$ itd. Prawdopodobieństwo warunkowe śmierci w czasie $t_{(j)}$, gdy produkt był żywy bezpośrednio przed $t_{(j)}$, jest szacowane jako d_j/n_j. Tak więc prawdopodobieństwo warunkowe przeżycia produktu wynosi $1-d_j/n_j$, co po przekształceniu daje $(n_j-d_j)/n_j$. Całkowite prawdopodobieństwo bezwarunkowe otrzymuje się mnożąc wszystkie prawdopodobieństwa warunkowe w stosownych momentach czasu aż do momentu czasu t.

Przykład

Mamy następującą listę produktów. Przyjmijmy następujące oznaczenie: niech 1 w kolumnie „Śmierć" oznacza śmierć produktu, a 0 oznacza ocenzurowanie produktu. W kolumnie „Czas" jest podane, kiedy nastąpiła śmierć lub ocenzurowanie produktu.

Tabela 1
Dane przykładowe dla estymatora Kaplan-Meier

Nr produktu	Czas	Śmierć
1	2	1
2	3	0
3	5	1
4	5	1
5	6	0
6	6	0
7	7	1
8	10	1

Rozwiązanie:

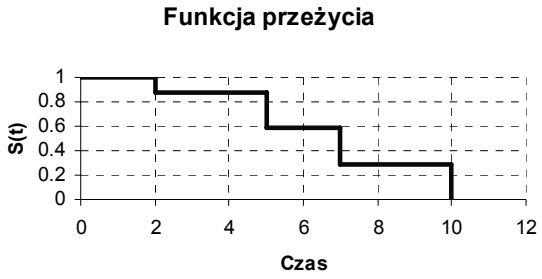
Tabela 2
Rozwiązanie przykładu estymatora Kaplan-Meier

j	$t_{(j)}$	n_j	d_j	$(n_j-d_j)/n_j$	$\hat{S}(t)$
0	0	8	0	1	1
1	2	8	1	0.875	0.875
2	5	6	2	0.666667	0.583333
3	7	2	1	0.5	0.291667
4	10	1	1	0	0

Śmierć pierwszego produktu następuje w $t_{(1)}=2$. W tym momencie czasu wszystkie produkty żyją. Dlatego $n_1=8$. Jeden produkt umiera, więc $d_1=1$ i $\hat{S}(t)=0.875$.

W momencie $t_{(2)}=5$, wartość $n_2=6$, gdyż umarł 1 produkt i 1 produkt został cenzurowany, a $\hat{S}(t)=0.875\cdot 0.666667=0.583333$. Obliczenia kontynuuje się dla kolejnych momentów czasu.

Na podstawie tych wyników można wykreślić schodkową funkcję przeżycia (rys. 3).



Rys. 3. Wykres funkcji przeżycia dla estymatora Kaplan-Meier
Fig. 3. Survival curve for Kaplan-Meier estimator

3.2. Dane

Dane używane do obliczeń w niniejszej pracy zostały zgromadzone przez korporację *Schlumberger*. Są to dane rzeczywiste.

Aplikacja do analizy statystycznej została rozwinięta w korporacji *Schlumberger*. Natomiast badania za pomocą algorytmów eksploracyjnych odbyły się poza przedsiębiorstwem. Korporacja *Schlumberger* udostępniła więc dane do dalszych celów badawczych, niekomercyjnych.

Dane reprezentują czas życia systemów ESP i ich charakterystykę. W jednym szybie naftowym może być zainstalowanych wiele systemów ESP. Grupa szybów naftowych jest rozumiana jako pole naftowe, które jest w posiadaniu pewnej firmy. Jedna firma może być właścicielem wielu pól naftowych.

Dane opisują dokładnie 23429 systemów ESP. Istnieją dwa rodzaje danych: nieocenzurowane i ocenzurowane. W pierwszym przypadku dla każdego systemu ESP znany jest czas poprawnego funkcjonowania tego systemu (aż do chwili istotnej usterki – przerwy pracy tego systemu). W drugim przypadku znany jest czas bezawaryjnego funkcjonowania systemu ESP do pewnego momentu czasu, w którym informacja o tym systemie została „ocenzurowana", czyli nie są znane dalsze losy tego systemu. Oba wspomniane czasy występują jako argument o nazwie *Instdays*. To, czy dane są ocenzurowane czy nie definiuje flaga *Censored*, wartość 'T' oznacza ocenzurowane dane, z kolei wartość 'F' oznacza dane nieocenzurowane. Liczba rekordów nieocenzurowanych wynosi 6452, pozostałe rekordy są ocenzurowane.

Dla każdego systemu ESP zdefiniowane są następujące pola:

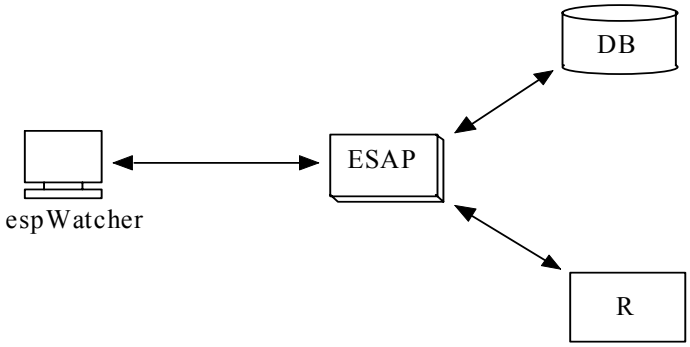
- Country – kraj, w którym znajduje się pole naftowe.
- Company – nazwa firmy będącej właścicielem pola naftowego.
- Field – nazwa pola naftowego.
- Well – nazwa szybu naftowego.
- Instyear – rok instalacji systemu ESP w szybie naftowym.

- Casing – ocementowana od zewnątrz rura o dużej średnicy, umieszczona w szybie naftowym, stanowiąca warstwę ochronną. W związku z faktem, iż wartość tego argumentu zapisywana jest w różny sposób (w różnym formacie) w danych, powstała kolumna Casing Double oznaczająca średnicę rury w calach.
- Applicn – zastosowanie. Wyróżnia się następujące wartości:
  - *Brine* – woda o bardzo dużym zasoleniu,
  - *CO2* – dwutlenek węgla,
  - *Horizontal* – pompowanie do poziomej rury,
  - *Mining* – płyn zawiera minerały, fragmenty skał,
  - *Off Shore* – zainstalowane poza terenem lądu,
  - *Oil* – ropa naftowa,
  - *Water flood* – woda.
- Deviated – wartość ‘T’ oznacza, że szyb jest zakrzywiony, wartość ‘F’ oznacza, że szyb jest pionowy.
- Openhole – wartość ‘T’ oznacza, że istnieją miejsca w rurze okładzinowej, które nie są pokryte cementem (chronione warstwą cementową), wartość ‘F’ oznacza, że takie miejsca nie istnieją.
- Pulltype – rodzaj sprzętu używanego do wyciągania pompy.
- Tblength – długość rury podana w stopach, przez którą transportowany jest płyn na powierzchnię.
- Corrosion – korozja, rdza – może występować w różnych metalowych. elementach zainstalowanych w szybie naftowym. Korozja występuje w systemie przy ustawionej wartości ‘T’.
- Scale – osady i zacieki pojawiające się na różnych metalowych elementach. Osady i zacieki występują w systemie przy ustawionej wartości ‘T’.
- Abrasion – abrazja – jest jednym z czynników erozyjnych, oznacza ścieranie się skały. Abrazja występuje w systemie przy ustawionej wartości ‘T’.
- Motthp – sumaryczna moc silników napędzających pompy, jest wyrażona w koniach mechanicznych.
- Inne pola, które nie zostały wzięte pod uwagę w rozważaniach.

