Prognozowanie i symulacje

Wykład 3:
Ekstrapolacja trendów za pomocą narzędzi graficznych programu STATISTICA

Tematyka wykładu

- **dopasowanie modelu trendu do danych;**
- **wybrane rodzaje modeli trendu i ich właściwości;**
- **dopasowanie modeli do danych za pomocą narzędzi wykresów liniowych (wykresów rozrzutu) programu *STATISTICA*;**
- **wyznaczanie prognoz w arkuszu danych;**
- **propozycja prezentacji graficznych;**
- **dodatkowe własności narzędzia *Dopasowanie* w programie *STATISTICA*.**

EGZAMIN!!!

Dopasowywanie modelu do danych

Intuicyjnie, stwierdzenie „dopasować model do danych jest oczywiste”. Chodzi o narysowanie pewnej funkcji (na przykład liniowej), tak aby pasowała ona „najlepiej” do pewnego zbioru punktów w układzie współrzędnych – punktów odpowiadających danym rzeczywistym.

Najczęściej chyba stosowanym kryterium optymalizacyjnym jest minimalizacja kwadratów różnic pomiędzy wartościami rzeczywistymi i dopasowanymi (będącymi wartościami optymalnej funkcji) i stąd nazwa *metoda najmniejszych kwadratów* (MNK).

Nie jest to jednak jedyny algorytm – w przypadku metody wyrównywania wykładniczego na przykład minimalizacji będzie podlegał błąd procentowy, zaś narzędzia dostępne w module *Estymacji nieliniowej* pozwalają w zasadzie w dowolny sposób określić sposób mierzenia różnicy pomiędzy danymi rzeczywistymi, a ich modelem.

Najprostsze modele trendów dla danych czasowych

Bez wątpienia najczęściej stosowanym modelem prognostycznym dla danych czasowych jest trend liniowy:

$$Y_t = b_0 + b_1 \cdot t$$

W modelu trendu liniowego zakłada się stałość różnic pomiędzy kolejnymi obserwowanymi wartościami (a więc stały trend spadkowy bądź rosnący).

Model w postaci funkcji kwadratowej pozwala na prognozowanie zjawisk, które charakteryzują się zmiennymi przyrostami kolejnych obserwowanych wartości, przy czym przyrosty te rosną bądź spadają liniowo:

$$Y_t = b_0 + b_1 \cdot t + b_2 \cdot t^2$$

Funkcja wykładnicza może dobrze oddawać przebieg zjawiska charakteryzującego się stałymi zmianami względnymi (a więc zmianą procentową):

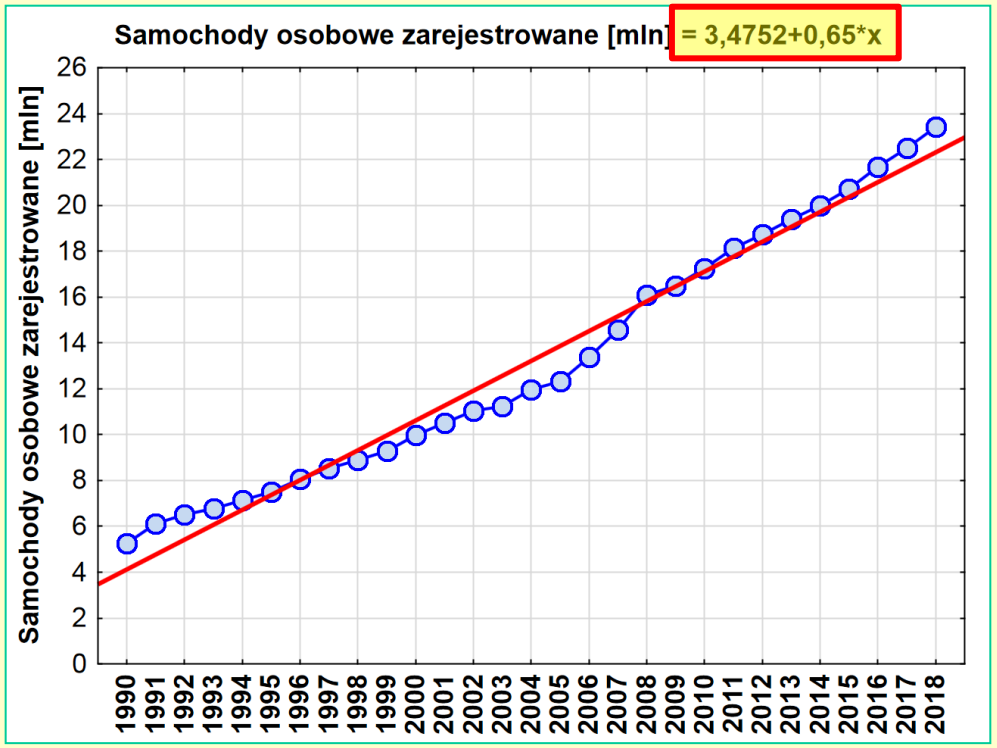
$$Y_t = b_0 \cdot b_1^t$$

Dopasowanie trendu liniowego w programie *STATISTICA*

Najprostszym sposobem dopasowania modelu liniowego (wielomianowego, wykładniczego) do danych w programie *STATISTICA* jest zastosowanie narzędzi graficznych podczas tworzenia wykresu liniowego.

