



WYKORZYSTANIE *STATISTICA DATA MINER* DO PROGNOZOWANIA W KRAJOWYM DEPOZYCIE PAPIERÓW WARTOŚCIOWYCH

Joanna Matych

Krajowy Depozyt Papierów Wartościowych S.A.

Funkcjonowanie firmy w dzisiejszych czasach jest w większości przypadków zależne od czynników zewnętrznych, często niezależnych od właściciela, czy zarządu firmy, takich jak: uwarunkowania polityczne, prawne, podatkowe, społeczne, ekonomiczne lub nawet technologiczne. Każdy z tych czynników jest zbiorem pewnych procesów (zmiennych losowych), mających wpływ na działalność przedsiębiorstwa i osiąganych wyników finansowych. Badanie tych procesów stanowi więc klucz do dobrego zarządzania firmą i jest niemalże niezbędne w podejmowaniu właściwych decyzji. Modelowanie ekonometryczne zjawisk mających wpływ na efekty działania przedsiębiorstwa jest więc niezastąpionym narzędziem wspomagającym procesy decyzyjne i ułatwiającym osiągnięcie sukcesu.

Podejmowanie właściwych decyzji zwykle ma znaczący wpływ na kształtowanie przyszłości i na odwrót – często dokonując wyboru chcielibyśmy wiedzieć, co będzie w przyszłości i jak zmiany warunków (np. ekonomicznych czy politycznych) mogą wpłynąć na nasz wybór. Innymi słowy, chcielibyśmy zbadać nie tylko, w jaki sposób badane zjawiska kształtują działanie firmy w chwili obecnej, ale także w przyszłości. W gospodarce wolnorynkowej dobre rozpoznanie otoczenia zewnętrznego w przyszłości, a więc stworzenie jednego lub kilku najbardziej prawdopodobnych scenariuszy dotyczących czynników zewnętrznych może często ułatwić planowanie związane z wynikami finansowymi, a nawet uratować firmę przed dużą stratą finansową. Wybór tego „najlepszego” scenariusza zazwyczaj jest bardzo trudny, wymaga dużej wiedzy i doświadczenia w danej dziedzinie. Pomocne jest więc użycie do tego celu profesjonalnych narzędzi do modelowania ekonometrycznego, które na podstawie danych historycznych o procesie, czy też zbiorze procesów, sporządzą ten „najlepszy” scenariusz (prognozę) na przyszłość zgodnie z regułami rachunku prawdopodobieństwa i statystyki. Mamy więc tutaj do czynienia z analizą zbioru zmiennych losowych zmieniających się w czasie, tj. szeregów czasowych.

Modele służących do analizowania i sporządzania prognoz szeregów czasowych jest wiele: począwszy od najprostszego modelu średniej ruchomej, poprzez modele wyrównywania wykładniczego, modele ARIMA, analizę widmową. Do analizy szeregów czasowych mogą służyć także sieci neuronowe. Każdy z tych modeli ma swoje założenia, pewne wymagania



odnośnie danych wejściowych i każdy z nich sprawdza się lepiej w modelowaniu szeregów o różnych strukturach. W niniejszym opracowaniu przedstawiona zostanie próba analizy i sporządzenia prognozy dla szeregów czasowych dotyczących rynku kapitałowego w Polsce za pomocą modeli wyrównywania wykładniczego. Zaletą tego sposobu modelowania jest prostota jego działania oraz duża dokładność generowanych prognoz dla szeregów czasowych, które zbyt szybko nie zmieniają swojej struktury.

Model wyrównywania wykładniczego

Definicja modelu

Prosty model wyrównywania wykładniczego zakłada, że każda obserwacja szeregu czasowego składa się ze stałej (b) i składnika losowego (ε), czyli:

$$X_t = b + \varepsilon_t,$$

gdzie b jest stałą, która powoli może zmieniać się w czasie, a ε ma rozkład $N(0,1)$.

Stałą b wyznacza się jako pewnego rodzaju średnią, w której większe wagi przypisuje się obserwacjom nowszym, a wagi te maleją wykładniczo, zgodnie z następującym wzorem:

$$S_t = \alpha * X_t + (1 - \alpha) * S_{t-1},$$

gdzie:

X_t - wartość obserwowana w czasie t ,

S_t - wartość wygładzonego szeregu w czasie t ,

α - współczynnik wygładzania.

Taka rekurencyjna procedura umożliwi obliczanie każdej kolejnej wartości wygładzonego szeregu jako średniej z poprzedniej obserwacji i poprzedniej wartości wygładzonej, która wyliczona była również według tej samej zasady. W wyniku takiej procedury każda wartość szeregu wygładzonego jest średnią wszystkich poprzednich obserwacji, przy czym wagi maleją wykładniczo, zależnie od parametru α . Zauważmy, że jeśli $\alpha=1$, to szereg wygładzony jest taki sam jak szereg obserwowany, a jeśli $\alpha = 0$, to szereg wygładzony jest stały i równy początkowej wartości obserwowanej S_0 . Należy więc założyć: $0 < \alpha < 1$.

