



## AUTOMATYCZNA BUDOWA MODELI PROGNOSTYCZNYCH

*dr Janusz Wątroba<sup>4</sup>*

W trakcie analizy rzeczywistych zjawisk i procesów, nawet w stosunkowo prostych sytuacjach, nie jesteśmy w stanie całkowicie ich wyjaśnić. Stąd też opisując współzależności zachodzące pomiędzy nimi, posługujemy się zazwyczaj pewnymi uproszczonymi modelami rzeczywistych współzależności. A zatem pod pojęciem modelu możemy rozumieć użyteczną postać przedstawienia danych empirycznych. Przystępując do procesu budowy modelu, musimy przyjąć pewien kompromis między zbytnim uproszczeniem rzeczywistości a chęcią uwzględnienia zbyt szczegółowych danych.

Końcowym efektem procesu poszukiwania najlepszego rozwiązania jest zazwyczaj dobra teoria. Rozpoczynamy od obszernego modelu zawierającego wszystkie potencjalne, podlegające testowaniu czynniki, wywierające wpływ na rozpatrywane przez nas zjawisko. Następnie poddajemy testowaniu składniki początkowego, obszernego modelu, aby zidentyfikować mniej obszerne podmodele, wyjaśniające w adekwatny sposób rozpatrywane zjawisko. W końcu, spośród tych potencjalnych podmodelei wybieramy najprostsz, który na zasadzie oszczędności traktujemy jako najlepiej opisujący badane zjawisko.

Proste modele preferujemy nie tylko z przyczyn filozoficznych, ale również z powodów czysto praktycznych. Proste modele są łatwiejsze do ponownego testowania w przypadku powtarzania badań lub poddawaniu ich ocenie krzyżowej. Proste modele wymagają zazwyczaj niższych kosztów w przypadku ich praktycznego wykorzystywania oraz kontroli wyników w przyszłości. Jednakże przyczyny filozoficzne nie powinny być niedoceniane. Proste modele są łatwiejsze do zrozumienia i docenienia i dlatego posiadają pewne piękno, którego brakuje bardziej skomplikowanym modelom.

Celem niniejszego wystąpienia jest przedstawienie wybranych opcji i narzędzi programu *STATISTICA*, pomocnych przy budowie i weryfikacji modeli regresyjnych. Najpierw zostaną przedstawione ogólne własności modułu *VGSR*, a następnie w module tym zostanie zaprezentowany przykład budowy i eksploracji modelu w oparciu o dane rzeczywiste dotyczące zużycia gazu ziemnego w wybranych miastach północnej części USA.

### Krótką charakterystyką właściwości modułu *VGSR*

Moduł ten jest implementacją ogólnego modelu liniowego. W procesie budowania modeli nawet dla bardzo złożonych układów, w tym układów zawierających efekty dla predyktorów jakościowych (zmiennych objaśniających), program pozwala użytkownikowi wykorzystywać metodę krokową oraz metodę wyboru najlepszego podzbioru. Określenie „ogólna” w nazwie *Visual General Stepwise Regression* odnosi się zarówno do stosowania technik ogólnego modelu liniowego, jak również do faktu, że w odróżnieniu od innych programów przeznaczonych do przeprowadzania regresji krokowej, *VGSR* nie ogranicza się wyłącznie do analizy układów, zawierających jedynie predyktory (zmienne objaśniające) o charakterze ciągłym. Metody te mogą być stosowane w przypadku

---

<sup>4</sup> StatSoft Polska Sp. z o.o.



układów, w których występują zarówno predyktory ciągłe, jak i jakościowe (tzn. układy ANOVA lub ANCOVA). W ramach metody krokowej *VGSR* umożliwia budowanie modeli poprzez jednorazowe dodawanie lub eliminowanie efektów (efekty mogą być wprowadzane lub usuwane tylko jeden raz w trakcie procesu selekcji) lub wielokrotne dodawanie i eliminowanie efektów (selekcja postępująca i wsteczna, tzn. efekty mogą być wprowadzane lub usuwane z modelu w każdym kroku, zgodnie z kryterium selekcji opartym na statystykach  $F$  lub wartości  $p$ ). Z kolei metoda najlepszego podzbioru regresji daje użytkownikowi wygodne opcje przeznaczone do weryfikacji modeli przyjmowanych w trakcie poszukiwania podzbioru (np. dla selekcji najlepszego podzbioru zmiennych można wybrać maksymalną lub minimalną wielkość podzbioru, statystykę  $C_p$  Mallowa,  $R$ -kwadrat oraz poprawiony  $R$ -kwadrat). Moduł *Visual General Stewise Regression (VGSR)* oferuje wszystkie standardowe opcje wyników wykorzystywanych zazwyczaj w procesie budowy i weryfikacji modeli prognostycznych (np. statystyki wartości prognozowanych i reszt dla próby przeznaczonej do analizy, próby do oceny krzyżowej lub próby do prognozowania, testy założeń modelu, wykresy średnich itd.). Ponadto są także dostępne unikalne, specyficzne dla regresji opcje wyników, w tym także wykresy Pareto ocen parametrów, podsumowanie (testy) dla pełnego modelu oraz opcje zawierające różne metody oceny modeli bez wyrazu wolnego, korelacje cząstkowe i semicząstkowe itd. W kontekście budowy i weryfikacji modeli prognostycznych szczególnie przydatne mogą być następujące ogólne własności:

