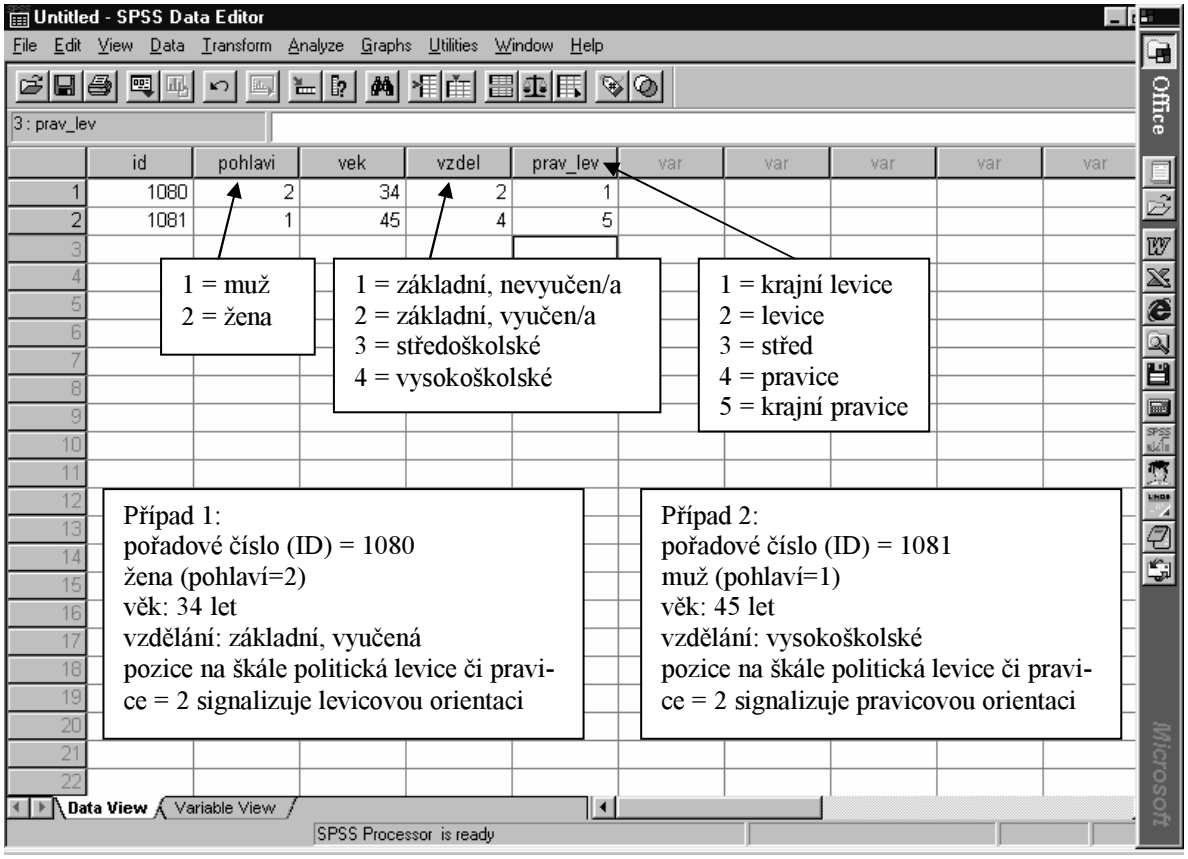
LEKCE 1
POVAHA HROMADNÝCH DAT A LOGIKA SURVEY. PRÁCE S HROMADNÝMI DATY PŘED JEJICH ANALÝZOU

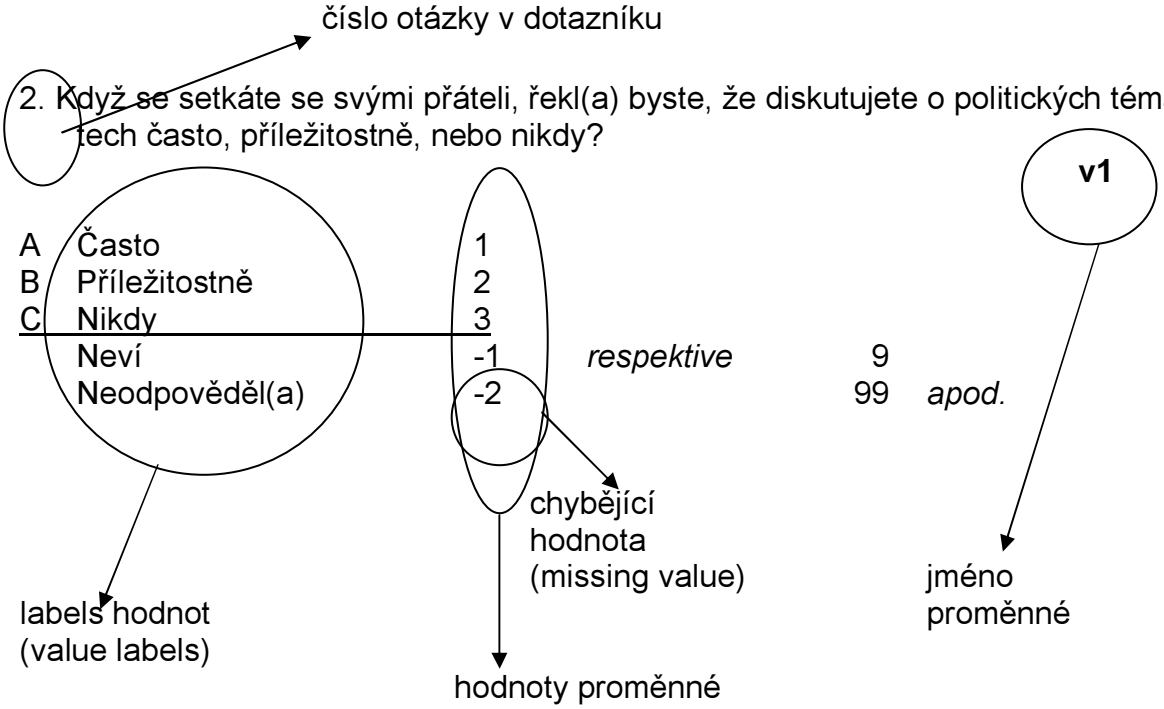
MATICE DAT

Protože jde o zpracování hromadných dat, pracujeme s kvantifikovanými charakteristikami případů (respondentů či jiných objektů, popřípadě aktů - charakterizovat můžeme například komunikaci, jednání apod.).

- Případy jsou popsány svými vlastnostmi (atributy) - variantami neboli hodnotami proměnných, které jsou jejich logickými uskupeními. Například proměnná vzdělání může být uskupením možných nejvyšších dosažených stupňů vzdělání: základní, středoškolské, vysokoškolské (které lze popřípadě dále členit: základní nedokončené, základní bez vyučení, základní s vyučením etc.).
- Každý případ tak představuje vektor obsahujících hodnoty příslušných proměnných (každá varianta každé proměnné má přiřazenu číslici).
- Vektory plníme do matice: co řádek, to případ (např. respondent) a co sloupec, to proměnná.



OTÁZKA V DOTAZNÍKU JAKO PROMĚNNÁ



přidáme:
label proměnné (variable label)

Co s variantami

- *Varianta „nevím“ a „neodpověděl/a“.*
- *Varianta „nevím“ a úroveň měření.*

BATERIE OTÁZEK V DOTAZNÍKU JAKO SADA PROMĚNNÝCH

1. Řekněte prosím o každé z následujících skutečností, jak je ve Vašem životě důležitá:

		Velmi důležitá	Dost důležitá	Ne příliš důležitá	Vůbec ne důležitá	Neví	Neodpověděl(a)	
A	Práce	1	2	3	4	-1	-2	v1a
B	Rodina	1	2	3	4	-1	-2	v1b
C	Přátelé a známí	1	2	3	4	-1	-2	v1c
D	Volný čas	1	2	3	4	-1	-2	v1d
E	Politika	1	2	3	4	-1	-2	v1a
F	Náboženství	1	2	3	4	-1	-2	v1f

Zde je každý řádek proměnnou s oborem hodnot <1;4>, záporné hodnoty představují missing value. Možná jména proměnných například: Q1_1 až Q1_6 napovídají, že všech 6 proměnných má něco společného.

DEFINICE JEDNOTLIVÝCH PROMĚNNÝCH

Abychom mohli matici naplnit, musíme ji nejprve definovat. Děje se tak v modu VARIABLE VIEW.