3.3. Architektura aplikacji ESAP

Po pierwsze, przy implementacji ESAP posłużono się wzorcem projektowym model-widok-kontroler MVC (ang. *Model-View-Controller*). Cechą tego wzorca jest odseparowanie warstwy prezentacji, logiki biznesowej i danych. Wpływa to na przejrzystość kodu źródłowego aplikacji i ułatwia jej utrzymywanie.

Po drugie, architektura aplikacji może być rozumiana jako architektura zorientowana na usługi SOA (ang. *Service Oriented Architecture*). ESAP korzysta przede wszystkim z usług espWatchera. Jako usługę można również rozumieć maszynę R i relacyjną bazę danych *firebird*, która jest „dostawcą danych" (rys. 4). Zaletą SOA jest duża atomowość. Serwisy mogą być używane niezależnie przez różne aplikacje. Cechą tych serwisów jest także to, że nie komunikują się między sobą.



Rys. 4. Architektura ESAP
Fig. 4. ESAP architecture

3.3.1. Baza danych

ESAP komunikuje z relacyjną bazą danych *firebird* w wersji *embedded superserver*. Baza danych jest zapisana w postaci jednego pliku na dysku, co umożliwia bardzo proste przenoszenie jej na inne komputery. Tak więc, jeżeli użytkownik zdecyduje się zainstalować ESAP na innym komputerze, to w bardzo prosty sposób może przywrócić stan aplikacji z poprzedniego komputera – wystarczy, że podmieni plik bazy danych.

Można wyszczególnić dwa rodzaje danych:

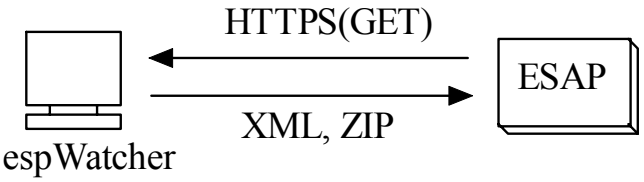
- dane historyczne,
- dane użytkownika.

Dane historyczne opisują systemy ESP, które przestały działać, albo dane o nich zostały ocenzurowane. Natomiast pozostałe dane związane z pracą aplikacji to dane użytkownika.

3.3.2. Komunikacja z espWatcherem

espWatcher przechowuje między innymi dane o szybach naftowych. Klienci korzystający z usług espWatchera mogą te dane obejrzeć na stronie internetowej. espWatcher zawiera najbardziej aktualne dane, dotyczące szybów naftowych i systemów ESP. Dane te są importowane do aplikacji ESAP i zostają zapisane lokalnie w bazie danych. Użytkownik ma więc możliwość wglądu do tych danych znacznie szybciej, gdy są one składowane lokalnie. Ponadto te dane są używane podczas pracy aplikacji ESAP.

espWatcher udostępnia dane za pomocą serwletów. ESAP wywołuje te serwlety przez bezpieczne połączenie na protokole HTTPS (rys. 5). W związku z tym akceptowany jest odpowiedni certyfikat. Wtedy następuje odpowiedź z espWatchera.



Rys. 5. Komunikacja ESAP z espWatcherem
Fig. 5. Communication between ESAP and espWatcher

3.3.3. Środowisko R

Środowisko R [9] pozwala na uruchamianie skryptów zapisanych w pliku. W korporacji *Schlumberger* został napisany skrypt czytający z odpowiednio sformatowanego pliku źródłowego listę wierszy opisujących systemy ESP, a następnie wywołujący estymator Kaplan-Meier. W wyniku wywołania skryptu zostaje wygenerowany plik wynikowy zawierający listę punktów, będących wynikiem wywołania estymatora. Także podane są punkty definiujące przedziały ufności.

ESAP przygotowuje listę systemów ESP, dla których ma być wywołany estymator. Ta lista zapisywana jest do pliku. Następnie tworzy się osobny wątek, w którym uruchamiany jest skrypt R. Po zakończeniu wykonywania skryptu ESAP czyta wygenerowane pliki z punktami i na ich podstawie rysuje wykres przeżycia.

3.3.4. Praca z aplikacją

ESAP wymaga komunikacji z espWatcherem. Przy pierwszym uruchomieniu aplikacji należy zaimportować dane o szybach naftowych wraz z systemami pomp ESP oraz zaktualizować dane o stanie szybu i alarmach szybu. Wszystkie te dane zostają zapisane lokalnie w bazie danych, a częstotliwość ich odświeżania zależy od użytkownika.

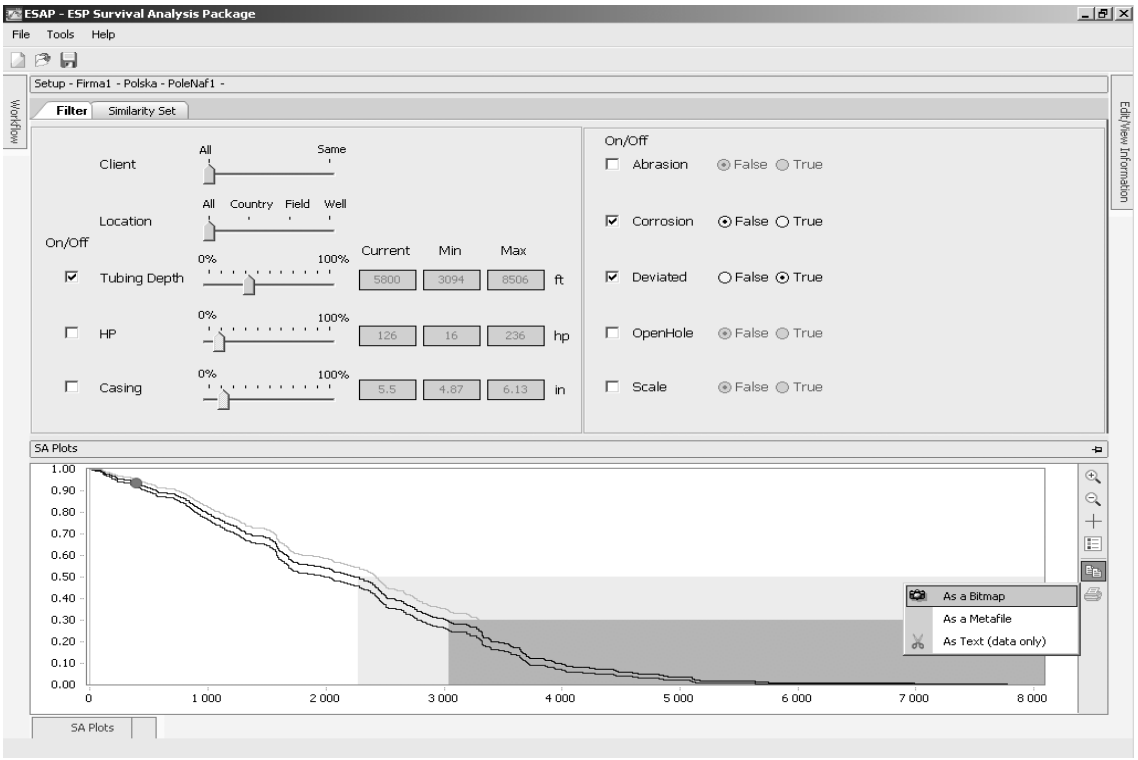
Następnie należy dla każdego (dla którego będzie przeprowadzana analiza) systemu ESP zdefiniować zestaw charakterystycznych cech. Na podstawie zestawu ważnych cech systemu ESP zostanie utworzony zbiór podobnych systemów ESP, dla których znany jest czas życia. Ten zbiór jest używany jako dane dla estymatora Kaplan-Meier. Na podstawie wyników estymacji wykreślona jest krzywa przeżycia dla historycznych systemów ESP.