W oknie *Wykresu liniowego* w zakładce *Więcej* wskazujemy rodzaj dopasowywanej funkcji i uzyskujemy zarówno graficzną prezentację jej przebiegu, jak i wzór modelu.

Obok przedstawiono trend liniowy dopasowany do danych o liczbie samochodów osobowych (plik: *Transport w Polsce 1990-2018*).



Konstrukcja prognozy za pomocą formuł arkusza danych

Aby skonstruować prognozę, wykorzystujemy wzór znajdujący się w nagłówku wykresu.

W arkuszu danych dodajemy kolumnę zawierającą numery okresów czasowych (*X*) i wypełniamy ją kolejnymi wartościami (formuła =v0) oraz zmienną *Prognoza liniowa liczby samochodów*…, której wartości wyliczamy za pomocą formuły skopiowanej z nagłówka wykresu.

Następnie dodajemy tyle przypadków, ile wynosi horyzont czasowy prognozy (przypadki warto przy tym nazwać kolejnymi latami, datami, miesiącami, etc.).

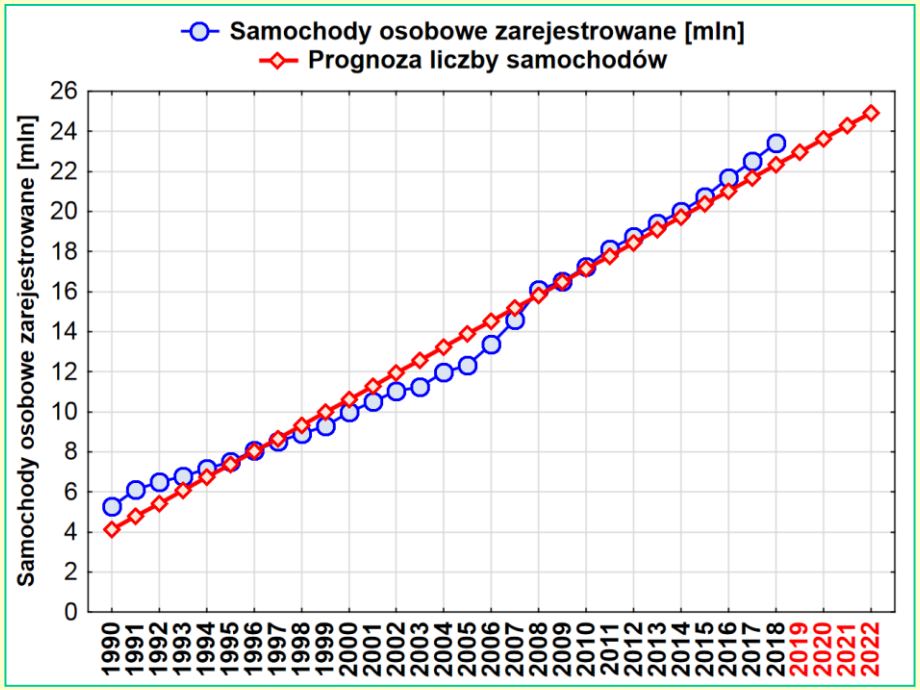
Jeżeli z jakichś powodów wyliczenie formuły nie powiodło się (w programie *STATISTICA* można je przez przypadek wyłączyć) należy wznowić automatyczne przeliczanie formuł za pomocą klawisza F9.

	16 X	17 Prognoza liczby samochodów
1996	7	8,025
1997	8	8,675
1998	9	9,325
1999	10	9,975
2000	11	10,625
2001	12	11,275
2002	13	11,925
2003	14	12,575
2004	15	13,225
2005	16	13,875
2006	17	14,525
2007	18	15,175
2008	19	15,825
2009	20	16,475
2010	21	17,125
2011	22	17,775
2012	23	18,425
2013	24	19,075
2014	25	19,725
2015	26	20,375
2016	27	21,025
2017	28	21,675
2018	29	22,325
2019	30	22,975
2020	31	23,625
2021	32	24,275
2022	33	24,925

