Kurs: 1DI1635– Podstawy sztucznej inteligencji

Opracował: dr hab. inż. Paweł Piotrowski, prof. uczelni, wersja 13.04.2022

Ćwiczenie 2 h– Prognozowanie krótkoterminowe generacji mocy ze wszystkich farm wiatrowych offshore zlokalizowanych na terenie Belgii

Środowisko: program Statistica wersja 13 (dostępna w zasobach CI PW, licencja roczna do końca września danego roku z możliwością przedłużenia na kolejne lata)

Celem ćwiczenia jest wykonanie analizy danych (dobór zmiennych do modeli) oraz prognozy z horyzontem 15 minut wprzód generacji mocy ze wszystkich farm wiatrowych typu offshore zlokalizowanych na terenie Belgii

1. Analiza danych

W pliku „B2_Belgia_offshore_wind_farms" w zakładce „dane_do_statistica" zgromadzono dane jako szereg czasowy. Kolumna C zawiera KOD_ZESTAWU. Ilustruje to rysunek poniżej.

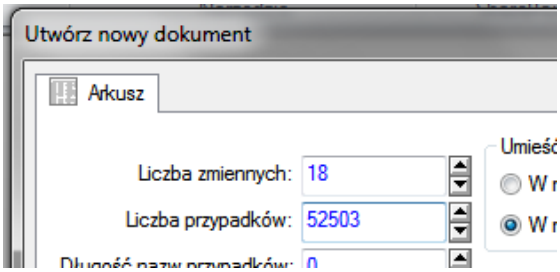
C	D	E	F	G	H
KOD_ZESTAWU	P	Pawel:			
3		1-TRENINGOWE (DO USTALENIA WAG CZYLI PARAMETRÓW DANEGO MODELU)			
3		2-WALIDACYJNE (DO SZUKANIA WŁAŚCIWYCH HIPERPARAMETRÓW			
3		3-TESTOWE (MODEL NIC O NICH NIE WIE)			

Kolumna D to wyjście dla modelu, natomiast kolumny od E do T to wejścia do modelu. Poniżej fragment danych.

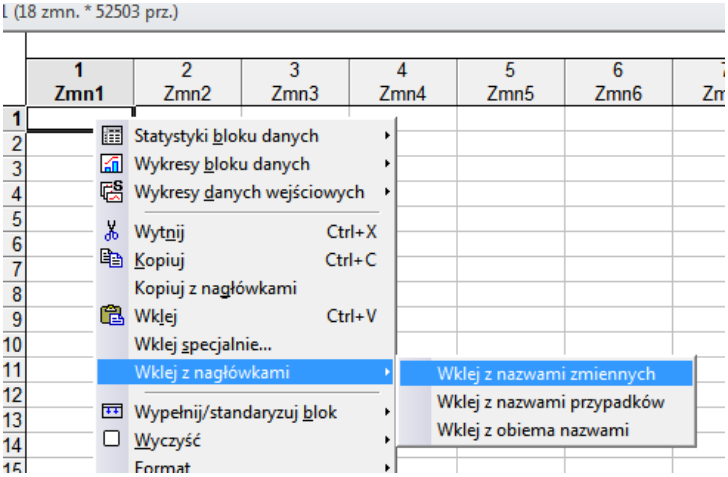
D	E	F	G	H	I	J	K	L	M
P_t+1[MW]_WYJSCIE	Miesiac	Godzina	P_t	P_t-1	P_t-2	P_t-3	P_t-4	P_t-5	P_t-6
454,53	9	4	459,47	476,04	495,11	526,16	563,76	624,42	625,44
432,41	9	4	454,53	459,47	476,04	495,11	526,16	563,76	624,42
338,93	9	6	365,81	400,26	419,57	419,23	432,41	454,53	459,47
381,23	9	6	371,15	349,59	338,93	365,81	400,26	419,57	419,23
368,47	9	7	371,94	381,23	371,15	349,59	338,93	365,81	400,26
337,92	9	7	332,52	368,47	371,94	381,23	371,15	349,59	338,93
261,07	9	9	270,66	297,16	326,93	349,9	337,92	332,52	368,47
163,73	9	12	164,54	151,74	168,02	202,89	230,15	274,11	273,12
116	9	14	104,22	111,58	110,57	106,07	132,35	163,73	164,54
153,22	9	14	140,56	127,81	116	104,22	111,58	110,57	106,07

Wykonujemy wykres szeregu czasowego P_t+1[MW]_WYJSCIE w celu wizualizacji zagadnienia

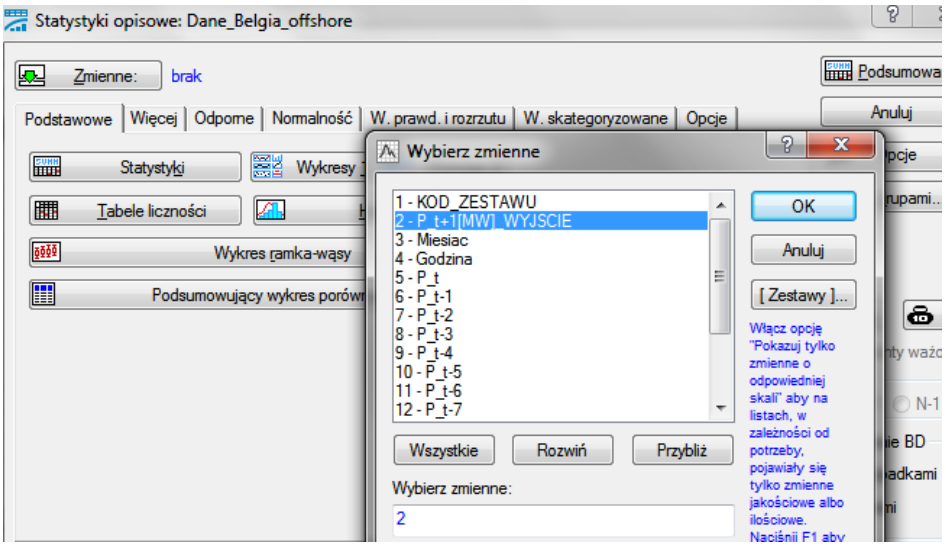
W programie Statistica wybieramy: Menu -> nowy -> arkusz wpisujemy liczba zmiennych 18, przypadków 52503