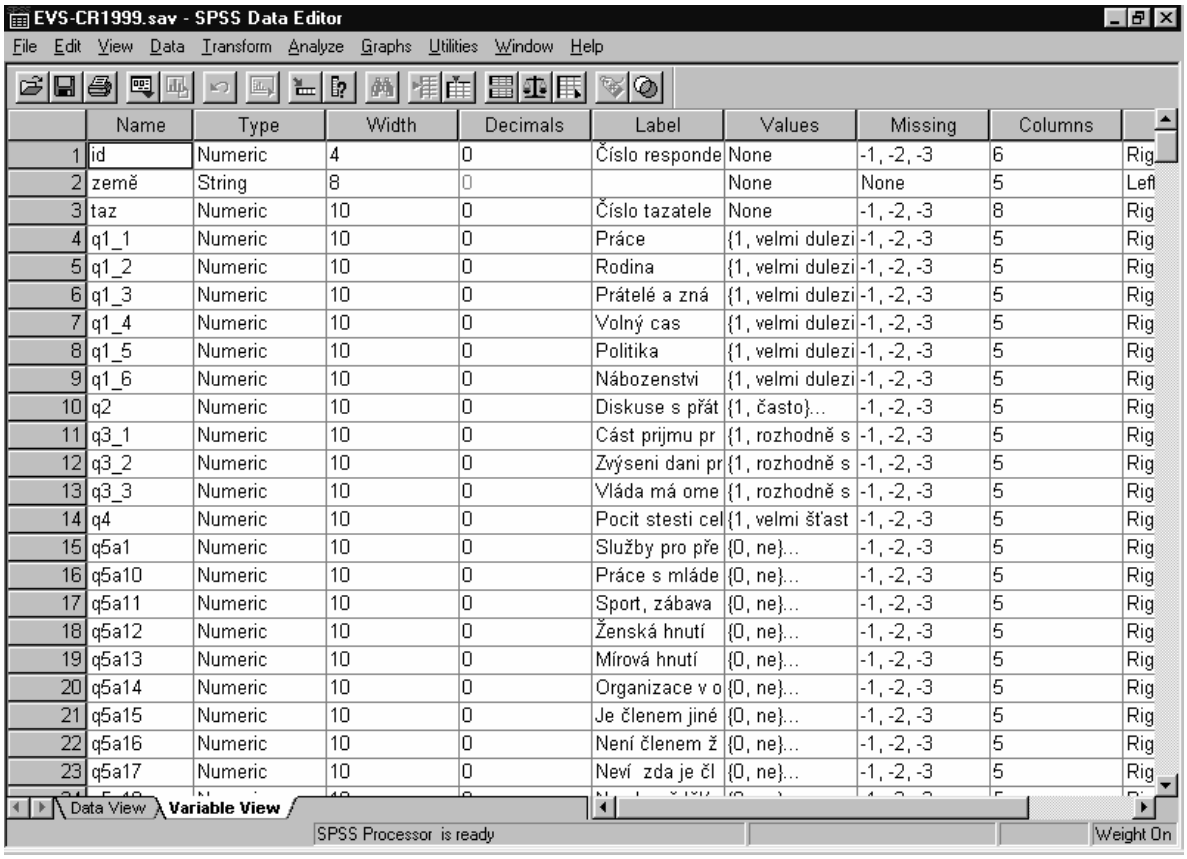
Jde o tyto úkony:

- Připsání jména proměnné, určení jejího místa v matici (sloupce/sloupců).
- Definice charakteru proměnné jako numerické či stringové (alfaznakové, kterou počítač chápe jako označení a neprovádí s ní početní operace) apd.
- Připsání širšího označení proměnné (variable labels).
- Připsání širšího označení jednotlivým hodnotám proměnné (value labels).

  Labels zpřehledňují tištěné výstupy, neboť přiřazují k jménům proměnných (jež mohou mít dle konvence pouze 8 znaků) i vysvětlující popis. Např. q1_2 (jméno proměnné neboli name) Význam rodiny v životě (label proměnné neboli value label).

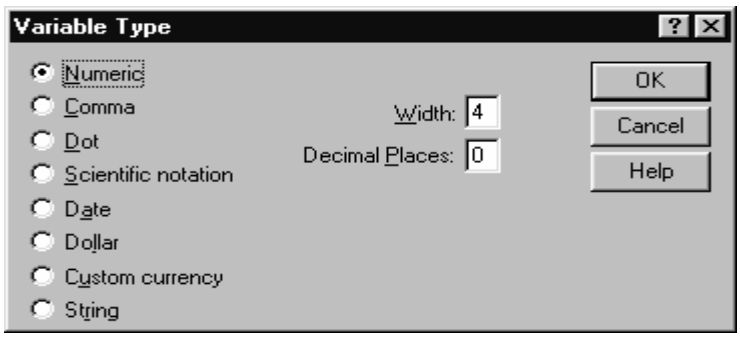
- Určení počtu desetinných míst.
  Pozor: souvisí s definicí počtu požadovaných sloupců v matici pro proměnnou.
- Definování tzv. missing value.

  Většinou se z analýzy (dočasně - jen pro danou operaci) případy s missing value vyřazují.



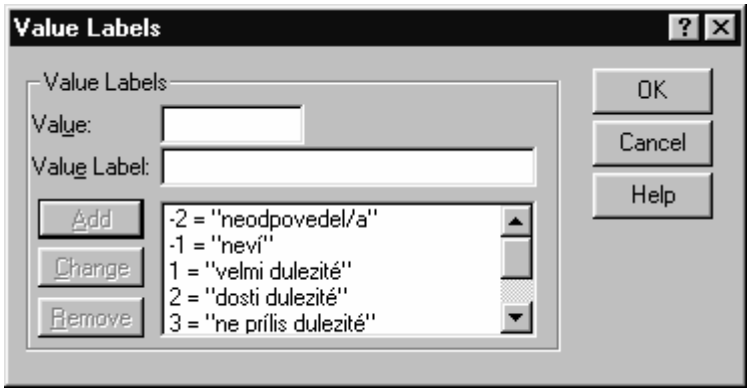
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	
1	id	Numeric	4	0	Číslo responde	None	-1, -2, -3	6	Rig
2	země	String	8	0		None	None	5	Left
3	taz	Numeric	10	0	Číslo tazatele	None	-1, -2, -3	8	Rig
4	q1_1	Numeric	10	0	Práce	{1, velmi dulezi	-1, -2, -3	5	Rig
5	q1_2	Numeric	10	0	Rodina	{1, velmi dulezi	-1, -2, -3	5	Rig
6	q1_3	Numeric	10	0	Prátelé a zná	{1, velmi dulezi	-1, -2, -3	5	Rig
7	q1_4	Numeric	10	0	Volný cas	{1, velmi dulezi	-1, -2, -3	5	Rig
8	q1_5	Numeric	10	0	Politika	{1, velmi dulezi	-1, -2, -3	5	Rig
9	q1_6	Numeric	10	0	Nábozenstvi	{1, velmi dulezi	-1, -2, -3	5	Rig
10	q2	Numeric	10	0	Diskuse s přát	{1, casto}...	-1, -2, -3	5	Rig
11	q3_1	Numeric	10	0	Část prijmu pr	{1, rozhodně s	-1, -2, -3	5	Rig
12	q3_2	Numeric	10	0	Zvýseni dani pr	{1, rozhodně s	-1, -2, -3	5	Rig
13	q3_3	Numeric	10	0	Vláda má ome	{1, rozhodně s	-1, -2, -3	5	Rig
14	q4	Numeric	10	0	Pocit stesti cel	{1, velmi šťast	-1, -2, -3	5	Rig
15	q5a1	Numeric	10	0	Služby pro pře	{0, ne}...	-1, -2, -3	5	Rig
16	q5a10	Numeric	10	0	Práce s mláde	{0, ne}...	-1, -2, -3	5	Rig
17	q5a11	Numeric	10	0	Sport, zábava	{0, ne}...	-1, -2, -3	5	Rig
18	q5a12	Numeric	10	0	Ženská hnutí	{0, ne}...	-1, -2, -3	5	Rig
19	q5a13	Numeric	10	0	Mírová hnutí	{0, ne}...	-1, -2, -3	5	Rig
20	q5a14	Numeric	10	0	Organizace v o	{0, ne}...	-1, -2, -3	5	Rig
21	q5a15	Numeric	10	0	Je členem jiné	{0, ne}...	-1, -2, -3	5	Rig
22	q5a16	Numeric	10	0	Není členem ž	{0, ne}...	-1, -2, -3	5	Rig
23	q5a17	Numeric	10	0	Neví zda je čl	{0, ne}...	-1, -2, -3	5	Rig

Vymezení typu proměnné a počtu desetinných míst
(v výjimkou kardinálních proměnných desetinných míst nepoužíváme).



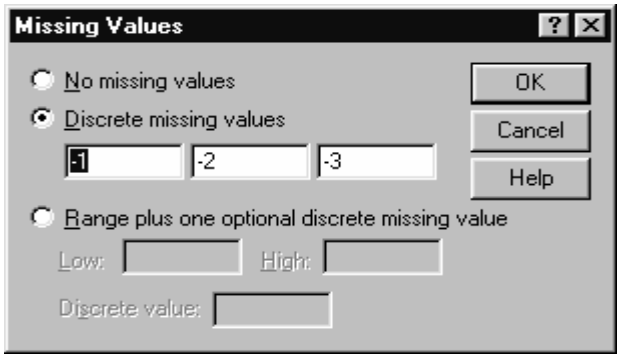
Vymezení labels

Variable label se píše do příslušného sloupce přímo, value labels zapíšeme do vyvolaného formuláře.