Konstrukcja prostego modelu wyrównywania wykładniczego sprowadza się więc do wyznaczenia parametru α oraz wartości wygładzanej startowej S_0 . Obie te wartości program *STATISTICA* wylicza automatycznie, chociaż istnieje też możliwość konstrukcji modelu przy zdefiniowanych przez użytkownika parametrach.

Powyższy model stosowany jest przy podstawowej procedurze wyrównywania wykładniczego. Jest ona użyteczna dla szeregów czasowych, w których nie występują trend i sezonowość. W rzeczywistości dość rzadko zdarzają się takie szeregi czasowe, dlatego też algorytm ten wzbogacono właśnie o te dwa składniki, które przy konstrukcji modelu dodaje się (lub mnoży) do wartości wygładzonej w prostym wyrównywaniu wykładniczym.



W przypadku gdy wartości szeregu co pewien stały okres p wzrastają o stałą wartość, np. 1 tys. PLN, to mamy do czynienia z sezonowością addytywną. W takim przypadku nasza prognoza w punkcie t przyjmuje postać:

$$\text{Prognoza}_t = S_t + I_{t-p},$$

gdzie

S_t to wartość wygładzona prostym algorytmem,

I_{t-p} – składnik sezonowości,

p – długość okresu sezonowości.

W przypadku sezonowości multiplikatywnej, tzn. kiedy wartości szeregu wzrastają co pewien stały okres p o równy % wartości, np. 20%, analogiczny wzór przyjmuje postać:

$$\text{Prognoza}_t = S_t * I_{t-p}.$$

Analogiczne wzory stosuje się w przypadku składnika trendu.

Ocena dobroci dopasowania modelu wyrównywania wykładniczego

Pierwszym i jednocześnie najbardziej elementarnym „testem” dobrego dopasowania wyrównanego szeregu do szeregu obserwacji jest zbudowanie zwykłego wykresu obu zmiennych. *STATISTICA* kreśli taki wykres automatycznie, dodatkowo pokazując jeszcze błędy (reszty).

Innym sposobem na sprawdzenie, czy model dobrze dopasowuje się do obserwowanych danych, jest analiza reszt. Model wyrównywania wykładniczego jako model autoregresyjny jest pewnym rodzajem regresji, a zatem reszty szeregu powinny spełniać założenie o normalności (patrz definicja modelu). Innymi słowy, należy sprawdzić, czy to, co zostaje po wyodrębnieniu wszystkich identyfikowalnych składników modelu, jest gaussowskim białym szumem.

Istnieje także cały szereg innych miar dopasowania, które dobiera się w zależności od postawionego zadania. Najczęściej używane to:

- ◆ błąd średni (*mean error*) – średnia arytmetyczna reszt,
- ◆ błąd średni bezwzględny (*mean absolute error*) – średnia arytmetyczna wartości bezwzględnych reszt,
- ◆ błąd procentowy (*percentage error*) – wartość reszty w stosunku do wartości obserwowanej, tj.

$$PE_t = (X_t - F_t) / X_t * 100\%$$

gdzie:

X_t – wartość obserwowana w czasie,

F_t – wartość prognozy w czasie t .



- ◆ średni błąd procentowy (*mean percentage error*) – średnia arytmetyczna wartości PE,
- ◆ średni bezwzględny błąd procentowy (analogicznie).

