



## SIECI NEURONOWE I REGRESJA WIELORAKA – CZYLI JAK OKIEŁZNAĆ ZŁOŻONOŚĆ W BADANIACH NAUKOWYCH?

*Maciej Szaleniec, Instytut Katalizy i Fizykochemii Powierzchni PAN w Krakowie*

Od wieków znajdowanie zależności było i jest jedną z podstawowych metod poznawania świata. Poprzez doszukiwanie się korelacji pomiędzy pewnymi procesami, zbiorami danych i obserwacji staramy się zrozumieć otaczający nas świat. Gwałtowny postęp nauki rozpoczął się dopiero wtedy, gdy badacze zaczęli stosować metodę naukową, która z oderwanych obserwacji tworzy spójne zbiory danych, a następnie odnajduje istniejące pomiędzy nimi zależności. W wyniku tego powstają empiryczne reguły, które po dogłębnej analizie dają podstawy do sformułowania Praw Przyrody.

Wspomniana wyżej metoda naukowa jest stosowana z powodzeniem od początku nowożytnej nauki. Badacze zawsze starają się sprowadzić badane zależności do jak najprostszej postaci matematycznej – z reguły do postaci równań liniowych, czasem hiperbolicznych czy parabolicznych. Takie podejście ma wiele zalet, wśród których należy wymienić możliwość przedstawienia zależności na dwuwymiarowym wykresie czy w postaci prostego równania, które łatwo sobie „wyobrazić”, a co za tym idzie i zrozumieć.

Niestety nie wszystkie badane zjawiska poddają się linearyzacji albo są na tyle proste, że jesteśmy w stanie przedstawić je w niewyszukanej formie matematycznej. Czasami odpowiednia formuła istnieje, ale jest nam nieznana i odkryciu jej moglibyśmy poświęcić wiele lat naszej pracy. Coraz częściej również badacze napotykać zjawiska i procesy zależne jednocześnie od wielu parametrów, które w różnym stopniu wpływają na obserwowane zjawisko (dobrym przykładem jest choćby pogoda, która zależy od bardzo wielu parametrów, jak: rozkład temperatury, wilgotności, ciśnienia itp.; pogody nie da się „obliczyć” prostym równaniem liniowym). Co więc ma zrobić chemik chcący opisać złożoną reakcję? Fizyk badający skomplikowane zjawiska jądrowe? Lekarz starający się opisać czynniki ryzyka zapadalności na chorobę? Socjolog starający się zrozumieć nietypowe zachowania zbiorowości ludzkiej? Na szczęście w sukurs przychodzą nowoczesne techniki statystyczne, takie jak regresja wieloraka i sieci neuronowe.

### **Podstawą są obserwacje – czyli dane**

Punktem wyjścia do poszukiwania zależności jest oczywiście rzetelne zgromadzenie danych. Z reguły jesteśmy w stanie dokonać fundamentalnego podziału zebranych danych



na te, które chcemy przewidywać (czyli tak zwane zmienne zależne), oraz te, które posłużą nam za przesłanki – czyli zmienne niezależne.

W skrajnie prostym przypadku, gdy mamy tylko jedną zmienną zależną i jedną niezależną, wybór ten nie jest bardzo istotny, gdyż obie zmienne mogą się wymieniać „rolami”. Na przykład prosta obserwacja, że „im grubsze drzewo (zmienna niezależna), tym starsze (zmienna zależna)” – daje się łatwo przekształcić na „im starsze tym grubsze”. Druga zależność jest jednak poprawniejsza, gdyż prawidłowo identyfikuje przyczynę i skutek – w końcu to w wyniku przyrostu kolejnych słoików w miarę życia drzewa (starzenia się) staje się ono grubsze, a nie na odwrót. Należy jednak bardzo uważać, gdyż z faktu korelacji nie wynika związek przyczynowo-skutkowy. Albo dokładniej, z faktu arbitralnego wyboru pewnej zmiennej jako niezależnej nie wynika związek przyczynowo-skutkowy ze zmienną zależną. Gdybyśmy zestawili ceny samochodów, jakimi jeżdżą rodziny, z cenami ich domów, najprawdopodobniej uzyskalibyśmy całkiem dobrą dodatnią korelację – to jednak nie wysoka cena samochodu powoduje wyższą cenę domu. Przyczyna korelacji tych parametrów zależy od zewnętrznego czynnika, a mianowicie od wielkości dochodów danej rodziny.

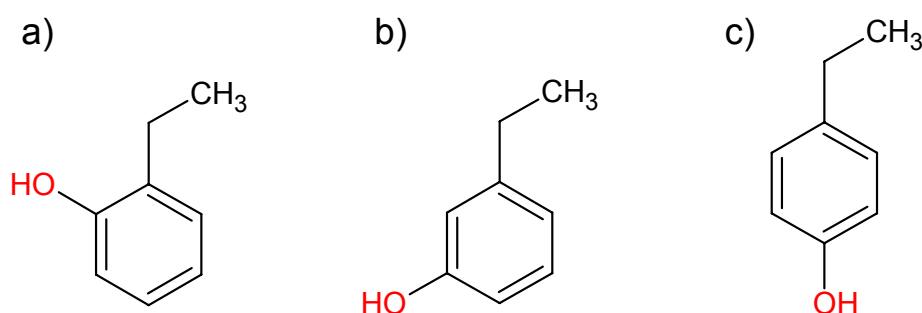
Podsumowując, znalezienie korelacji mówi tylko o współzmienności. Innymi słowy – statystyka nigdy nie zastąpi myślącego człowieka, bo to ostatecznie na drodze rozumowania i zewnętrznych przesłanek możemy zrozumieć korelacje.

Wróćmy jednak do naszych danych. Najczęściej przedstawiamy je w tabeli, gdzie poszczególne kolumny reprezentują różne zmienne, wiersze zaś przypadki. Należy bowiem zaznaczyć, że do znalezienia prawidłowej zależności potrzebujemy wielu przykładów. Im większa złożoność zjawiska (im więcej zmiennych niezależnych będziemy musieli użyć), tym więcej przypadków należy zgromadzić. Niestety nie ma jasnej reguły określającej, ile takich przypadków mieć należy. Im więcej, tym lepiej. Rygorystyczna reguła zakłada, że powinniśmy mieć 10 razy więcej przypadków niż zmiennych (w naszym przykładzie moglibyśmy więc konstruować modele tylko z jedną zmienną). Minimalna liczba przypadków musi być większa o 1 niż szacowanych zmiennych (w przeciwnym razie nie ma możliwości matematycznego rozwiązania regresji wielorakiej).

Ponieważ jestem biochemikiem, posłużę się dla zilustrowania mojego artykułu przykładem z mojej pracy badawczej – modelowaniem aktywności biologicznej enzymu. Pytanie, na jakie chciałbym znaleźć odpowiedź za pomocą statystyki, jest następujące: co powoduje, że enzym reaguje z jednymi substancjami chemicznymi szybciej, a z innymi wolniej? Postanowiłem zbadać ten problem, gromadząc dane eksperymentalne (a więc mierząc szybkości reakcji dla różnych substratów - czyli związków chemicznych, z którymi reaguje enzym) oraz opisując związki chemiczne różnymi fizycznymi parametrami pochodzącymi z obliczeń kwantowo-chemicznych oraz ze zwyczajnego opisu ich kształtu (np. licząc ilość atomów). Dla jasności tego wywodu nie będę wnikał w chemiczne szczegóły tego przykładu – dodam tylko, że sposób uzyskania zmiennej zależnej (a więc szybkości reakcji) i zmiennych niezależnych (teoretycznych stałych oraz parametrów kwantowo-chemicznych i topologicznych) był zupełnie niezależny – jedyną cechą wspólną była tożsamość danego związku chemicznego.



Do prowadzenia modelowania metodami regresyjnymi musimy operować na zmiennych, które są liczbami. Nie wszystko jednak da się opisać parametrami liczbowymi – czasem dysponujemy opisem (np. kolorem, kształtem) albo skalą jakościową (np. w wyniku badań ankietowych popularności polityka). Wtedy dane należy odpowiednio zakodować. Najlepiej posłużyć się metodą zero-jedynkową, gdyż arbitralne przypisanie większych wartości jakiejś kategorii cech (np. kolorom: czarnemu – 0, różowemu – 1, a białemu – 2) może prowadzić do artefaktów wynikających z kodowania (np. możemy uzyskać mylny związek ilościowy, w którym kolory o wyższym kodzie będą powodowały większą zmianę parametru niezależnego – a przecież wartość kodu zależy od naszego „widzimisię”). W moim przykładzie do opisu rozmieszczenia podstawnika w pierścieniu aromatycznym posłużyłem się tzw. metodą 1zN. Możliwe było obsadzenie jednej z trzech pozycji (patrz rys. 1) – każdemu potencjalnemu miejscu lokalizacji przypisałem jedną zmienną, która przyjmuje wartość zero, jeżeli w danym miejscu jest tylko atom wodoru, lub 1 jeżeli przyłączony jest tam jakiś inny, dowolny ciężki atom. W ten sposób podstawnik w miejscu para (pozycja trzecia) opisują trzy liczby: 0 (dla pierwszej pozycji), 0 (dla drugiej pozycji) i 1 (dla trzeciej pozycji). Wadą takiej metody jest mnożenie zmiennych – do prawidłowego opisu jednej cechy musiałem stworzyć trzy zmienne liczbowe.