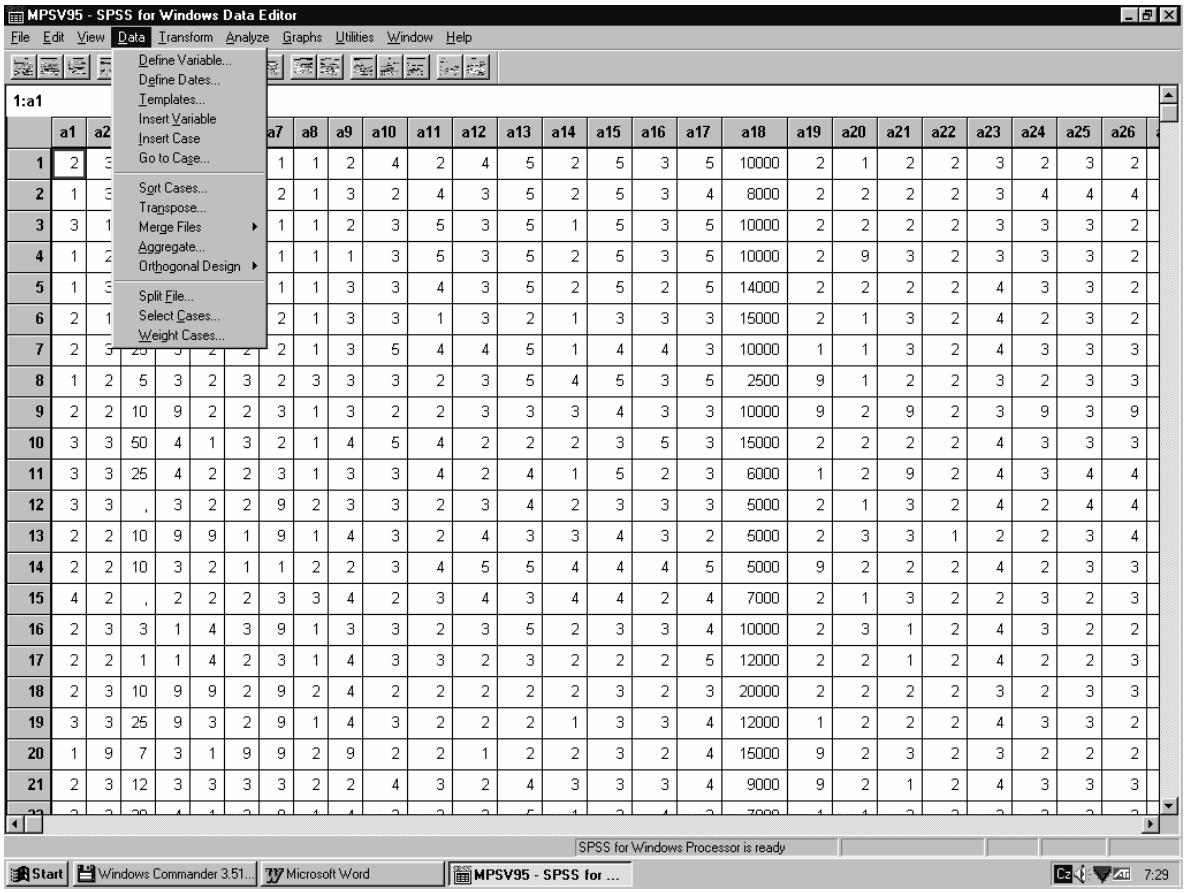
Vymezení missing value

Missing value jsou hodnoty, které nevcházejí (pokud si to výslovně nepřejeme a nezadáme) do analýzy. Jsou to kódy například pro případ, že respondent na otázku neodpověděl, odpověděl variantou nevím etc.



PLNĚNÍ MATICE DATY

Děje se tak zatím nejčastěji vkládáním jednotlivých hodnot (navedení jednotlivých dotazníků) do prázdné definované matice (definujeme ji popisem proměnných – viz). Výsledkem je matice dat, která může být dále upravována (například pomocí transformací proměnných nebo výběrem případů) a analyzována.

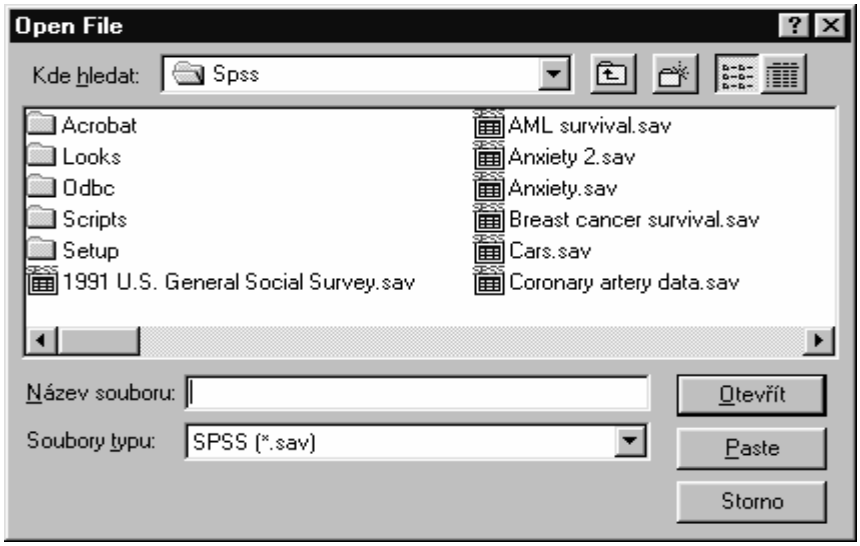


Data ovšem můžete dostat do matice i jinými způsoby. Důležité jsou pro nás zejména:

- Otevřeni již existujícího souboru. V SPSS již dříve vytvořené a uložené matice dat neboli systémové soubory mají příponu .sav,, soubory vytvořené ještě v době, kdy program pracoval pod operačním systémem DOS mohou mít přílohu .sys (tyto soubory lze také otevřít, je však třeba při jejich otevírání tuto možnost nastavit). Systémové soubory s příponou sav. Lze ve Wincommandru často spustiti zakliknutím (poku mají definovánu vazbu na SPSS jako prohlížeč (pokud tomu tak není, nezbývá než nejprve spustit SPSS a teprve v něm pomocí FILE $\rightarrow$ OPEN $\rightarrow$ DATA soubor natáhnout.
- Import dat ze souboru jiného typu (z textového editoru, databáze či spredsheetového programu jako je Excel).

OTEVŘENÍ SYSTÉMOVÉHO SOUBORU

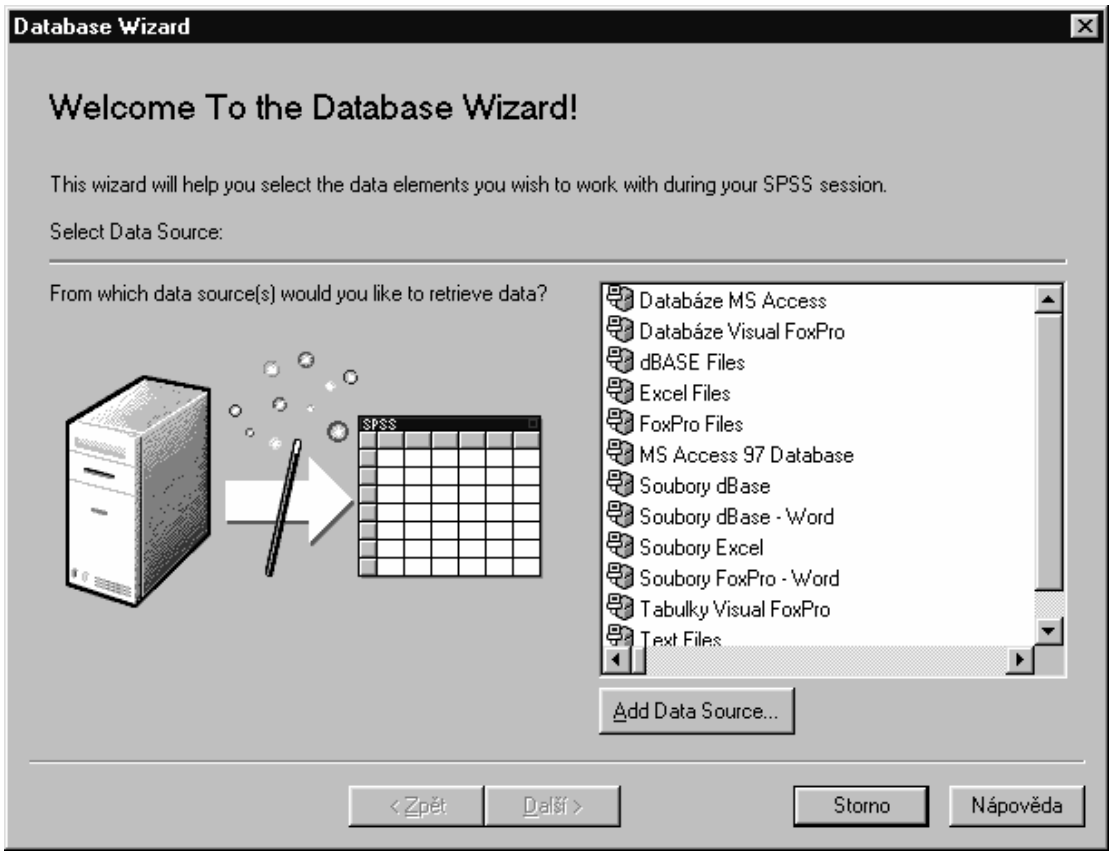
FILE → OPEN → DATA



Program si pamatuje soubory, s nimiž naposledy pracoval, lze je spustit přímo z FILE.