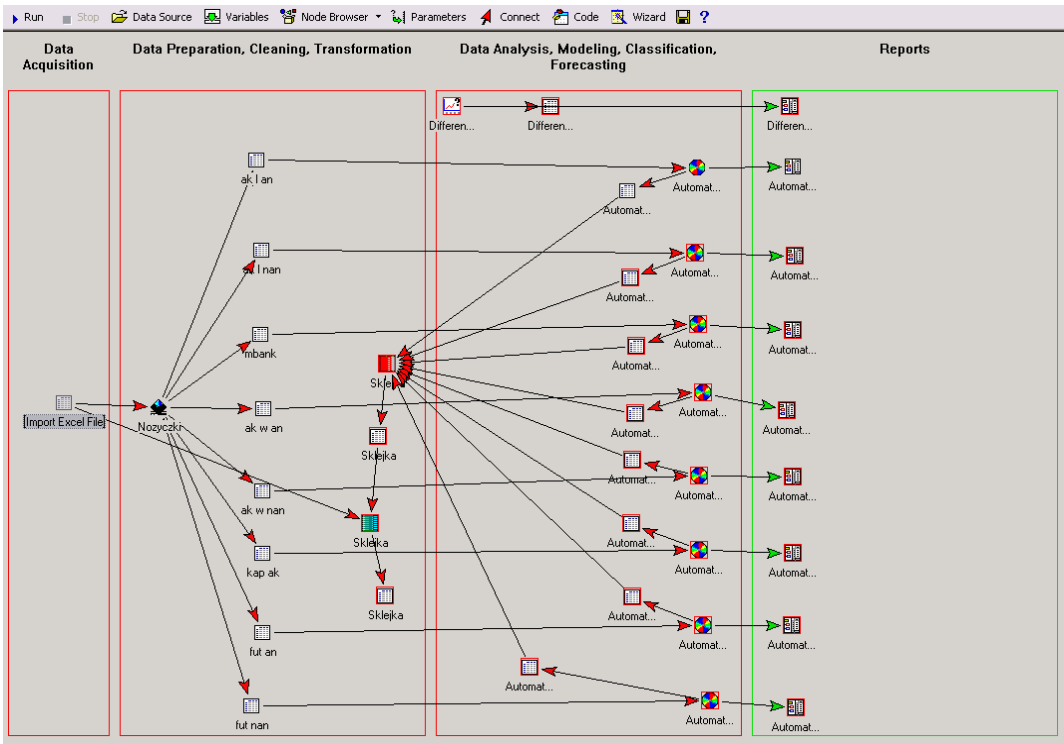
Przykład zastosowania zestawu modeli wyrównywania wykładniczego w *STATISTICA Data Miner* do prognozowania danych związanych z rynkiem kapitałowym w Polsce

KDPW S.A. jest jedną z głównych instytucji polskiego rynku kapitałowego. Pełni on funkcje depozytu papierów wartościowych oraz izby rozliczeniowej dla wszystkich transakcji rynku regulowanego w Polsce, zarówno kasowego, jak i terminowego. KDPW rozlicza także transakcje zawarte poza rynkiem regulowanym. Zadania KDPW obejmują swoim zakresem szeroką gamę usług związanych z rynkiem kapitałowym, co sprawia, że czynników mających wpływ na jego funkcjonowanie jest wiele. Przeprowadzane analizy dotyczą różnych aspektów działalności i zazwyczaj wymagają bardzo szczegółowych danych, na podstawie których wykonuje się różnorodne symulacje, np. przychodów KDPW.

Opis projektu

W trakcie przeprowadzanych analiz wyodrębniono kilkadziesiąt zmiennych, które odzwierciedlają swoim zakresem czynniki mające wpływ na przychody firmy. Każda z tych zmiennych musi być uwzględniona w analizie przychodów jako oddzielny szereg czasowy. Prognozowanie tak dużej liczby szeregów czasowych o odmiennych strukturach za pomocą złożonych narzędzi ekonometrycznych byłoby bardzo czasochłonne i skomplikowane, przede wszystkim ze względu na potrzebę częstych aktualizacji. Ze zbioru zmiennych wyodrębniono więc kilka najbardziej istotnych, pod względem wielkości generowanego przez nie przychodu. Wybrane, najbardziej istotne dla analizy zmienne są danymi wyjściowymi dla systemu modeli wyrównywania wykładniczego. Pozostałe, mniej znaczące zmienne prognozowane są za pomocą prostych modeli średniej ruchomej.

Poniższe okno przedstawia całość projektu w programie *STATISTICA Data Miner* tworzącego 8 modeli dla wybranych zmiennych.



Wszystkie dane historyczne są przechowywane i aktualizowane w formatach arkuszy *MS Excel*. Przed przystąpieniem do tworzenia prognoz w programie *STATISTICA* sporządzany jest arkusz zawierający 8 wybranych zaktualizowanych szeregów czasowych. W części *Data Acquisition* przestrzeni roboczej *STATISTICA Data Miner* umieszczony jest węzeł, umożliwiający import arkusza o ścieżce dostępu podanej jako parametr węzła.



Takie rozwiązanie umożliwia zapamiętanie źródła danych, na podstawie których tworzone są prognozy w danym projekcie. Przy częstych aktualizacjach prognoz jest to bardzo wygodne, ze względu na to, że pozwala dokładnie określić, z jaką datą aktualizowane były dane.