3.4. Przykładowa analiza systemu ESP

Przedmiotem analizy jest system ESP zamontowany w zakrzywionym szybie o średnicy 5.5 cala. System ESP pracuje na głębokości 5800 stóp. Sumaryczna moc silników wynosi

126 koni mechanicznych. Aktualnie system pracuje poprawnie 389 dni (nie ma problemu występowania zjawiska korozji).

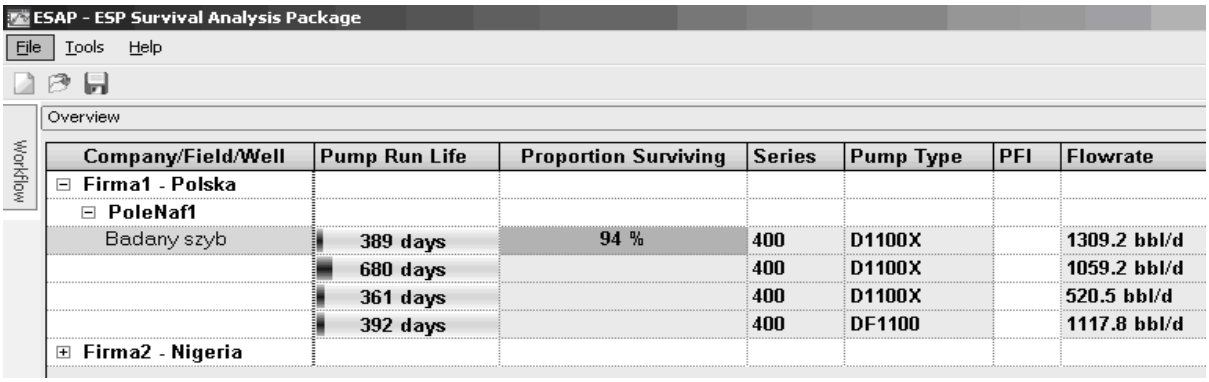
Przyjmuje się, że istotnym kryterium jest brak zakrzywienia szybu, brak korozji i głębokość, na jakiej pracuje system ESP. Dla głębokości przyjmowany jest przedział 3094-8506 stóp. W danych historycznych ESAP ma do dyspozycji 1237 podobnych systemów ESP do badanego. Po uruchomieniu analizy otrzymuje się następujący wykres (rys. 6).



Rys. 6. Ekran Setup z wykresem
Fig. 6. Setup screen with a plot

Oś OX reprezentuje dni, a oś OY przedstawia, jaki procent podobnych systemów ESP do badanego systemu nie miał awarii. Oprócz wykresu głównego są dwa wykresy stanowiące przedziały ufności górny i dolny. Punktem zaznaczono na wykresie badany system ESP. Jest on zaznaczony w takim miejscu, że widać, iż ponad 90% podobnych systemów ESP wciąż pracowało bezawaryjnie.

Informacje podsumowujące są zebrane na głównym ekranie aplikacji (rys. 7).



Rys. 7. Ekran Overview podczas analizy
Fig. 7. Overview screen during analysis

Na ekranie (rys. 7) została umieszczona informacja, że dokładnie 94% podobnych systemów ESP pracowało bezawaryjnie przez 389 dni. W związku z tym możliwość wystąpienia awarii badanego systemu jest niewielka.

Następny etap stanowi porównanie otrzymanego wyniku ze wskazaniem PFI. Jednakże dla badanego szybu ten wskaźnik nie jest wyznaczany.

4. Analiza eksploracyjna danych systemów ESP środowisku *MS SQL Server 2005*

Do eksploracji danych (ang. *data mining*) potrzebne są dane i metody eksploracji udostępnione przez środowisko, np. *MS SQL Server 2005* [6].

Dostępne dane zostały przygotowane w celu utworzenia bazy danych. W tym przypadku należało odrzucić dane, które zostały ocenzurowane, ponieważ żaden z eksploracyjnych algorytmów w *MS SQL Server* nie ma wbudowanego mechanizmu, który analizuje takie dane. Natomiast pozostałe dane są danymi rzeczywistymi, zatem nie wymagają dalszej obróbki.

Do eksploracji zostały wybrane następujące kolumny: *Abrasion, Casing Double, Corrosion, Deviated, Instdays, Motthp, Openhole, Scale, Tblength.* Wszystkie wybrane kolumny są używane w aplikacji ESAP. Ograniczenie liczby analizowanych kolumn pozwala na dokładniejszą ich analizę.

W celu zapewnienia możliwości oceniania skuteczności algorytmów eksploracyjnych dane zostały podzielone losowo na dwie części: dane testowe (1337 rekordów) i dane treningowe (5115 rekordów).

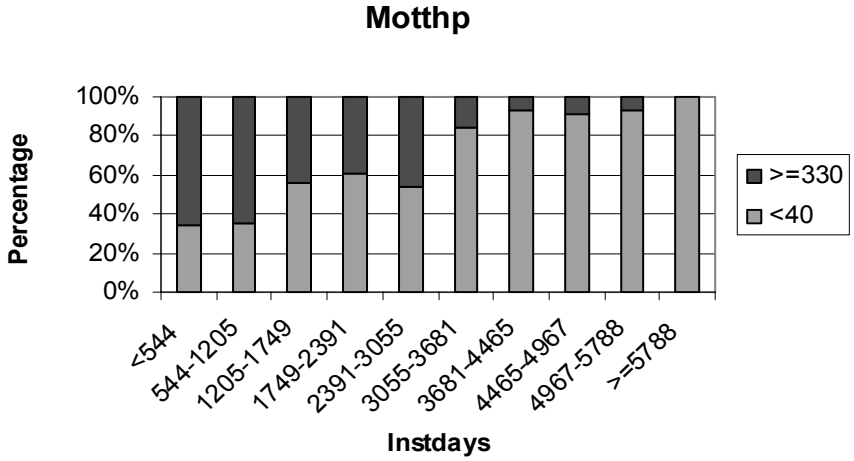
4.1. Naiwny algorytm bayesowski

Naiwny algorytm bayesowski jest algorytmem klasyfikacji używanym do predykcyjnego modelowania. Algorytm oblicza prawdopodobieństwo warunkowe pomiędzy kolumnami wejściowymi i przewidywanymi. Algorytm nazywany jest naiwnym ze względu na fakt, iż nie bierze pod uwagę zależności, jakie mogą występować między kolumnami wejściowymi. Tak więc każda kolumna jest traktowana w obliczeniach niezależnie [8].