- ◆ **Porównywanie i modyfikacja analiz;** moduł *VGSR* jest programem w pełni realizującym mechanizm przetwarzania wielowejściowego, co oznacza, że można jednocześnie przeprowadzić wiele analiz na tych samych lub różnych zbiorach danych. Jest to własność niezwykle użyteczna przy porównywaniu wyników pochodzących z różnych analiz tego samego zbioru danych lub analiz tego samego typu, przeprowadzanych w oparciu o różne dane. Modyfikacja analizy nie wymaga ponownego jej określania - wystarczy tylko określić żądane zmiany. Wyniki uzyskiwane w oparciu o różne modyfikacje analizy mogą być w łatwy sposób porównywane.
- ◆ **Automatyczna aktualizacja wyników;** moduł *VGSR* może być uruchamiany w trybie automatycznej aktualizacji. Po każdorazowej zmianie danych zawartych w pliku danych wyniki są przeliczane, tzn. wszystkie arkusze wyników i wykresy wyświetlane aktualnie na ekranie są automatycznie aktualizowane. Własność ta może być wykorzystywana do przeprowadzania analiz typu „Co-jeśli”, w celu badania skutków usuwania ze zbioru danych określonych obserwacji (np. obserwacji odstających). Usunięcie obserwacji powoduje automatyczną korektę wszystkich wyników, natomiast naciśnięcie kombinacji klawiszy Ctrl-Z automatycznie cofa wszystkie zmiany.
- ◆ **Automatyczne generowanie poleceń składniowych;** program pozwala na jednoczesne (niejako w tle) generowanie pełnego zbioru poleceń składniowych dla dowolnego układu definiowanego za pomocą **Okien szybkiego definiowania** lub za pomocą **Kreatora układu**. Polecenia składni równoważne najbardziej złożonym i niestandardowym układom pozostają „aktywne”, tzn. mogą być ponownie uruchamiane, zapisywane do wykorzystania w przyszłości, modyfikowane, a także umieszczane w programach wsadowych pisanych w języku *SCL*.

## Przykład budowy i weryfikacji modelu prognostycznego

Celem analizy jest zbudowanie modelu opisującego kształtowanie się poziomu zużycia gazu ziemnego (ZUZYC; zmienna objaśniana) przez firmy dostarczające ciepło w zależności od różnego rodzaju czynników (np. temperatury, od której w dużej mierze zależy zużycie gazu). Dane do



przykładu dotyczą jednego sezonu grzewczego (zima) w kilku głównych miastach położonych w północnej części USA. Jako zmienne objaśniające wzięto pod uwagę: wskaźnik temperatury (TEMP; wyrażony jako 65 - średnia temperatura dobową; duża wartość oznacza chłodny dzień), prędkość wiatru (P\_WIATRU; średnia dobową), dzień wolny (D\_WOLNE; w dni wolne od pracy zużycie gazu było zazwyczaj niższe; była to zmienna zero-jedynkowa) oraz wartości wskaźnika temperatury opóźnione o jeden dzień (TEMP\_OP). Przyjmujemy liniową postać modelu regresji wielokrotnej.

Wspomniany wcześniej proces poszukiwania najlepszego rozwiązania został uwzględniony w technikach budowania modelu regresji dostępnych w module *VGSR*. Użytkownik ma bowiem do wyboru metodę krokową i metodę najlepszego podzbioru. Wykorzystajmy najpierw tę pierwszą metodę. W tym celu po uruchomieniu modułu *VGSR* w panelu początkowym jako *Rodzaj analizy* należy wskazać opcję *Regresja wielokrotna*, natomiast jako *Sposób definiowania analizy* należy wybrać opcję *Szybkie definiowanie*. Po kliknięciu przycisku *OK* program przejdzie do okna dialogowego *Visual GSR - Regresja wielokrotna*. W oknie tym należy wskazać zmienne do analizy. Na liście zmiennych zależnych (objaśnianych) wskazujemy zmienną ZUŻYC, a na liście predyktorów (zmiennych objaśniających) wskazujemy zmienne: TEMP, TEMP\_OP, P\_WIATRU oraz D\_WOLNE. Z kolei w polu *Budowanie modelu* wybieramy opcję *Krokowa postępująca* i klikamy przycisk *OK*, aby program przeprowadził analizę. Na ekranie pojawi się okno wynikowe *Visual GSR - Wyniki analizy 1*.

Analizę uzyskanych wyników rozpoczniemy od obejrzenia wyników budowania modelu metodą krokową umieszczonych w polu *Wyniki budowania modelu* pod przyciskiem *Regresja krokowa*. Proces budowania modelu w poszczególnych krokach przedstawiono poniżej.

Podsumowanie regresji wielokrotnej dla zmiennej ZUŻYC							
VISUAL Krokowa postępująca							
GSR P do wprowadzenia: .05, P do usunięcia: .05							
Efekt	Kroki	Stopnie Swobody	F do usunięc.	P do usunięc.	F do wprowadz.	P do wprowadz.	Efekt (stan)
TEMP	Krok 1	1			681.569	0.0000	Wszedł
TEMP OP		1			86.387	.0000	Poza
P WIATRU		1			3.076	.0845	Poza
D WOLNE		1			.944	.3351	Poza
TEMP	Krok 2	1	681.569	0.0000			W modelu
TEMP OP		1			23.963	.0000	Wszedł
P WIATRU		1			1.205	.2766	Poza
D WOLNE		1			9.003	.0039	Poza
TEMP	Krok 3	1	363.026	0.0000			W modelu
TEMP OP		1	23.963	.0000			W modelu
P WIATRU		1			2.685	.1066	Poza
D WOLNE		1			5.864	.0186	Wszedł
TEMP	Krok 4	1	398.281	0.0000			W modelu
TEMP OP		1	19.926	.0000			W modelu
D WOLNE		1	5.864	.0186			W modelu
P WIATRU		1			5.180	.0266	Wszedł
TEMP	Krok 5	1	401.034	0.0000			W modelu
TEMP OP		1	21.838	.0000			W modelu
D WOLNE		1	8.436	.0052			W modelu
P WIATRU		1	5.180	.0266			W modelu

Jak widać, przy zastosowaniu metody krokowej postępującej końcowy model zawiera wszystkie zmienne objaśniające. Oceniając ogólne dopasowanie modelu do danych, możemy posłużyć się opcjami wyników umieszczonymi pod przyciskiem *R pełnego modelu*. W otrzymywanym oknie arkusza wyników program podaje wartość współczynnika korelacji wielokrotnej (R), wartość współczynnika determinacji ( $R^2$ ), wartość współczynnika determinacji skorygowaną ze względu na liczbę stopni swobody oraz wyniki testu F, służącego do oceny dopasowania pełnego modelu. Na podstawie przedstawionych w tabeli mierników dopasowania modeli liniowych możemy stwierdzić, że otrzymany model wyjaśnia około 95% całkowitej zmienności zmiennej prognozowanej.