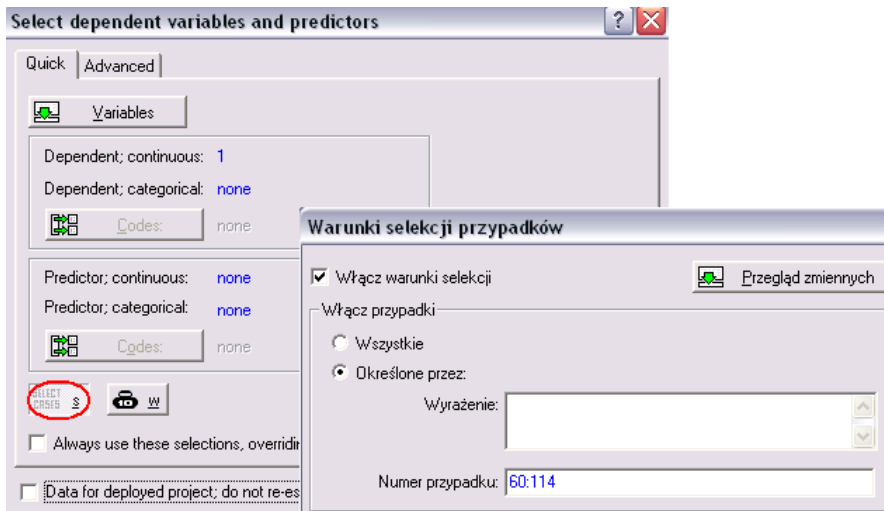


Zaimportowany plik zawiera 8 zmiennych, związanych ze specyfiką Tabeli Opłat KDPW:

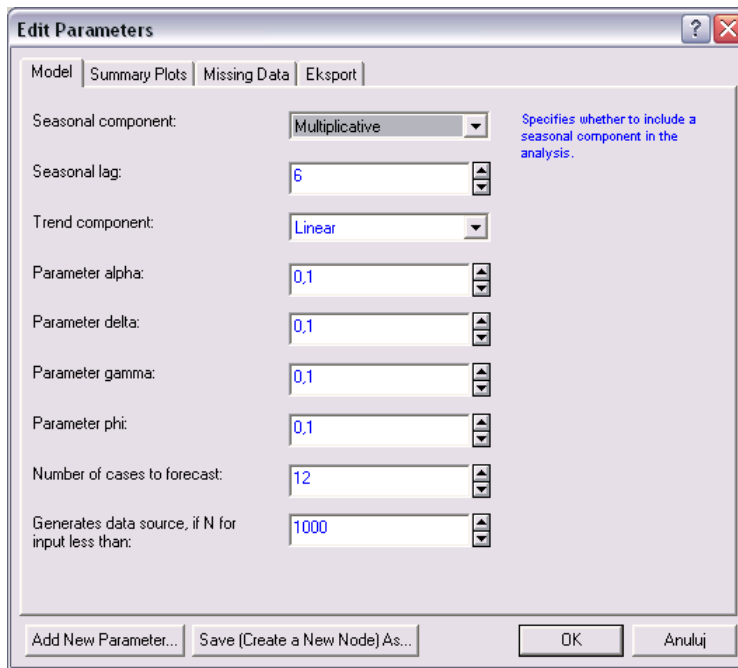
- ◆ kapitalizacja akcji (w mln PLN) – zmienna 1,
- ◆ liczba transakcji giełdowych na akcjach w podziale na transakcje zawierane przez animatorów i nieanimatorów – zmienne 2 i 3,
- ◆ wartość transakcji giełdowych na akcjach w podziale na transakcje zawierane przez animatorów (w tys.) i nieanimatorów (w mln) – zmienne 4 i 5,
- ◆ liczba transakcji zawieranych na rynku międzybankowym – zmienna 6,
- ◆ liczba transakcji giełdowych na kontraktach terminowych w podziale na transakcje zawierane przez animatorów i nieanimatorów – zmienne 7 i 8.

Zmienne zawierają dane miesięczne o różnych długościach: od 78 do 114 przypadków, tzn. najdłuższa zmienna (kapitalizacja akcji) zawiera dane od stycznia 1996 roku. Szeregi różnią się też strukturą, dlatego dla każdej zmiennej budowany jest w projekcie oddzielny model. Węzeł *Nożyczki* (modyfikacja węzła *Multiple Copies of Data Source*) umieszczony w części *Data Preparation, Clearing, Transformation* „rozcina” wejściowy arkusz zawierający 8 zmiennych na 8 arkuszy o jednej zmiennej, które następnie podłączane są do węzłów analitycznych tworzących modele wyrównywania wykładniczego.

W każdym z arkuszy wejściowych zmienną zależną ciągłą będzie zmienna 1, ponieważ takie wymagania stawia nam budowa modelu wyrównywania wykładniczego. Za pomocą opcji *Select cases* można ustawić wybór odpowiednich przypadków, w zależności od zakresu danych w szeregu.



Wszystkie węzły wyrównywania wykładniczego w projekcie zostały specjalnie zmodyfikowane. Są to węzły, które automatycznie wyliczają odpowiednie współczynniki modelu w zależności od zadanych parametrów trendu i sezonowości. Węzeł posiada także parametr, który wskazuje, jaką liczbę przypadków ma wygenerować dany model w prognozie na przeszłość.