Przy użyciu opisanego algorytmu wygenerowano różne modele i wybrano spośród nich taki, który został najwyżej oceniony przez mechanizm oceniający wbudowany w SQL Server 2005. Dane dla tego modelu przedstawiają się następująco. Kolumna predykcyjna *Instdays* została podzielona na 10 dyskretnych przedziałów metodą klastrową, a kolumny *Motthp, Tblength* metodą automatyczną także na 10 przedziałów. Pozostałe kolumny (*Abrasion, Casing Double, Corrosion, Deviated, Openhole, Scale*) mają dyskretne wartości. Dla tego modelu wykonano analizę deskrypcyjną i predykcyjną.

4.1.1. Deskrypcja

Naiwny algorytm bayesowski przez obliczenie prawdopodobieństw warunkowych pokazuje, jak kształtują się dane. Dla każdego z przedziałów kolumny predykcyjnej można odczytać, jaki jest udział przy danej wartości dla każdej kolumny. Niżej dokonana jest deskrypcja dla kolumny *motthp.*



Rys. 8. Wykres prawdopodobieństwa mocy silników pomp >=330 i <40
Fig. 8. Plot – probability for motors with horse power >=330 and <40

W badanym modelu moc silnika systemu ESP (*motthp*) została zdyskretyzowana na 10 przedziałów. W celu uproszczenia analizy wzięto pod uwagę skrajne przedziały mocy silnika. Wartości te zostały przeskalowane, aby obie wartości dawały w sumie 100% (rys. 8).

Wyraźnie widać, że silniki o małej mocy (<40) częściej występują w dłużej żyjących systemach ESP niż silniki mocne (>=330). Jednak silniki mocne występują częściej w przypadku systemów ESP „żyjących krótko”.

Inną cenną informacją, jaką można otrzymać z deskrypcji, jest klasyfikacja, które kolumny wejściowe mają największy wpływ na wartość kolumny predykcyjnej. W badanym modelu przedstawia się to następująco (zaczynając od najważniejszej): *Casing Double, Abrasion, Motthp, Corrosion.*

4.1.2. Predykcja

Wygenerowany model może być użyty do przewidzenia szukanej wartości. W tym celu zostały użyte wszystkie dostępne dane, tzn. zarówno treningowe, jak i testowe. Dla każdego z rekordów jest wykonywana predykcja. Dodatkowo, można podejrzeć rzeczywiste dane czasu życia systemu ESP i porównać z przewidzianymi wartościami.

W celu oceny predykcji napisano zapytanie w języku DMX [3, 4] dla przedziału 1749-2391 kolumny *instdays.*

```
SELECT
  [B1].[Instdays] as Predicted_days,
  t.[instdays] as Actual_days
From
  [B1]
PREDICTION JOIN
  OPENQUERY([ESP],
    'SELECT
      [instdays],
      [instyear],
      [deviated],
      [openhole],
      [tblength],
      [corrosion],
      [scale],
      [abrasion],
      [motthp],
      [casing_double]
    FROM
      [dbo].[ESP_DATA]
    ') AS t
ON
  [B1].[Instdays] = t.[instdays] AND
  [B1].[Deviated] = t.[deviated] AND
  [B1].[Openhole] = t.[openhole] AND
  [B1].[Tblength] = t.[tblength] AND
  [B1].[Corrosion] = t.[corrosion] AND
  [B1].[Scale] = t.[scale] AND
  [B1].[Abrasion] = t.[abrasion] AND
  [B1].[Motthp] = t.[motthp] AND
  [B1].[Casing Double] = t.[casing_double]
WHERE
    [B1].[Instdays]>=1749
    AND
    [B1].[Instdays]<=2391
ORDER BY
t.[instdays]
```

W wyniku wykonania zapytania otrzymano listę rekordów, dla których wartość *instdays* została przewidziana w przedziale 1749-2391 oraz rzeczywistą wartość *instdays* dla tych

rekordów. Na podstawie otrzymanych wyników utworzono wykres pokazujący rzeczywiste wartości i przedziały wartości *instdays*, wynikające z predykcji (rys. 9).

Rys. 9. Wykres rzeczywistych wartości i przedziałów wartości instdays (dla instdays z przedziału 1749-2391)
Fig. 9. Plot – real values and range values of instdays (for instdays in range 1749-2391)

Dla przedziału 1749-2391 zostało sklasyfikowanych poprawnie 34,3% rekordów. Wynik ten nie jest satysfakcjonujący.

4.2. Algorytm drzew decyzyjnych

Drzewa decyzyjne są bardzo popularną metodą eksploracji danych. W celu ich budowy można wykorzystać wiele algorytmów szeroko opisywanych w literaturze stosowanej statystyki i maszyn uczących się. Jednym z dobrze znanych algorytmów jest algorytm *Quinlan'a ID3* oparty na podziałach zależnych od jednej zmiennej losowej. Rozszerzeniem tego algorytmu jest algorytm *C4.5*. Innym stosowanym algorytmem jest algorytm *CART* (ang. *Classification and Regression Tree*). Algorytm drzew decyzyjnych zaimplementowany w *MS SQL Server 2005* jest hybrydą wymienionych algorytmów [2, 5, 7].

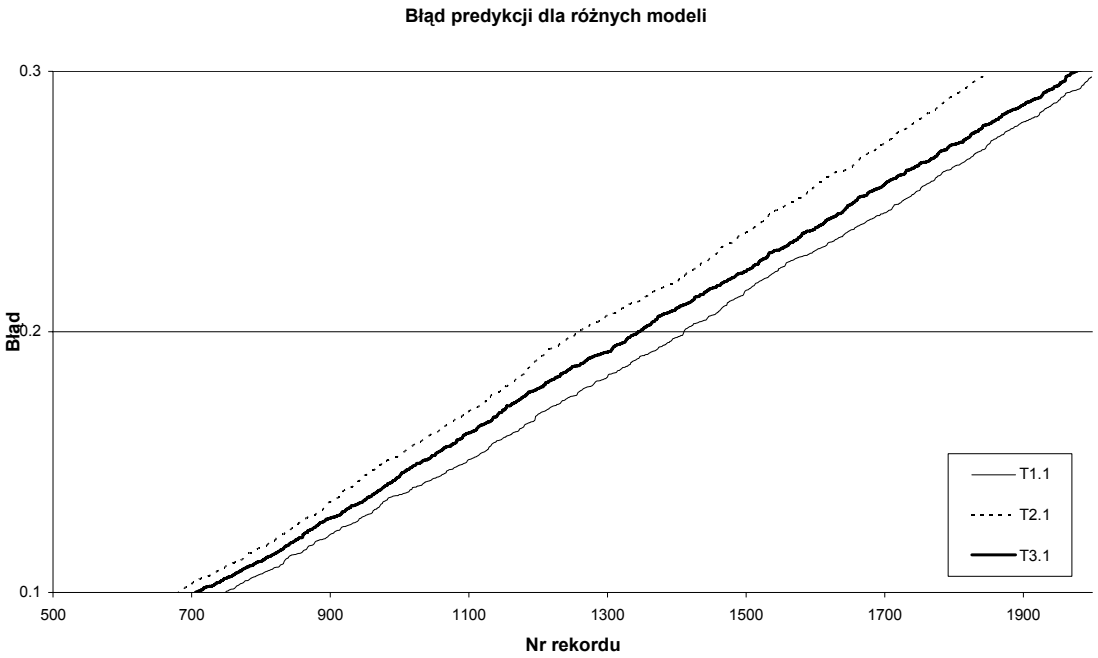
Utworzono kilka różnych modeli, każdy z nich traktuje kolumnę predykcyjną *Instdays* jako kolumnę ciągłą. Model T1.1 używa następujących kolumn wejściowych *Motthp*, *Tblength* (ciągłe) oraz *Abrasion*, *Corrosion*, *Deviated*, *Openhole*, *Scale* (dyskretne). Model T2.1 jest taki sam jakiego użyto przy analizie algorytmem naiwnym bayesowskim. Z kolei model T3.1 w odróżnieniu od modelu T1.1 traktuje kolumnę *Casing Double* jako ciągłą.