W arkuszu excel zaznaczamy zakres C1:T52504 czyli z nagłówkami wybrane dane z zakładki "dane_do_statistica"
W statistica wklejamy te dane do arkusza -> wklej z nagłówkami -> wklej z nazwami zmiennych



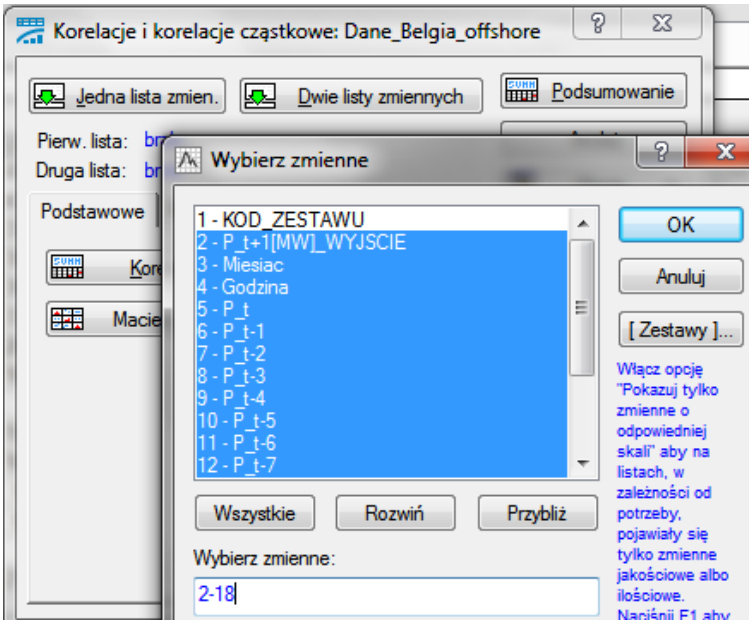
Wybieramy->Statystyka ->statystyki podstawowe->statystyki opisowe
Zmienne-> wybieramy P_t+1[MW]_WYJSCIE
Wybierając kolejne zakładki możemy wykonać wybrane analizy statystyczne wybranego szeregu czasowego.



Wybór zmiennych wg wartości współczynnika korelacji liniowej Pearsona

Wybieramy->Statystyka ->statystyki podstawowe-> macierze korelacji
Wybieramy->jedna lista zmiennych->

Wpisujemy 2-18 w polu wyboru zmiennych lub ręcznie wybieramy zmienne z przyciśniętym CTRL.



W zakładce podstawowe wybieramy -> korelacje. Otrzymujemy wartości korelacji liniowej Pearsona pomiędzy wyjściem modelu oraz wejściami modelu. Im wartość jest bliższa 1 lub –1 tym korelacja jest silniejsza (dodatnia lub ujemna). Uwaga: są to korelacje liniowe a zależności we-wy bywają nieliniowe i tego może nie wykazać wartość R.

- Korelacje (Dane_Belgia_offshore)

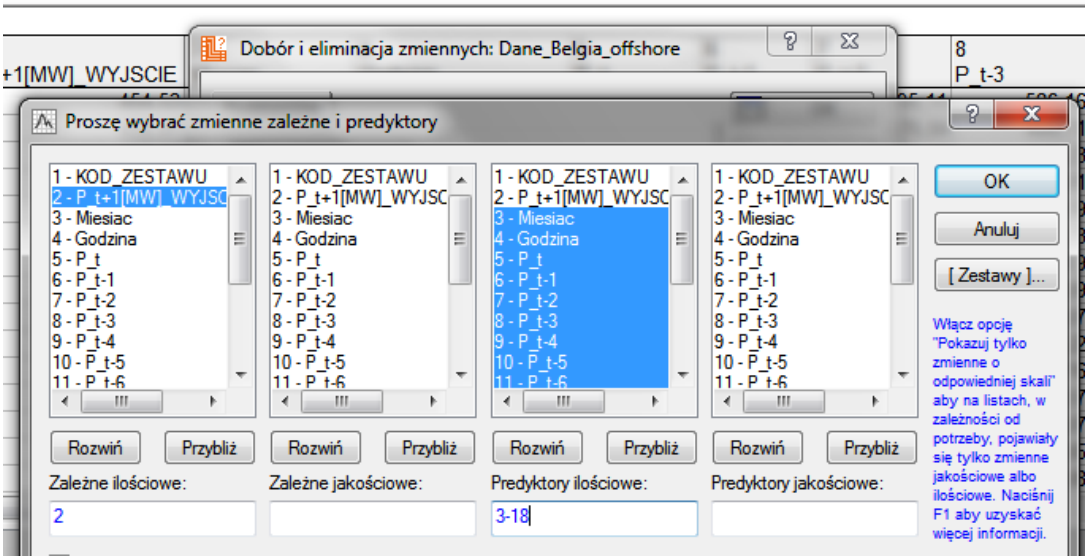
Korelacje (Dane_Belgia_offshore)
Oznaczone wsp. korelacji są istotne z p < ,05000
N=52503 (Braki danych usuwano przypadkami)