VISUAL	Wielokr.	Skorygow	SS	SS	df	MS	F	p
GSR	R2	R2	Model	Reszta	Reszta	Reszta		
ZUZYC	.93098	.94760	386382	19913.	58	343.33	281.31	0.00

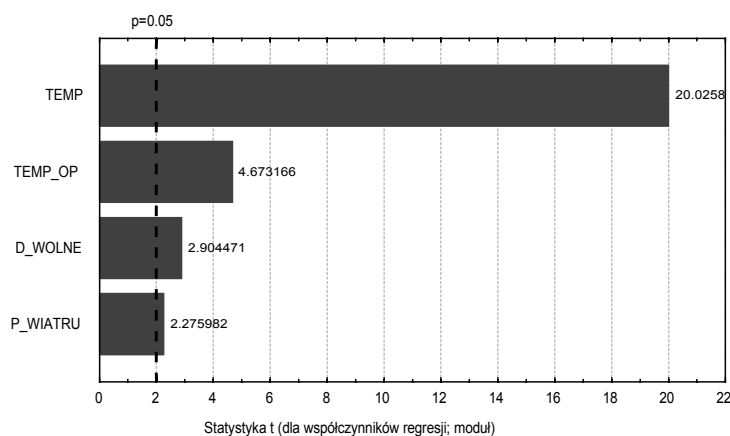
Ocenę parametrów strukturalnych modelu uzyskujemy po kliknięciu przycisku *Współczynniki*. W oknie arkusza wyników oprócz ocen wartości niestandardyzowanych i standardyzowanych współczynników regresji program podaje również oceny ich statystycznej istotności oraz 95% przedziały ufności dla tych ocen. Fragment wyników zamieszczono poniżej.

Efekt	Parametryzacja z sigma-ograniczeniami						
	ZUZYC Param.	ZUZYC Bł. std.	ZUZYC t	ZUZYC p	-95.00% Gr.ufn.	+95.00% Gr.ufn.	ZUZYC Beta (B)
Wyraz wolny	1.021	11.694	.156	.8768	-21.59	25.228	
TEMP	5.891	.294	20.026	0.0000	5.302	6.480	.8136
TEMP_OP	1.386	.296	4.673	.0000	.792	1.979	.1902
P_WIATRU	1.332	.585	2.276	.0266	.160	2.503	.0689
D_WOLNE	-15.66	5.393	-2.904	.0052	-26.46	-4.869	-.0881

Na podstawie wyników podanych w powyższej tabeli możemy w ostatecznym modelu pominąć wyraz wolny. Wartości statystyk t, służących do oceny statystycznej istotności oszacowanych współczynników regresji, można również przedstawić graficznie za pomocą wykresu Pareto (dostępny w oknie wyników na karcie *Podsumowanie*, pod przyciskiem *Pareto*).

Wykres Pareto statystyki t dla współczynników regresji

Zmienna: ZUZYC



Budując model bez wyrazu wolnego, wystarczy w oknie wyników analizy kliknąć przycisk *Zmień*, a następnie w oknie dialogowym *Visual GSR - Regresja wielokrotna* zaznaczyć opcję *Bez wyrazu wolnego*. Po kliknięciu *OK* w oknie wyników analizy klikamy przycisk *Współczynniki*, aby otrzymać na ekranie okno arkusza danych z ocenami parametrów strukturalnych modelu bez wyrazu wolnego. Dzięki możliwości porównywania wielu analiz przeprowadzanych na tych samych lub różnych zbiorach danych możemy na ekranie jednocześnie łatwo obejrzeć oceny parametrów strukturalnych dla modelu uwzględniającego wyraz wolny oraz bez wyrazu wolnego.

Korzystając jeszcze raz z tej samej własności, możemy przeprowadzić kolejną analizę wykorzystując tym razem jako metodę budowania modelu metodę najlepszego podzbioru i model bez wyrazu wolnego. Ponadto w charakterze kryterium, według którego zostaną przedstawione modele, wskażmy statystykę  $C_p$  Mallowa. Ta miara jakości dopasowania modelu wykazuje mniejsze

uzależnienie (w stosunku do statystyki  $R^2$ ) od liczby efektów zawartych w modelu i dlatego też pozwala znaleźć najlepszy podzbiór, który zawiera tylko predyktory ważne dla odpowiedniej zmiennej zależnej. W opcji *Wyświetl podzbiory* wstawmy liczbę 8 (która oznacza, że program wyświetli 8 najlepszych podzbiorów wg wielkości statystyki  $C_p$  Mallowa).