Rys. 1. Kodowanie zmiennej jakościowej – lokaliza podstawnika OH w molekułe etylofenolu. Kodowanie poszczególnych zmiennych liczbowych metodą 1-z-N: a) 1-0-0, b) 0-1-0, c) 0-0-1.

## Macierz korelacji

Skoro mamy już zbiór danych (który powstał w wyniku naszej ciężkiej naukowej pracy) i mamy jasny pogląd na to, którą zmienną chcemy przewidywać (zmienna zależna – szybkość reakcji), należy rozpocząć analizę naszego zbioru. Bardzo często badania naukowe mogą dostarczyć nam bardzo wielu różnych parametrów, z których w jakiś sposób musimy wybrać te, które będą przydatne w naszym modelu. Najłatwiej jest zlokalizować zależności liniowe – i zgodnie z regułą Brzytwy Ockhama (*Bytów nie mnożyć, fikcyj nie tworzyć, tłumaczyć fakty jak najprościej*) właśnie od najprostszych modeli należy zaczynać. Dzięki mojej chemicznej wiedzy wiem również, że zmienną zależną muszą przedstawić w postaci logarytmicznej – będę wtedy pracować zgodnie z paradygmatem ilościowych zależności między strukturą a aktywnością. Gdybym jednak takiej wiedzy nie miał, to oczywiście poszukiwania rozpocząłbym od danych nieprzekształconych. Dla pewności więc umieszczam w swojej tabeli zarówno zmienną zależną zlogarytmizowaną, jak i zupełnie



„gołą” – taką, jaką otrzymałem w eksperymencie. Analiza liniowych korelacji za pomocą macierzy pozwala w szybki i prosty sposób spojrzeć na wszystkie (liniowe) zależności w naszym zbiorze danych, na podstawie których często możemy wiele wywnioskować na temat naszego problemu badawczego.

Macierz korelacji można obliczyć, korzystając z polecenia *Macierz korelacji* dostępnego w module *Statystyki podstawowe i tabele*. W oknie definiowania analizy *Macierz korelacji* wybieramy przycisk *Jedna lista zmiennych* i zaznaczamy wszystkie zmienne. Chcemy zobaczyć wszystkie zależności w naszym zbiorze – nie tylko korelacje poszczególnych zmiennych niezależnych z zależną. Kliknięcie przycisku *Podsumowanie: macierz korelacji* zwraca nam tabelę, gdzie nazwy zmiennych są widoczne zarówno w nagłówkach kolumn, jak i wierszy (rys. 2). Wartości w komórkach przedstawiają współczynniki korelacji liniowej Pearsona. *STATISTICA* na czerwono zaznacza tylko te zależności, które są istotne statystycznie (a więc te, gdzie test Studenta wykazał  $p < 0.05$ ). Im wartość korelacji jest bliższa 1 (lub -1), tym silniej liniowo związane są dane zmienne. Ujemny znak współczynnika korelacji oznacza zależność przeciwną – to znaczy im jedna zmienna jest większa, tym druga mniejsza (np. im dłużej się opalam, tym mniej jestem błądy). Na przekątnej tabeli mamy oczywiście rząd jedynek – to współczynniki korelacji zmiennych samych ze sobą.

Korelacje (MLR nowy model z nowymi związkami 060328.sta)														
Oznaczone wsp. korelacji są istotne z $p < .05000$														
N=12 (Braki danych usuwano przypadkami)														
Zmienna	log kcat	kcat	sigma	Es	Pi	Sr	Fh2o	Foct	Vm	MolRef	Najmniejszy Mulliken	Największy mulliken	D Mulliken	Ład met m
log kcat	1.00	<b>0.87</b>	-0.56	<b>0.62</b>	<b>-0.72</b>	-0.26	<b>-0.65</b>	<b>-0.63</b>	-0.43	-0.34	-0.56	0.36	0.53	
kcat	<b>0.87</b>	1.00	-0.55	0.44	<b>-0.59</b>	-0.12	-0.57	-0.55	-0.39	-0.33	-0.47	0.29	0.44	
sigma	-0.56	-0.55	1.00	-0.05	0.34	-0.12	0.39	0.38	-0.16	-0.22	0.45	-0.21	-0.39	
Es	<b>0.62</b>	0.44	-0.05	1.00	<b>-0.62</b>	-0.46	-0.41	-0.36	<b>-0.73</b>	<b>-0.74</b>	-0.27	0.57	0.45	
Pi	<b>-0.72</b>	<b>-0.59</b>	0.34	<b>-0.62</b>	1.00	-0.17	<b>0.92</b>	<b>0.87</b>	0.54	0.41	<b>0.89</b>	<b>-0.64</b>	<b>-0.88</b>	
Sr	-0.26	-0.12	-0.12	-0.46	-0.17	1.00	-0.33	-0.36	0.24	0.23	-0.44	0.37	0.46	
Fh2o	<b>-0.65</b>	-0.57	0.39	-0.41	<b>0.92</b>	-0.33	1.00	<b>0.99</b>	0.31	0.23	<b>0.94</b>	<b>-0.66</b>	<b>-0.92</b>	
Foct	<b>-0.63</b>	-0.55	0.38	-0.36	<b>0.87</b>	-0.36	<b>0.99</b>	1.00	0.25	0.20	<b>0.92</b>	<b>-0.66</b>	<b>-0.91</b>	
Vm	-0.43	-0.39	-0.16	<b>-0.73</b>	0.54	0.24	0.31	0.25	1.00	<b>0.96</b>	0.16	-0.32	-0.26	
MolRef	-0.34	-0.33	-0.22	<b>-0.74</b>	0.41	0.23	0.23	0.20	<b>0.96</b>	1.00	0.04	-0.41	-0.22	
Najmniejszy Mulliken	-0.56	-0.47	0.45	-0.27	<b>0.89</b>	-0.44	<b>0.94</b>	<b>0.92</b>	0.16	0.04	1.00	-0.57	<b>-0.92</b>	
Największy mulliken	0.36	0.29	-0.21	0.57	<b>-0.64</b>	0.37	<b>-0.66</b>	<b>-0.66</b>	-0.32	-0.41	-0.57	1.00	<b>0.85</b>	
D Mulliken	0.53	0.44	-0.39	0.45	<b>-0.88</b>	0.46	<b>-0.92</b>	<b>-0.91</b>	-0.26	-0.22	<b>-0.92</b>	<b>0.85</b>	1.00	
Ładunek na metynowym mulliken	-0.18	-0.20	-0.03	-0.43	0.49	-0.20	0.30	0.24	<b>0.69</b>	<b>0.64</b>	0.37	-0.34	-0.41	
Sym rozciągające [cm-1] protony aktywne	-0.31	-0.13	-0.06	-0.11	0.47	-0.11	0.36	0.29	-0.01	-0.20	0.48	-0.07	-0.35	
Najmniejszy NBO	-0.43	-0.32	0.36	-0.08	<b>0.69</b>	-0.24	<b>0.58</b>	0.50	0.13	-0.13	<b>0.71</b>	-0.10	-0.50	
Największy NBO	0.51	0.47	-0.18	0.50	<b>-0.77</b>	0.38	<b>-0.84</b>	<b>-0.85</b>	-0.48	-0.52	<b>-0.68</b>	<b>0.86</b>	<b>0.85</b>	
D NBO	<b>0.69</b>	0.51	-0.31	0.43	<b>-0.93</b>	0.40	<b>-0.93</b>	<b>-0.90</b>	-0.44	-0.35	<b>-0.86</b>	<b>0.72</b>	<b>0.90</b>	
Metynowy	0.55	<b>0.59</b>	<b>-0.75</b>	0.28	-0.28	-0.02	-0.30	-0.30	0.07	0.05	-0.30	0.34	0.36	
H-aktywne	-0.21	-0.22	0.15	-0.05	0.13	-0.02	0.03	-0.01	-0.13	-0.22	0.13	-0.05	-0.11	
C-aktywne	0.08	0.17	-0.14	-0.01	0.31	-0.16	0.34	0.32	0.03	-0.05	0.38	-0.10	-0.30	
SCF	0.08	0.10	0.17	-0.03	0.11	-0.52	0.26	0.30	-0.33	-0.17	0.24	<b>-0.75</b>	-0.51	
Dipol	0.39	0.33	-0.15	0.57	<b>-0.82</b>	0.40	<b>-0.75</b>	<b>-0.71</b>	-0.46	-0.44	<b>-0.71</b>	<b>0.90</b>	<b>0.89</b>	
LUOMO	0.16	0.01	-0.22	-0.19	-0.25	-0.09	-0.28	-0.25	0.31	0.52	-0.44	-0.38	0.10	
Homo	0.57	0.45	<b>-0.58</b>	0.19	<b>-0.81</b>	0.40	<b>-0.84</b>	<b>-0.81</b>	0.02	0.17	<b>-0.94</b>	0.51	<b>0.85</b>	
Gap	<b>-0.59</b>	-0.51	0.57	-0.30	<b>0.83</b>	-0.50	<b>0.85</b>	<b>0.82</b>	0.09	0.01	<b>0.90</b>	<b>-0.73</b>	<b>-0.93</b>	
ZPE	-0.41	-0.38	-0.13	<b>-0.76</b>	0.55	0.10	0.36	0.32	<b>0.96</b>	<b>0.98</b>	0.19	-0.54	-0.38	
MW	-0.27	-0.27	-0.22	-0.38	0.20	0.49	-0.03	-0.09	<b>0.78</b>	<b>0.66</b>	-0.10	0.33	0.23	

Rys. 2. Macierz korelacji. Zmienne korelujące w sposób istotny statystycznie są zaznaczone na czerwono. Im wyższa liczba, tym silniejsza korelacja.