Zmienna	Średnia	Odch.std	P_t+1[MW]_WYJSCIE
P_t+1[MW]_WYJSCIE	785,733	643,5461	1,000000
Miesiac	6,861	3,8019	-0,067495
Godzina	11,502	6,9214	0,038285
P_t	785,735	643,5457	0,996181
P_t-1	785,735	643,5456	0,989026
P_t-2	785,734	643,5459	0,981291
P_t-3	785,732	643,5463	0,973280
P_t-4	785,730	643,5468	0,964994
P_t-5	785,726	643,5476	0,956442
P_t-6	785,721	643,5493	0,947693
P_t-7	785,713	643,5521	0,938747
P_t-8	785,703	643,5562	0,929582
P_t-9	785,692	643,5606	0,920299
P_t-10	785,681	643,5646	0,910808
P_t-11	785,671	643,5684	0,901154
P_t-12	785,662	643,5720	0,891345
P_znam_dostepna_[MW]	1769,878	255,6543	0,212368

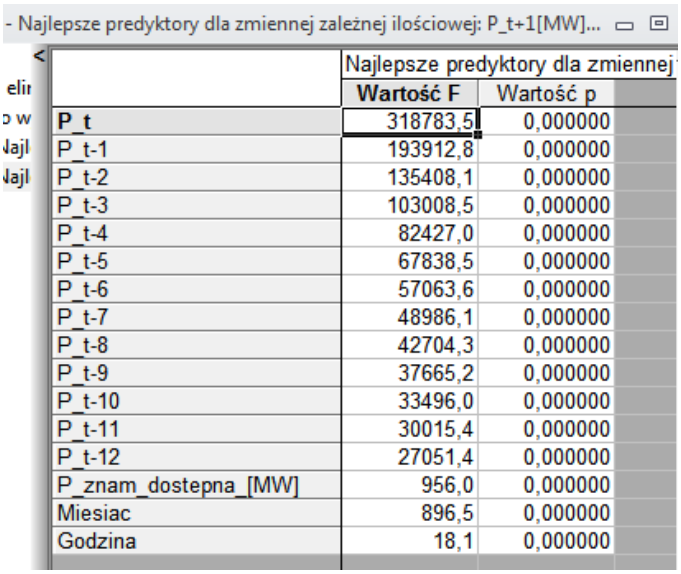
Notujemy wyniki kopiujemy (wybierz wszystko, kopiuj z nagłówkami) do arkusza kalkulacyjnego
Dokonujemy eksperckiego wyboru zmiennych wejściowych analizując współczynniki korelacji zmiennej wyjściowej do potencjalnych zmiennych wejściowych (dane istotne statystycznie zaznaczone są na czerwono)

Wybór zmiennych jako problem regresji wielokrotnej liniowej (statystyka F Fischera Snedecora – eliminacja zmiennych z równania regresji

Z menu górnego wybieramy „data mining" , wybieramy ->dobór zmiennych -> dobór zmiennych
Zależna ilościowa = zmienna wyjściowa czyli 2, predyktory ilościowe = potencjalne zmienne wejściowe 3-18.



Liczba klas dla predyktorów ilościowych np. 10
Wybieramy pokaż 16 (czyli wszystkie) najlepszych predyktorów (komentarz: Pokaż k najlepszych predyktorów. Ta opcja oznacza wyświetlanie k najlepszych predyktorów. W problemach typu regresyjnego (dla ilościowych zmiennych zależnych), jest to k predyktorów o najwyższej wartości statystyki F)
Klikamy w podsumowanie najlepsze predyktory
Notujemy w arkuszu kalkulacyjnych wyniki – sugestie – wybór zmiennych wejściowych


- Najlepsze predyktory dla zmiennej zależnej ilościowej: P_t+1[MW]...

	Najlepsze predyktory dla zmiennej	
	Wartość F	Wartość p
P_t	318783,5	0,000000
P_t-1	193912,8	0,000000
P_t-2	135408,1	0,000000
P_t-3	103008,5	0,000000
P_t-4	82427,0	0,000000
P_t-5	67838,5	0,000000
P_t-6	57063,6	0,000000
P_t-7	48986,1	0,000000
P_t-8	42704,3	0,000000
P_t-9	37665,2	0,000000
P_t-10	33496,0	0,000000
P_t-11	30015,4	0,000000
P_t-12	27051,4	0,000000
P_znam_dostepna_[MW]	956,0	0,000000
Miesiac	896,5	0,000000
Godzina	18,1	0,000000

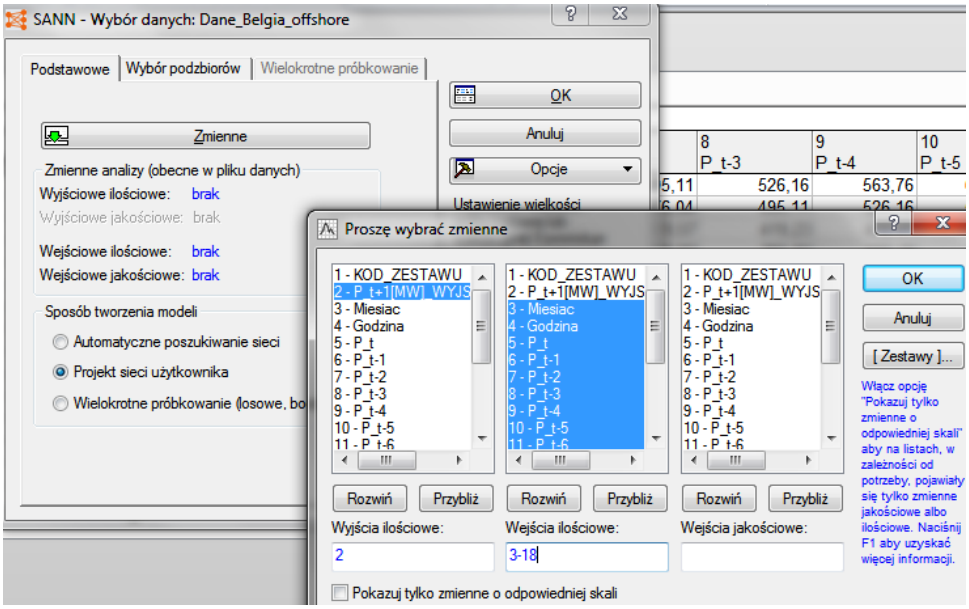
Dokonujemy końcowego eksperckiego wyboru zmiennych wejściowych analizując ranking predyktorów wg statystyki F oraz wcześniej obliczone współczynniki korelacji zmiennej wyjściowej do potencjalnych zmiennych wejściowych.