Efekt	Param.	Bł. std.	zużyc t	zużyc p	-95.00% Śr. ufn.	+95.00% Śr. ufn.	Beta (B)
TEMP	5.901	.285	20.726	0.0000	5.33	6.470	.7772
TEMP OP	1.403	.273	5.144	.0000	.86	1.949	.1829
P WIATRU	1.376	.506	2.720	.0086	.36	2.389	.0576
D WOLNE	-15.57	5.313	-2.930	.0048	-26.20	-4.937	-.0253

Bardzo ważnym etapem przy dopasowaniu modeli, które mają być wykorzystywane do przewidywania przyszłych obserwacji, jest ocena krzyżowa wyników, tzn. zastosowanie bieżących wyników do nowego zbioru obserwacji, który nie był wykorzystywany przy obliczaniu (estymacji parametrów) tych wyników. Moduł *VGSR* oferuje bardzo elastyczne metody do obliczania szczegółowych wartości przewidywanych i statystyk resztowych dla obserwacji, które nie były wykorzystywane w trakcie obliczeń występujących przy dopasowaniu bieżącego modelu i posiadają zaobserwowane wartości dla zmiennych zależnych (próba do oceny krzyżowej). Te niezastąpione narzędzia do oceny trafności progностycznej modelu często są niedostępne w mniej kompletnych implementacjach ogólnego modelu liniowego. Aby pokazać opcję umożliwiającą przeprowadzanie sprawdzianu krzyżowego, należy kliknąć przycisk *Sprawdzian krzyżowy*, umieszczony w dolnej części opisywanego okna dialogowego. Na ekranie pojawi się okno dialogowe o tej samej nazwie. W oknie tym należy wybrać opcję *Włączona* oraz (po naciśnięciu przycisku *Zmienna identyfikująca próby*) wskazać zmienną *S\_KRZYŻ*, identyfikującą przynależność obserwacji do próby obliczeniowej i próby do sprawdzianu krzyżowego.

Po kliknięciu *OK* program przeprowadzi odpowiednie obliczenia. Przeglądanie wyników rozpoczniemy tym razem od obejrzenia zestawienia najlepszych 8 podmodeli uzyskanych metodą najlepszego podzbioru. Program sugeruje wybranie modelu zawierającego wszystkie 4 zmienne objaśniające. Użytkownik, kierując się zasadą dobierania modeli najprostszych, może rozważyć kolejno modele z trzema, dwoma i jedną zmienną objaśniającą.

Nr podzb.	Mallowa Cp	Liczba Efektów	TEMP	TEMP OP	P WIATRU	D WOLNE
1	4.000	4	.7833	.1567	.0802	-.0274
2	6.776	3	.7506	.1943	.0596	
3	9.482	2	.8022	.1988		
4	9.620	3	.8321	.1775		-.0162
5	10.991	3	.9322		.0928	-.0397
6	19.402	2	1.0131			-.0284
7	20.266	2	.9370		.0642	
8	23.730	1	.9974			

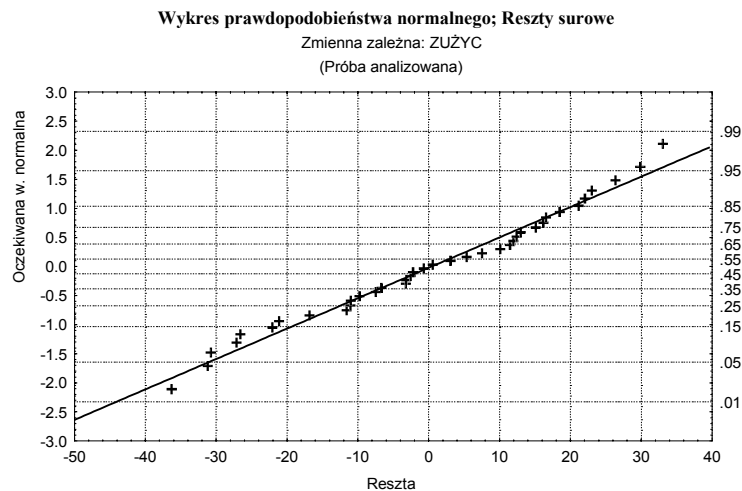
Oceny parametrów strukturalnych modelu zawierającego wszystkie cztery zmienne objaśniające (dla próby przeznaczonej do analizy) możemy uzyskać po kliknięciu przycisku *Współczynniki*. Przedstawia je poniższa tabela.



Oceny parametrów [zużycie_gazu_sta]							
VISUAL							
GSR							
Parametryzacja z sigma-ograniczeniami							
Efekt	Param.	BR. std.	ZUŻYC t	ZUŻYC p	-95.00% Gr. ufn.	+95.00% Gr. ufn.	ZUŻYC Beta (B)
TEMP	5.996	.43460	13.8011	.00000	5.115	6.8011	.78333
TEMP OP	1.157	.30583	2.9905	.00504	.373	1.9410	.15671
P WIATRU	1.945	.70475	2.7605	.00923	.513	3.3777	.08021
D WOLNE	-15.513	7.09854	-2.1854	.03585	-29.939	-1.0070	-.02736

Ważnym elementem weryfikacji modelu jest analiza reszt. Moduł *VGSR* umożliwia badanie najrozmaitszych charakterystyk i postępowań pozwalających na ocenę modelu regresyjnego z punktu widzenia zgodności obserwacji z próby ze wskazaniami modelu. Większość z nich opiera się na resztach, które są obliczane za pośrednictwem pewnych transformacji danych, takich jak standaryzacja, studentyzacja i inne. Opcje analizy reszt są dostępne w oknie wyników analizy na karcie *Reszty*.

Dla przykładu sprawdzimy obecnie normalność rozkładu reszt w obrębie próby, na podstawie której przeprowadzana była analiza. W tym celu na karcie *Reszty* należy kliknąć przycisk *Normalność reszt*. Poniżej zamieszczono odpowiedni wykres normalności rozkładu reszt.