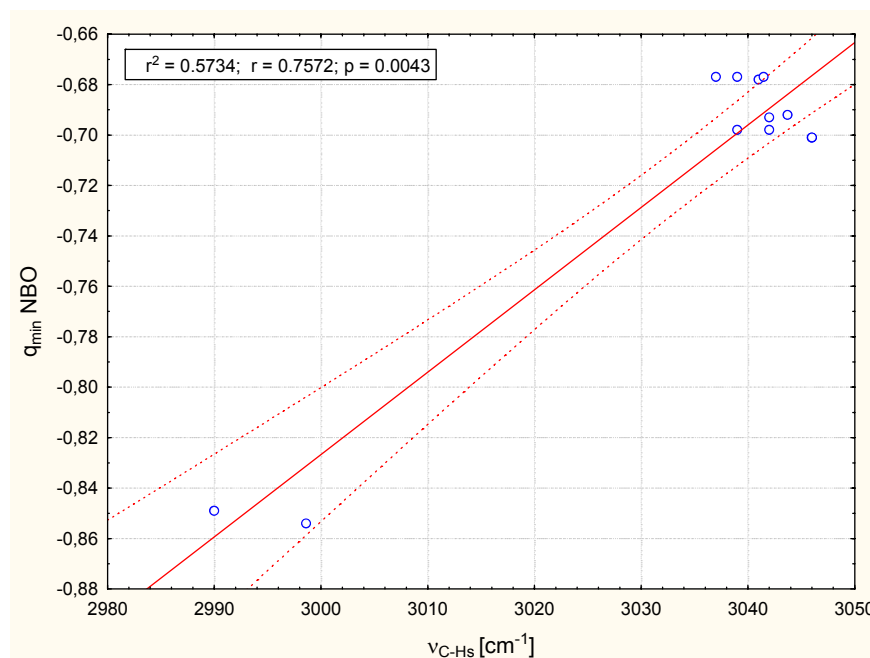
Pierwszym krokiem analizy jest lokalizacja tych zmiennych, które w istotnie statystyczny sposób korelują ze zmienną zależną. Są to parametry przydatne do budowy liniowego modelu regresyjnego. Już na tym etapie będziemy w stanie ustalić, czy problem jest prosty czy złożony – jeżeli prosty, to może znajdziemy JEDNĄ zmienną, która wysoko koreluje ze



zmienną zależną (np. 0.95). Wtedy być może warto ograniczyć się do prostego modelu regresji z jedną zmienną niezależną. Gdy jednak nie mamy tyle szczęścia (tak jak w moim przypadku, gdzie najwyższe R wynosi -0.72), wybieramy te zmienne, które wydają się nam najbardziej interesujące, i przechodzimy do dalszej analizy.

Kolejnym krokiem jest bowiem sprawdzenie, jakie są korelacje pomiędzy wybranymi przez nas zmiennymi. Może się bowiem zdarzyć, że w naszym zbiorze są zmienne liniowo zależne. Oznacza to, że w wyniku prostych przekształceń matematycznych jesteśmy w stanie uzyskać jedną z drugiej (np. wiek przedstawiony w latach i dniach) albo powiązane bardzo mocną zależnością liniową (np. cena produktu z wielkością opakowania). Ponieważ w przypadku, gdy zmienna niezależna koreluje dobrze z jakimś parametrem, to będzie również świetnie korelować ze zmiennymi zależnymi od niej liniowo, zanim przystąpimy do budowy modelu regresyjnego musimy wybrać, którą z nich się posłużymy. Błędem bowiem jest użycie ich obu, gdyż prowadzi to bardzo często do wyeliminowania obu zmiennych (ze względu na sposób testowania istotności statystycznej zmiennych niezależnych; ponadto jest to błąd metodologiczny).

Na koniec za pomocą przycisku *Jedna lista zmiennych* (w oknie analizy *Macierz korelacji*) wybieramy tylko te parametry, które na podstawie powyższej analizy uznaliśmy za potencjalnie interesujące. Wykorzystując przycisk *Dwie listy zmiennych*, wprowadzamy naszą zmienną zależną i na karcie *Więcej, wykresy* klikamy przycisk *2W Rozrzutu*. Otrzymamy serię wykresów rozrzutu (tzw. *scatter plot*), które w graficzny sposób pokazują nam otrzymane wcześniej korelacje.



Rys 3. Przykład „falszywej” wysokiej korelacji liniowej. Parametr R jest istotny statystycznie i wskazuje na dobrą korelację ( $R=0.7572$ ), tymczasem powstały trend jest artefaktem powstałym z połączenia dwóch skupisk trendem liniowym.



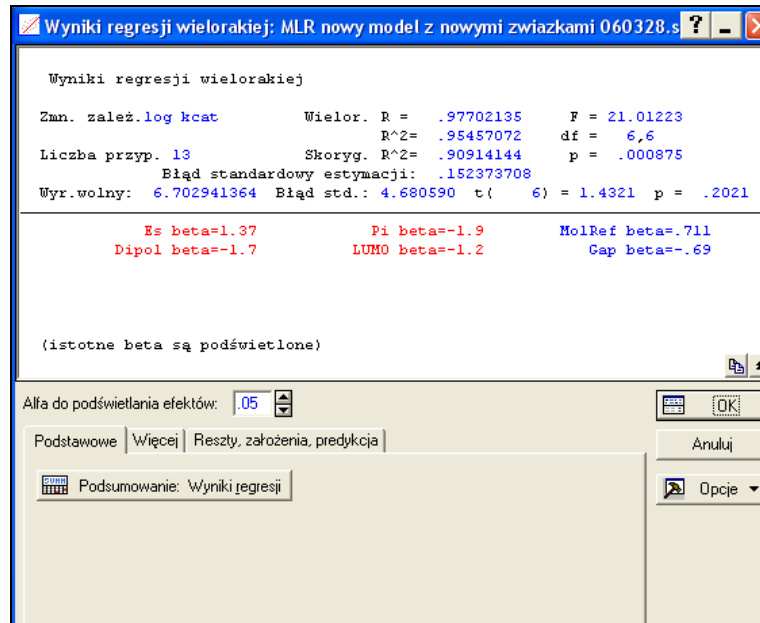
Należy teraz gołym okiem dokonać sprawdzenia, czy przypadkiem nie uzyskaliśmy istotnej wartości korelacji jedynie przypadkiem. Bardzo często lekko wygięte zależności paraboliczne lub logarytmiczne są dość dobrze opisywane również przez prostą – musimy ocenić, czy zależność jest faktycznie liniowa. Czasem wyjątkowo dobre parametry korelacji liniowej otrzymujemy w wyniku wystąpienia dwóch skupisk punktów nietworzących prawdziwej korelacji liniowej (patrz rys 3). Wszystkie powyżej wspomniane przypadki eliminują zmienną z dalszej analizy.

## Regresja wieloraka

Mając wybrane parametry, które nieźle korelują z naszą zmienną zależną, możemy przystąpić do budowy modelu regresyjnego. Zakładamy na wstępie, że wszystkie zebrane przez nas przypadki będą dobrze pasowały do modelu końcowego. Musimy się jednak liczyć z tym, że nie uda nam się znaleźć zależności doskonale opisujących wszystkie przypadki (dla złożonych problemów jest to raczej pewne).