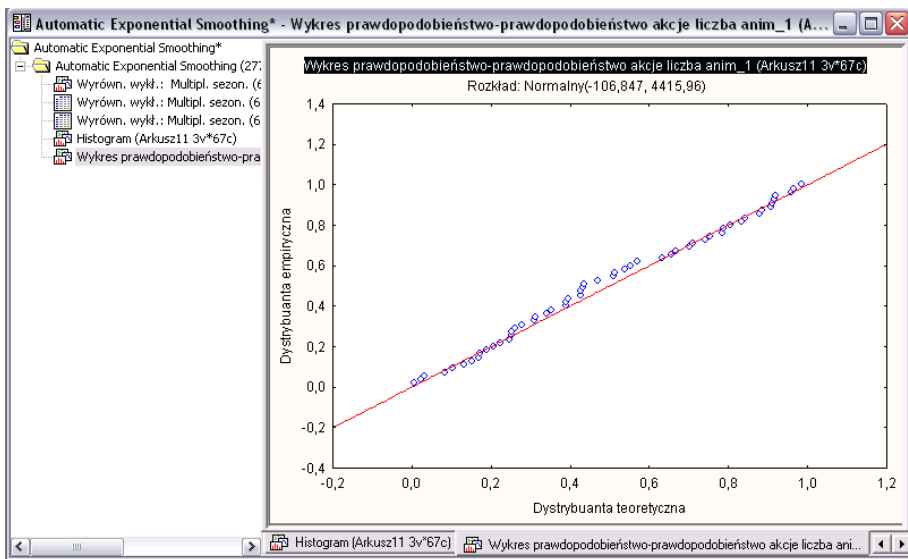
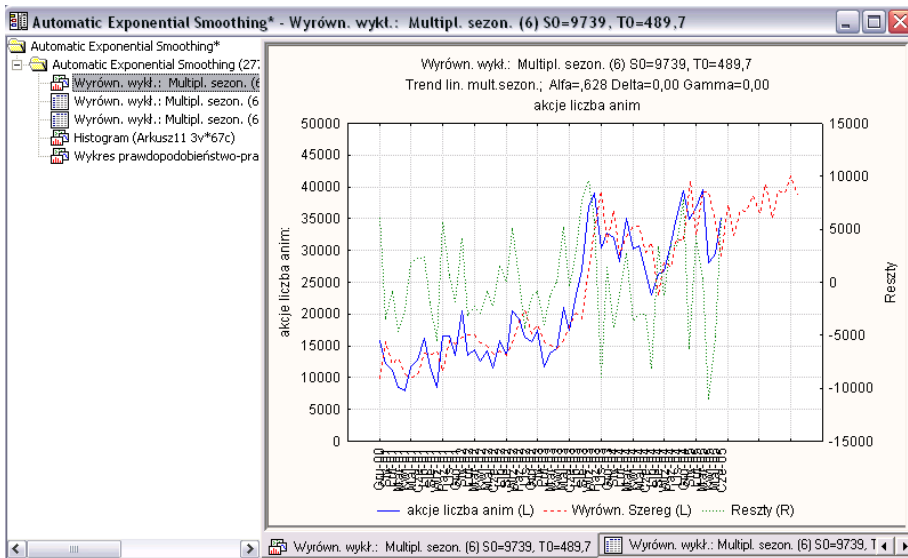


Wybór parametrów trendu i sezonowości poprzedzają zazwyczaj badania funkcji korelacji i autokorelacji oraz różnicowanie poszczególnych szeregów. W pewnych przypadkach parametry te wybiera się tzw. „metodą ekspercką”, tzn. jeśli z pewnych pozastatystycznych przesłanek wiadomo, że np. trend liniowy procesu może zmienić się w gasnący. Za pomocą węzła wyrównywania wykładniczego łatwo jest wówczas sterować parametrami modelu. W praktyce często okazuje się, że np. zmiana parametru trendu z liniowego na gasnący powoduje, że wyliczony przez węzeł szereg tworzy prognozy, które bardziej odpowiadają naszym przypuszczeniom, a parametry dobroci dopasowania modelu do danych rzeczywistych zmieniają się nieznacznie.

Wyniki analiz tworzonych w węzłach projektu

Każdy zmodyfikowany węzeł automatycznego wyrównywania wykładniczego generuje arkusz wynikowy zawierający wyrównany szereg, wartości rzeczywiste i wartość błędu oraz raport w formie skoroszytu zawierający standardowo ten sam arkusz wynikowy, wykres obu szeregów i błędów, wyniki błędów średnich i procentowych. W dobrze dopasowanym modelu błędy mają rozkład normalny. Do standardowych elementów raportu dodano więc dwa wykresy błędów: histogram i wykres prawdopodobieństwo-prawdopodobieństwo w celu sprawdzenia normalności reszt.

