Analiza szeregów czasowych

Szereg czasowy to zbiór wartości zarejestrowanych w funkcji czasu y=f(t|**к**), często szereg czasowy może być również funkcją innych parametrów (oznaczono je jako **к**).

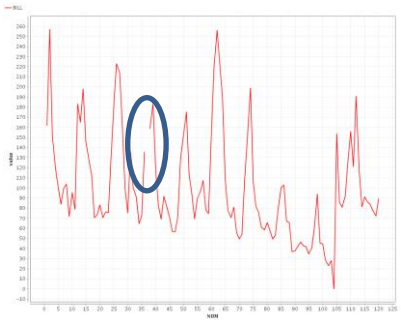
1) Wczytaj zbiór danych *electricbill.dat* jako plik CSV (uwaga włącz opcję *trim lines* podczas importu danych), którego wartości oddzielone są spacjami. Wczytany zbiór danych stanowi przykład szeregu czasowego opisującego wielkość rachunków za energię elektryczną wyrażoną w $ wraz z dodatkowymi zmiennymi. Są nimi:

Opis	Nazwa
Numer obserwacji	Num
Rok	Year
Miesiąc	Month
Wielkość rachunku miesięcznego za energię elektryczną wyrażona w $	Bill
Średnia temperatura za miesiąc	Temp
Liczba dni z dogrzewaniem	HDD
Liczba dni z chłodzeniem	CDD
Liczba członków rodziny	Size
Nowy miernik (zmienna binarna)	Meter
Nowa pompa ciepła 1 (zmienna binarna)	Pump1
Nowa pompa ciepła 2 (zmienna binarna)	Pump2
Łączna opłata (za kWh)	RIDER TOTAL
Obliczona knsumpcja w kWh	Consumption

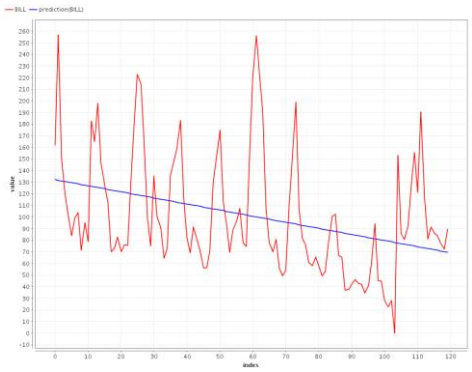
2) Za pomocą operatora *Select attrbutes* wybierz dwa atrybuty: Num,Year,Bill Wczytane dane przedstaw na rysunku. W tym celu wykorzystaj widok „plot" -> Plotter -> Series Multiple. Na wykresie przedstaw zmienne Bill oraz Year
3) Spróbuj wyznaczyć chwile zmian roku, tak aby lepiej było widać kiedy następuje zmiana roku (styczeń). W tym celu zastosuj operator Differentiate. Operator ten wyznacza pochodną licząc różnicę między sąsiednimi wartościami y(t)-y(t-1). Narysuj wykres wielkości rachunku i zmian roku. Dla lepszej wizualizacji spróbuj podzielić serię czasową na lata, tak aby wiersz stanowił jeden rok. W tym celu odfiltruj i pozostaw jedynie atrybut Bill, a następnie wykorzystaj operator Windowing ustawiając odpowiednie wartości *window size* oraz *step size*. Parametr *windows size* określa rozmiar okna czyli liczbę wierszy które będą stanowiły jeden wektor, natomiast *step size* to wielkość kroku jaki należy wykonać aby zrobić nowy wektor. W wyniku wykonanych operacji powinieneś dostać zbiór danych składający się z 10 rekordów, każdy składający się z 12 kolumn odpowiadających kolejnym miesiącom.
   Spróbuj uzupełnić dane o rok – w tym celu możesz zastosować operator *windowing* na zbiorze danych składającym się ze zmiennych *Bill* oraz *Year*, a następnie odfiltrowując – za pomocą operatora *Select attributes* niepotrzebne i nadmiarowe atrybuty typu *Year-?* Pozostaw tylko *Year-11*. Na koniec ustaw rolę atrybutu *Year-11 na* Label (za pomocą operator *Set Role*)
   Zaprezentuj uzyskany wynik w postaci Parallel Plot. Parallel plot przedstawia na wykresie na osi X kolejne kolumny ze zbioru danych, a pojedyncza krzywa przedstawia pojedynczy wiersz ze zbioru danych. Ustaw *Color Column* na *Year-11* Zastanów się nad interpretacją uzyskanych wyników.
4) Jak zbadać źródło zaobserwowanych zmian – czy ich wynikiem jest działanie człowieka (przeprowadzone inwestycje), czy może czynniki zewnętrzne jako pogoda.