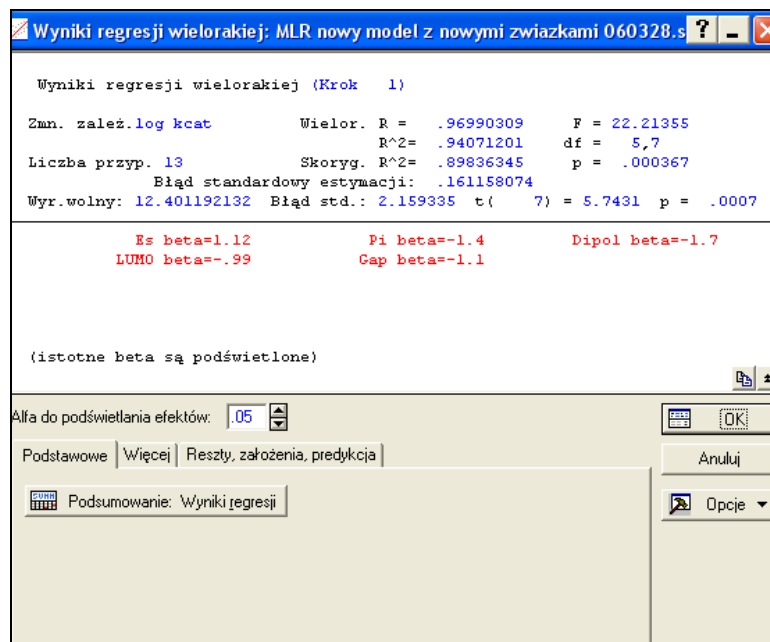
W programie *STATISTICA* dla przeprowadzenia analizy regresji wielorakiej korzystamy z modułu *Regresja wieloraka* dostępnego w menu *Statystyki*. W celu zdefiniowania zmiennej zależnej i zmiennych niezależnych korzystamy z przycisku *Zmienne* umieszczonego w oknie definiowania analizy. Następnie aby móc przeprowadzić bardziej wnikliwą selekcję zmiennych, przechodzimy na kartę *Więcej* i zaznaczamy pole wyboru *Więcej opcji*. Dzięki temu będziemy mogli nie tylko dokonywać regresji ze wszystkimi wybranymi przez nas zmiennymi, ale również uzyskać dostęp do regresji krokowej. W pierwszym etapie dokonujemy zawsze standardowej regresji z uwzględnieniem wszystkich wybranych przez nas parametrów – wszystkie zadane zmienne zostaną użyte w tworzeniu modelu liniowego. Dla przykładu wybrałem kilka zmiennych, które w istotny statystycznie sposób korelują z szybkością reakcji (tutaj oznaczonej jako  $\log k_{cat}$ ) oraz jedną, która z nią zupełnie nie koreluje. Okno *Wyniki regresji wielorakiej* (rys. 4) zwraca nam parametry statystyczne naszego modelu. Zmienne zaznaczone na czerwono są istotne statystycznie (ich związek liniowy ze zmienną zależną spełnił przyjęte kryteria alfa), natomiast niebieskie są nieistotnie statystycznie. Okno podaje również parametr  $R$  (opisujący, jak mocno dane przewidywane przez model korelują z danymi eksperymentalnymi),  $R^2$  oraz skorygowane  $R^2$ . Wiemy dobrze, że parametr  $R^2$  opisuje zasób zmienności opisaną przez model. Z tym że w regresji wielorakiej parametr ten obniżony jest w skorygowanym  $R^2$  ze względu na dodatkowe stopnie swobody wprowadzane przez kolejne zmienne (dla modelu z jedną zmienną  $R^2$  jest równe skorygowanemu  $R^2$ ). Oznacza to, że nie powinniśmy zbyt się cieszyć, widząc wysokie  $R^2$  – czasem wprowadzanie zbyt wielu zmiennych w stosunku do liczby przypadków prowadzi do nadmiernego dopasowania (*over-fitting*) i możemy uzyskać nieprawdziwy model (o  $R^2$  równym nawet 1).

Natomiast parametrem, który pozwala nam porównywać modele, jest wynik testu F Fishera – im większy, tym lepszy model.



Rys. 4. Wyniki standardowej regresji wielorakiej.

Jaki jest kolejny krok postępowania? Otóż uzyskaliśmy całkiem niezły model regresji, ale mamy nieistotne zmienne, które trzeba wyeliminować. Ponieważ jednak parametr GAP istotnie statystycznie korelował z szybkością reakcji, a teraz jest nieistotny w całym równaniu, należy zrobić to ostrożnie, posługując się regresją krokową (wsteczną). W tym celu w polu wyboru *Metoda*, zamiast *Standardowa*, wybieramy *Krokowa wsteczna*. Ta technika budowania modelu zaczyna od kompletu zmiennych i będzie wyrzucała te, które są nieistotnie skorelowane, za każdym razem testując istotność statystyczną pozostawionych w równaniu zmiennych. Pozwoli nam ona wyeliminować nieistotną zmienną MolRef (po jej usunięciu zmienna GAP stanie się istotna).



Rys. 5. Wyniki krokowej regresji wielorakiej (wszystkie parametry istotne statystycznie).



Jak widać, uzyskaliśmy równanie (rys. 5), w którym wszystkie zmienne są istotne. Parametry R są nieznacznie tylko niższe, ale F jest wyższe (co wskazuje na większą dobroć modelu).

Uzyskane równanie ma postać:

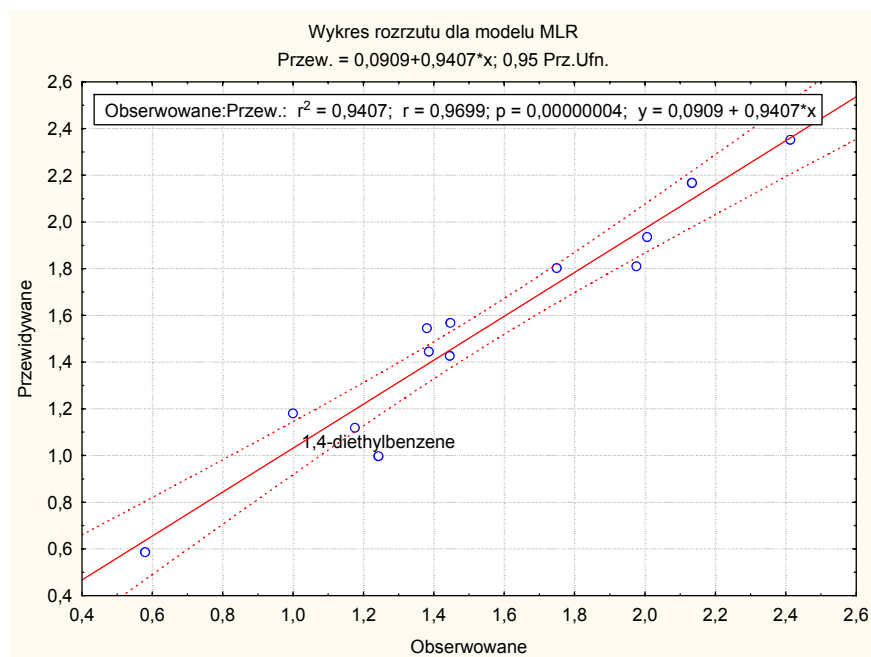
$$\log k_{\text{cat}} = 1.12 \text{ Es} - 1.4 \text{ Pi} - 1.7 \text{ dipol} - 0.99 \text{ LUMO} - 1.1 \text{ GAP} + 12.4$$

Współczynniki przed zmiennymi nie są jednak współczynnikami regresji, które znamy z normalnej regresji liniowej – są to tak zwane współczynniki beta, które zostały znormalizowane tak, abyśmy mogli porównywać wzajemną wagę parametrów. Z naszego równania wynika, że największy wpływ na szybkość reakcji ma moment dipolowy oznaczony zmienną dipol i największym parametrem beta (-1.7). Współczynniki przed którymi występuje znak minus oznaczają, że dana zmienna wpływa ujemnie na zmienną zależną (im są większe, tym w naszym przypadku wolniej reaguje dany związek), te zaś, które mają wartość dodatnią, wpływają pozytywnie na zmienną zależną (gdyby w naszym przypadku była taka zmienna, to jej większa wartość dla danego przypadku przyspieszałaby reakcję). Aby uzyskać prawdziwe stałe kierunkowe, które pozwolą nam samodzielnie obliczyć wartość  $\log k_{\text{cat}}$ , klikamy przycisk *Podsumowanie: Wynik regresji* i otrzymujemy arkusz (tabela 1), zawierający zarówno współczynniki beta, jak i współczynniki kierunkowe B.

Tabela 1. Podsumowanie regresji wielorakiej.

R= .96990309 R2= .94071201 Skoryg. R2= .89836345 F(5,7)=22.214 p						
	BETA	Błąd st.	B	Błąd st.	t(7)	poziom p
W. wolny			12.4012	2.15934	5.74306	0.000703
Es	1.12136	0.322895	0.6025	0.17349	3.47282	0.010366
Pi	-1.44882	0.376871	-0.7941	0.20656	-3.84435	0.006339
Dipol	-1.72375	0.239854	-1.3495	0.18778	-7.18666	0.000180
LUMO	-0.99386	0.203289	-58.1563	11.89563	-4.88888	0.001776
Gap	-1.14729	0.249937	-40.7784	8.88358	-4.59031	0.002513

Jak widać z poziomów p, wszystkie zmienne i wyraz wolny są istotne statystycznie. Kolejnym krokiem jest przedstawienie graficzne naszego modelu. Nie możemy wykonać klasycznego wykresu typu  $y(x)$ , bo mamy wiele zmiennych niezależnych – zamiast tego sporządzamy wykres rozrzutu wartości przewidywanej przez eksperyment (szybkości reakcji) z wartością przewidywaną przez model. Klikamy przycisk OK., przechodząc do *Analizy reszt* wybieramy opcję *Podsumowanie: reszty i przewidywane*. Z uzyskanego arkusza wykonujemy wykres rozrzutu wartości przewidywanych od obserwowanych, koniecznie zaznaczając wyliczenie parametrów statystycznych na karcie *Więcej*. Warto również wyrysować pas regresji (95% ufność), by zobaczyć, które z przypadków pozostają poza zakresem (rys. 6). W naszym przykładzie mamy jeden przypadek solidnie odstający (1,4-dietylobenzen)



Rys. 6. Korelacyjny wykres rozrzutu zestawiający wyniki eksperymentalne z danymi obliczonymi przez model regresyjny.