Poniżej przedstawiono wybrane elementy raportu wynikowego dla zmiennej „liczba transakcji giełdowych na akcjach zawieranych przez animatorów”: wykres wartości rzeczywistych, wyrównanego szeregu i błędów oraz wykres prawdopodobieństwo-prawdopodobieństwo dla reszt modelu.

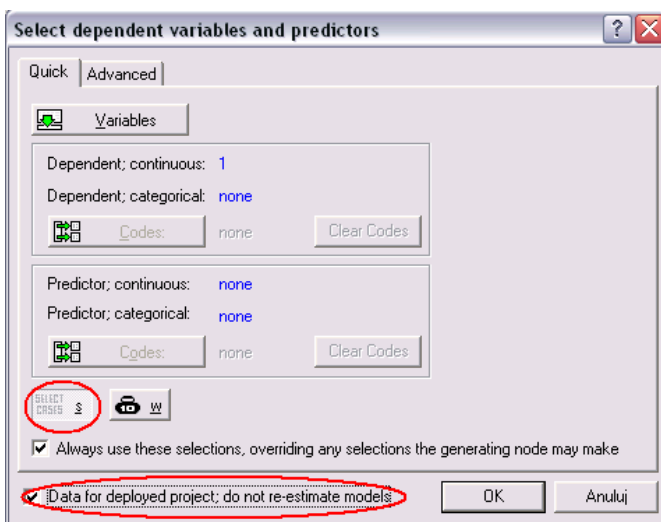


Wszystkie wygenerowane prognozy wykorzystywane są do dalszej analizy za pomocą MS Excel. Węzeł Sklejka z poszczególnych arkuszy wyników węzłów analitycznych wyrównywania wykładniczego wybiera zmienne zawierające wyrównane szeregi i skleja w jeden arkusz o 8 zmiennych i liczbie przypadków równych długości prognozy (zazwyczaj 12 miesięcy). Taki arkusz zawierający wszystkie prognozy zapisywany jest w postaci pliku *.xls i wykorzystywany do dalszych analiz, dotyczących np. przychodów firmy.

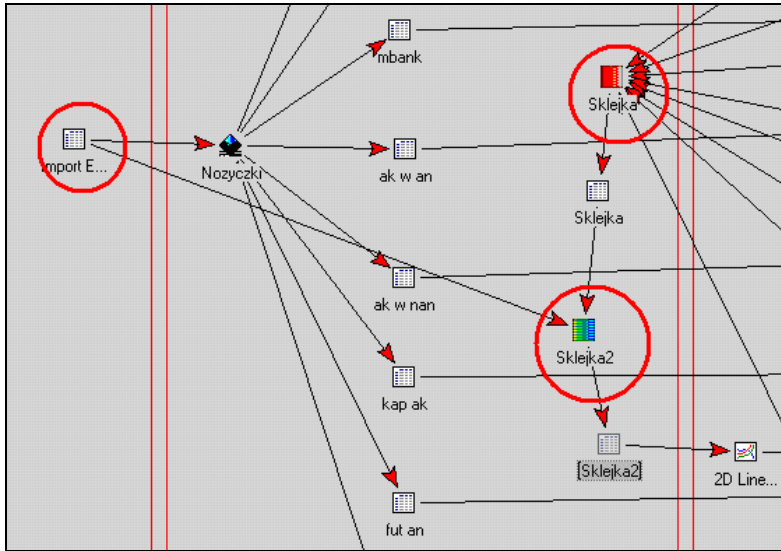


Sprawdzenie jakości prognoz generowanych przez modele

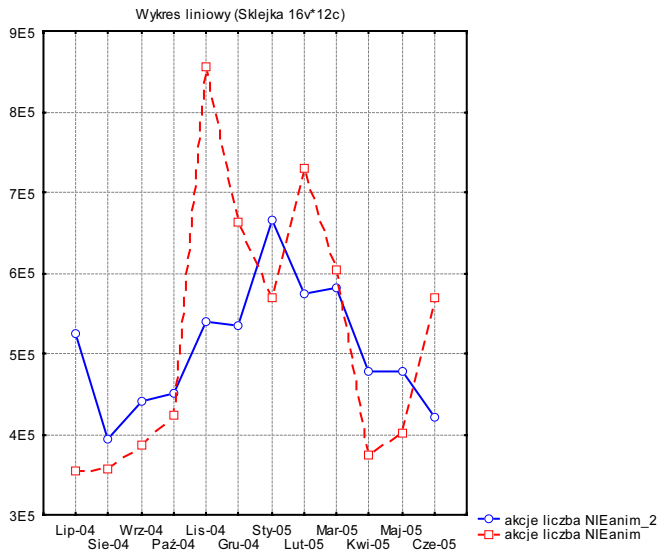
Jednym ze sposobów na sprawdzenie dobroci dopasowania modelu i jakości sporządzanych przez niego prognoz jest wygenerowanie prognozy na podstawie zmiennej uciętej o kilka lub kilkanaście ostatnich przypadków i sprawdzenie, czy wygenerowane przez model prognozy sprawdzają się z wartościami rzeczywistymi. W projekcie *STATISTICA Data Miner* robi się to bardzo łatwo. Wystarczy w oknie wyboru zmiennych zaznaczyć opcję *Data for Deployed Project*, co spowoduje, że węzeł nie będzie wyliczał nowych, dopasowanych do skróconego szeregu parametrów modelu, a jedynie zastosuje wyliczone wcześniej modele przy użyciu nowego szeregu. Za pomocą opcji *Select Cases* można wybrać odpowiednio mniejszą liczbę przypadków do analizy.