4.2.1. Predykcja modeli

Dla każdego z modeli zostało wykonane zapytanie predykcyjne w języku DMX na wszystkich dostępnych rekordach (czyli zarówno danych testowych, jak i treningowych). Następnie dla każdego wiersza odpowiedzi został wyznaczony moduł różnicy przewidzianej i rzeczywistej wartości *Instdays*, czyli o ile dni algorytm przewidujący pomylił się. Wartość tę porównano z rzeczywistą wartością *Instdays*, w wyniku czego wyznaczono, o ile procent

różni się rzeczywista wartość *Instdays* od wartości przewidzianej. Wyniki zostały posortowane. Na tej podstawie można stwierdzić, dla jakiej części danych wejściowych, jaki był błąd predykcji. Można przyjąć, iż akceptuje się pewien błąd, jest on traktowany jako tolerancja błędu.

Rysunek 10 przedstawia fragment wykresu obrazującego zmianę błędu predykcji w funkcji liczby rekordów wejściowych. Wyraźnie widać, że największy błąd występuje dla modelu T2.1, a najmniejszy dla T1.1.



Rys. 10. Błąd predykcji dla różnych modeli drzew decyzyjnych
Fig. 10. Prediction error for different decision tree models

Dla najlepszego modelu (T1.1) predykcja jest poprawna jedynie dla 21.9% spośród rekordów, jeśli przyjmie się 20% tolerancję błędu. Natomiast akceptując 50% błędów algorytm przewidział poprawną wartość dla 49.6% rekordów. Te wartości nie są satysfakcjonujące.

4.2.2. Deskrypcja

Deskrypcję przeprowadzono dla modelu, który cechuje najlepsza predykcja (T1.1). Dla niego algorytm wygenerował drzewo decyzyjne, w którym pierwsze rozgałęzienie występuje dla *Casing Double* równego i różnego od 8.625.

Ponadto, za najważniejsze zostały uznane kolumny wejściowe (począwszy od najważniejszej): *Casing Double*, *Motthp*, *Deviated*, *Tblength*, *Abrasion*, *Scale*, *Corrosion*, *Openhole*.

4.3. Algorytm klastrowania

Algorytm klastrowania jest algorytmem segmentacyjnym, polegającym na grupowaniu obiektów. Grupy obiektów (klastry) zawierają przypadki o podobnej charakterystyce.

Wyróżnia się dwa rodzaje algorytmów klastrujących – twarde i miękkie. Przykładem twardego algorytmu jest algorytm k-średnich (ang. *k-means*). Wybiera się środki klastrów i grupuje przypadki na podstawie odległości od tych środków. W tego rodzaju klastrowaniu jeden przypadek należy do jednego klastra, stąd nazwa klastrowanie twarde. Natomiast w klastrowaniu miękkim klastry zachodzą na siebie. Przykładem takiego algorytmu jest EM (ang. *Expectation Maximization*). Ten algorytm używa probabilistycznej metody do obliczenia prawdopodobieństwa, że dany przypadek należy do klastra. Dalsza analiza jest przeprowadzona przy zastosowaniu algorytmu skalowalnego EM [7, 8].

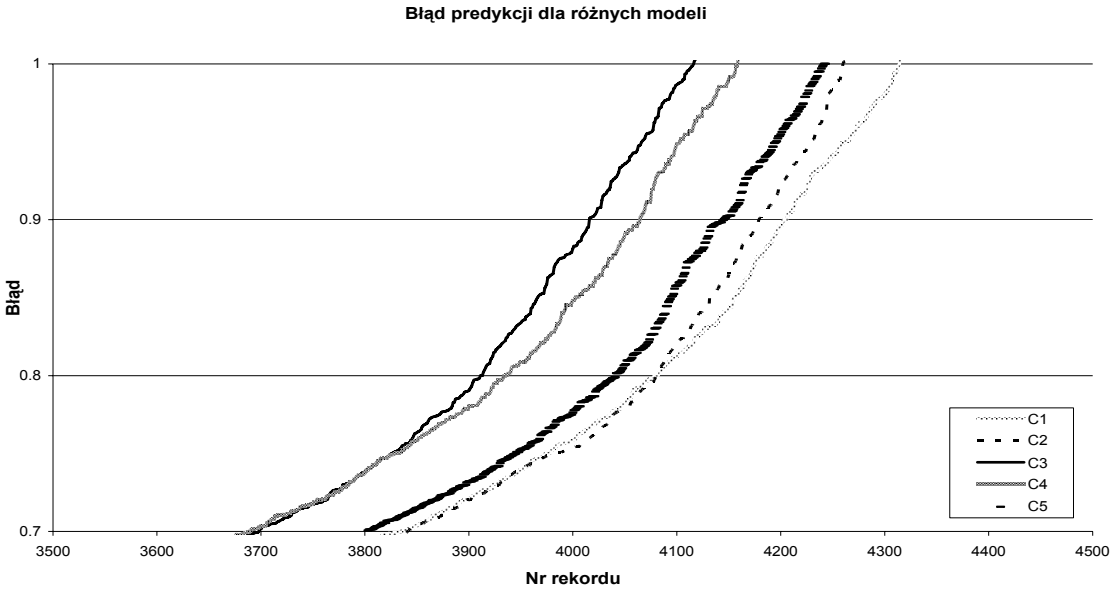
Wygenerowano kilka różnych modeli. We wszystkich kolumną predykcyjną jest ciągła kolumna *Instdays*. Kolumny wejściowe *Motthp* i *Tblength* są także traktowane jako ciągłe, natomiast pozostałe kolumny jako dyskretne.

W pierwszej kolejności został wygenerowany model C1, używający wszystkich kolumn wejściowych. W kolejnym modelu C2 zostały wyłączone wszystkie ciągłe kolumny wejściowe (niepredykcyjne), to znaczy *Motthp* i *Tblength*. Natomiast w modelu C3 na odwrót, wszystkie kolumny dyskretne zostały wyłączone. Model C4 powstał z modelu C2, z wyłączeniem kolumny *Casing Double*. Ta kolumna przyjmuje wiele wartości dyskretnych. Dlatego warto „odseparować" ją od pozostałych dyskretnych (a dwustanowych kolumn). Z kolei model C5 jest „komplementarny" do modelu C4, czyli opiera się na kolumnach ciągłych i na kolumnie *Casing Double*.