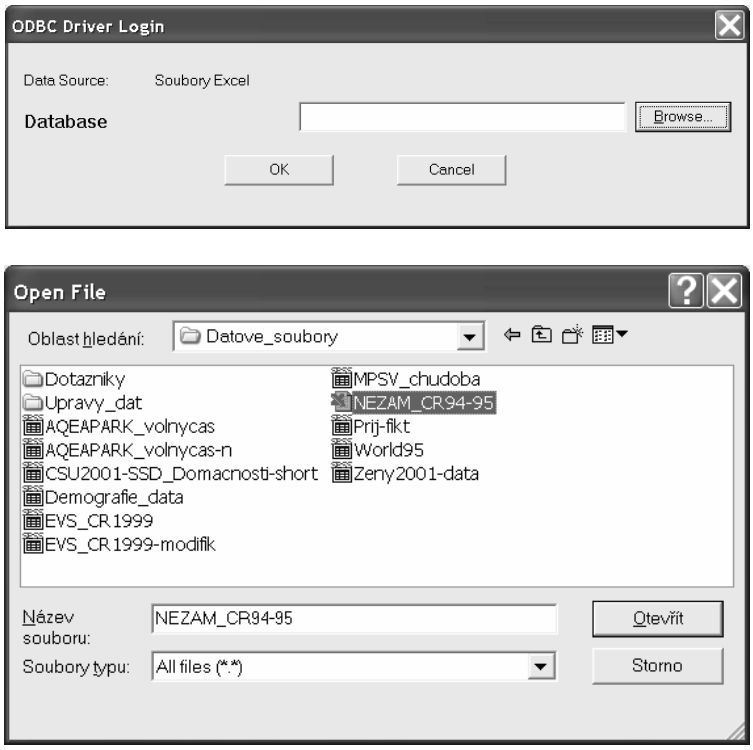
PŘEVOD DATABÁZOVÉHO SOUBORU

FILE → OPEN DATABASE → NEW QUERY

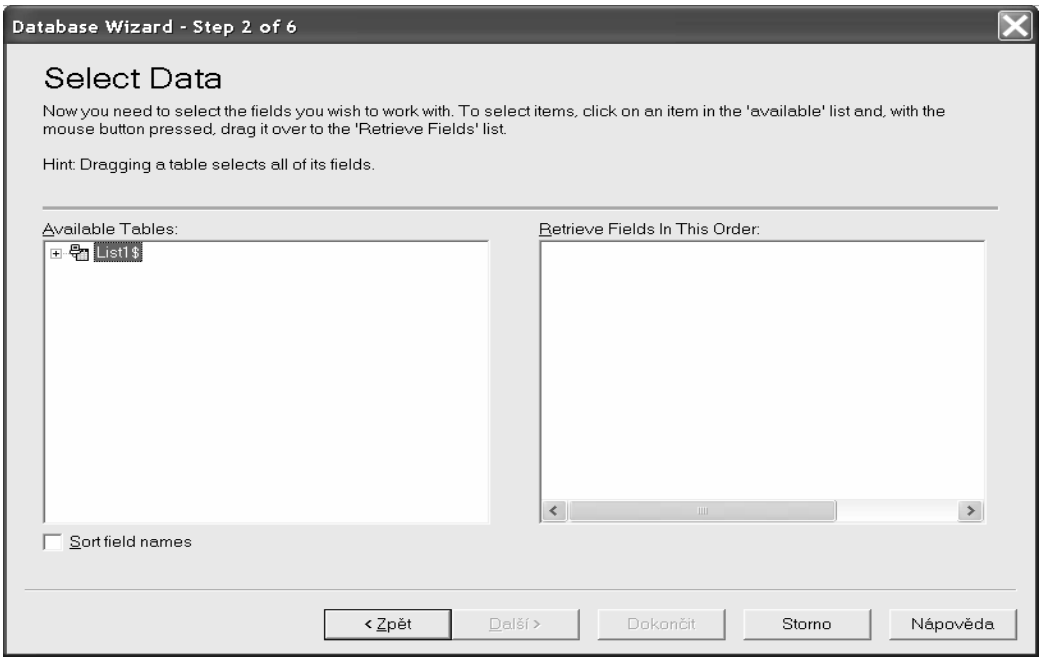


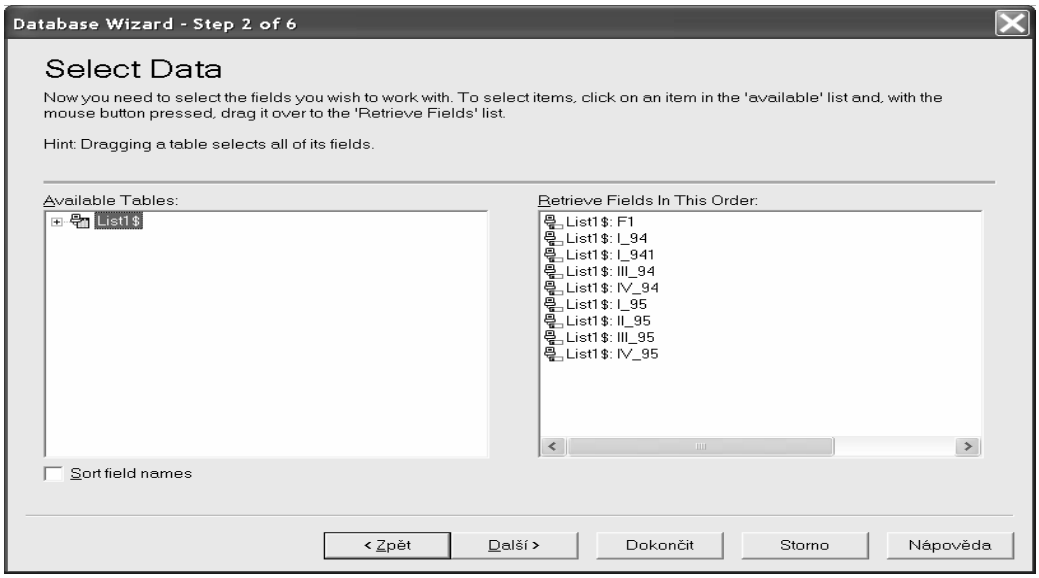
Zvolíme typ souboru (např. EXCEL files).

Najdeme příslušný soubor pomocí *Browse*:

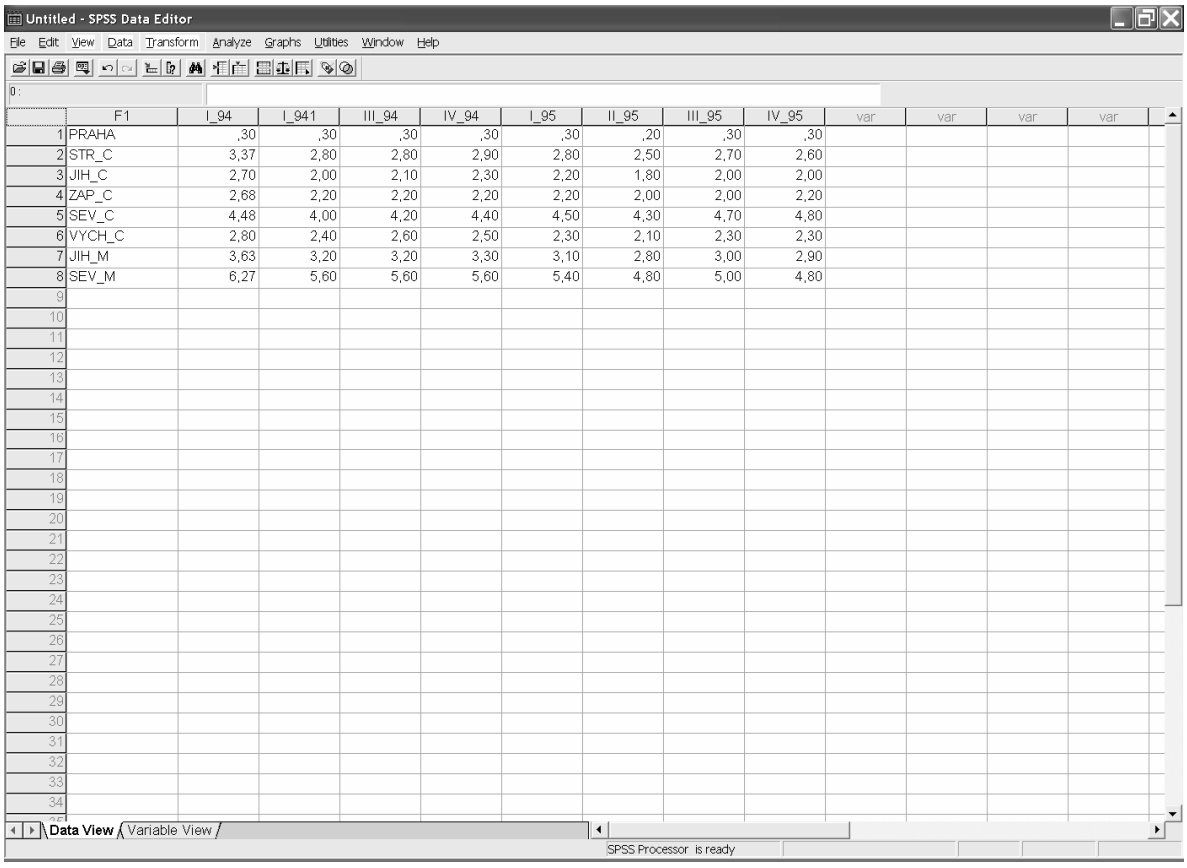


Otevřeme ho a odsouhlasíme (OK v ODBC Driver Login). Pak přetáhnu pomocí myši List z levého do pravého okna.





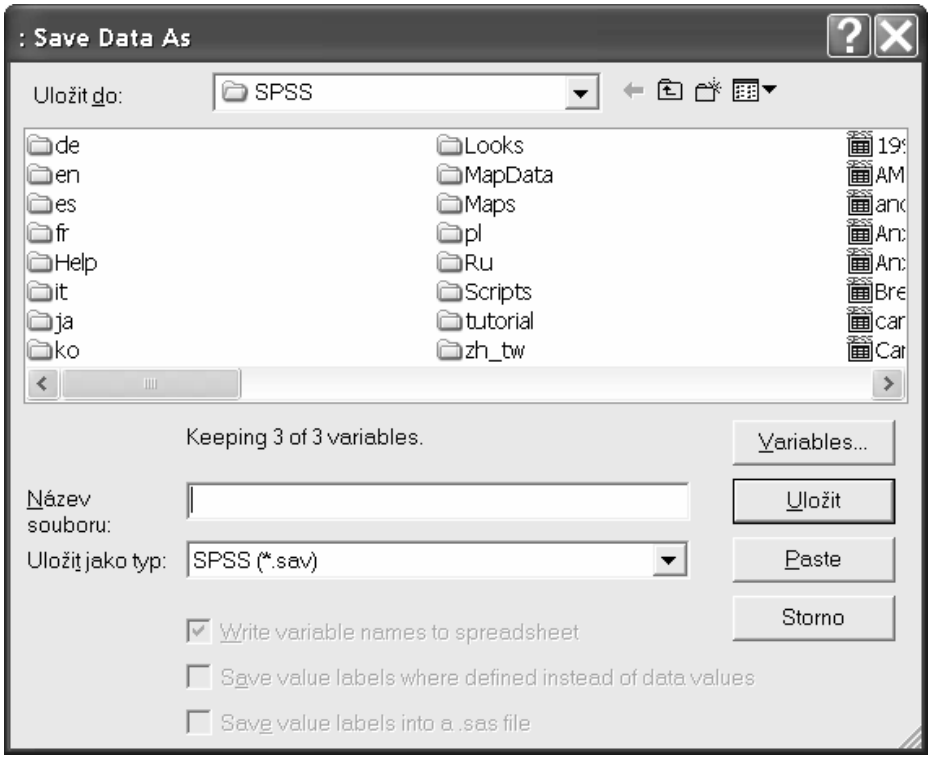
Pomocí *Další* mohu omezit přetahované případy, nebo mohu *Dokončit.* Obsah Excelového souboru je přetažen do systémového souboru SPSS. Je to matice dat i se sloupcem představujícím jména bývalých krajů (proměnnou F1 mohu v okně *VARIABLES VIEW* přejmenovat)) a jmény proměnných (jednotlivá čtvrtletí let 1994 a 1995). Data v matici představují příslušné míry nezaměstnanosti v daných krajích (kraje jsou případy) v těchto čtvrtletích (čtvrtletí jsou proměnnými a data v dané kolonce vždy hodnotou dané proměnné – svou povahou jsou to kardinální/spojité proměnné).



	F1	I_94	I_941	III_94	IV_94	I_95	II_95	III_95	IV_95	var	var	var	var
1	PRAHA	,30	,30	,30	,30	,30	,20	,30	,30				
2	STR_C	3,37	2,80	2,80	2,90	2,80	2,50	2,70	2,60				
3	JIH_C	2,70	2,00	2,10	2,30	2,20	1,80	2,00	2,00				
4	ZAP_C	2,68	2,20	2,20	2,20	2,20	2,00	2,00	2,20				
5	SEV_C	4,48	4,00	4,20	4,40	4,50	4,30	4,70	4,80				
6	VYCH_C	2,80	2,40	2,60	2,50	2,30	2,10	2,30	2,30				
7	JIH_M	3,63	3,20	3,20	3,30	3,10	2,80	3,00	2,90				
8	SEV_M	6,27	5,60	5,60	5,60	5,40	4,80	5,00	4,80				

UKLÁDÁNÍ SOUBORŮ

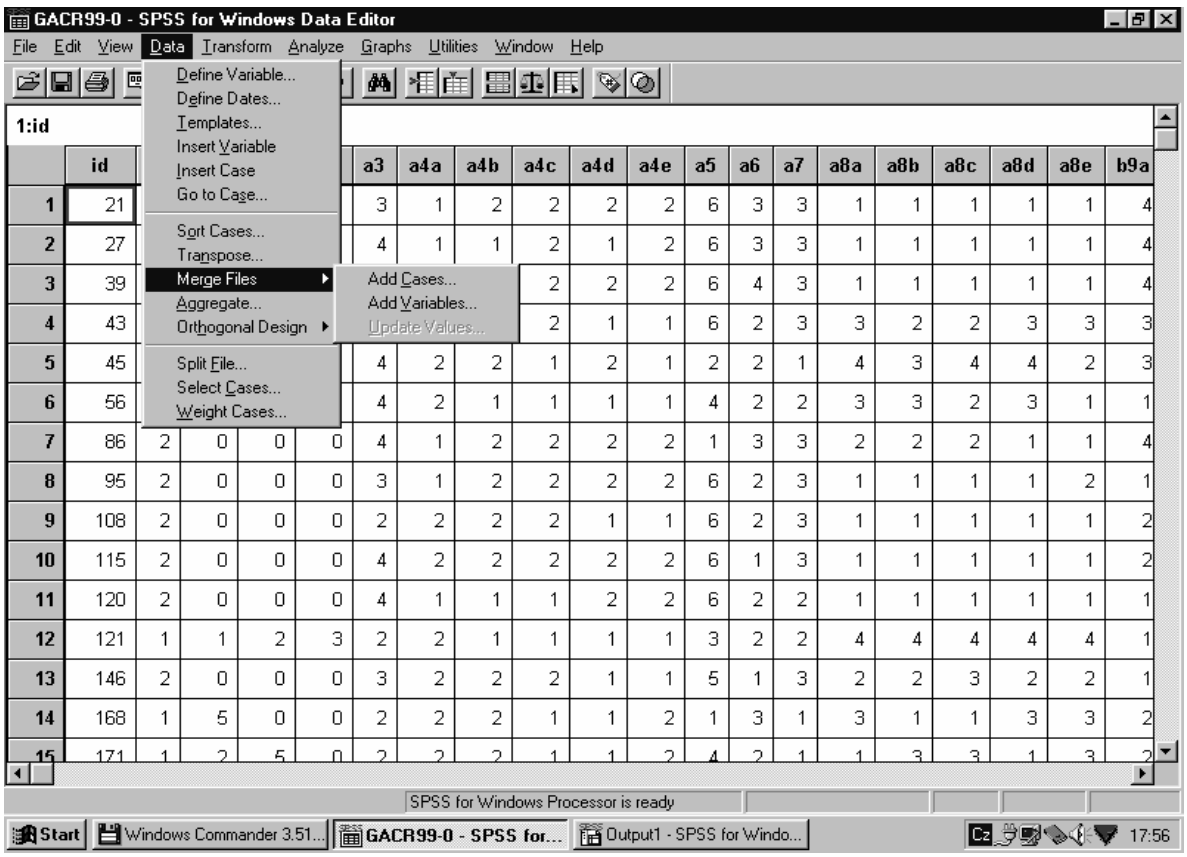
Data je třeba uložit (jako soubor s příponou .sav, což je systémový soubor, obsahující popsanou matici neboli definované a popsané proměnné a jejich hodnoty, naplněnou daty).