Propozycja wizualizacji danych i prognozy

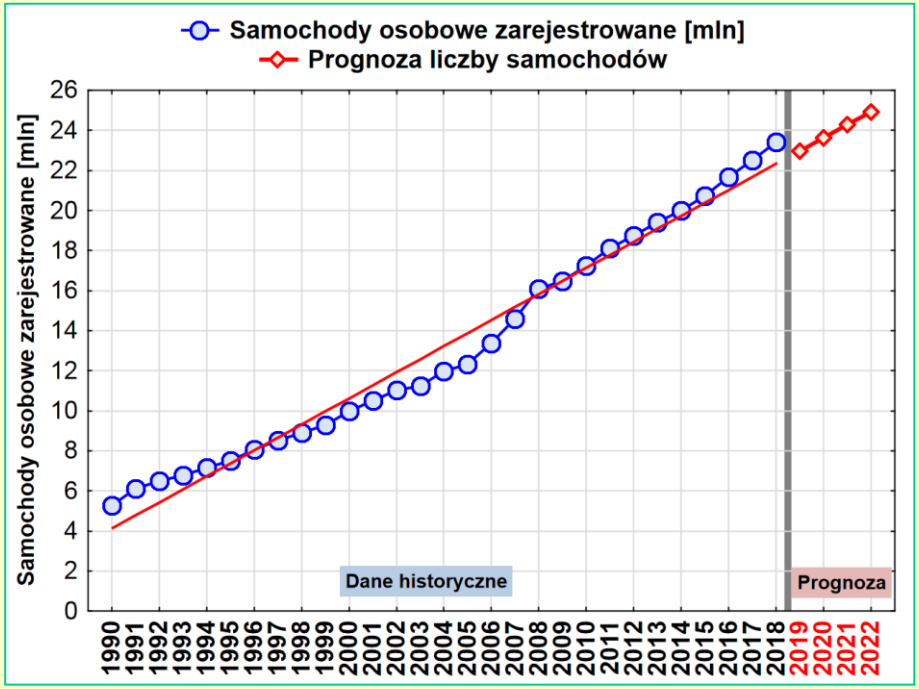
Jeżeli w arkuszu mamy już kolumny z oryginalnymi danymi oraz prognozowanymi wartościami, to poprzez zastosowanie wykresu liniowego w postaci wielokrotnej możemy zobrazować przebieg analizowanego szeregu i prognozę. Oto dwie propozycje takiego wykresu.

W wersji najprostszej…



…i w wersji nieco bardziej skomplikowanej, choć tylko pozornie. Aby stworzyć taki wykres jak poniżej, w arkuszu danych „rozdzielamy" wartości modelowane (1990-2018) od prognoz, które kopiujemy do nowej kolumny, po czym sporządzamy wykres wielokrotny dla trzech zmiennych.

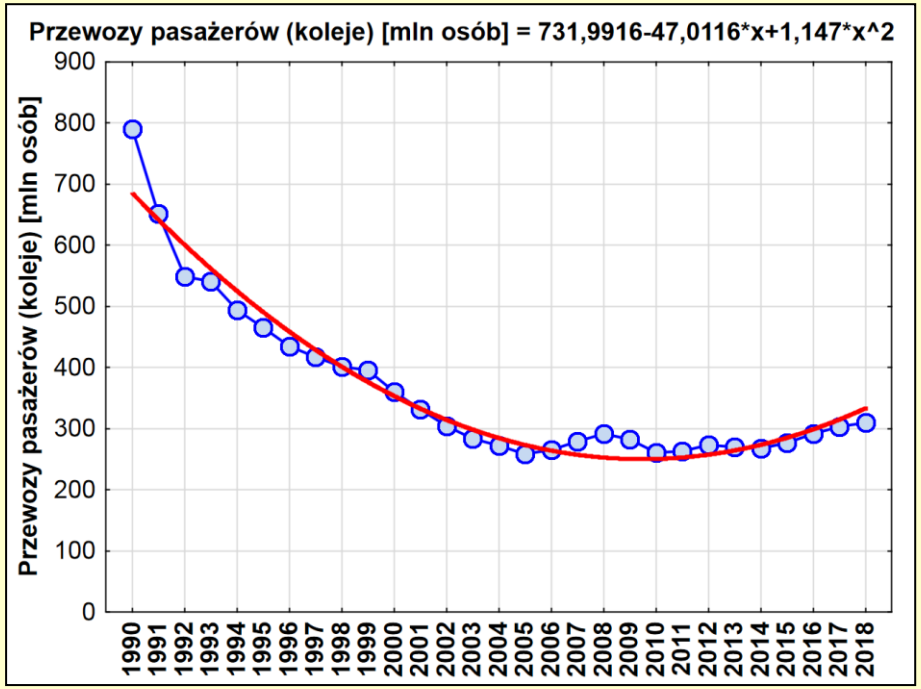
Reszta polega na dodaniu linii odniesienia, wstawienia pola tekstowego „Prognoza" i „Dane historyczne"



Model kwadratowy

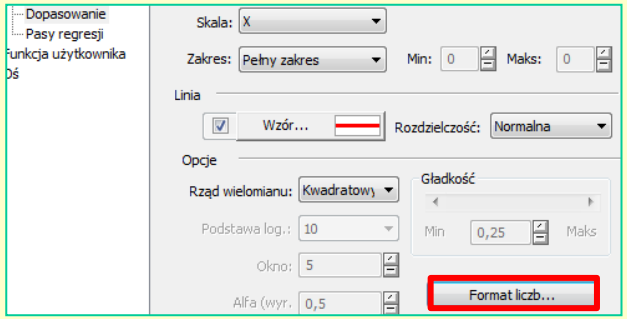
Aby zastosować model kwadratowy w zakładce *więcej* wybieramy dopasowanie w postaci wielomianu. Domyślnie jest on stopnia drugiego, natomiast można „skomplikować” postać funkcji aż do postaci wielomianu piątego stopnia.