---

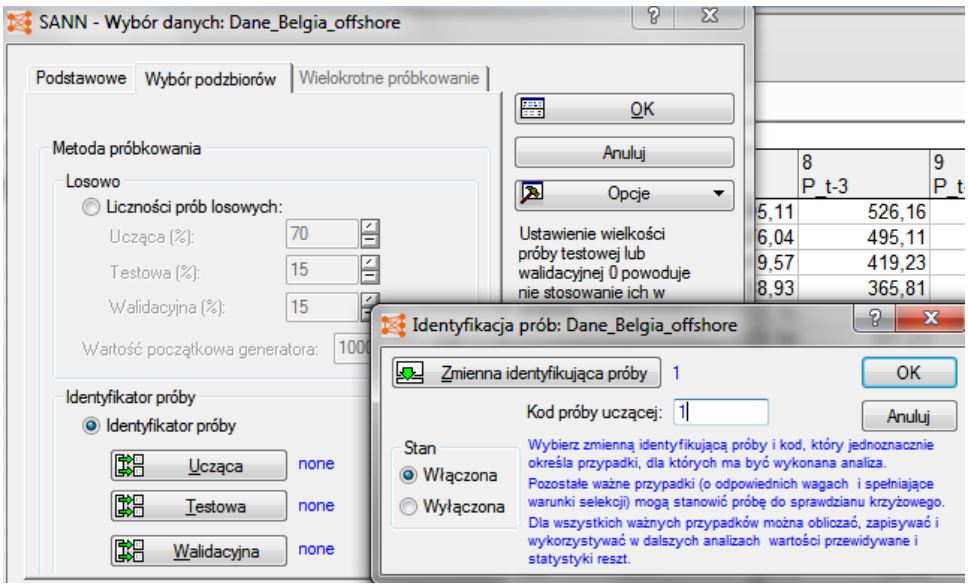
2. Prognozy z wykorzystaniem sieci neuronowej MLP oraz RBF dla horyzontu 15 minut wprzód

Wybieramy menu -> Data Mining -> Sieci neuronowe
wybieramy -> nowa analiza -> regresja

okno SANN wybór danych
zakładka "podstawowe"
wybieramy ->projekt sieci użytkownika
klikamy w pole "zmienne"
wybieramy "wyjscia ilościowe" – P_t+1[MW]WYJSCIE
wybieramy "wejścia ilościowe" - tutaj wskazujemy (trzymając naciśnięty klawisz CTRL) wybrane przez nas zmienne objaśniające (sugerowałbym najpierw komplet zmiennych od 3 do 18)



klikamy w zakładkę "wybór podzbiorów" i zaznaczamy pole "identyfikator próby" klikamy w pole "ucząca" - klikamy w "stan" włączona - klikamy w pole "zmienna identyfikująca próby" wybieramy zmienną "KOD_ZESTAWU", w polu "kod próby uczącej" wpisujemy "1", klikamy ok.



klikamy w zakładkę "wybór podzbiorów" i zaznaczamy pole "identyfikator próby" klikamy w pole "testowa" - klikamy w "stan" włączona - klikamy w pole "zmienna identyfikująca próby" wybieramy zmienną " KOD_ZESTAWU ", w polu "kod próby testowej" wpisujemy "2", klikamy ok.

klikamy w zakładkę "wybór podzbiorów" i zaznaczamy pole "identyfikator próby" klikamy w pole "walidacyjna" - klikamy w "stan" włączona - klikamy w pole "zmienna identyfikująca próby" wybieramy zmienną " KOD_ZESTAWU ", w polu "kod próby walidacyjnej" wpisujemy "3", klikamy ok.

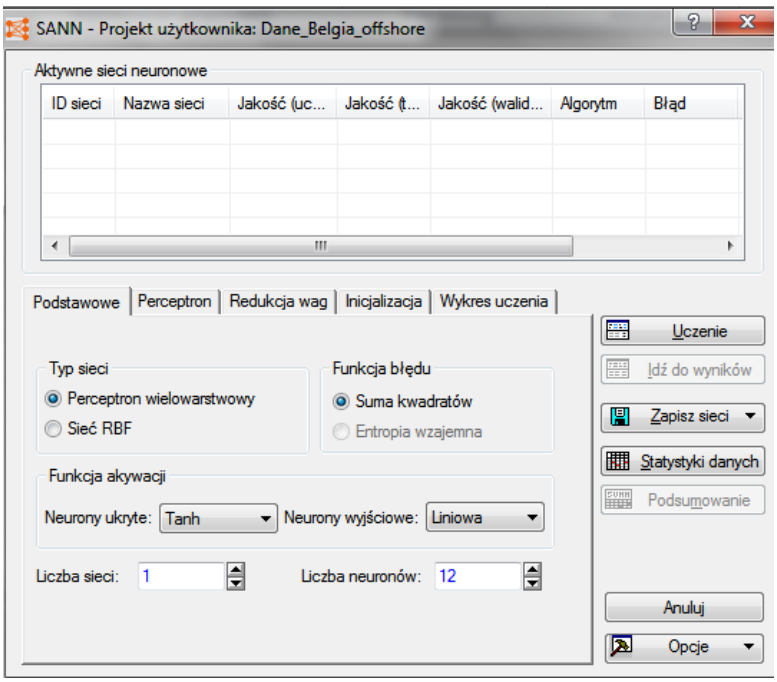
UWAGA: w konwencji Statistica

Testowa to klasycznie walidacyjna do ustalenia hiperparametrów modelu (liczba neuronów, funkcje aktywacji, algorytm uczący)
Walidacyjna to klasycznie testowa czyli końcowo sprawdza się jakość ostatecznego najlepszego modelu na zakresie walidacyjnym klasycznym.