Ukládejte soubor po každé změně (přidání případu nebo vytvoření nových proměnných – viz lekce věnovaná transformaci proměnných). Ponechávejte (samozřejmě pod různými názvy):

- Pramenný soubor (naplněná a zkontrolovaná původní matice, v níž nebyly provedeny žádné další změny).
- Předposlední podobu souboru.
- Poslední podobu souboru.

SLUČOVÁNÍ SOUBORŮ - ADD CASES



	id					a3	a4a	a4b	a4c	a4d	a4e	a5	a6	a7	a8a	a8b	a8c	a8d	a8e	b9a
1	21					3	1	2	2	2	2	6	3	3	1	1	1	1	1	4
2	27					4	1	1	2	1	2	6	3	3	1	1	1	1	1	4
3	39								2	2	2	6	4	3	1	1	1	1	1	4
4	43								2	1	1	6	2	3	3	2	2	3	3	3
5	45					4	2	2	1	2	1	2	2	1	4	3	4	4	2	3
6	56					4	2	1	1	1	1	4	2	2	3	3	2	3	1	1
7	86	2	0	0	0	4	1	2	2	2	2	1	3	3	2	2	2	1	1	4
8	95	2	0	0	0	3	1	2	2	2	2	6	2	3	1	1	1	1	2	1
9	108	2	0	0	0	2	2	2	2	1	1	6	2	3	1	1	1	1	1	2
10	115	2	0	0	0	4	2	2	2	2	2	6	1	3	1	1	1	1	1	2
11	120	2	0	0	0	4	1	1	1	2	2	6	2	2	1	1	1	1	1	1
12	121	1	1	2	3	2	2	1	1	1	1	3	2	2	4	4	4	4	4	1
13	146	2	0	0	0	3	2	2	2	1	1	5	1	3	2	2	3	2	2	1
14	168	1	5	0	0	2	2	2	1	1	2	1	3	1	3	1	1	3	3	2
15	171	1	2	5	0	2	2	2	1	1	2	4	2	1	1	3	3	1	3	2

ÚLOHA

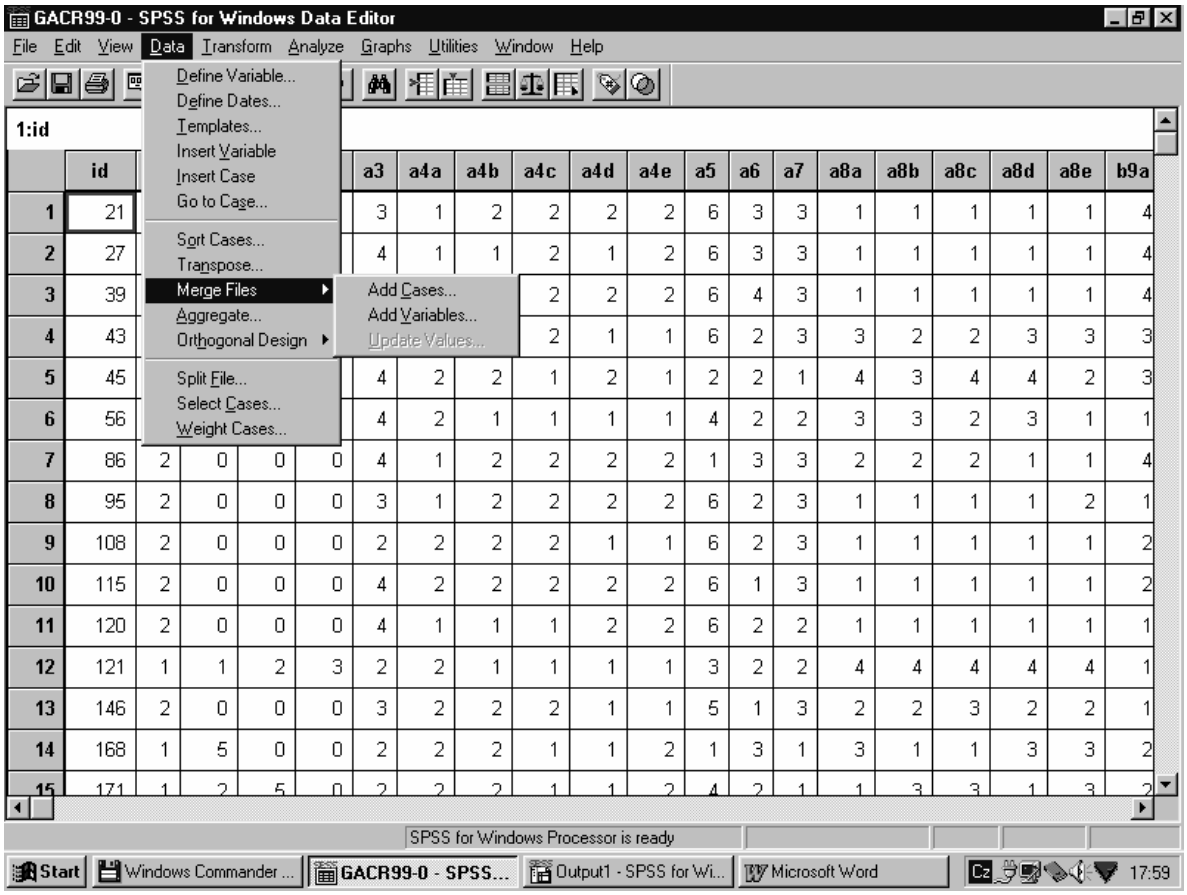
Máme personální databáze jednotlivých imatrikulačních ročníků studentů (každý ročník je samostatná matice dat) a chceme vytvořit jednotnou databázi studentů všech ročníků (jednu matici). Struktura matice je stejná: sledují se stejné proměnné (charakteristiky studentů) a v maticích jsou uvedeny ve stejném pořadí. K případům jednoho souboru se přidají případy druhého souboru.

	A1	A2	A3	A4	A5	A6	..... Ai	An
Adamec								
Blahá								
.....								
Zemina								

+

	A1	A2	A3	A4	A5	A6	..... Ai	An
Deml								
Stará								
.....								
Vechtr								

SLUČOVÁNÍ SOUBORŮ - ADD VARIABLES



	id					a3	a4a	a4b	a4c	a4d	a4e	a5	a6	a7	a8a	a8b	a8c	a8d	a8e	b9a
1	21					3	1	2	2	2	2	6	3	3	1	1	1	1	1	4
2	27					4	1	1	2	1	2	6	3	3	1	1	1	1	1	4
3	39								2	2	2	6	4	3	1	1	1	1	1	4
4	43								2	1	1	6	2	3	3	2	2	3	3	3
5	45					4	2	2	1	2	1	2	2	1	4	3	4	4	2	3
6	56					4	2	1	1	1	1	4	2	2	3	3	2	3	1	1
7	86	2	0	0	0	4	1	2	2	2	2	1	3	3	2	2	2	1	1	4
8	95	2	0	0	0	3	1	2	2	2	2	6	2	3	1	1	1	1	2	1
9	108	2	0	0	0	2	2	2	2	1	1	6	2	3	1	1	1	1	1	2
10	115	2	0	0	0	4	2	2	2	2	2	6	1	3	1	1	1	1	1	2
11	120	2	0	0	0	4	1	1	1	2	2	6	2	2	1	1	1	1	1	1
12	121	1	1	2	3	2	2	1	1	1	1	3	2	2	4	4	4	4	4	1
13	146	2	0	0	0	3	2	2	2	1	1	5	1	3	2	2	3	2	2	1
14	168	1	5	0	0	2	2	2	1	1	2	1	3	1	3	1	1	3	3	2
15	171	1	2	5	0	2	2	2	1	1	2	4	2	1	1	3	3	1	3	2