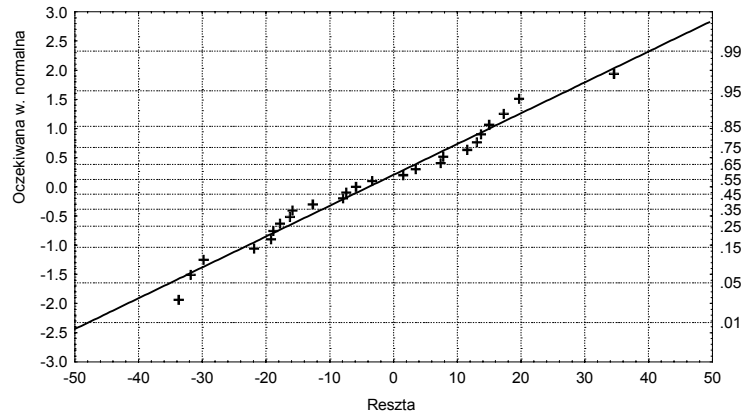


Jak widać, założenie normalności rozkładu reszt jest raczej spełnione. Podobnie wygląda sytuacja dla próby do oceny krzyżowej.



### Wykres prawdopodobieństwa normalnego; Reszty surowe

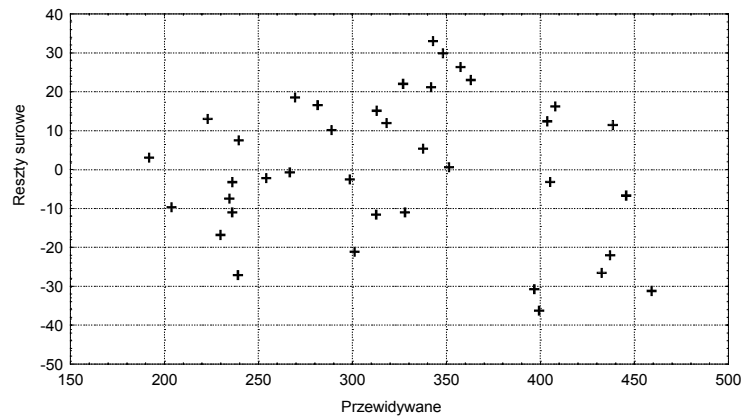
Zmienna zależna: ZUZYC  
(Próba do oceny krzyżowej)



Możemy jeszcze sprawdzić model pod kątem odstających obserwacji. Dla modelu oszacowanego na podstawie próby do analizy wykres wartości przewidywanych i reszt surowych wygląda jak poniżej

### Przewidywane względem wartości resztowych

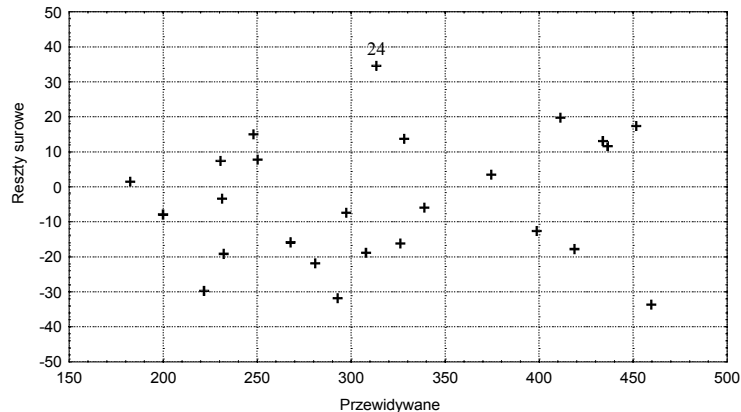
Zmienna zależna: ZUZYC  
(Próba analizowana)



Analogiczny wykres dla próby przeznaczonej do oceny krzyżowej wskazuje na jeden „podejrzany” punkt dla obserwacji o numerze 24.



## Przewidywane względem wartości resztowych

Zmienna zależna: ZUŻYC  
(Próba do oceny krzyżowej)

Kończąc prezentację możliwości modułu *VGSR*, pokażemy jeszcze możliwość zapisania w pliku tekstowym (np. dla szybkiego przeprowadzania podobnych analiz w przyszłości) składni analizy. W tym celu należy w oknie analizy kliknąć przycisk *Zmień*, a następnie przycisk *Edytor składni*. Poniżej przedstawiono składnię analizy dla przeprowadzanego przykładu.

```
GSR;  
DEPENDENT = ZUŻYC;  
GROUPS = none;  
COVARIATE = TEMP TEMP_OP P_WIATRU D_WOLNE;  
DESIGN = TEMP + TEMP_OP + P_WIATRU + D_WOLNE;  
INTERCEPT = exclude;  
LACKOFFIT = no;  
ESTIMATE = none;  
SDELTA = 7;  
IDELTA = 12;  
SURFACE = none;  
MIXTURE = none;  
SAMPLE = S_KRZYŻ ( 1.);  
MBUILD = bestsubset;  
FORCE = 0;  
BESTCRIT = mallowscp;  
START = 1;  
STOP = 4;  
MAXSUB = 8;  
OUTPUT = none;
```

Zbudowany model regresji można wykorzystywać do prognozowania przyszłych wartości lub (jeśli charakter danych na to zezwala) na symulowanie zachowania się zmiennej prognozowanej przy danych wartościach zmiennych objaśniających - jeśli zmienne objaśniające mają charakter zmiennych sterujących.



## SYSTEM INFORMATYCZNY WSPOMAGAJĄCY STOSOWANIE PROGNOZOWANIA W PRZEDSIĘBIORSTWIE

*mgr Jerzy Gurycz<sup>5</sup>*

### **Wymagania, jakie użytkownicy stawiają systemowi informatycznemu wspomagającemu stosowanie prognozowania w przedsiębiorstwie**

Firma StatSoft, w trakcie implementacji systemów analizy danych, odnotowuje jakie w opinii klientów istotne warunki musi spełniać dobry system informatyczny wspomagający prognozowanie w przedsiębiorstwie. Wśród najważniejszych cech takiego systemu znajdują się:

#### **Aktualność**

Warunkiem sukcesu w obecnym konkurencyjnym świecie jest wykonywanie analiz na bieżąco i szybkie reagowanie na zdarzenia. System musi zapewniać możliwość wykonania na żądanie, w krótkim czasie, prognoz z uwzględnieniem najświeższych danych, najlepiej w sposób jak najbardziej zautomatyzowany. Analiza danych przebiega zazwyczaj w cyklicznie następujących po sobie etapach: eksploracji danych, budowania modelu, weryfikacji poprawności modelu. Prognozowanie nie różni się pod tym względem od innych analiz statystycznych. Wynika z tego dodatkowe wymaganie - zapewnienie możliwości modyfikacji modelu prognozy w oparciu o dane bieżące. Oprócz automatycznej korekty modelu system powinien umożliwiać jego modyfikację przez analityka. Zadanie to (trudne lub wręcz niemożliwe do zrealizowania w niektórych programach i zamkniętych systemach informacyjnych) powinien ułatwiać dobry system prognozowania.