Na poniższym wykresie przedstawiono przebieg modelu kwadratowego dla danych dotyczących przewozu pasażerów koleją w latach 1990-2018.



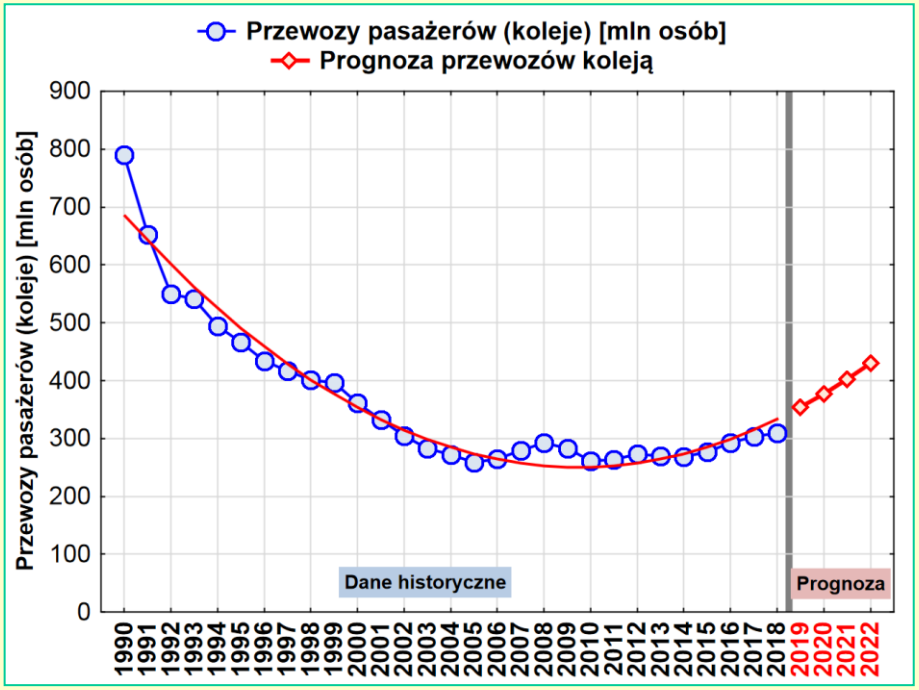
Aby zmienić dokładność wyświetlania współczynników we wzorze należy w zakładce *Dopasowanie* ustawić odpowiednio ich format.

Dla estetyki warto zmniejszyć dokładność wyświetlania, ale w przypadku złożonych wzorów większa liczba miejsc po przecinku da dokładniejsze wyniki prognoz.



Prognoza za pomocą modelu kwadratowego

Na poniższym wykresie zaprezentowano prognozę wykonaną za pomocą modelu kwadratowego dla liczby pasażerów przewiezionych kolejami na lata 2019-2022. A także wartości obliczone w arkuszu danych.



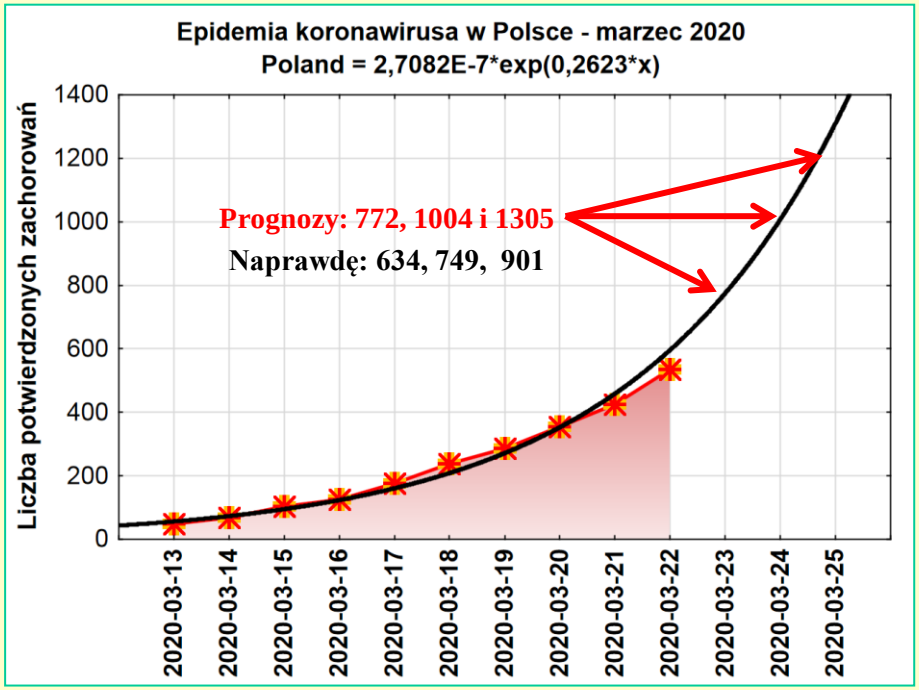
	16 X	17 Prognoza przewozów pasażerów koleją
2014	25	273,6
2015	26	285,1
2016	27	298,8
2017	28	314,9
2018	29	333,3
2019	30	353,9
2020	31	376,9
2021	32	402,1
2022	33	429,7