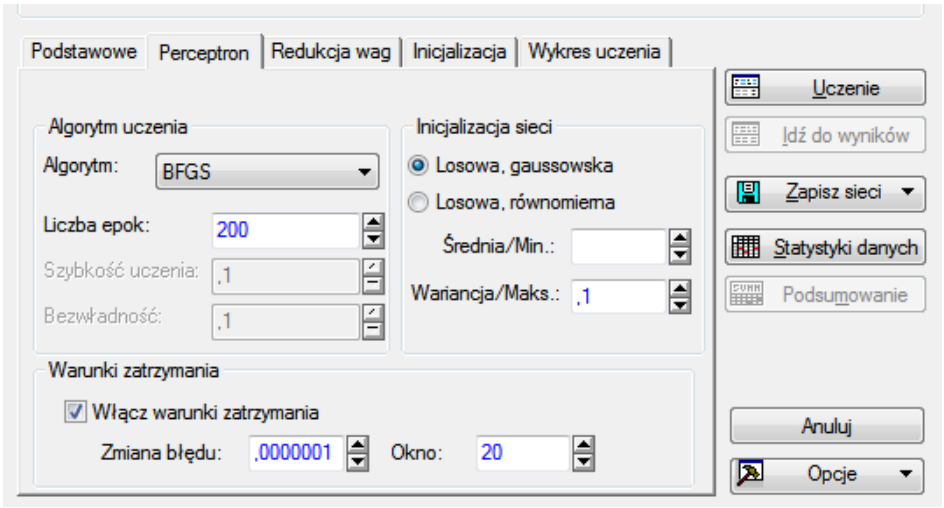
zamykamy okno "SANN wybór danych klikając pole OK.

pojawia się zakładka "SANN projekt użytkownika"

w zakładce "podstawowe"
wybieramy -> typ sieci -> perceptron wielowarstwowy
wybieramy funkcję aktywacji dla neuronów ukrytych oraz neuronu wyjściowego
wybieramy liczbę sieci oraz liczbę neuronów w warstwie ukrytej

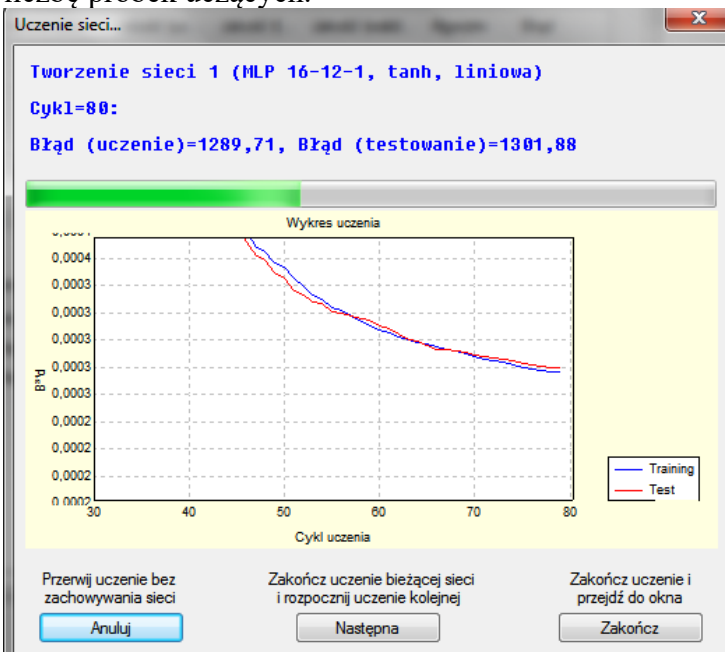


w zakładce "perceptron"
wybieramy algorytm uczenia (BFGS), liczbę epok oraz rodzaj inicjalizacji sieci



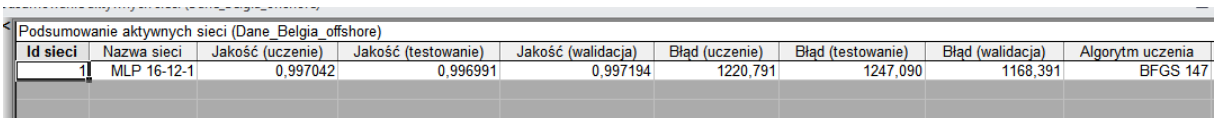
w zakładce "wykres uczenia'
wybieramy - wykres uczenia dla błędów w próbie uczącej oraz testowej.

w oknie głównym "SANN projekt użytkownika" klikamy w przycisk "uczenie" Rozpocznie się proces uczenia sieci neuronowej. Niestety – kilka minut trwają obliczenia z uwagi na dużą liczbę próbek uczących.