#### **Prostota i zaawansowanie**

Nie ulega wątpliwości, że wyniki analiz muszą być przedstawione w sposób przejrzysty, nieskomplikowany i możliwy do interpretacji dla każdego. Same analizy mogą (a często muszą) być zaawansowane. Proste zestawienia typu tygodniowa wartość sprzedaży są użyteczne (a system musi zapewniać możliwość ich wykonania), często już jednak nie wystarczają. Klienci widzą potrzebę korzystania w prognozowaniu z pełnego arsenału zaawansowanych technik analitycznych. Innymi słowy dane są na tyle złożone, że potrzeba zaawansowanych narzędzi, aby wydobyć z nich użyteczną wiedzę. Często konieczne jest stosowanie takich technik jak: analizy szeregów czasowych (np. modele wygładzania wykładniczego Holta i Wintersa, ARIMA), sieci neuronowe, drzewa klasyfikacyjne, analiza regresji wielokrotnej. Te metody analityczne (i wiele innych) dostępne są dzięki narzędziom analitycznym z rodziny *STATISTICA*.

Analizy stosowane w prognozowaniu są często zaawansowane, a do ich przygotowania i wykonania potrzebna jest specjalistyczna wiedza, którą posiada wąska grupa analityków, czasami konsultantów z zewnątrz. System musi zapewniać możliwość przechowania tej wiedzy i korzystania z niej przez

---

<sup>5</sup> StatSoft Polska Sp. z o.o.



osoby, które zainteresowane są jedynie wynikami końcowymi. Na tej właśnie zasadzie działa system *SENS*.

### Centralizacja i jednolitość

Wszystkie definicje analiz powinny być przechowywane centralnie, w sposób umożliwiający zainteresowanym osobom dostęp do raportów i bezproblemowe wykonanie odpowiednich analiz. Gwarantuje to zgodność wyników uzyskiwanych przez różnych osoby, ułatwia wymianę informacji i uporządkowanie analiz i raportów (znika znany doskonale każdemu problem „Raport którego potrzebuję musi być w którymś z segregatorów... Tylko w którym?”). Centralne przechowywanie szablonów analiz ułatwia również zarządzanie wiedzą korporacyjną: to jak tworzy się raport zapisane jest w bazie danych, a nie w głowie któregoś z analityków.

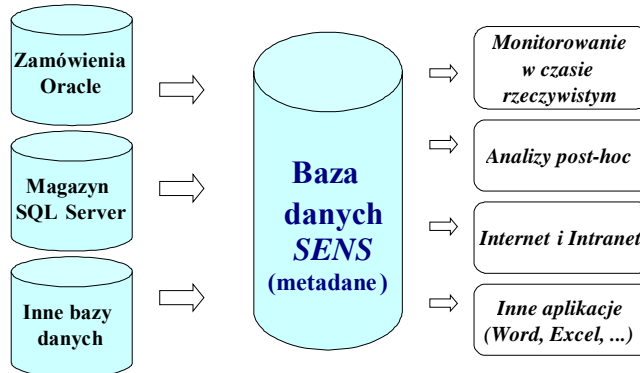
### Dostosowanie do użytkownika

Użytkownicy systemu różną się znacznie wiedzą i potrzebami. Powoduje to konieczność zapewnienia każdej grupie użytkowników odmiennego środowiska pracy. Oczywiście jest, że inne środowisko jest odpowiednie dla analityka, a inne dla osoby korzystającej tylko i wyłącznie z wyników jednej analizy. Dostyc często stawiany jest postulat, aby system umożliwiał odbiorcom analiz korzystanie z wyników analiz w swoim dla nich środowisku np. przeglądarki internetowej, edytora tekstu czy arkusza kalkulacyjnego. Podsumowując, to system powinien dostosowywać się do użytkownika, a nie użytkownik do systemu.

## Architektura systemu *SENS*

StatSoft jako kompleksowy system analityczny wspomagający wnioskowanie na podstawie danych, a w szczególności prognozowanie proponuje *STATISTICA Enterprise-Wide System (SENS)*. System ten został zaprojektowany tak, aby spełnić wymienione wyżej wymagania.

*STATISTICA Enterprise Wide System (SENS)* to kompleksowy system analizy danych. W jego skład wchodzi narzędzia analityczne *STATISTICA*, mechanizmy dostępu do zewnętrznych baz danych, generator raportów oraz narzędzia pracy grupowej zorganizowane w oparciu o dedykowaną hurtownię danych. Dzięki przejrzystemu, graficznemu środowisku *SENS* umożliwia łatwe tworzenie zapytań do nawet bardzo złożonych, niejednorodnych baz danych i nie wymaga od użytkownika wiedzy informatycznej. Jako narzędzie analityczne *SENS* wykorzystuje program *STATISTICA*, wielokrotnie uznawany za najlepszy program statystyczny.