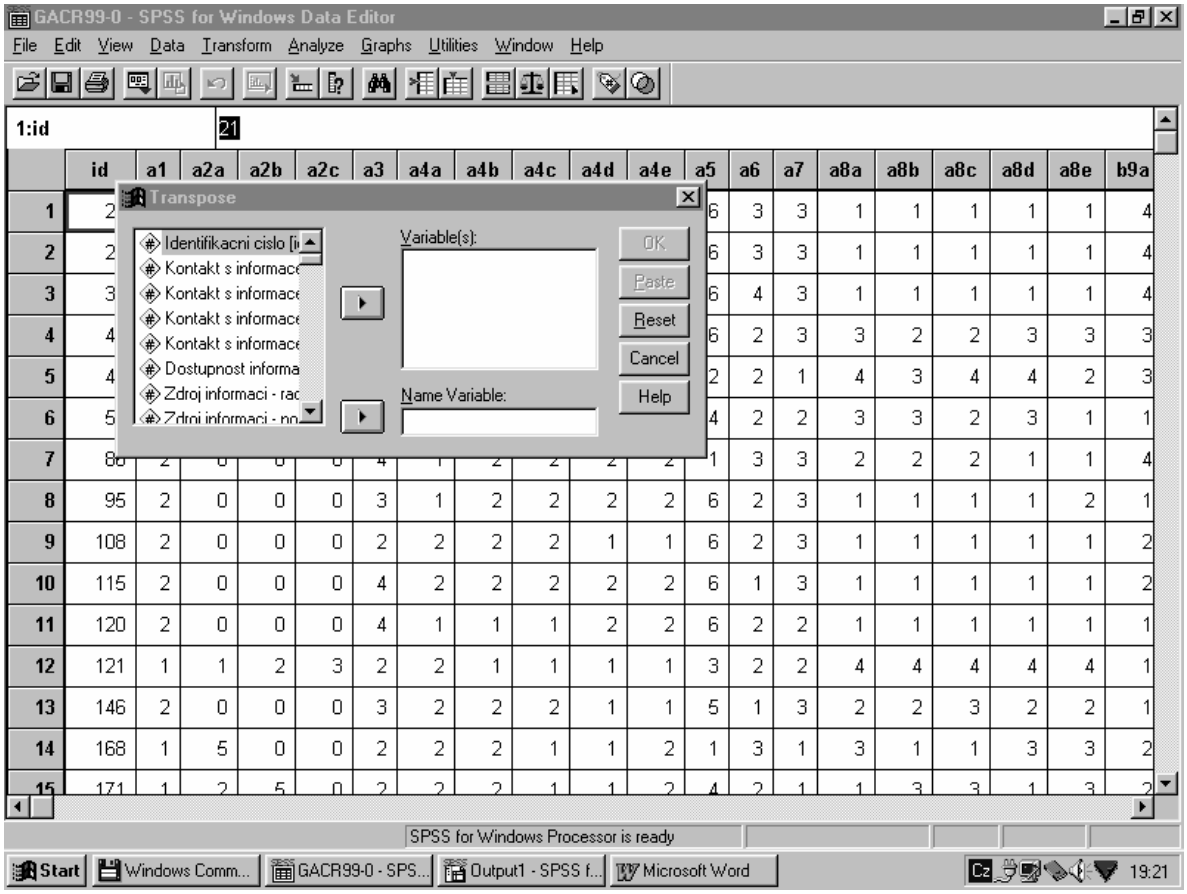
ÚLOHA
Máme v jedné databázi (matici) údaje o osobních charakteristikách studentů a v druhé databázi (matici) údaje o jejich prospěchu. Chceme je dostat do jedné matice všech údajů o studentech. Pořadí studentů musí být ve slučovaných maticích shodné, nebo musíme mít znak, který každého studenta jednoznačně definuje. K proměnných jednoho souboru se přidají proměnné dalšího souboru.

	O1	O2	.....	Oi	...	On
Adamec						
Blahá						
.....						
Zemina						

+

	P1	P2	....	Pi	...	Pn
Adamec						
Blahá						
.....						
Zemina						

TRANSPOSE



	id	a1	a2a	a2b	a2c	a3	a4a	a4b	a4c	a4d	a4e	a5	a6	a7	a8a	a8b	a8c	a8d	a8e	b9a
1	2											6	3	3	1	1	1	1	1	4
2	2											6	3	3	1	1	1	1	1	4
3	3											6	4	3	1	1	1	1	1	4
4	4											6	2	3	3	2	2	3	3	3
5	4											2	2	1	4	3	4	4	2	3
6	5											4	2	2	3	3	2	3	1	1
7	86	2	0	0	0	4	1	2	2	2	2	1	3	3	2	2	2	1	1	4
8	95	2	0	0	0	3	1	2	2	2	2	6	2	3	1	1	1	1	2	1
9	108	2	0	0	0	2	2	2	2	1	1	6	2	3	1	1	1	1	1	2
10	115	2	0	0	0	4	2	2	2	2	2	6	1	3	1	1	1	1	1	2
11	120	2	0	0	0	4	1	1	1	2	2	6	2	2	1	1	1	1	1	1
12	121	1	1	2	3	2	2	1	1	1	1	3	2	2	4	4	4	4	4	1
13	146	2	0	0	0	3	2	2	2	1	1	5	1	3	2	2	3	2	2	1
14	168	1	5	0	0	2	2	2	1	1	2	1	3	1	3	1	1	3	3	2
15	171	1	2	5	0	2	2	2	1	1	2	4	2	1	1	3	3	1	3	2

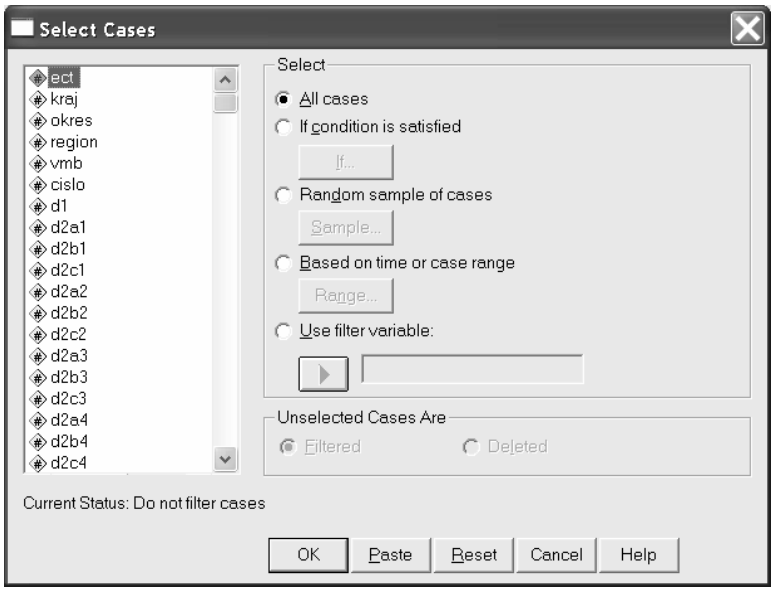
Toto je matice před provedením příkazu TRANSPOSE

Příkaz TRANSPOSE vytváří nový datový soubor ve kterém jsou:
- původní řádky (případy) sloupci (proměnnými)
- původní sloupce (proměnné) řádkami (případy)

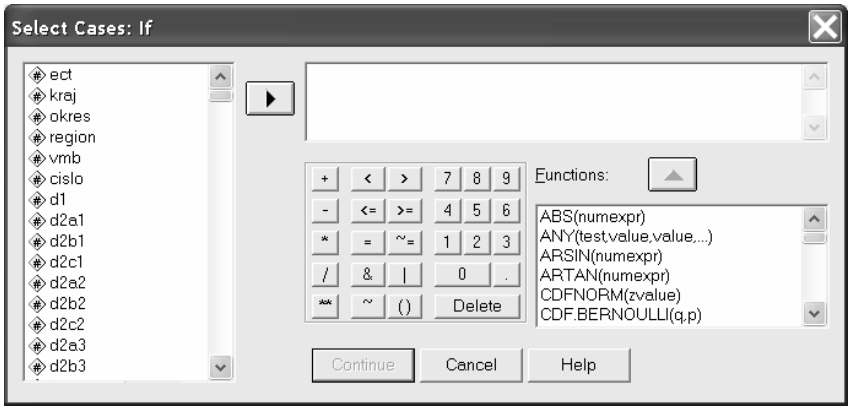
Automaticky se vytvářejí nová jména proměnných

VÝBĚR PŘÍPADŮ

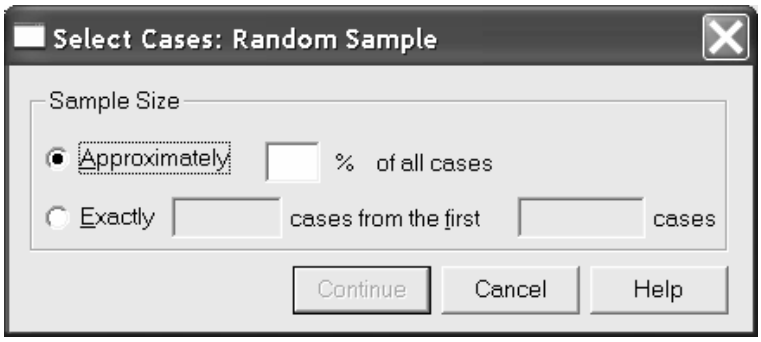
Nemusíme vždy pracovat s celým výběrovým souborem, ale pomoci procedury SELECT CASES si z můžeme vybrat jen určitým způsobem definovaný podsoubor.