Model wykładniczy

W analogiczny sposób można do danych dopasować model wykładniczy, który jak to przypomniano wcześniej, jest formułą, w której zakłada się stałe tempo wzrostu (spadku) względnego badanej wielkości.

Analiza dotyczy liczby potwierdzonych przypadków zarażenia wirusem COVID-19 do dnia 22-03-2020

UWAGA: Dane pochodzą ze strony ourworldindata.org – daty dotyczą momentu wpłynięcia raportu z danego kraju do WHO, stąd ok. 2-dniowe opóźnienie w stosunku do faktycznych wartości)



Sposób zapisu równania modelu wykładniczego nie jest może zbyt szczęśliwy, dlatego do interpretacji zalecamy podniesienie liczby *e* do odpowiedniej potęgi (można to uczynić za pomocą formuły *Excela* lub funkcji *Exp* kalkulatora.

Powyższy wzór przybierze wtedy postać (exp(0,2623) = 1,300):

$$Y = 2{,}7 * 10^{-7} * 1{,}300^{x}$$

Oznacza to, że jeśli liczba przypadków będzie zgodna z powyższym modelem, dzienny przyrost wynosić będzie ok. 30%.

Wątpliwości...

1. **Po pierwsze, należy pamiętać, że modelowanie jakiegokolwiek zjawiska tylko na podstawie przebiegu historycznych danych nie umożliwia wglądu w mechanizm rozwoju tego zjawiska i jest bardzo dużym uproszczeniem.**

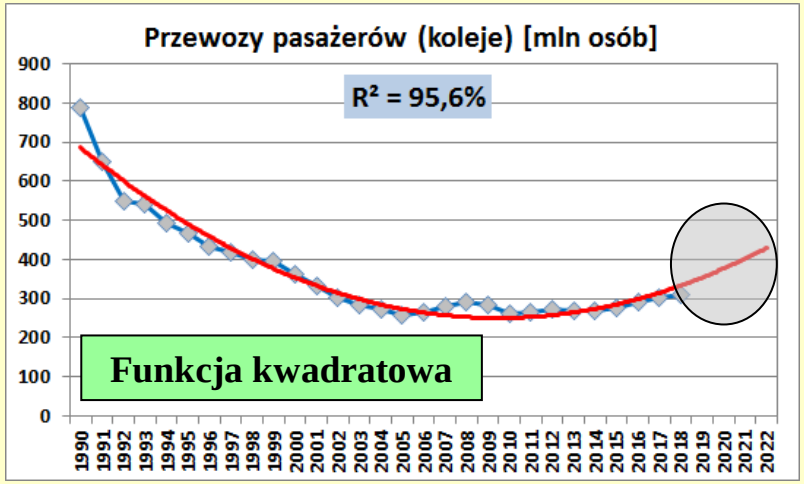
2. **Po drugie, nawet jeśli mechanizm rozwoju modelowanego zjawiska daje się przybliżyć pewną funkcją, należy pamiętać, że trafność prognoz zależy od stabilności modelu w czasie, czego nikt nie może zagwarantować – to można tylko założyć.**

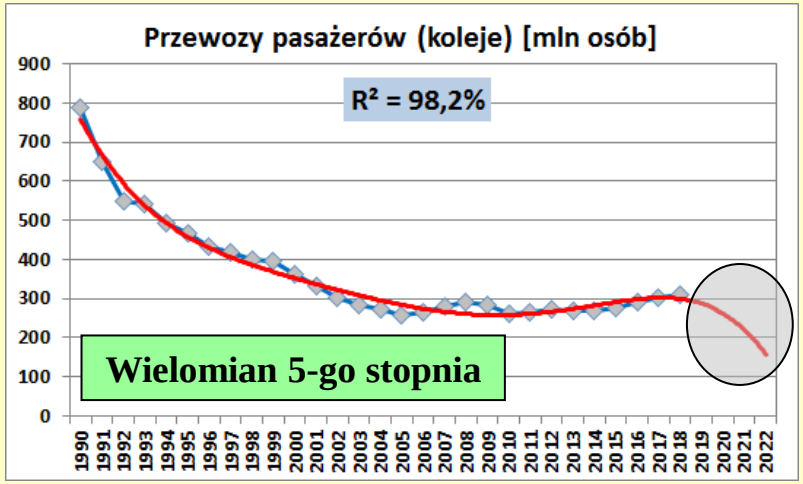
3. **Po trzecie, nie znając mechanizmu rozwoju danego zjawiska, możemy dobrać błędnie typ funkcji, a jak zilustrowano na kolejnym slajdzie wyniki modelowania zmieniają się w bardzo nieoczekiwany sposób wraz z pozornie niewielkimi zmianami typu funkcji.**