## Walidacja zewnętrzna

Uzyskanie modelu istotnego statystycznie nie oznacza automatycznie, że jest on prawidłowy. Czasem mimo zachowania ostrożności następuje zbyt dokładne dopasowanie modelu do danych, przez co traci on zdolność poprawnego przewidywania dla przypadków nim nie objętych. Dlatego dobrą metodą jest stosowanie tak zwanej zewnętrznej walidacji. W tym celu jeszcze przed rozpoczęciem modelowania dobrze jest wybrać (najlepiej losowo) pewną grupę kontrolną – czyli kilka przypadków, których nie uwzględnimy w zbiorze danych do budowania modelu regresji, a następnie obliczymy dla tych przypadków zmienną zależną (szybkość reakcji) za pomocą uzyskanego modelu i porównamy z wynikiem eksperymentalnym. Możemy tego dokonać w oknie *Wyniki regresji wielorakiej*, karta *Reszty, założenia, predykcje*, klikając przycisk *Predykcja zmiennej zależnej*. Wystarczy wprowadzić wartości wszystkich zmiennych, by uzyskać wynik. Jeżeli przewidywania modelu są zadowalające, możemy go wykorzystywać w naszej pracy. Należy jednak pamiętać, że liniowe modele regresyjne mają bardzo kiepskie dokonania jeśli chodzi o ekstrapolację przewidywań poza zakresy użyte w tworzeniu modelu. Tak więc jeżeli będziemy chcieli użyć naszego modelu do przewidywania szybkości reakcji jakiejś egzotycznej cząsteczki chemicznej (o bardzo odmiennych parametrach użytych w modelu), to najprawdopodobniej wynik  $\log k_{\text{cat}}$  będzie odległy od rzeczywistości.



## Interpretacja

Wartość naukowa modeli regresyjnych polega nie tyle na ich możliwościach predykcji (choć i ta jest bardzo często ważna, np. w projektowaniu leków), ale przede wszystkim w informacji naukowej, jaką możemy wyekstrahować z uzyskanego równania. Regresja wieloraka pozwala nam wyszukiwać i opisywać ilościowo złożone zależności. Dzięki nim możemy ilościowo opisać wpływ poszczególnych czynników na badane zjawisko. Musimy jednak zawsze pamiętać, że pojawienie się jakiejś zmiennej w gronie parametrów niezależnych nie implikuje związku przyczynowo skutkowego. W naszym przykładzie z dość dużym prawdopodobieństwem możemy przyjąć, że zmiana własności molekularnych związków chemicznych wpływa na ich aktywność. Sęk w tym, że nie zawsze jest to takie oczywiste – często jakaś istotna statystycznie zmienna koreluje bardzo mocno z jakimś innym parametrem, który jest prawdziwą przyczyną. Musimy pamiętać, że model regresyjny informuje tylko o współzmienności – dowód na związek przyczynowo skutkowy musi opierać się na naszym rozumowaniu lub danych zewnętrznych.

## Generalizacja, ekstrapolacja i nieliniowość – czyli dlaczego warto używać sztucznych sieci neuronowych

Jak już wspomniano we wstępie, nie wszystkie zależności są liniowe (wbrew pozorom tych liniowych jest raczej niewiele). Dzięki regresji nieliniowej jesteśmy w stanie pokonać ten problem, dopasowując różne funkcje do zbioru danych i porównując parametry statystyczne (test Fishera). Najlepszym sposobem jest ocena graficznego dopasowania trendu do punktów – nasze oko i mózg są w tym względzie daleko lepsze niż najbardziej zaawansowana statystyka. Możemy to jednak robić tylko dla wykresów dwuwymiarowych (a więc w funkcji tylko od jednego parametru niezależnego). Gdy trzeba wprowadzić więcej parametrów, poszukiwanie odpowiedniej funkcji staje się procesem bardzo żmudnym i wymagającym dużej wiedzy na temat natury badanego procesu.

Ponadto regresyjne modele liniowe zazwyczaj wykazują się słabymi zdolnościami ekstrapolacji i generalizacji. Dobrze opisują zjawiska w granicach parametrów użytych do ich budowy. Im bardziej złożony model, tym słabiej będzie sobie radził z odległą ekstrapolacją (m.in. ze względu na propagację błędów).

Istnieją jednak narzędzia modelowania, które charakteryzują się zaskakująco dobrymi zdolnościami generalizacji i ekstrapolacji (jeśli weźmie się pod uwagę ich stosunkowo prostą budowę), a ponadto nie wymagają od nas założenia a priori formuły modelowanej funkcji. Chodzi oczywiście o sztuczne sieci neuronowe (SSN). SSN są tworem informatycznym, zespołem liczb zapisanych w komórkach (tzw. neuronach), które tworzą kombinację liniową z wprowadzonych danych, poddają obliczony wynik modyfikacji przez nieliniową funkcję (bardzo często o charakterze sigmoidalnym) i następnie na tej podstawie wyliczają wynik. Ich struktura jest modułarna (możemy zestawiać neurony w różne układy – z reguły jednak stosujemy warstwy neuronów), a działanie poszczególnych neuronów jest matematycznie bardzo proste. Jednak wyróżniającą cechą sieci neuronowych jest

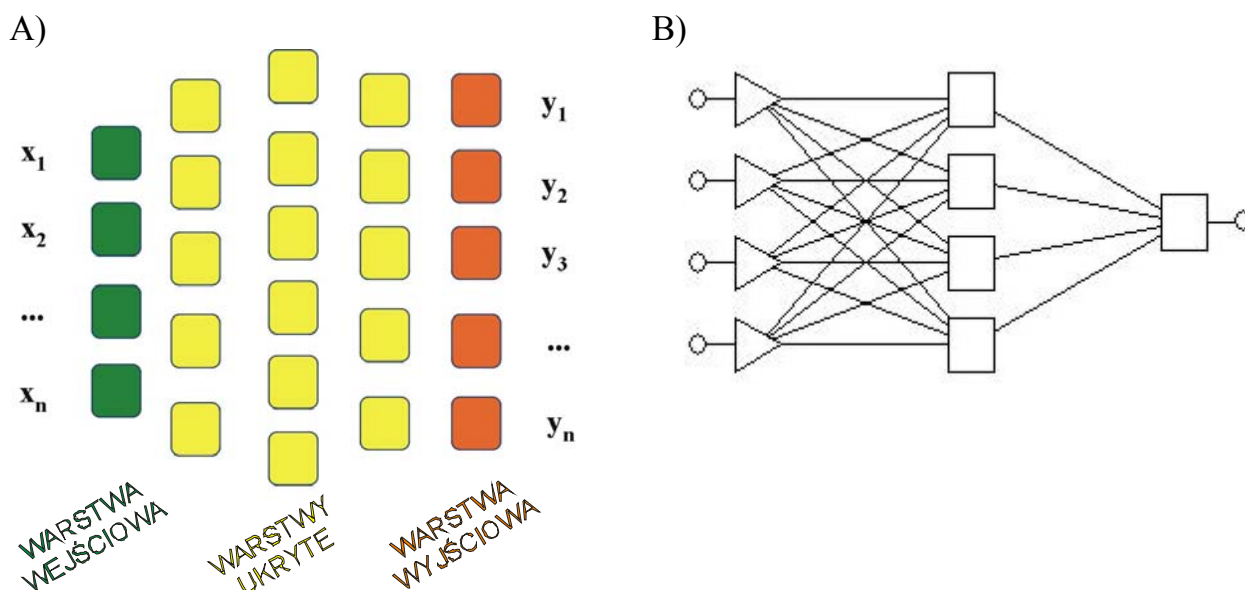


ich zdolność uczenia się, która przypomina pod wieloma względami iteracyjne dopasowywanie funkcji nieliniowej (tak samo musimy wybrać jakieś początkowe wartości parametrów, które krok po kroku dopasowywane są do danych eksperymentalnych). Sieci neuronowe „uczą się” na przykładach – tzn. tak zmieniają swoje wewnętrzne parametry (tzw. wagi), aby obliczony przez nie wynik (w naszym przypadku szybkość reakcji) był jak najbliższy rzeczywistości (a więc zupełna analogia do modelu regresyjnego). Różnica polega tylko na tym, że neurony przetwarzające nasze dane zawierają nieliniową funkcję (choć istnieją również sieci liniowe będące w dużej mierze analogami regresji liniowej wielorakiej).

Wszystko to może się wydawać trochę skomplikowane na pierwszy rzut oka, ale prawda jest taka, że nie trzeba być specjalistą od sieci neuronowych, by wykorzystywać je do modelowania. Trzeba jednak znać kilka podstawowych reguł, które uchronią nas przed błędami metodologicznymi.