If condition is satisfied:
Zajímají nás jen menší podsoubory (například jen ženy nebo jen muži, nebo jen osoby s vysokoškolským vzděláním, nebo jen osoby bydlící v Praze, nebo jen osoby deklarující se jako příslušníci střední třídy, nebo jen nezaměstnané osoby apod.) a proto si je vybíráme, abychom další analytické výpočty prováděli jen s těmi případy, které do nich patří. Je pochopitelné, že je můžeme vybírat jen podle známých – zjištěných – charakteristik: pokud jsme například v dotazníku nezjišťovali místo bydliště respondenta, nemůžeme obyvatele Prahy vybrat, pokud jsme nerozlišili v dotazníku mezi osobami se základním vzděláním vyučené a nevyučené, nemůžeme ani s jedním takto vymezeným souborem pracovat a musíme se spokojit s podsouborem osob se základním vzděláním.
Podsoubory s nimiž chceme pracovat určujeme pomocí podmínky: do okénka vyklikáme nebo vypíšeme podmínku, např. SEX =1 (chceme-li pracovat jen s muži a víme, že v proměnné SEX 1=muž), OBEC=15 (chceme-li pracovat jen s obyvateli Prahy a víme, že v proměnné OBEC Praha=15), VZDEL > 2 (chceme-li pracovat s osobami, jež mají středoškolské a vysokoškolské vzdělání a víme, že v proměnné VZDEL osoba se středoškolským vzděláním=3 a osoba s vysokoškolským vzděláním = 4.



Random sample of cases:
Dovoluje nám vytvořit z našeho souboru náhodný výběr (omezit počet jeho jednotek při zachování reprezentativity souboru – samozřejmě, pokud byl reprezentativní původní soubor).



Můžete nechat vybrat přibližný podíl z původního souboru, který stanovíme, nebo určitý počet případů (do *from the first cases* vypíšeme celkový počet jednotek původního souboru nebo někdy – spíše výjimečně – výběr omezíme jen na určitý počet případů).

Co se týče **rozhodnutí co s nevybranými případy**, používejte raději variantu:

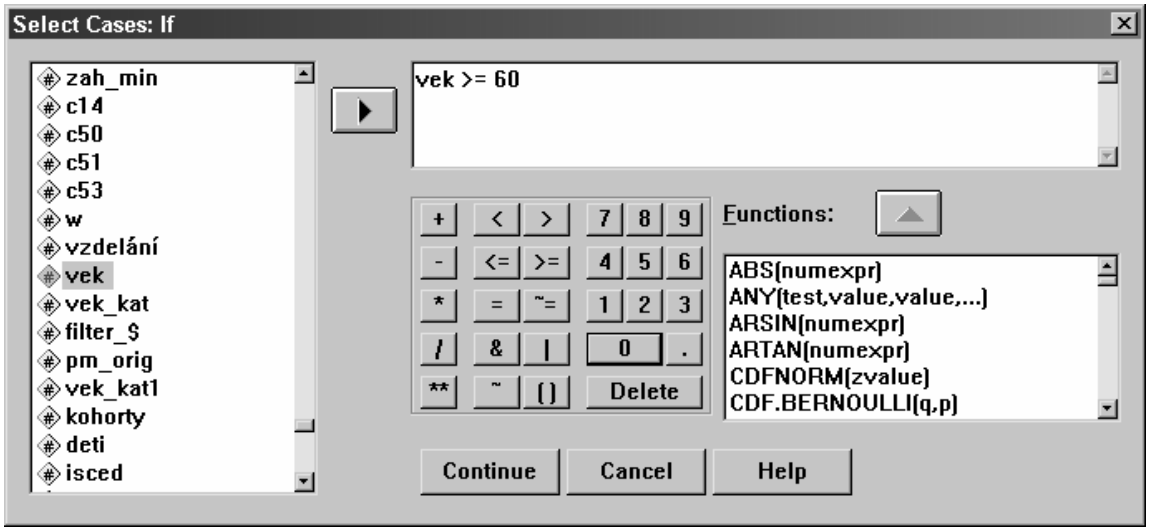
Unselect cases are filtered. Filtr lze odstranit a dále pracovat s celým souborem, pokud použijete variantu *Unselect cases are deleted*, musíte být velmi opatrní: nesmíte si takto upravený soubor uložit pod stejným jménem – přepsal by původní soubor a zůstal by Vám jen soubor s vybranými jednotkami (a právem také jen oči pro pláč, pokud byste neměli poslední podobu souboru zálohovanou).

Manipulace s datovým souborem

K transformačním procedurám lze také přiřadit manipulaci s datovým souborem – je možné pracovat pouze s podsouborem případů. Např. nás může zajímat analýza lidí ve věku 60 let a starších. K vývěru takového podsouboru použijeme proceduru

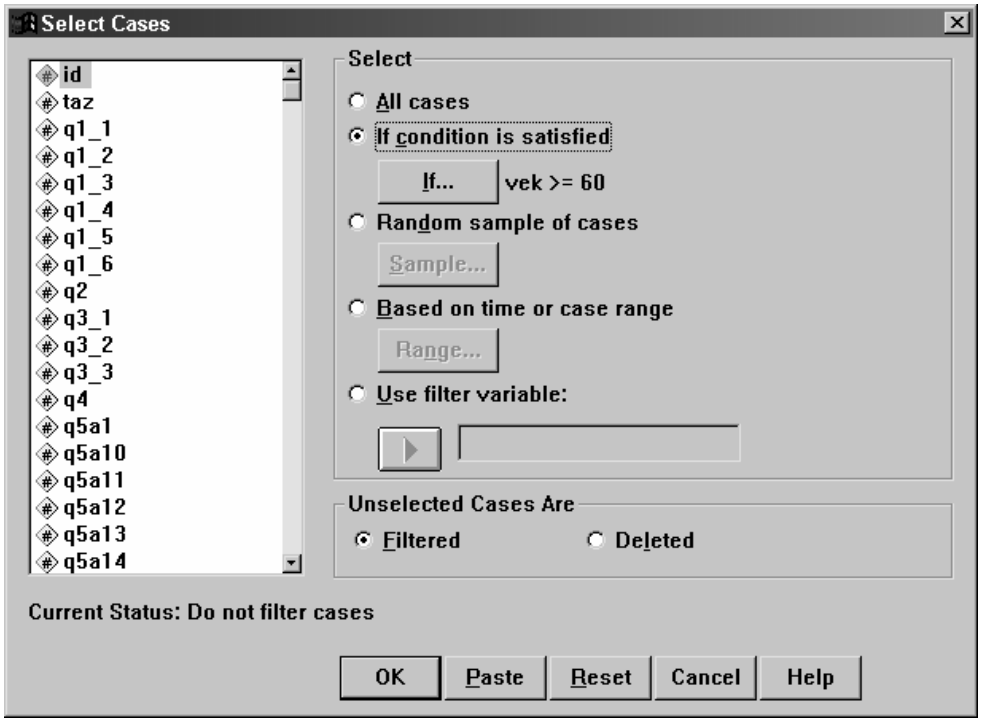
Data – Select Cases – If condition is satisfied

Po kliknutí na tlačítko **If…** se objeví dialogové okno, do nějž vepíšeme příslušnou podmínku pro výběr (viz).



Po kliknutí na tlačítko **Continue** si dejte pozor, aby v dialogovém okně, které se objeví, bylo nastaveno, že případy, které nesplňují podmínku (to jsou tedy nevybrané případy neboli **Unselected Cases**) jsou **Filtered** – filtorvány a nikoliv **Deleted** – vymazány (viz obr. níže). Jak napovídá název, filtrované případy zůstávají dále v souboru, pouze se s nimi nepracuje, vymazané případy jsou smazány a zůstávají pouze případy splňující podmínku.

Když si pro kontrolu necháme udělat rozložení takto redukovaného souboru, získáme výsledek, který je uveden dole v tabulce (viz). Podmínky lze samozřejmě různě kombinovat, např. bylo by možné získat podsoubor mužů ve věku 60+ let, kteří ještě pracují apod. Někdy mají tyto operace analytický smysl.



VEK

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	60	23	5,0	5,0	5,0
	61	16	3,5	3,5	8,5
	62	33	7,4	7,4	15,8
	63	21	4,6	4,6	20,4
	64	24	5,4	5,4	25,8
	65	26	5,7	5,7	31,5
	66	44	9,7	9,7	41,2
	67	32	7,0	7,0	48,2
	68	27	6,0	6,0	54,2
	69	26	5,7	5,7	60,0
	70	19	4,3	4,3	64,3
	71	22	5,0	5,0	69,2
	72	21	4,6	4,6	73,8
	73	20	4,4	4,4	78,3
	74	26	5,8	5,8	84,0
	75	27	5,9	5,9	90,0
	76	11	2,5	2,5	92,5
	77	7	1,5	1,5	93,9
	78	8	1,9	1,9	95,8
	79	5	1,0	1,0	96,8
	80	4	,8	,8	97,7
	81	2	,5	,5	98,2
	82	1	,3	,3	98,5
	84	4	,9	,9	99,3
	85	2	,4	,4	99,7
	87	1	,2	,2	99,9
	88	0	,1	,1	100,0
	Total	**453**	**100,0**	**100,0**	