4.3.1. Predykcja modeli

Podobnie jak w przypadku drzew decyzyjnych zostało wykonane zapytanie w języku DMX dla każdego z modeli (oparte na danych testowych i uczących), a następnie zostały wykonane obliczenia błędu predykcji.

Rysysunek 11 przedstawia fragment wykresu, obrazującego zmianę błędu predykcji w funkcji liczby rekordów wejściowych.



Rys. 11. Błąd predykcji dla rożnych modeli algorytmu klastrowego
Fig. 11. Prediction error for different clustering algorithm models

Różnica błędów dla różnych modeli rośnie w funkcji liczby rekordów wejściowych.

Dla „małej liczby rekordów" (np. 10%) największy błąd generują modele (w kolejności malejącej): C3, C2/C4, C1, C5. Dla „średniej liczby rekordów" (np. 50%) modele: C3, C4, C2, C1/C5. Natomiast dla „dużej liczby rekordów" (np. 75%): C3, C4, C5, C2, C1.

Najmniejszy błąd w całym spektrum dostępnych rekordów jest generowany przez model C5. Z kolei, największy błąd generują modele C3 i C4.

Tabela 3

Procent poprawnie sklasyfikowanych rekordów dla tolerancji błędu 20% i 50% dla wszystkich modeli algorytmu klastrowego

Tolerancja\Model	C1	C2	C3	C4	C5
20.0%	17.7%	17.4%	17.0%	17.4%	18.1%
50.0%	44.4%	43.9%	41.7%	42.3%	43.9%

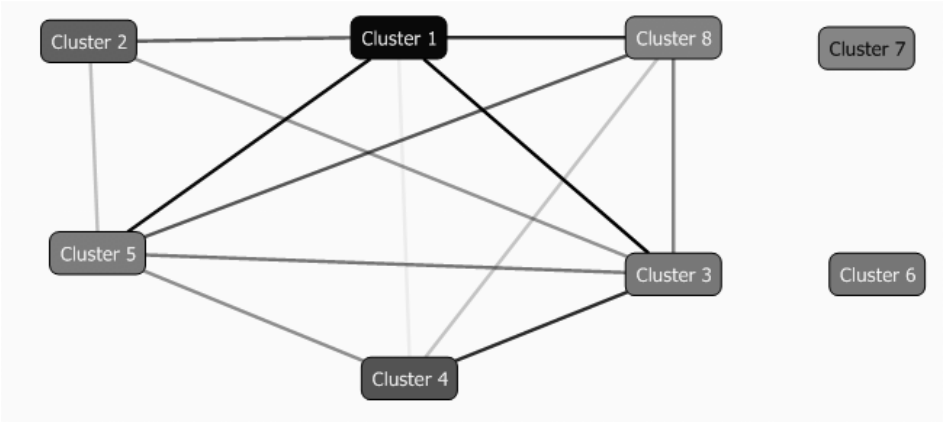
Są to wyniki gorsze niż te, które zostały otrzymane w modelu drzew decyzyjnych.

4.3.2. Deskrypcja

Przy predykcji ustalono, że jednym z najlepiej klasyfikującymi modelami klastrowymi są modele C1 i C5. Do deskrypcji użyto modelu C1 ze względu na fakt, iż korzysta on z większej liczby kolumn wejściowych niż model C5. Dla modelu C1 algorytm wygenerował osiem klastrów.

Każdy przypadek z danych treningowych został przydzielony do jednego klastra. Najbardziej liczny jest klaster 1 z populacją 1028. Klaster 4 ma 714 przypadków, a klaster 2 ma 658. Pozostałe klastry klasyfikują liczbę przypadków zgodną z przedziałem <507,569>.

Sieć zależności pomiędzy klastrami (rys. 12) obrazuje, jakie jest powiązanie między nimi. Brak powiązania sugeruje większą niezależność klastra.



Rys. 12. Sieć zależności między klastrami dla modelu C1 (wygenerowane przez MS SQL Server 2005)
Fig. 12. Cluster dependency network for C1 model (generated by MS SQL Server 2005)

Najbardziej niezależnymi klastrami są klastry 7 i 6.

Klaster 6 charakteryzują przypadki o czasie życia 567-1381 z prawdopodobieństwem 46%, a dla czasu życia 7-566, prawdopodobieństwo wynosi 25%. Tak więc sklasyfikowane są tutaj „krótko żyjące" przypadki. Klaster cechuje występowanie dużej wartości *Tblength* (5345-11592). Moc silników wynosi 246-599. Charakterystyczna jest także wartość *Casing Double* wynosząca 9.625. Istotne jest również, iż z dużym prawdopodobieństwem dwuwartościowe dyskretne argumenty są ustawione na wartość 'T', czyli na przykład występuje korozja.

Klaster 7 charakteryzują przypadki o czasie życia 567-1381 z prawdopodobieństwem 46%, a dla czasu życia 7-566, prawdopodobieństwo wynosi 24%. Tak więc podobnie jak dla klastra 6, klaster 7 klasyfikuje „krótko żyjące" przypadki. *Casing Double* ma najczęściej wartość 7 i 5.5. Silniki przeważnie mają moc 42-144. Wartość *Tblength* mieści się głównie w przedziale 3533-11592. Parametry *scale* i *corrosion* (przy ustawionej wartości 'T') występują z blisko 100% prawdopodobieństwem. Także *abrasion* i *openhole* są bardzo prawdopodobne dla tego klastra.

Wspólnym wnioskiem wynikającym z charakterystyki obu klastrów jest to, że w przypadku głębokich szybów zachodzą niszczące zjawiska, takie jak zacieki czy korozja. Czas życia takiego systemu ESP jest krótki.

Analiza odróżniania się klastra od innych klastrów również może doprowadzić do interesujących wniosków. Pewne cechy mogą bardziej charakteryzować jeden klaster niż inny lub wszystkie inne razem wzięte.

Taka analiza została przeprowadzona dla każdego z klastrów. Najciekawsze rezultaty daje porównanie klastra 2 z pozostałymi klastrami. Klaster 2 klasyfikuje silniki o małej mocy 41-69, te o większych mocach z większym prawdopodobieństwem trafiają do innych klastrów. Inna charakterystyczną cechą jest *Casing Double* o wartości 8.625, który bardzo

rzadko cechuje pozostałe klastry. Istotne jest również, iż *openhole* (które jest ściśle powiązane z *casing*) jest ustawione na 'F'. Podsumowując, klaster 2 grupuje przypadki szybów nietkniętą rurą okładzinową o średnicy 8.625, w której umieszczono silnik małej mocy. Dla takiego systemu czas życia jest większy niż dla pozostałych systemów i jest sklasyfikowany w przedziale 3893-8051.

4.4. Podsumowanie eksploracji danych systemów ESP

Zostały dostrzeżone pewne charakterystyczne cechy wspólne dla modeli powstałych przy użyciu różnych algorytmów eksploracji.

Zarówno dla modelu drzewa decyzyjnego, jak i naiwnego bayesowskiego najważniejszą kolumną jest *Casing Double*. Ponadto, kolumna *Motthp* jest również uważana za istotną przez oba modele.

W modelu naiwnym bayesowskim zauważono, że systemy ESP z silnikami o małej mocy potrafią więcej przeżyć niż te o większej mocy. Opisany klaster 2 dla modelu klastrowego klasyfikuje przypadki o małej mocy silnika dla długo żyjących systemów ESP.