## Co trzeba wiedzieć o sieciach neuronowych

W programie *STATISTICA* mamy do dyspozycji kilka rodzajów sieci. Dla jasności wywodu posłużę się tylko najczęściej używanym w zagadnieniach modelowych, tzw. perceptronem wielowarstwowym. Jak już wspomniałem sieci neuronowe składają się z warstw (a przynajmniej te, z którymi mamy do czynienia w *STATISTICA Sieci Neuronowe*), które możemy podzielić na: wejściową, ukrytą(e) oraz warstwę wyjściową (patrz rys. 7A). Każdy parametr, którego używamy w modelowaniu, zostaje zakodowany w jednym neuronie. Jeżeli więc chcemy podobnie jak w regresji zbudować model zawierający 4 zmienne, nasza sieć będzie mieć 4 neurony warstwy wejściowej. Neurony warstwy ukrytej zajmują się przetwarzaniem informacji – w każdym z nich tworzona jest kombinacja liniowa ze wszystkich neuronów wejściowych, a jej wynik poddawany jest dalszej obróbce przez funkcję nieliniową. Z praktycznego punktu widzenia najlepiej jest stosować sieci z jedną warstwą ukrytą. Liczba neuronów ukrytych zależy od stopnia komplikacji zadania. Im jest ono trudniejsze, tym musi być ich więcej. Ale uwaga – gdy jest ich za dużo, sieć będzie uczyć się „na pamięć” zamiast prawidłowo modelować naszą funkcję. Ich liczbę najczęściej dobieramy eksperymentalnie. Na koniec są neurony wyjściowe. Ich liczba zależy od tego, w jakiej postaci sieć zwraca wynik. Jeżeli używamy sieci do regresji (analogicznie jak w regresji liniowej wielorakiej) – będziemy mieć tylko jeden neuron, który dostarczy nam liczby przewidywanej przez model (w naszym przypadku szybkości reakcji). Możemy jednak postawić przed siecią problem klasyfikacyjny (np. aby sieć dzieliła nasze związki na takie, które w ogóle nie reagują, reagują słabo i dobrze) – definiując kategorie, do których zaliczane będą poszczególne przypadki. Wtedy na każdą kategorię przypada jeden neuron.



Rys. 7. A) Schemat przedstawiający typy neuronów w perceptronach wielowarstwowych; B) Schemat budowy perceptronu trójwarstwowego wraz z połączeniami między neuronami.

Jak już wspomniałem, sieci uczy się na przykładach – a to oznacza, że musimy z naszego zbioru wyznaczyć podgrupę przypadków uczących (na nich będziemy trenować sieć), przypadków walidacyjnych (na nich będziemy sprawdzali, czy sieć się nie uczy na pamięć i kiedy przerwać uczenie) oraz przypadki testowe (stanowiące ostateczny test jakości sieci). *STATISTICA* zaproponuje nam automatycznie dość dobry podział na trzy podzbiory (z reguły w proporcji 2:1:1), ale zasada jest taka, że najwięcej przypadków umieszczamy w zbiorze uczącym, natomiast pozostałe zbiory muszą mieć na tyle dużą liczbę przypadków, by test wyszedł reprezentatywny. Oczywiście, gdy mamy bardzo mało przypadków (co zdarza się w badaniach naukowych dość często), pozostaje nam tak dobrać proporcje, by system neuronowy potrafił się zoptymalizować, a nasze grupy testowe i walidacyjne były wciąż wiarygodne. W skrajnym przypadku można zrezygnować z grupy testowej.

Parametrami, które mówią nam o jakości sieci, są błędy predykcji dla grupy przypadków uczących, walidacyjnych i testowych – a więc: *Błąd ucz.*, *Błąd walid.* oraz *Błąd test.* Generalnie im mniejsze są błędy, tym lepsza jest sieć i w dobrze dopasowanych sieciach z reguły *Błąd ucz.* będzie mniejszy niż pozostałe dwie kategorie błędów (bo sieć się dopasowuje właśnie do przypadków uczących). Istotne jest, aby błędy walidacyjne i testowe nie były znacząco większe od błędów uczących (np. o rząd wielkości), gdyż taki objaw wskazuje na słabą zdolność sieci do generalizowania (sieć nauczyła się na pamięć wszystkich przypadków uczących i nie radzi sobie z nowymi problemami). Ważne jest również, aby błąd testowy porównywalny był z błędem walidacyjnym (jeśli różnią się znacząco, oznacza to, że przypadki zostały dobrane w sposób niereprezentatywny).

Na koniec należy wiedzieć, że sieci uczone są za pomocą różnych algorytmów przez różną ilość epok (to po prostu takie iteracje). Np. w automatycznym projektancie sieci standardową procedurą jest 100 epok „wstecznej propagacji błędów” (najczęściej używany



algorytm), po którym następuje 20 epok optymalizacji metodą „gradientów sprzężonych”. W miarę nauki (kolejnych epok) sieć poprawia swoje wyniki, uzyskując coraz mniejsze błędy w zbiorze uczącym i jednocześnie walidacyjnym (jeżeli jest dobrany w sposób reprezentatywny), co możemy graficznie śledzić na wykresie (włączając *Interakcyjne uczenie* w zakładce *Interakcyjne w Projektancie sieci użytkownika*). Gdy sieć zaczyna się przeuczać (zbyt dobrze dopasowywać do danych uczących), następuje wzrost błędu walidacji przy jednoczesnym dalszym obniżaniu błędu uczenia. Wtedy należy przerwać proces uczenia, a *STATISTICA* automatycznie cofnie się do optymalnej epoki, w której nasza sieć była najlepsza.

Oprócz błędów *STATISTICA* opisuje jakość sieci za pomocą *Jakości* (odpowiednio *ucz.*, *wal.* oraz *test.*), jednak parametry te są inaczej zdefiniowane dla sieci klasyfikacyjnych i regresyjnych, przez co dla niespecjalisty mogą być mylące – lepiej oprzeć się więc na błędach, które zasadniczo możemy bezpośrednio porównywać pomiędzy różnymi typami sieci.

## Jak ugryźć problem – czyli jak wyhodować sobie sieć neuronową?

Generalnie aby uzyskać model neuronowy badanego przez nas zjawiska postępujemy według bardzo podobnego schematu jak w przypadku modelu regresyjnego. W pierwszym kroku bowiem musimy wybrać zmienne, których użyjemy do modelowania. Ponieważ jednak zakładamy, że model jest nieliniowy (w końcu jeżeli potrzeba sieci, oznacza to, że liniowy model regresyjny nie był zadowalający), nie ma bezpośredniego sposobu na przetestowanie przydatności parametrów (takiego jak macierz korelacji). Powinniśmy jednak starać się wycofać ze zbioru zmienne liniowo zależne – czyli będące kombinacją innych zmiennych, lub takie, które są bardzo silnie skorelowane – a przez to nie niosą istotnie nowych informacji na temat badanego zjawiska. W ten sposób ułatwimy pracę naszej sieci.

Program *STATISTICA Sieci Neuronowe* dostarcza nam dwóch narzędzi służących do wyboru zmiennych. Pierwszym jest *Dobór cech*, zaś drugim *Automatyczny projektant sieci* wraz z opcją *Wybierz podzbiór spośród zmiennych niezależnych*. Moduł *Dobór cech* udostępnia nam znane z regresji wielorakiej algorytmy krokowego (wstecznego i postępującego) doboru zmiennych wejściowych oraz algorytm genetyczny. Wszystkie trzy metody zwracają nam tabelę, w której zmienne polecane do wykorzystania są oznaczone literą T. Ponadto moduł zwraca nam błąd (im mniejszy tym lepiej), który pozwoli nam zdecydować, który algorytm dał nam bardziej wiarygodne wyniki. W sekcji *Zakończenie* możemy ustawić opcję, która uruchomi nam *Automatycznego projektanta* dla wybranych przez *Dobór cech* zmiennych wejściowych.

Inną strategią jest eksperymentalna eliminacja zbędnych zmiennych. Otóż sieci neuronowe mają zdolność marginalizacji zmiennych wejściowych, które nie wpływają w istotny sposób na rozwiązanie problemu (po prostu wagi od ich neuronów są bliskie zera). Twórcy programu wykorzystali tę cechę w *Automatycznym projektancie*, dla możliwości wyboru podzbioru zmiennych niezależnych. Otóż możemy przebadać po prostu wiele sieci z różnym zestawem parametrów i wybrać takie, jakie będą najlepsze. Parametrem decydującym



o „dobroci sieci” jest w tym przypadku błąd dla zbioru walidacyjnego. Automatyczny projektant może przetestować tysiące sieci, za każdym razem losując początkowe wartości wag, losując rozkład przypadków, stosując zunifikowaną metodologię uczenia (zawsze taki sam algorytm) przy jednoczesnym modyfikowaniu wektora wejściowego (czyli zestawu zmiennych niezależnych).