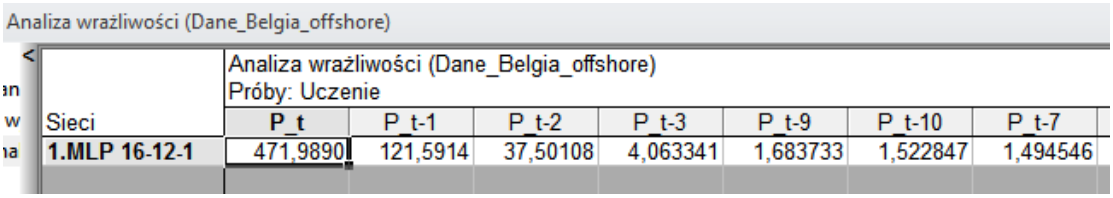
Aby zobaczyć liczbę epok uczących należy rozszerzyć kolumnę "algorytm"
klikamy w "wybór sieci" i wskazujemy najlepszą sieć - z najmniejszym błędem w zakresie "jakość uczenia" - ta miara błędu to po prostu współczynnik korelacji liniowej Pearsona (uwaga – przy pierwszym badaniu mamy tylko jedną sieć neuronową na liście)

Klikając w pole "podsumowanie" zobaczymy wartości miar błędów dla zakresów uczenia i testowania (błąd SOS to średnia z sum kwadratów odchyleń dla próby uczącej i testowej)


Podsumowanie aktywnych sieci (Dane_Belgia_offshore)

Id sieci	Nazwa sieci	Jakość (uczenie)	Jakość (testowanie)	Jakość (walidacja)	Błąd (uczenie)	Błąd (testowanie)	Błąd (walidacja)	Algorytm uczenia
1	MLP 16-12-1	0,997042	0,996991	0,997194	1220,791	1247,090	1168,391	BFGS 147

Kilkamy na zakładkę”Szczegóły”
Klikając w pole "globalna analiza wrażliwości" - uzyskujemy informacje o **ważności poszczególnych zmiennych objaśniających** w modelu prognostycznym –
opis analizy wrażliwości w "dymku"


Analiza wrażliwości (Dane_Belgia_offshore)

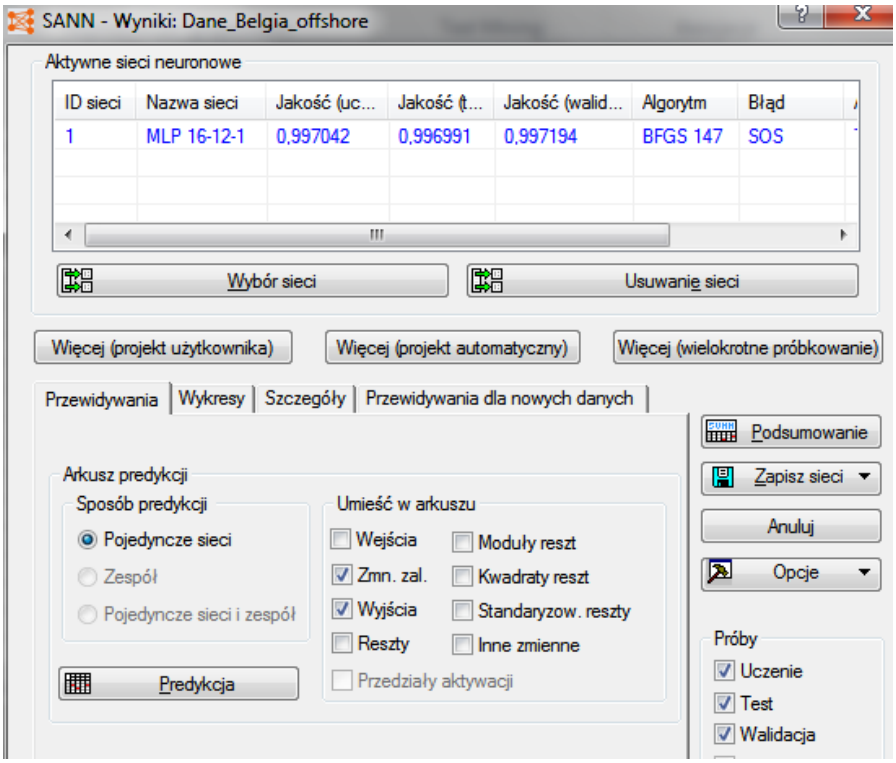
Sieci	Analiza wrażliwości (Dane_Belgia_offshore) Próby: Uczenie						
	P_t	P_t-1	P_t-2	P_t-3	P_t-9	P_t-10	P_t-7
1.MLP 16-12-1	471,9890	121,5914	37,50108	4,063341	1,683733	1,522847	1,494546