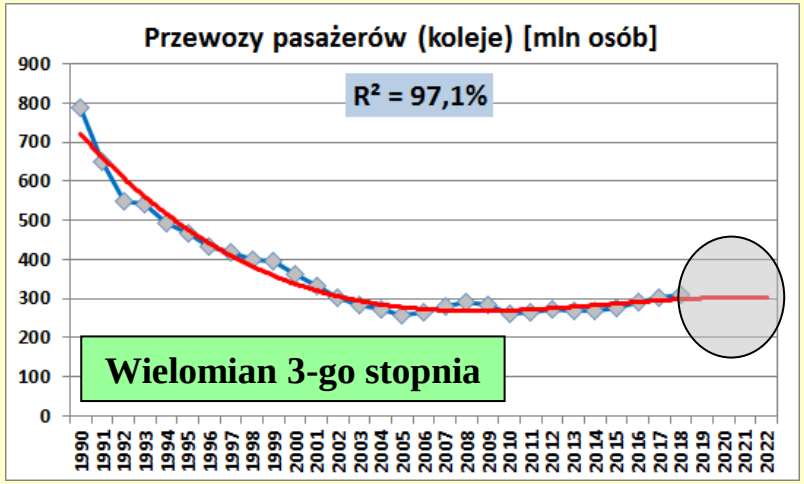
4. **Należy starać się stosować jak najprostsze funkcje, ponieważ dopasowanie bardziej złożonych funkcji do danych jest zawsze lepsze, ale prognozy zwykle nie – to również pokazano na następnym slajdzie.**

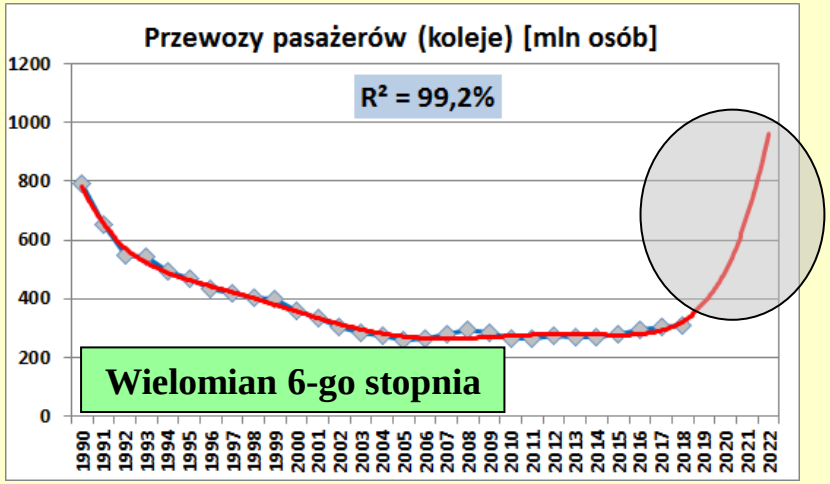
Uwaga na złożoność dobieranej funkcji!

Poniżej przedstawiono jak mocno różnią się prognozy dokonywane dla tych samych danych za pomocą wybranych modeli wielomianowych różnych stopnia (od funkcji kwadratowej do wielomianu stopnia szóstego). Jakość dopasowania do danych, mierzona tzw. współczynnikiem determinacji (R^2) wzrasta wraz ze złożonością funkcji – co jest zresztą dość oczywiste – ale wiarygodność prognoz raczej nie!









Modele oparte na części danych

Nie zawsze wszystkie obserwacje danego zjawiska w czasie są istotne dla wychwycenia trendu i sporządzenia prognozy. Można pominąć dane odległe w czasie, czasem z uzasadnionych powodów można pominąć pewne obserwacje odstające.