Wspomniany klaster 2 wskazuje na zależność, iż systemy ESP z *Casing Double* o wartości 8.625 są „długowieczne". Także dla modelu drzewa decyzyjnego pierwsze rozgałęzienie w drzewie jest właśnie na tym atrybucie z taką samą wartością.

Powyższe zależności nie są oczywiste. Natomiast fakt, iż zjawiska niszczące, takie jak abrazja, korozja, …, powodują skrócenie systemu ESP, są proste do przewidzenia. Niemniej jednak modele te także potwierdzają tę tezę.

Każdy z algorytmów wygenerował modele, na podstawie których można odczytać ukryte zależności, wpływające na czas życia systemów ESP. Dlatego stosowanie tych algorytmów przynosi korzyści.

Każdy z algorytmów został użyty do próby predykcji czasu życia systemów ESP. Najlepiej spisał się model drzewa decyzyjnego, najgorzej model naiwny bayesowski. Niestety, jakość predykcji nawet dla najlepiej spisującego się modelu nie jest satysfakcjonująca. Jeśli przyjmie się, że dopuszczalny jest 50% błąd przy predykcji, to tylko dla, w przybliżeniu 50%, wszystkich danych otrzyma się poprawną klasyfikację. 50% szansa, że predykcja jest poprawna, może mieć zastosowanie jedynie w celach orientacyjnych.

Najszybciej model generował algorytm bayesowski. Natomiast najwięcej czasu potrzebował algorytm klastrujący. Niemniej jednak dla wykonywanych obliczeń nie było to odczuwalne, gdyż najdłuższe obliczenia na komputerze z procesorem 1600 MHz trwały co najwyżej kilka sekund.

5. Podsumowanie analizy

W artykule zostały użyte dwa podejścia: statyczne i eksploracyjne w celu próby analizy czasu życia systemów ESP.

5.1. Porównanie zastosowania

Aplikacja ESAP ma zastosowanie dla istniejących, zainstalowanych systemów ESP. Użytkownik wybiera sobie system ESP, dla którego przy użyciu metody statystycznej może zobaczyć, jak kształtował się czas życia podobnych systemów ESP. W rezultacie użytkownik jest informowany, jaki procent podobnych systemów przeżył tyle, ile wynosi wiek badanego systemu. Dodatkowe informacje pochodzą z espWatchera. Na bazie tych informacji użytkownik decyduje, czy należy wykonać jakieś działania związane z pracą systemu ESP w szybie naftowym.

Zastosowanie algorytmów eksploracyjnych jest szersze. Po pierwsze, można wykonać zapytanie predykcyjne, które jako wynik wskaże, jaka będzie długość życia danego systemu ESP. Jednakże, jak wcześniej pokazano, otrzymany wynik jest bardzo często błędny lub cechuje się małą dokładnością. Podejście statystyczne daje dużo bardziej wiarygodną sugestię. Po drugie, dzięki zastosowaniu algorytmów eksploracyjnych można zauważyć pewne ukryte właściwości, cechy systemu ESP, które mają wpływ na długość jego życia. Te informacje można wykorzystać podczas budowy szybu naftowego czy instalowania w nim systemu ESP.

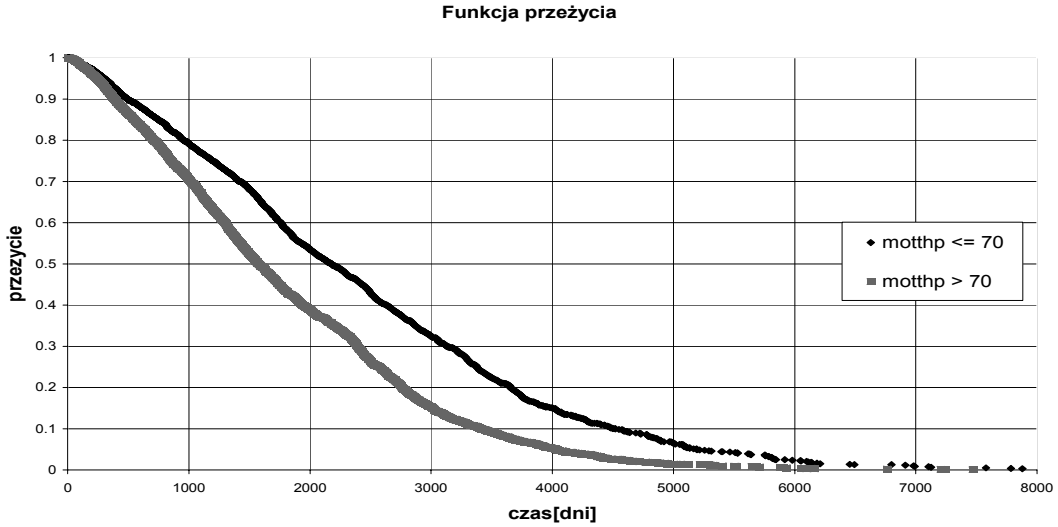
5.2. Porównanie wyników

Jako rezultat analizy eksploracyjnej otrzymano kilka informacji, które systemy ESP są bardziej żywotne od innych. Należy zaznaczyć, że analiza eksploracyjna odbyła się przy użyciu około 6 500 rekordów. Natomiast przy analizie statystycznej dysponowano około 23 500 rekordami. Warto więc (posiadając więcej danych-rekordów) sprawdzić, czy prawdziwe są spostrzeżenia przy analizie eksploracyjnej.

Stwierdzono, że silniki o małej mocy występują częściej w dłużej żyjących systemach ESP niż silniki o większej mocy.

Dane zostały podzielone na dwie części: pierwsza z silnikami o mocy ≤ 70 koni mechanicznych, druga o mocy > 70 koni mechanicznych. Dla każdych danych zastosowano estymator Kaplan-Meier.

Wyniki przedstawiono na poniższym wykresie (rys. 13).

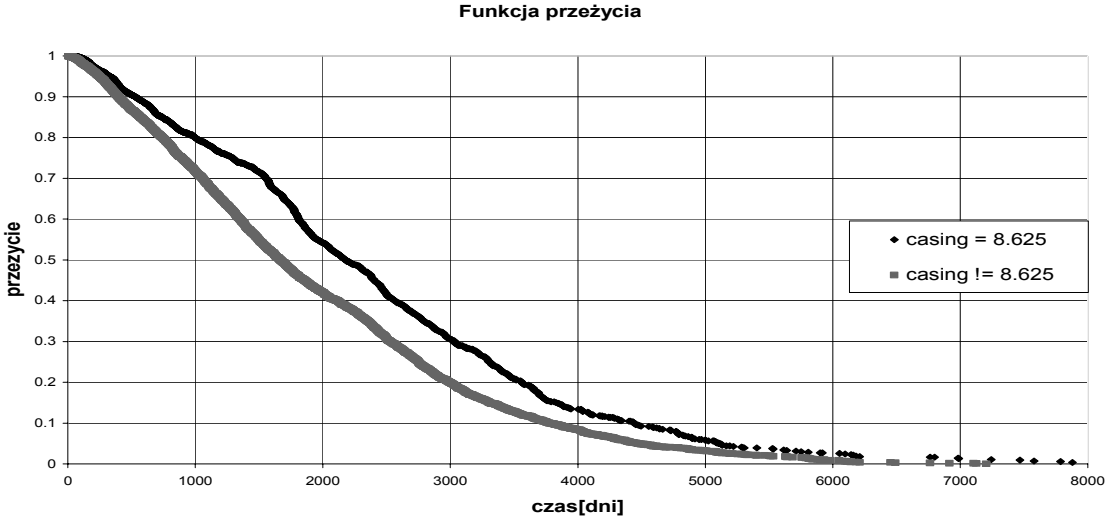


Rys. 13. Funkcje przeżycia dla systemów ESP z silnikami o różnej mocy
Fig. 13. Survival curves for ESP systems with various horse power values

Kolejne spostrzeżenie mówi, że rury (ang. *casing*) o średnicy 8.625 występują częściej w dłużej żyjących systemach ESP niż rury o innych średnicach.