Na rysunku powyżej przedstawiona jest architektura systemu *SENS*. W systemowej bazie danych *SENS* (centralna część schematu) przechowywane są metadane, czyli dostosowany do celów analitycznych opis danych zewnętrznych, zapis modeli prognoz. Baza *SENS* stanowi warstwę tłumaczącą. W bazie tej przechowywane są również informacje o tym, jakie elementy systemu będą dostępne dla poszczególnych użytkowników. Warto zaznaczyć, iż w bazie *SENS* nie są przechowywane dane surowe (źródłowe), można w niej natomiast przechowywać wyniki analiz.

## Źródła danych do analiz

W organizacjach zbierane są różnorodne dane w różnych dedykowanych systemach informacyjnych. Do prognoz można wykorzystać te już istniejące dane. Właściwe modele prognostyczne mogą jednak wymagać uwzględnienia dodatkowych istotnych informacji o zmianach wartości czynników zewnętrznych takich jak stan pogody (jak wpływa ona na sprzedaż doskonale wiedzą producenci lodów czy piwa), zmiany kursów walut, czy danych o natężeniu akcji promocyjnych. W trakcie implementacji systemu może zająć konieczność uzupełnienia istniejących baz danych o te informacje ewentualnie utworzenia odświeżanych okresowo pomocniczych baz danych uwzględniających dodatkowe dane zewnętrzne.

Źródłami danych dla systemu informatycznego wspomagającego stosowanie prognozowania w przedsiębiorstwie są:

- ◆ Dane z istniejących baz danych (transakcyjnych) w działających systemach informatycznych (np. dane z systemu finansowo-księgowego czy systemu ERP).
- ◆ Dane z tematycznych baz danych
- ◆ Dane z pośrednich baz danych
- ◆ Dane z hurtowni danych

## Obiekty systemu *SENS*

### Użytkownicy i grupy użytkowników

System użytkowników i grup umożliwia określanie uprawnień poszczególnych osób korzystających z systemu do odpowiednich analiz, raportów i definicji połączeń ze źródłami danych. Użytkownik po zalogowaniu widzi tylko te obiekty w schemacie systemu (monitory i profile), które powinien widzieć (do których ma uprawnienia). Warto dodać, iż ilość użytkowników systemu jest praktycznie nieograniczona, a sposób licencjonowania oparty jest na maksymalnej liczbie użytkowników korzystających z systemu równocześnie, a nie na ogólnej liczbie użytkowników.

### Profile OLE DB

Profile definiują, jakie dane mają być odczytywane z hurtowni danych lub zewnętrznych źródeł danych i w jaki sposób interpretowane w *SENS* przy wykonywaniu analiz. Profile OLEDB zawierają tworzone w interfejsie graficznym zapytania oraz różnorodne opcje np. format danych, określenie względem jakich wymiarów (np. region, czas) dane będą mogły być filtrowane przez użytkowników przed wykonaniem na nich analizy.

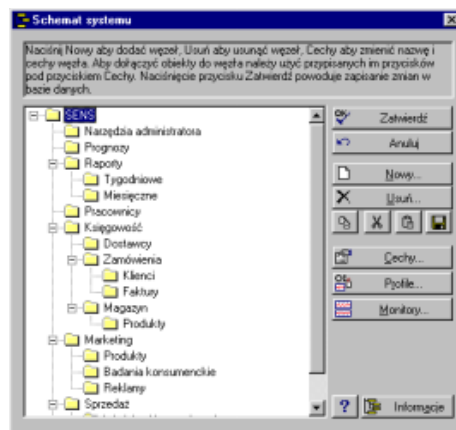


## Monitory (aktywne raporty)

Monitory zawierają definicje wykonywanych analiz, wykresów i raportów oraz informację, jakich danych mają one dotyczyć (innymi słowy z jakiego profilu będą korzystać). Monitory mogą być aktualizowane na bieżąco, po każdej modyfikacji danych. Monitory są aktywnymi raportami, tzn. tworzone są na podstawie aktualnych danych i mogą być modyfikowane przez użytkownika (np. można zmienić skalę wykresu, tak aby uwypuklić interesujące nas zależności).

## Schemat systemu

Schemat systemu to hierarchiczna struktura, która porządkuje wszystkie obiekty systemu.



Umożliwia on uporządkowanie informacji w systemie, współdzielenie wiedzy (użytkownicy mogą np. za jego pomocą udostępniać przygotowane przez siebie monitory innym użytkownikom systemu) i łatwe odnajdywanie potrzebnych profili, monitorów, właściwości i innych obiektów systemu.

Schemat całego systemu może być nawet bardzo rozbudowany, ważne jest jednak że użytkownik widzi tylko tę jego część, w której znajdują się interesujące go analizy i raporty (mogą być to np. tylko trzy najistotniejsze wykresy dla danych za ostatni kwartał, a może być cała gama analiz).

## Aspekty techniczne

### Infrastruktura informatyczna

*SENS* instalowany jest na serwerze z systemem zarządzania bazami danych i stanowiskach roboczych. Jako system operacyjny stanowisk roboczych wykorzystywana jest platforma Windows (95, 98, NT, 2000). Oczywiście niezbędna jest sieć komputerowa.

Systemowa baza danych *SENS* może być zainstalowana w oparciu o system zarządzania relacyjną bazą danych Oracle, MS SQL Server, IBM DB2, Informix, Sybase (i każdy inny zgodny z ODBC). Zazwyczaj w implementacji *SENS* wykorzystywany jest system już działający w firmie, którego obsługa nie stanowi problemu dla działu informatyki danej firmy. Dla niewielkich instalacji system zarządzania nie jest konieczny, *SENS* może korzystać z własnej bazy danych.

Użytkownicy potrzebujący jedynie dostępu do aktualnych wyników analiz mogą korzystać z *SENS* przez przeglądarkę internetową. Na takim stanowisku nie ma konieczności instalowania



oprogramowania *SENS*, wystarcza odpowiednie skonfigurowanie przeglądarki i dostępu do Internetu/Intranetu.