W celu ułatwienia porównania wartości rzeczywistych z prognozami wygenerowanymi przez modele skonstruowany został węzeł *Sklejka2*. Arkusz wynikowy tego węzła zawiera 16 zmiennych o liczbie przypadków równych długości prognozy. Pierwsze 8 zmiennych to prognozy wygenerowane przez 8 modeli podłączonych z arkusza *Sklejka*. Pozostałe 8 zmiennych zawiera 12 ostatnich przypadków z wejściowego arkusza zaimportowanego z *MS Excel*. Poniższy fragment projektu pokazuje, w jaki sposób łączone są dane, o których mowa powyżej.



Dzięki tak skonstruowanemu arkuszowi łatwo jest prześledzić, jak sprawdzają się wygenerowane prognozy, np. podłączając wynikowy arkusz Sklejka2 do węzła sporządzającego wykres liniowy wielu zmiennych. Poniżej pokazano przykładowe wyniki dla zmiennej „akcje liczba nieanimatorzy”. Linia ze znacznikami w kształcie kwadratów to wartości rzeczywiste. Linia ze znacznikami w kształcie kółek to prognozy wygenerowane przez model.





Informacje o błędach i aktualizacje prognoz

Raporty generowane przez węzły automatycznego wyrównywania wykładniczego w projekcie zawierają arkusz, w którym zawarte są informacje o błędach modeli, a w szczególności wartości średniego błędu procentowego i średniego bezwzględnego błędu procentowego. Przykładowy arkusz podsumowania błędu pokazano poniżej.

Podsumowanie błędu	Błąd
Błąd śred.	3805,61
Średni błąd bezwzg.	71691,59
Suma kwadratów	927618482590,98
Średni kwad.	10662281409,09
Średni błąd procent.	-3,99
Średni bezwz. bł. proc.	18,79

Spośród sześciu wynikowych wartości dwa ww. błędy są najłatwiejsze w interpretacji. Pierwszy z nich mówi, o ile procent średnio wartości wyrównanego szeregu odstawiały od wartości rzeczywistych. Różnice te mogą być zarówno dodatnie, jak i ujemne. Przy wyliczaniu średniej mogą się one wzajemnie znosić, dlatego lepszą miarą wydaje się średni bezwzględny błąd procentowy, liczony na podstawie wartości bezwzględnych różnic. Prognozy na przyszłość tworzone są zazwyczaj na 12 miesięcy. Naszym celem jest więc stworzenie modelu, który będzie lepiej prognozował nie pojedyncze wartości szeregu, ale pewne (12 miesięczne) okresy. Innymi słowy wymagamy, aby model sprawdzał się lepiej w całym okresie, a więc bardziej interesuje nas suma wartości niż pojedyncze punkty szeregu. Lepszym parametrem błędu dla tego zadania wydaje się więc błąd średni procentowy.

Należy pamiętać, że błędy wyświetlane w powyższym arkuszu dotyczą jednak tylko dopasowania modelu do wartości rzeczywistych do momentu sporządzenia prognoz. Są to tzw. mierniki dokładności *ex ante*. Prawdziwą miarą dokładności dla wnioskowania wprzód są tzw. mierniki dokładności *ex post*, czyli różnice pomiędzy prognozą wygenerowaną np. 12 miesięcy wprzód i jej 12-miesięczną realizacją. Błędy *ex ante* są zazwyczaj znacząco mniejsze od rzeczywistych błędów *ex post*. Poniższa tabela przedstawia porównanie błędów wyrównanego szeregu *ex ante* i błędów związanych z realizacją prognozy sporządzonej w lipcu 2004 roku na 12 następujących miesięcy dla wszystkich zmiennych w projekcie.