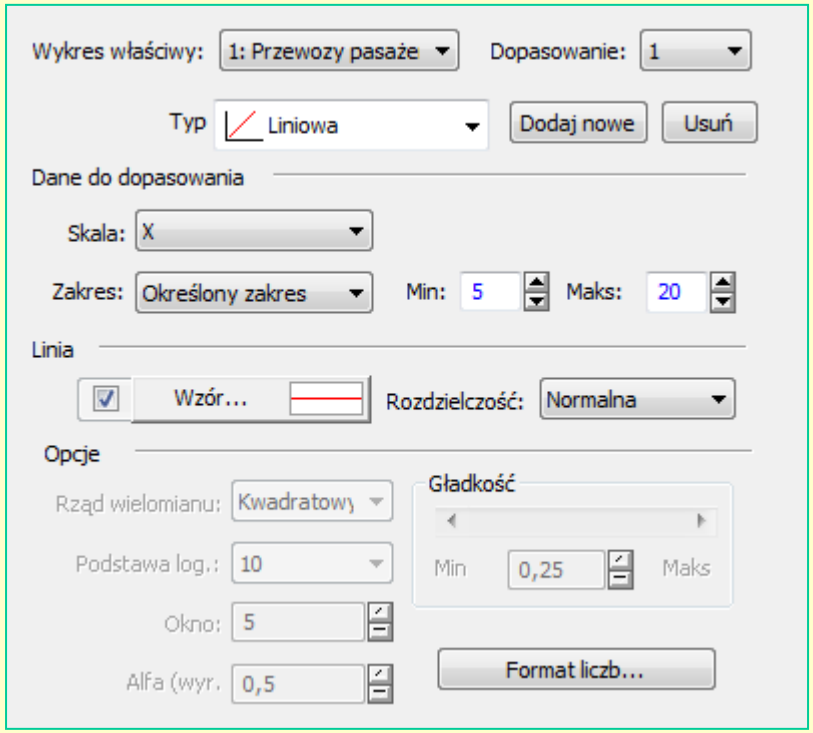
Oczywiście można dokonać takiego „ucięcia” modelowanych danych za pomocą odpowiednio skonstruowanego *warunku selekcji*, jednak istnieją też inne, bardziej intuicyjne sposoby.

Poniżej przedstawiono dwie takie propozycje:

- **wykorzystanie możliwości dopasowania modelu do pewnego przedziału (opcje narzędzia *dopasowanie*);**
- **interaktywną eksplorację wykresu za pomocą narzędzia wyróżniania.**

Zmienianie zakresu danych za pomocą narzędzie *dopasowanie*

Po sporządzeniu wykresu liniowego (i nie tylko) istnieje możliwość modyfikacji postaci sposobu dopasowania modelu do danych.

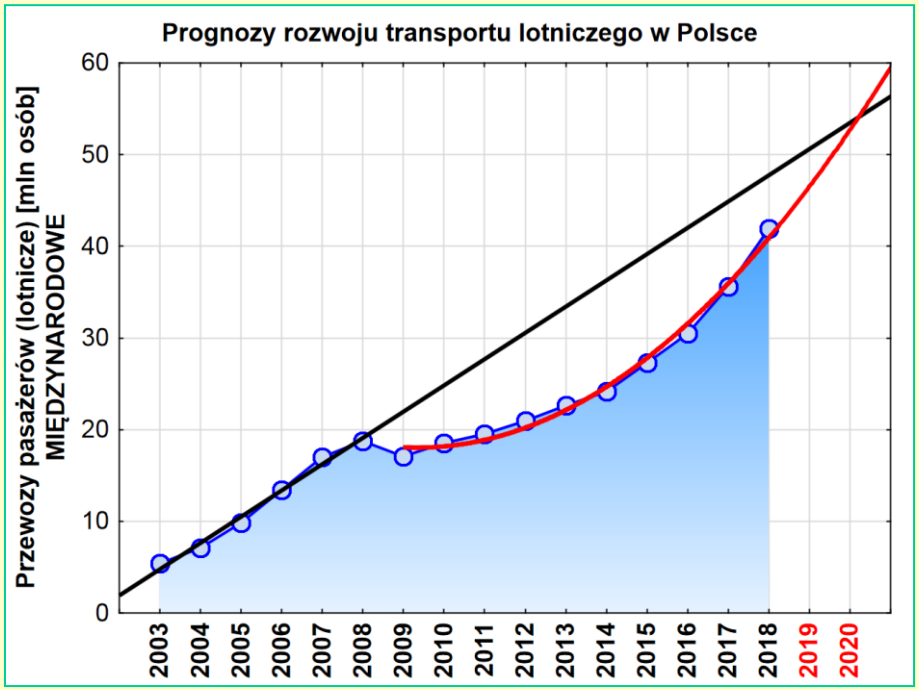


W szczególności można ustalić zakres danych (numery przypadków – od … do …), który będzie uwzględniany podczas wyznaczania modelu.

Można także do istniejącego już wykresu dodawać nowe modele, co umożliwia ich bezpośrednie porównanie.

Prognoza liczby pasażerów w transporcie lotniczym

Na wykresie przedstawiono prognozy krajowego ruchu lotniczego w Polsce sporządzane z perspektywy roku 2008 (na podstawie trendu liniowego dla danych z lat 2003-2008) i roku 2018 (na podstawie trendu kwadratowego dla danych z lat 2009-2018).