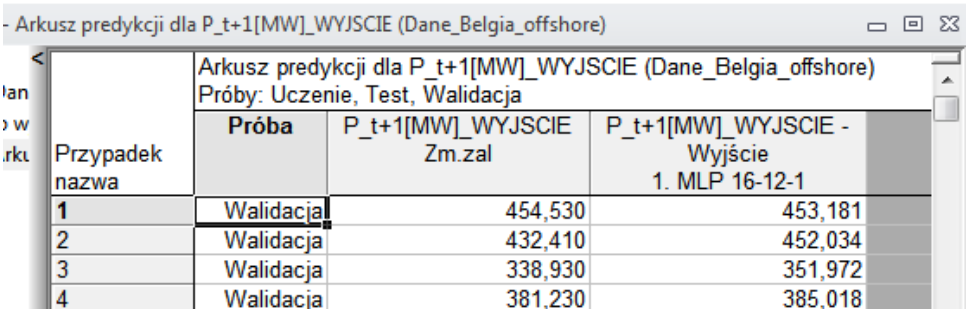
Fragment help programu Statistica

Globalna analiza wrażliwości. Globalna analiza wrażliwości daje pojęcie o tym, jak ważne są poszczególne zmienne wejściowe sieci. Wykonanie analizy wrażliwości polega na sprawdzeniu jak zachowuje się błąd sieci w przypadku gdy coś złego dzieje się ze zmiennymi niezależnymi. Konkretnie, po kolei dla każdej zmiennej wejściowej jej wartości zamieniane są na średnią (ze zbioru uczącego). Tak więc zmienna przestaje wnosić jakąkolwiek informację. Po podaniu tak zmodyfikowanych danych na wejście sieci sprawdza się końcowy błąd predykcji. Błąd ten może poważnie wzrosnąć, albo wzrosnąć nieznacznie lub wcale. Oznacza to, że sieć jest albo bardzo wrażliwa na daną zmienną wejściową, albo też sieci na tej zmiennej zupełnie nie zależy.
W arkuszu, dla każdej sieci podany jest iloraz wskazujący przyrost błędu przy usunięciu danej zmiennej wejściowej. Jeżeli wartość jest 1 lub mniejsza to sieć działa lepiej bez danej zmiennej - znak, ze należy ją usunąć na stałe.
Jednak pamiętać trzeba, że analiza dotyczy konkretnej sieci. Tymczasem zmienne bywają na różne sposoby powiązane, skorelowane i wykazują redundancje. Dlatego różne sieci mogą "wybrać" jako ważne różne zmienne. Dopiero wykonanie analizy wrażliwości dla wielu modeli i powtarzalność wyników powinny być podstawą do wyciągania praktycznych wniosków na temat zmiennych.

Ewentualnie korygujemy później model wybierając nowy, lepszy zestaw zmiennych objaśniających wejściowych.

Wybieramy zakładkę „Przewidywania”
w prawym dolnym rogu okna głównego - opcja "próby" zaznaczamy pozycję „test” oraz "walidacja”
klikamy w przycisk "predykcja"





- Arkusz predykcji dla P_t+1[MW]_WYJSCIE (Dane_Belgia_offshore)

Arkusz predykcji dla P_t+1[MW]_WYJSCIE (Dane_Belgia_offshore)
Próby: Uczenie, Test, Walidacja

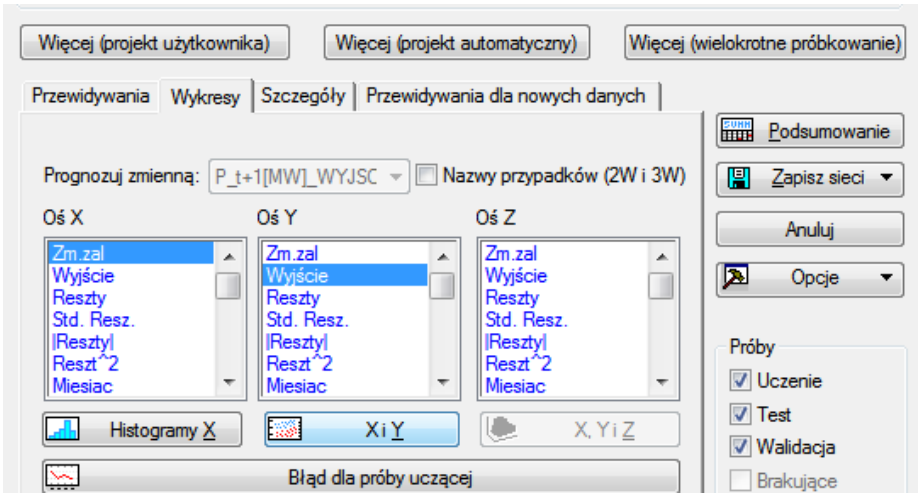
Przypadek nazwa	Próba	P_t+1[MW]_WYJSCIE Zm.zal	P_t+1[MW]_WYJSCIE - Wyjście 1. MLP 16-12-1
1	Walidacja	454,530	453,181
2	Walidacja	432,410	452,034
3	Walidacja	338,930	351,972
4	Walidacja	381,230	385,018

kolumna „P-t+1[MW]_WYJSCIE-Wyjscie” - wyjscie to prognozy dla trzech zakresów, wklejamy ten blok danych do excel -zakładka "MLP_prognozy" - kolumna C

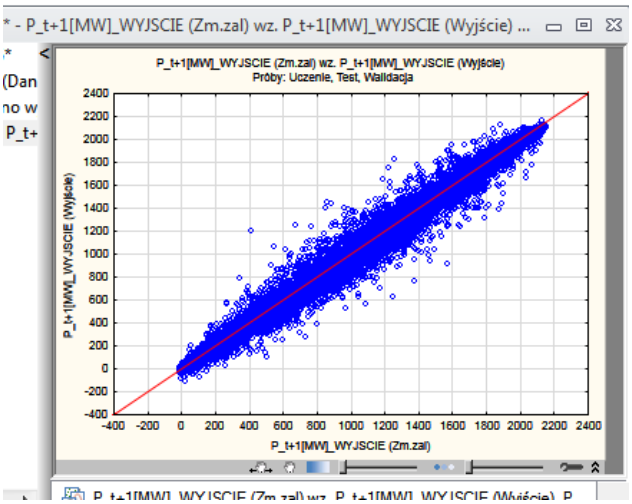
W kalkulatorze błędów pojawią się obliczone miary błędów– kopiujemy wyniki do tabeli z wariantami oraz opisujemy w tabeli parametry danego wariantu. Kolumnę prognoz wklejamy do archiwum wyników – kolumna AF itd.