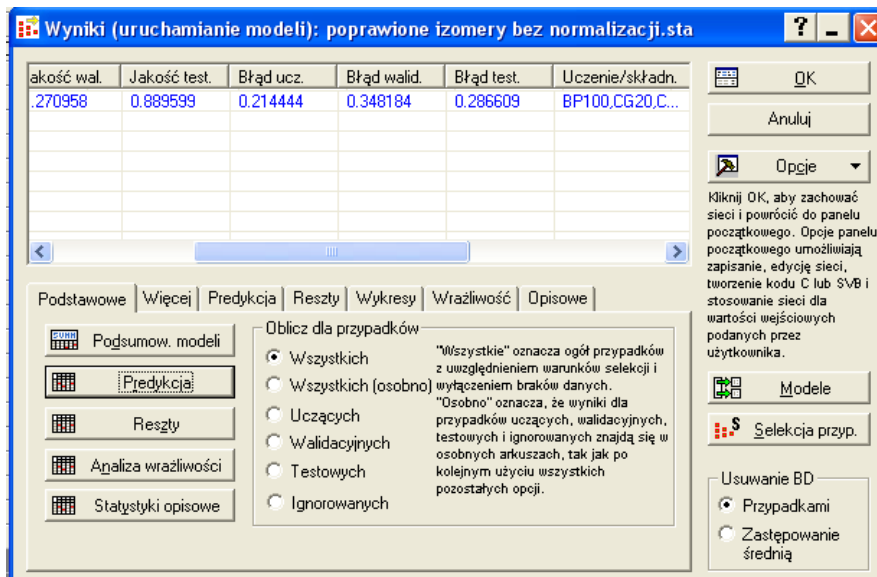
Drugim problemem jest architektura sieci neuronowej. Tutaj oczywiście najlepiej posłużyć się doświadczeniem osoby obytej w sieciach, ale z braku takiej możliwości pozostaje nam również wykorzystać *Automatycznego projektanta*. Narzędzie to można wykorzystać do wstępnej selekcji optymalnej architektury. Program *STATISTICA* oferuje nam sieci liniowe, radialne, percepcyjny trój- i czterowarstwowe oraz sieci typu GRNN i PNN. Generalnie w zadaniach regresyjnych warto rozważyć sieci liniowe (dla sprawdzenia, czy jednak model regresyjny nie byłby możliwy), percepcyjne trójwarstwowe (tzw. MLP) oraz czasem sieci radialne (RBP). Idea budowy tych ostatnich jest zupełnie inna i wdawanie się w szczegóły wykracza poza granice tego artykułu. Przeciętny użytkownik nie musi jednak znać zasady działania sieci danego typu, by wykorzystać go do testowego modelowania. Jeżeli dany typ sprawdzi się w jego problemie badawczym, zawsze będzie czas na niezbędne poszerzenie swej wiedzy w tym aspekcie.

Kolejnym krokiem jest wybranie liczby testowanych sieci (warto na początek dać np. 100, żeby zobaczyć, ile czasu potrzebuje nasz komputer do uporania się z testami, a później wpisać kilka tysięcy – z reguły nowoczesne komputery radzą sobie z takim problemem w kilka do kilkunastu minut – w zależności od wielkości zbioru przypadków). Należy również na karcie *Zachowywanie* zwiększyć liczbę zapisywanych sieci, np. do 50 lub 100, tak abyśmy mogli samodzielnie zdecydować, które spośród najlepszych sieci są rzeczywiście najlepsze. Pamiętajmy bowiem, że program kieruje się tylko ostrym kryterium minimalizacji błędu walidacyjnego – my z naszą biologiczną siecią neuronową jesteśmy lepszym sędzią.

Automatyczny projektant zwróci nam zadany zbiór sieci, który warto zapisać w zakładce *Sieci i zespoły sieci* (w głównym menu programu *STATISTICA Sieci Neuronowe*). Kolejnym krokiem jest wybór najbardziej obiecujących modeli. Należy wybrać oczywiście takie, które mają najmniejsze błędy we wszystkich grupach, ale ich błąd walidacyjny nie jest znacząco większy od błędu uczącego. Od teraz należy zacząć udoskonalać model neuronowy ręcznie (oczywiście pod warunkiem, że uzyskana jakość modelu wciąż nas nie zadowala). W tym celu wykorzystujemy opcję *Ponowne uczenie sieci* (co pozwala nam ponownie wylosować wagi oraz modyfikować algorytm uczenia, który jesteśmy w stanie śledzić na interakcyjnym wykresie uczenia) oraz opcję *Edytor modelu*, który pozwala nam ręcznie eliminować zmienne wyjściowe i neurony warstwy ukrytej. Ta ostatnia opcja jest szczególnie istotna w sytuacji, gdy mamy naprawdę dobry zestaw parametrów wejściowych, a sieć uzyskuje znacząco lepsze wyniki dla grupy uczącej przy niezbyt dobrych parametrach dla grupy walidacyjnej (co objawia się na wykresie uczenia niemożnością sprowadzenia krzywej walidacyjnej w pobliże krzywej uczącej, bez względu na stosowany algorytm). W takim przypadku należy sieć neuronową „ogłupić” poprzez wycięcie części jej neuronów w warstwie ukrytej (bo ma za dużą pojemność i „uczy się na pamięć”). Należy usuwać wewnętrzne neurony i ponownie przeprowadzać uczenie (najlepiej na takiej



samej metodzie uczenia oraz zestawie przypadków, by mieć dokładne porównanie efektów zabiegów neurochirurgicznych – taką opcję uzyskujemy po kliknięciu przycisku *Uczenie*, dostępnego na karcie *Ponownym uczeniu*, w polu *Podzbiory* (przycisk opcji *Przypisanie do podzbiorów: Takie jak w istniejącej sieci...*).



Rys. 8. Okno wyników uczenia sieci dostępne po zakończeniu uczenia lub za pomocą przycisku *Uruchom istniejący model* – umożliwia przeprowadzenie analizy wrażliwości i uzyskanie predykcji.

Racjonalnego doboru zmiennych wejściowych możemy też dokonać na podstawie analizy wrażliwości. Analizę tę dla konkretnej wybranej sieci możemy uzyskać dzięki opcji *Uruchom istniejący model* (wystarczy wybrać na karcie *Sieci i zespoły sieci* interesujący nas model, a następnie uruchomić go ponownie). Analiza wrażliwości zwraca nam dla każdej zmiennej niezależnej ranking oraz iloraz. Ranking układa nam ważność zmiennych od najważniejszej (1 miejsce) do najmniej ważnej (ostatnie miejsce). Iloraz informuje nas o tym, jak zmienia się wydajność sieci definiowana przez błąd predykcji, gdy usuniemy daną zmienną. Z reguły błąd rośnie, co jest uwidocznione przez iloraz większy od jedności. Jeżeli jednak po usunięciu danej zmiennej błąd predykcji maleje, to iloraz jest mniejszy od 1. Jest to wyraźny sygnał, że warto przeuczyć sieć bez danej zmiennej (co oczywiście spowoduje nowy rozkład wag i zupełnie inny rozkład ilorazów i rang).

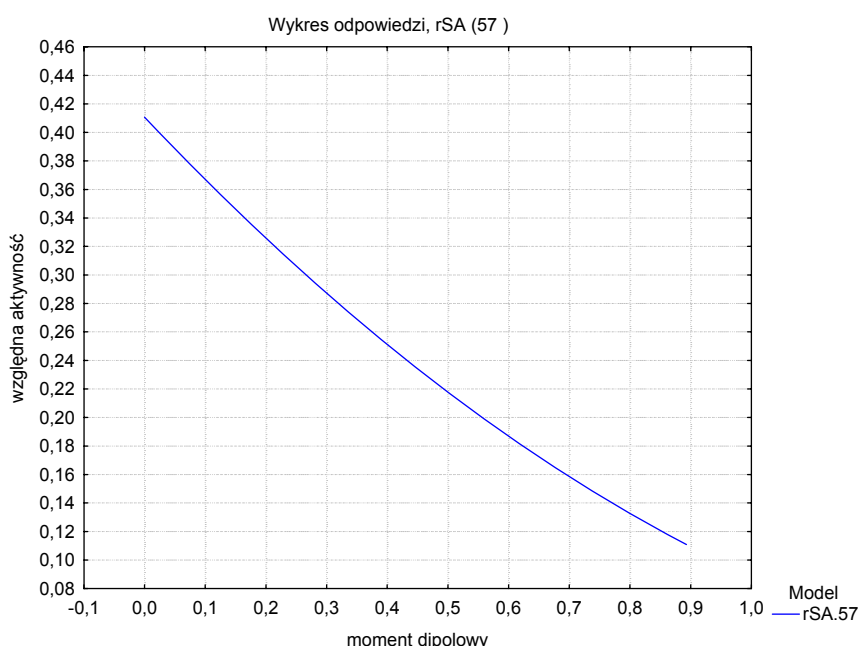
Ostateczną weryfikacją naszego modelu nie są jakieś „tajemnicze” błędy tylko konkretna predykcja (klikamy *Predykcja*). Warto na karcie *Predykcja* wybrać opcję *Przypisanie do zbioru*, co pozwoli nam zorientować się, jak sieć radzi sobie w zbiorze testowym i walidacyjnym. Najlepiej jest wykonać wykres rozrzutu analogiczny do tego, jaki stosowaliśmy przy modelach regresyjnych. Jeżeli otrzymujemy wysoką korelację oznacza to, że nasz model jest dobry. Jednocześnie jeżeli poza przedziałem ufności nie leżą punkty ze zbioru walidacyjnego i testowego, mamy przeprowadzoną walidację zewnętrzną (gdyż model działa dla przypadków, których nie użyliśmy do jego optymalizacji).