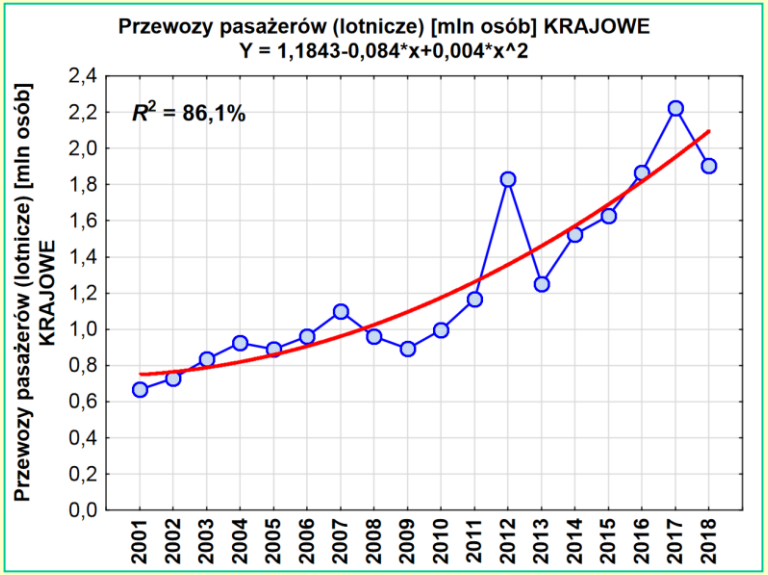


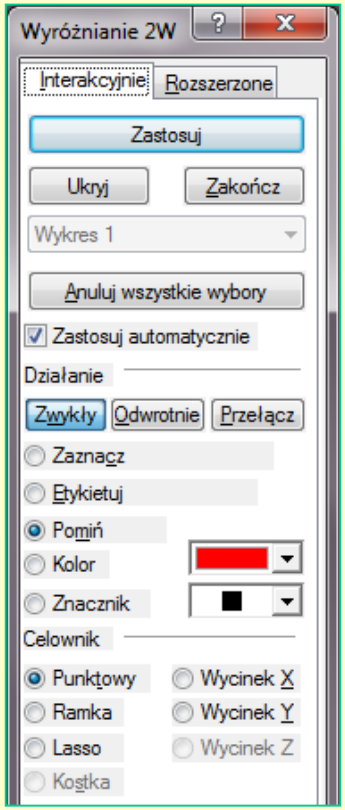
Eliminowanie obserwacji odstających

Wykorzystując narzędzie wyróżniania możemy usunąć interaktywnie z analizy dowolne obserwacje. Oczywiście, muszą za tym iść odpowiednie przesłanki merytoryczne, gdyż usuwając kolejne nie pasujące do modelu obserwacje, doprowadzilibyśmy do tego, że do każdych danych wyjściowych pasowałby w końcu model liniowy (czy innego dowolnego typu).

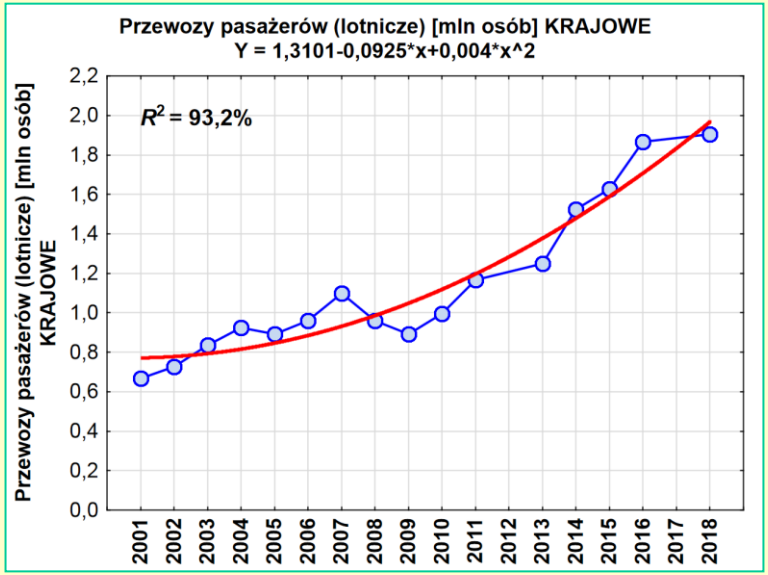
Na przebieg funkcji kwadratowej wpływają dwie obserwacje odstające – z roku 2012 (EURO) i 2017.

Aby polepszyć dopasowanie modelu do danych usuwamy te dwie obserwacje.





Po usunięciu danych z roku 2012 i 2017 polepsza się dopasowanie modelu danych (wzrost współczynnika R^2), a prognoza nie jest przeszacowana przez wystąpienie dwóch lat z nietypowo wysokim poziomem przewozów.



Inne funkcje

W programie *STATISTICA* można znacznie rozszerzyć klasę funkcji, spośród których poszukuje się modelu szeregów czasowych.

Dowolną, ogólną postać funkcji można podać w modele *Estymacja nieliniowa* i w zupełnie podobny sposób jak to było w przypadku pracy z wykresami, wykorzystać znaleziony wzór do sporządzenia prognozy.

Opis wykorzystania modułu *Estymacja nieliniowa* do prognozowania będzie tematem jednego z kolejnych wykładów.