Dane zostały podzielone na dwie części: pierwsza z casing double równym 8.625 cali, druga z *casing double* różnym od 8.625 cali. Dla każdych danych zastosowano estymator Kaplan-Meier.

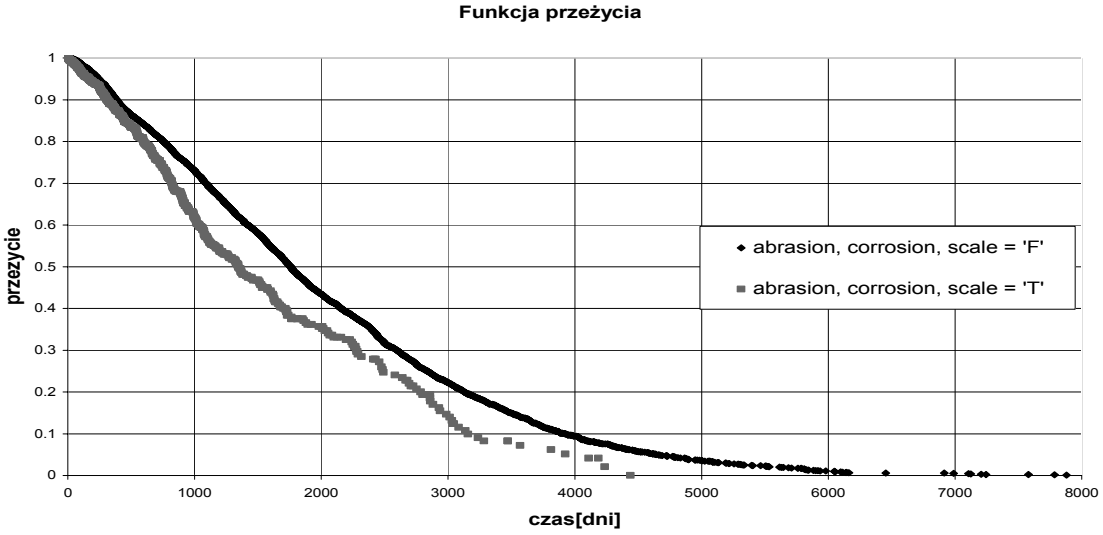
Wyniki przedstawiono na poniższym wykresie (rys. 14).



Rys. 14. Funkcje przeżycia dla systemów ESP z różną średnicą rury okładzinowej
Fig. 14. Survival curves for ESP systems with various diameter of a casing

Potwierdzono także, że różne zjawiska niszczące wpływają na krótsze życie systemu ESP.

Dane zostały podzielone na dwie części: pierwsza bez niszczących zjawisk (*abrasion*, *corrosion*, *scale* z wartością 'F'), druga z niszczącymi zjawiskami (*abrasion*, *corrosion*, *scale* z wartością 'T'). Dla każdych danych zastosowano estymator Kaplan-Meier. Wyniki przedstawiono na poniższym wykresie (rys. 15).



Rys. 15. Wykresy przeżycia dla systemów ESP w zależności od występowania lub nie zjawisk niszczących
Fig. 15. Survival curves for ESP systems with and without abrasion, corrosion, scale

Na wszystkich powyższych przykładach widać, że spostrzeżenia i wnioski z analizy eksploracyjnej znajdują potwierdzenie w analizie statystycznej przy użyciu większej ilości danych.

LITERATURA

1. Kaplan E. L., Meier P.: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, Vol. 53, No. 282, 1958, s. 457÷481.
2. Kantardzic M.,Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, 2003.
3. Harinath S., Quinn S. R.: Professional SQL Server Analysis Services 2005 with MDX, Wrox Press, 2006.
4. Brust A. J., Forte S.: Programming Microsoft SQL Server 2005. Microsoft Press, 2006.
5. Ville B.: Microsoft Data Mining: Integrated Business Intelligence for e-Commerce and Knowledge Management, Digital Press, 2001.
6. Gorawski M.: Laboratorium hurtowni danych poziomu MS SQL SEVER 2005. Krajowa konferencja naukowa „Metody i narzędzia wytwarzania oprogramowania", Szklarska Poręba 2007, s. 583÷596.

7. Motyka A., Tuzinkiewicz L.: Ocena implementacji algorytmów predykcji w Microsoft SQL Server 2005. Krajowa konferencja naukowa „Metody i narzędzia wytwarzania oprogramowania", Szklarska Poręba 2007, s. 3007÷317.
8. Microsoft – Dokumentacja Microsoft SQL Server 2005.
9. Dokumentacja środowiska R – http://www.r-project.org/.
10. Schlumberger – broszura espWatcher, http://www.slb.com/media/services/artificial/submersible/espwatcher.pdf.

Recenzent: Prof. dr hab. inż. Antoni Ligęza

Wpłynęło do Redakcji 7 listopada 2007 r.

Abstract

Oil is arguably the most important natural resource in the modern era. The demand for petrol continues to grow at unparalleled rates; and thus, the process oil production must be inevitably improved. Such achievements can be realized by widely understood planning and management regimens. In particular, this refers to regularly replacing or properly maintaining equipment. Clearly, taking a proactive approach before serious mechanical failure occurs is of paramount importance. Therefore, knowledge regarding estimated lifespan of equipment is critical. This literature review shows various paradigms of survival analysis for particular types of equipment, referred to as ESP systems.

This paper begins by describing the basics of oil production processes and focuses on an improvement system called the artificial lift. An example of such a system is the electrical submersible pump (ESP) system (Fig. 2), which is the subject of analysis in this paper. The analysis consists of two parts: statistical and data mining approaches. The calculations are based on "historical" data from ESP systems. There are more than 20,000 ESP systems defined with information about the system installation time and properties of the system's equipment.

Statistical analysis is done by means of the nonparametric Kaplan-Meier estimator, which takes into account censored data. In order to simplify the process of analysis, an application has been developed (Fig. 3). This application chooses a subset of data about historical ESP systems; then sends the data to a statistical R engine, and subsequently produces a survival plot as a final result.

The latter part of this paper focuses on data mining algorithms: naïve bayes, decision trees and clustering algorithms. These algorithms are used for descriptive and predictive

analysis. The analysis leads to conclusions about increased probability of failure within the ESP system, as well as predicting the lifespan of a system.

This paper concludes by a comparison of the statistical and data mining methodologies.

Adresy

Marcin GORAWSKI: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, Marcin.Gorawski@polsl.pl.
Jarosław ŻYCIŃSKI: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, Jaroslaw.Zycinski@polsl.pl.