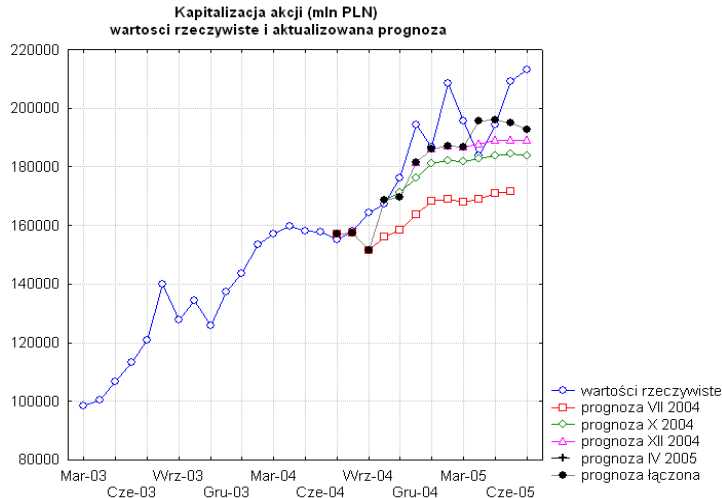
nr zmienniej	nazwa zmiennej	błąd średni %		błąd średni bezwzględny %	
		ex ante	ex post	ex ante	ex post
1	kapitalizacja akcji	-0,1%	-9%	7%	11%
2	akcje liczba animatorzy	-4,4%	5%	18%	15%
3	akcje liczba nieanimatorzy	-3,2%	3%	18%	21%
4	akcje wartość animatorzy	-3,3%	-12%	24%	23%
5	akcje wartość nieanimatorzy	-4,5%	-25%	24%	40%
6	liczba rynek międzybankowy	-5,8%	-17%	19%	22%
7	wartość futures animatorzy	-4,1%	6%	25%	15%
8	wartość futures nieanimatorzy	-4,9%	5%	25%	20%

Różnice pomiędzy błędami sięgają nawet 15-20%. Oznacza to, że nie zawsze model, który dobrze dopasowuje się do danych rzeczywistych, będzie generował prognozy o dużej sprawdzalności. Zazwyczaj najlepiej sprawdza się prognoza kilku początkowych miesięcy. Dobrym sposobem na poprawę jakości prognoz są więc częste aktualizacje prognoz na najbliższe miesiące. Poniżej przedstawiono porównanie błędów procentowych dla prognozy zmiennej „kapitalizacja akcji” sporządzonej na 12 miesięcy oraz tej samej prognozy aktualizowanej w X i XII 2004 r. oraz w IV 2005 r. Dzięki tym aktualizacjom średni błąd prognozy *ex post* zmniejszył się z 10,1% do 2,5%, a średni błąd bezwzględny *ex post* z 10,3% do 4,1%.

	błąd prognozy (%)		błąd bezwzględny prognozy %	
	Wersja pierwotna prognozy	Wersja aktualizowana prognozy	Wersja pierwotna prognozy	Wersja aktualizowana prognozy
lip-04	1%	1%	1%	1%
sie-04	0%	0%	0%	0%
wrz-04	-8%	-8%	8%	8%
paź-04	-7%	1%	7%	1%
lis-04	-10%	-3%	10%	3%
gru-04	-16%	-7%	16%	7%
sty-05	-10%	0%	10%	0%
lut-05	-19%	-10%	19%	10%
mar-05	-14%	-5%	14%	5%
kwi-05	-8%	6%	8%	6%
maj-05	-12%	1%	12%	1%
cze-05	-18%	-7%	18%	7%
błąd średni	-10,1%	-2,5%	10,3%	4,1%



Na poniższym wykresie przedstawiono szereg „kapitalizacja akcji” oraz prognozy generowane w kolejnych aktualizacjach, a także prognozę łączoną z kolejnych aktualizacji.



Podsumowanie

Sporządzanie analiz finansowych dotyczących przyszłości firmy nie jest zagadnieniem prostym, zwłaszcza w przypadku, gdy mamy do czynienia z mnogością czynników mających wpływ na działalność, tak jak w przypadku KDPW. Wykorzystanie narzędzi matematycznych do tych celów znacznie ułatwia rozpoznanie procesów rządzących rynkiem, co pozwala na sporządzenie prognoz przychodów firmy. Programy statystyczne, takie jak *STATISTICA Data Miner*, nawet przy użyciu prostych matematycznie modeli są bardzo pomocne w tego typu zagadnieniach. Dzięki nim zmniejsza się pracochłonność wykonywanych analiz, a same analizy przybierają formę przejrzystych i przyjaznych dla użytkownika procedur.

Literatura i źródła danych:

1. Sokołowski A., Materiały kursowe „Prognozowanie i analiza szeregów czasowych”, Statsoft Polska 2003.
2. „*STATISTICA Data Miner*”, Statsoft 2003.
3. „*STATISTICA PL* dla Windows (Tom III): Statystyki II”, Statsoft Polska 1997.
4. Dane pochodzące z systemu depozytowo-rozliczeniowego KDPW.