### Dostęp do danych

Dostęp do danych realizowany jest przez mechanizm OLE DB. Umożliwia to korzystanie z baz relacyjnych dla których istnieje odpowiedni provider OLE DB lub sterownik ODBC oraz innych źródeł danych (np. plików arkuszy kalkulacyjnych). Dzięki zastosowaniu OLE DB przedmiotem analiz w *SENS* mogą być dane przechowywane w praktycznie wszystkich aktualnie wykorzystywanych systemach zarządzania bazą danych (np. MS Access, Oracle, IBM DB2, MS SQL Server, Informix, Sybase,...) pracujących na dowolnej platformie nie tylko PC/Windows. Ponieważ standard OLE DB jest otwarty, to dla nietypowych źródeł danych można stworzyć własne sterowniki.

### Analiza danych

Motor analityczny systemu stanowi *STATISTICA*, w której analizy statystyczne prowadzone są na danych z plików lokalnych. Istnieje możliwość ekstrakcji danych w *SENS* i dalszej analizy za pomocą *STATISTICA*, na przykład na komputerze przenośnym.

### Udostępnianie wyników

Raporty w *SENS* (zawierające wykresy i arkusze wyników analiz) mogą być udostępniane zainteresowanym osobom na *różne* sposoby:

- ◆ publikacja w Internecie/Intranecie w formacie HTML do przeglądania za pomocą przeglądarki internetowej,
- ◆ zapis w postaci dokumentów w formacie RTF, które mogą być odczytane przez prawie wszystkie współczesne edytory tekstu
- ◆ przeglądanie, drukowanie, wykorzystanie w innych aplikacjach uruchamianych na żądanie monitorów.

Raporty *można* zapisać w plikach zewnętrznych, mogą one być również archiwizowane w systemowej bazie *SENS*.

### Instalacja

Instalacja systemu jest stosunkowo prosta - sprowadza się do założenia na serwerze z systemem zarządzania bazami danych (SZBD) schematu hurtowni danych oraz do instalacji plików programu na stanowiskach (zazwyczaj jest to instalacja sieciowa na serwerze, z którego następnie może być wykonana instalacja na stanowiskach, nawet przez samych użytkowników). W przypadku niewielkich instalacji systemy mogą działać bez SZBD, baza systemu jest wtedy instalowana automatycznie w postaci pliku MS Access.

Mechanizmy dostępu do danych są rozbudowane i wykorzystują najnowsze technologie informatyczne (dzięki czemu możliwy jest dostęp do przeróżnych typów danych). Jeżeli na danym stanowisku mają być wykonywane analizy w oparciu o dane zewnętrzne przechowywane w specjalnym formacie może być konieczne zainstalowanie odpowiedniego sterownika (providera OLE DB). Przy instalacji systemowej bazy danych *SENS* wykorzystywane jest natomiast niezbędne minimum standardu ODBC i można ją wykonać na większości SZBD zgodnych z tym standardem.



## Konfiguracja

*SENS* jest systemem z półki. Wymagane jest zwykle pewne przystosowanie go do konkretnych potrzeb (wstępna konfiguracja), jednak jest to proces prosty i szybki. Ponieważ sytuacja każdej firmy jest inna, zdarza się iż w fazie implementacji potrzebne jest tworzenie dodatkowej hurtowni danych, zbudowanie właściwych mechanizmów czyszczenia czy określanie poziomów agregacji danych. Warto też poświęcić czas na przygotowanie odpowiednich modeli analitycznych.

System jest rozwiązaniem sprawdzonym i unika się wielu pułapek związanych z aplikacjami tworzonymi na zamówienie (długie czasy wdrożenia, problemy z utrzymaniem się w harmonogramie, odkrywanie błędów konstrukcyjnych w trakcie pracy z systemem i ciągłe ich usuwanie). Narzędzia administracyjne pozwalają łatwo rozbudowywać system o nowe analizy, raporty, dodawać nowych użytkowników.

## Administracja

Zadania administracyjne w systemie *SENS* sprowadzają się do niezbędnego minimum. Wiele czynności, które często spadały na barki informatyka (na przykład tworzenie odpowiednich zapytań), może być wykonanych przez zainteresowanych użytkowników już po krótkim przeszkoleniu. Dobrze skonfigurowany system wymaga bardzo rzadkich ingerencji administratora. Jego skalowalność zapewnia, że przy dodawaniu nowych użytkowników, działów, analiz, nie jest potrzebna przebudowa systemu.

Zmiany zapisane w *SENS* (wprowadzone przez administratora) są od razu uwzględniane na stacjach roboczych użytkowników. Wykorzystanie kreatorów przy budowie profili i monitorów pozwala bardziej zaawansowanym użytkownikom tworzyć różnorodne raporty i analizy i udostępniać je osobom z odpowiedniego działu czy grupie roboczej. Zadanie administratora sprowadza się do zapewnienia dostępu właściwym osobom do właściwych danych (lub zapewnienia braku dostępu do tych danych innym użytkownikom). Przechowywanie w *SENS* definicji połączeń z danymi i ustalanie uprawnień do korzystania oraz modyfikowania tych połączeń przy budowie profili zapewniają właściwą kontrolę i bezpieczeństwo.

## Licencjonowanie

Warto zwrócić uwagę na korzystny dla użytkownika sposób licencjonowania systemu *SENS*. Licencje dotyczą elementów systemu wykorzystywanych w danej chwili – przykładowo w systemie można zdefiniować 50 różnych operatorów, ale jeżeli jednocześnie będzie pracowało 5, to wystarczy wykupić licencje na 5 operatorów. Licencje dotyczą możliwości analitycznych, odpowiadających konkretnemu programowi z rodziny *STATISTICA*, np. samej *STATISTICA* lub *Quick STATISTICA*.