## Jak interpretować model neuronowy

Interpretacja modelu nieliniowej sieci neuronowej jest znacznie trudniejsza niż w przypadku wieloczynnikowego modelu regresyjnego. Praktycznie rzecz biorąc, nie sposób bezpośrednio odgadnąć roli poszczególnych neuronów w uzyskanym wyniku. Należy sobie zdawać sprawę, że modele regresyjne mają w tym względzie dużą przewagę nad neuronowymi – są łatwe w interpretacji, choć mają mniejsze zdolności predykcji. Nie oznacza to jednak, że nie możemy starać się badać wpływu poszczególnych zmiennych na predykcje modelu neuronowego.

Założenie, jakie musimy zrobić na wstępie, jest następujące: „dobra sieć neuronowa odzwierciedla modelowane zjawisko – a więc badając model neuronowy, badamy zjawisko” (oczywiście takie założenie stawiamy również w przypadku modelu regresyjnego). Następnym krokiem jest przebadanie, jak wpłynie na predykcję naszej zmiennej niezależnej podanie przypadków o zmienionych parametrach – tutaj możemy pofolgować naszej wyobraźni badawczej. Wystarczy bowiem zapisać sieć, przygotować nowy zestaw przypadków (nowe dane) i ponownie uruchomić moduł sieci. Sieć wykona predykcję, używając zmodyfikowanych danych wyjściowych, dostarczając nam wiedzy na temat naszego zjawiska. W naszym przykładzie można np. ocenić wpływ momentu dipolowego poprzez stworzenie klonów takiej samej cząsteczki (wszystkie takie same parametry), ale różne momenty dipolowe (od małego do dużego) i porównanie zmian szybkości reakcji.

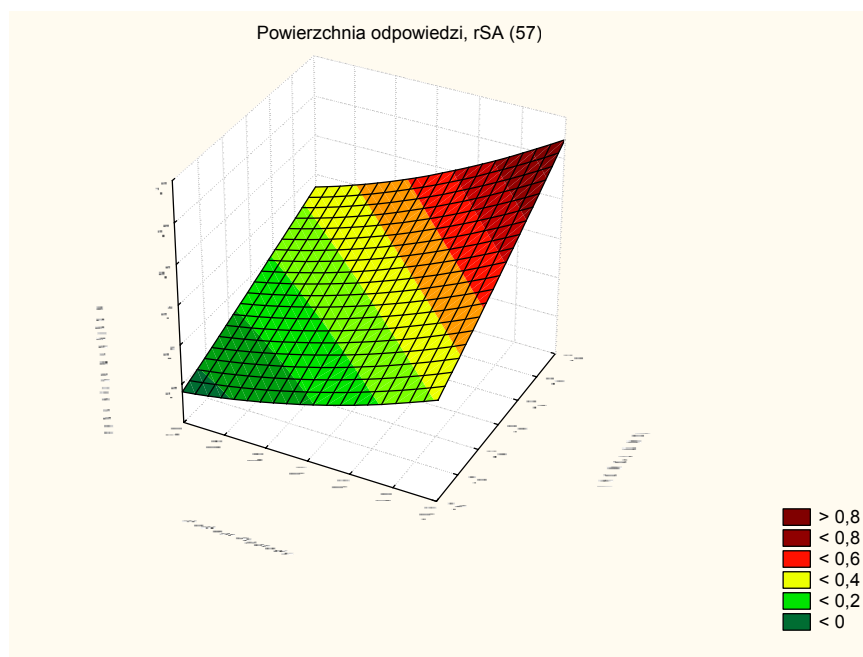


Rys. 9. Wykres odpowiedzi dla momentu dipolowego. Dla małych wartości momentu dipolowego model przewiduje wyższe względne aktywności.

*STATISTICA* dostarcza też innych, trochę bardziej wygodnych metod analizy modelu neuronowego, a mianowicie *Wykresów odpowiedzi* i *Powierzchni odpowiedzi* (dostępnych na karcie *Więcej* w oknie *Wyniki: uruchamianie modeli*). Za pomocą wykresów odpowiedzi możemy przedstawić, jak zmienia się wielkość przewidywana w funkcji wybranej



zmiennej niezależnej (w naszym przykładzie przedstawionym na rys. 9 jako zmienna niezależna został wykorzystany moment dipolowy – gdy jest mały, mamy większą aktywność, gdy jest duży, szybkość reakcji spada). Z kolei powierzchnie odpowiedzi pozwalają nam obserwować zmianę wielkości przewidywanej w funkcji dwóch wybranych parametrów (rys. 10 – na osi z względna aktywność, na osi  $x$  moment dipolowy, na osi  $y$  różnica między energiami orbitali HOMO i LUMO).



Rys. 10. Powierzchnia odpowiedzi dla dwóch zmiennych (moment dipolowy oraz  $GAP=HOMO-LUMO$ ).

Jeżeli zestawimy wykresy odpowiedzi z rangami dostarczanymi przez analizę wrażliwości, jesteśmy w stanie ustalić, które parametry, w jakim stopniu, i w jaki sposób wpływają na wartość naszej zmiennej zależnej. Uzyskujemy więc informację komplementarną do tej, jakiej dostarczała nam analiza modelu regresyjnego.

Niezależnie jednak od tego, czy stosujemy sieci neuronowe czy modele regresyjne, należy pamiętać, że w rzeczywistym zjawisku rzadko istnieje możliwość izolowanej zmiany jednego parametru. Jeżeli np. zmieniamy moment dipolowy cząsteczki (w wyniku zmian jej budowy, wprowadzając bardziej elektroujemne atomy), następuje zmiana wielu innych parametrów uwzględnianych przez nasz model (np. rozkładu ładunku, poziomów orbitalnych itd.). Dlatego w wieloczynnikowym modelu należy dokonywać analizy w sposób ostrożny. Tylko jeśli model opiera się o zmienne w stu procentach niezależne (nie ma żadnej istotnej korelacji między nimi, ani liniowej, ani nieliniowej), możemy dyskutować wpływ ich zmienności na parametr przewidywany w sposób całkowicie niezależny.



## Podsumowanie

Dzięki zaawansowanym metodom statystycznym możemy modelować znacznie bardziej skomplikowane procesy niż miało to dotąd miejsce. Nie musimy się ograniczać do modeli liniowych z jedną zmienną. Im bardziej model jest skomplikowany, tym w oczywisty sposób dokładniej jest w stanie odzwierciedlać złożoność otaczającej nas przyrody.

Należy jednak równocześnie pamiętać, iż im bardziej złożony jest model, tym trudniejsza jest jego interpretacja, a przez to przydatność dla naukowca. Sieci neuronowe są tu doskonałym przykładem – ich złożoność uniemożliwia objęcie całego zjawiska z rzutu oka i przedstawienia ich interpretacji np. w krótkim wystąpieniu konferencyjnym. Dlatego modelując zjawiska, zawsze należy szukać najprostszego modelu, czasem poświęcając dokładność predykcji na rzecz jasności. Z drugiej strony bardzo często, w szczególności w dziedzinie nauk stosowanych, istnieje zapotrzebowanie na dobre systemy przewidywania złożonych zjawisk i procesów – w takich przypadkach sieci neuronowe spisują się doskonale dzięki swoim zdolnościom poprawnej interpretacji przypadków nowych i niestandardowych. Dlatego tam, gdzie potrzebny nam jest sprawny system wspomagający naszą decyzję, warto zastosować sieć neuronową – zgodnie z zasadą, że „lepsza inteligencja sztuczna niż żadna”.

## Literatura

1. M. Szaleniec, M. Witko, R. Tadeusiewicz, J. Goclon, 2006, Application of artificial neural networks and DFT-based parameters for prediction of reaction kinetics of ethylbenzene dehydrogenase, *J. Comp-Aid. Mol. Des.* 20, 145-157.
2. M. Szaleniec, R. Tadeusiewicz, A. Skoczowski, 2006, *Optymalizacja modeli neuronowych na przykładzie oceny aktywności biologicznej związków chemicznych*, *Computer Methods Mater. Sci.*, 6 65-80.
3. M. Szaleniec, R. Tadeusiewicz, M. Witko, 2008, *The selection of optimal neural models for forecasting of biological activity of chemical compounds*. *Neurocomputing*, in press, doi:10.1016/j.neucom.2008.01.003.
4. A. Stanisław, 2007, *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach medycyny*, tom II. Modele liniowe i nieliniowe, Statsoft Polska, Kraków.
5. R. Tadeusiewicz, 1993, *Sieci neuronowe*, Akademicka Oficyna Wydawnicza Warszawa.