B	C	D	E	F	G	H	I	J	K
P_t+1[MW]_WYJSCIE	PROGNOZA_MLP	BŁĄD AE	BŁĄD E						
454,53	453,181	0,779719	1,349						
432,41	452,034	17,71021	-19,624		ZAKRES DANYCH TRENINGOWYCH				
338,93	351,972	9,089583	-13,042						
381,23	385,018	3,781404	-3,788		WARIANT	MAE	RMSE	R	BIAS
368,47	366,577	4,764201	1,893		kalkulator	27,67	49,54	0,9970	0,2905
337,92	314,603	26,25874	23,317		1				
261,07	261,169	1,313698	-0,099		2				
163,73	175,659	12,57056	-11,929		3				
116	100,368	10,82892	15,632		4				

W zakładce „wykresy” wybieramy Os X – Zm.zal, Os Y – Wyjscie i klikamy w pole „X i Y”



Uzyskujemy wykres rozrzutu pomiędzy prognozami a wartości rzeczywistymi. Gdyby pary liczb były równe powstała by krzywa Y=X ale niebieskie punkty tworzą „chmurę” wskazującą na różnice pomiędzy prognozami a wartościami rzeczywistymi.

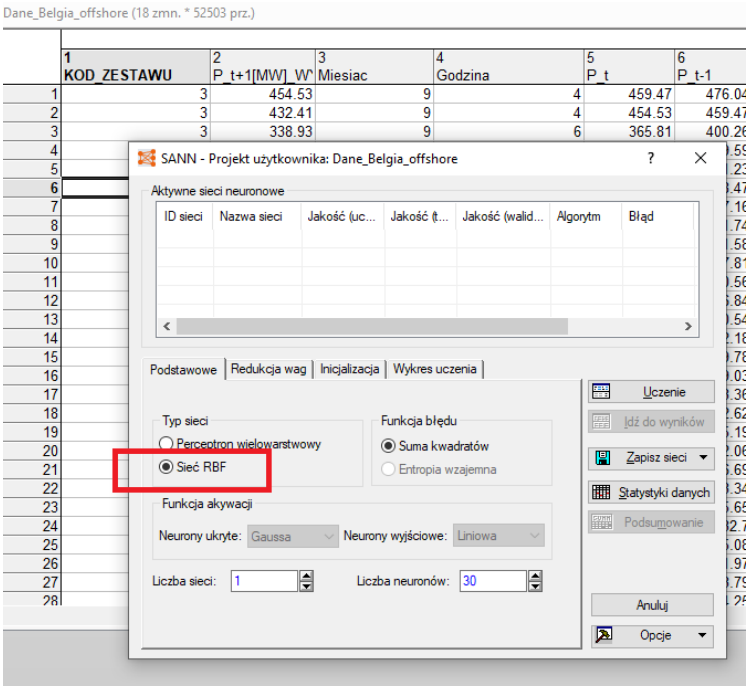


kolejne kroki:

1. Szukamy właściwych hiperparametrów sieci neuronowej dla których wartości miar błędów w zakresie danych testowych będą najmniejsze-można wykorzystać narzędzie "więcej (projekt automatyczny). Do hiperparametrów należą m.in.:
   - funkcje aktywacji w poszczególnych warstwach,
   - liczba nuronów w warstwie ukrytej
   - typ algotymu uczącego
   - typ inicjalizacji wag
   - liczba epok uczących
2. Poszukiwanie właściwego modelu to również manipulowanie doborem zmiennych

Budowa nowej sieci neuronowej to wybór – „Więcej projekt użytkownika" – wtedy możemy zmienić hiperparametry modelu. Szukając najlepszego modelu patrzymy na miary błędów w zakresie walidacyjnym w arkuszu excel i staramy się zmieniając hiperparametry redukować błąd RMSE oraz MAE. Błędy pomocnicze: błąd R – współczynnik korelacji liniowej Pearsona powinien być jak największy, BIAS powinien być jak najmniejszy.

3. Warto zbadać wariant najprymitywniejszy tzn. tylko z ostatnią wartością cofniętą prognozowanego szeregu – można go wtedy przyjąć jako model odniesienia.
4. Wykonanie prognoz siecią neuronową typu RBF w celu porównania wyników z siecią neuronową typu MLP. Poniżej zrzut ekranu z wyborem tej sieci neuronowej.



Należy zwrócić uwagę, że funkcja aktywacji w warstwie ukrytej to Gaussa a w warstwie wyjściowej to liniowa i nie ma możliwości jej zmiany. Manipulować można hiperparametrem:

- liczba neuronów w warstwie ukrytej.

UWAGA!!!!
Końcowa ocena jakości to odczytanie miar błędów dla zakresy testowego dla najlepszego modelu na danych walidacyjnych. Nie można !!!!! patrzeć na błąd na zakresie testowym szukając najlepszego modelu – to byłby błąd – te dane są niewidoczne dla modelu i dla nas - zapobiega to „wyciekowi" informacji do modelu (podpowiedz).

Miary błędów - charakterystyka

Mean Absolute Error is calculated by formula (1)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

In the process of forecasting of electric energy production in a wind turbine, the changes of RMSE and MAE have the same trend, and the smaller the two error values are, the more accurate the prediction results are. MAE is related to the first order of error moment while RMSE is related to the second order.

Root Mean Square Error which is sensitive to large error values is calculated by formula (2)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

where, $\hat{y}_i$ is the predicted value, y_i is the true value, and n is the number of prediction points.

Pearson linear correlation coefficient of the observed and predicted data is calculated by formula (3)

$$R = \frac{C_{y\hat{y}}}{std(y)\cdot std(\hat{y})} \tag{3}$$

where, $C_{y\hat{y}}$ is the covariance value between the really observed and predicted data and std denotes standard deviation of the appropriate variable.
The bigger the error R value is (range from -1 to 1), the more accurate the prediction results is.

Mean Bias Error (MBE) captures the average bias in the prediction and is calculated by formula (4)

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{4}$$

Value of single i-th Absolute Error (AE) needed for calculation of percentiles of AE errors is calculated by formula (5).

$$AE_i = |y_i - \hat{y}_i| \tag{5}$$

Sprawozdanie: 1 wstęp, 2. Tabelaryczna i graficzna prezentacja wyników, 3. Wnioski