Jakie narzędzie należałoby użyć aby dokonać weryfikacji uzyskanych wyników?

5) Analizując wykres z zadania 2 zwróć uwagę że w wykresie brakuje jednej wartości. Aby ją uzupełnić wykorzystaj operator Replace Missing znajdujący się w pakiecie Series

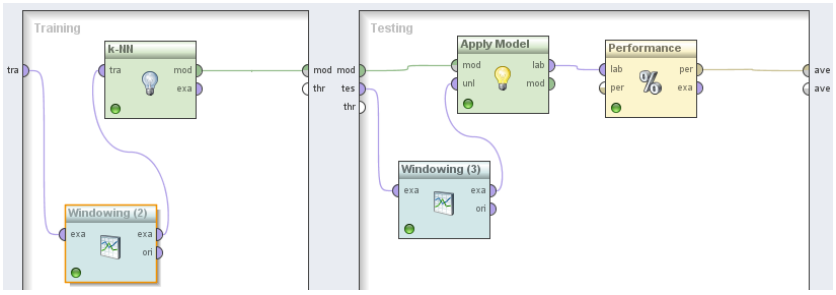


6) Wygładzanie i trend. Typowym problemem w analizę serii danych jest wyznaczenie linii trendu. Zastanów się jak wyznaczyć linię trendu dla serii danych Bill, korzystając ze standardowych operatorów RapidMinera. Dla ułatwienia zastanów się jak można tutaj wykorzystać regresję liniową. Wyniki przedstaw na wykresie korzystając z Plotter -> series
Uzyskany wynik powinien przedstawiać:



7) Dla ułatwienia powyższe zadanie można też wykonać korzystając z operatora Fit Trend. Sprawdź czy obydwa wyniki sa identyczne
8) Budowa modeli predykcyjnych
Aby zbudować model predykcyjny dla szeregu czasowego najlepiej jest skorzystać z operatora Windowing (patrz wyżej), jednakże zmieniając ustawienia, tak iż *window size* powinno odpowiadać rozmiarowi historii którą chcemy analizować, a *step size* najlepiej ustawić na wartość 1, wówczas okno będzie przesuwało się na kolejny element w serii. Uwaga, budowa modelu predykcyjnego wymaga kolumny *Label* w tym celu zaznacz opcję *create label* wybierając odpowiedni atrybut który ma stanowić etykietę oraz horyzont czasowy czyli na ile w przód chcielibyśmy dokonać przewidywania -> liczbę próbek w przód
Na tak przygotowanym zbiorze można zbudować model predykcyjny np. k-NN, sieć neuronową itp.
Zbuduj model predykcyjny oparty o algorytm kNN (k=3) i spróbuj dobrać rozmiar okna tak, aby zapewnić największą dokładność predykcji
9) Testowanie modeli predykcyjnych
Do testowania modeli predykcyjnych służy operator Validation. Operator Validation znajdujący się w selkcji *Series -> Evaluation -> Validation -> Sliding window Validation.* Operator ten tworzy okno dla danych treningowych o rozmiarze *training window width* (uwaga tutaj można użyć dużego okna, aby początkowy model miał się na czym uczyć) np. w naszym przypadku można tą

wartość ustalić na 24, co oznacza że dane co najmniej z 2 lat zostaną wykorzytsane do uczenia. *Training window step size* oznacza krok z jakim okno ma się przesuwać *test window szie* to rozmiar okna używany do testowania. Np. można tą wartość ustawić na pożądany rozmiar wektora – dla przykładu wartość 6.
Pole *horizon* oznacza horyzont czasowy w którym wygenerowany zostanie testowy przypadek (po części treningowej). Dla zrozumienia działania operatora Validation wybierz jedynie atrybut NUM, a następnie zbuduj proces jak na obrazku:



Ustaw parametry operatora Windowing jak na obrazku:

w sekcji tranining wstaw breakpoint za operatorem Windowing i zobacz jak będzie wyglądał zbiór danych dostarczonych na wejście. Training oraz porównaj z wynikiem na wyjściu operatora Windowing
Następnie wybierz w zbiorze danych atrybut Bill i powtórz obliczenia.