NADSTAVBOVÝ MODUL MONHSA V1

Nadstavbový modul pro nehierarchické shlukování se jmenuje Clustering_NH (MONHSA V1) je součástí souboru Shluk_Nehier.xls. Tento soubor je přístupný na http://jonasova.upce.cz, a je určen pro nekomerční účely, výuku a jako podpora pro zpracování studentských závěrečných prací.

1.1 Struktura Nadstavbového modulu

Nadstavbový modul se nachází v souboru typu sešit aplikace Excel s podporou maker *Shlukova_analyza_nehierarchicke_metody.xlsm*. Všechny hlavní procedury, které provádí jednotlivé kroky procesu shlukování, jsou uloženy v modulu „Clustering_NH“. Všechny podpůrné procedury a funkce jsou umístěny v modulu „Routines“. K interakci s uživatelem slouží ovládací prvky typu ActiveX, umístěné na příslušných listech sešitu. Všechny rutiny, prováděné jako reakce na události způsobené interakcí uživatele s ovládacími prvky, se nachází v modulech kódu příslušných objektů třídy *Sheet*. Názvy všech procedur, proměnných a všech autorem definovaných názvů, které se nachází přímo v kódu, jsou v angličtině. Dále je pomocí příkazu *Option explicit* zapnuta povinná deklarace všech proměnných, které jsou v programu použity. Jednotlivé funkční celky kódu, nebo místa s netypickým řešením nějakého problému jsou opatřeny komentáři.

K ochraně zdrojového kódu je použito uzamčení celého projektu pomocí hesla. Aplikace obsahuje velké množství podpůrných procedur vykonávajících specifické činnosti během jejího chodu, které by se v tomto listu za běžných podmínek objevily, ale není žádoucí, aby k nim měl uživatel přístup. Proto je do každé procedury, která nevyžaduje žádný argument, přidán falešný volitelný argument *Optional FakeArgument As String*. Díky tomu se již takové procedury v listu maker nenabízí k použití.

1.2 Procedury

Pro každou hlavní proceduru, provádějících jednotlivé kroky procesu shlukování, obsahuje tato práce popis jejího účelu, vývojový diagram a popis funkčnosti.

1.2.1 Globální hodnoty

K některým hodnotám potřebuje program přistupovat z více procedur a dokonce i z různých modulů. Navíc je nutné, aby byl jejich obsah uchován i po ukončení aplikace a jejím opětovném spuštění. K tomu je využit skrytý list č. 7 s názvem „Control“. List je skrytý prostřednictvím kódu VBA, nastavením vlastnosti *Visible* objektu *Sheet*

na *xlSheetVeryHidden*. Tím je zamezeno možnosti odkrytí listu uživatelem prostřednictvím panelu nástrojů. Na tomto listu jsou vytvořené pojmenované buňky, které obsahují řídící proměnné, a pomocí těchto jmen je k nim z kódu přistupováno.

Při programování jednotlivých modulů jsou použity hodnoty, které se mnohokrát opakují. Aby nebylo nutné je při jejich změně přepisovat na mnoha místech, jsou uloženy do tzv. globálních konstant, na které se jednotlivé procedury odkazují. Na jednotlivé listy je naopak odkazováno přímo prostřednictvím jejich *CodeName* aby nebyl program narušen ani při případném přejmenování některých listů uživatelem. Dále jsou vytvořeny konstanty definující barvy písma a pozadí buněk, aby byl udržen jednotný vzhled a zajištěna možnost jej později snadno z jednoho místa upravit.

1.2.2 Pomocné rutiny

Protože jsou některé dílčí postupy používány opakovaně, je vhodné vytvořit pro ně samostatné procedury. Je tak zamezeno několikanásobnému opakování stejné sekvence příkazů na různých místech a významným způsobem se usnadňují případné následné modifikace. Jedná se například o smazání listu, deaktivaci všech ovládacích prvků, obnovení ovládacích prvků, skrytí a zobrazení listů, vložení hlášení atd.

Dalšími jsou například:

Sub RandomCenters(centersCount As Single) – náhodný výběr typických bodů

Kromě manuálního výběru typických bodů má uživatel možnost využít náhodný výběr. Tato procedura náhodně volí jednotlivé objekty v prvku *ListBox* a označuje je. Pokud je náhodně vybrán již označený objekt nebo objekt duplicitní k již vybranému, dochází k dalšímu výběru, dokud není dosaženo počtu předaného argumentem.

Sub RecountCenterPoints() – přepočet vybraných typických bodů

Při výběru typických bodů uživatelem prostřednictvím prvku *ListBox* je zobrazován jejich aktuální počet. Procedura RecountCenterPoints() je tedy mapována na událost *Change* daného objektu a vždy při nějaké změně aktualizuje počet vybraných bodů.

Sub RestoreSelections () – obnovení výběru v prvku *ListBox*

Pomocí této procedury dochází k obnově označených objektů ze zálohy jejich indexů uložené na listu „Control“ v hodnotě *listBoxIndexes*. Je volána při znovuotevření aplikace s uloženou prací.

Function CheckListbox()As Boolean – kontrola výběru v prvku *ListBox*

Před zahájením shlukování je nutné ověřit, zda je vybrán alespoň jeden objekt v prvku *ListBox* jako typický bod. Funce vrací hodnotu *True*, pokud je vybrán alespoň jeden a *False* pokud není vybrán žádný.

Function CheckListboxForDuplicities() As Boolean – kontrola duplicit ve vybraných objektech

Aplikace umožňuje při úvodní kontrole dat provádět i detekci duplicitních objektů. Duplicitní objekty mohou za určitých okolností a při nevhodně zvolených typických bodech narušit průběh shlukování. Uživatel má proto možnost takové objekty odhalit a případně odstranit. Pokud ale tuto možnost nevyužije, musí aplikace zajistit, že takové objekty nebudou jako typické body vybrány. K tomu slouží tato funkce. Postupně prochází všechny objekty vybrané v ovládacím prvku *ListBox,* a pokud se mezi nimi nachází duplicity, vrátí hodnotu *False*. V opačném případě vrátí hodnotu *True*.

Sub RemoveDuplicateObjects() – odstranění duplicitních objektů

Tato procedura provede odebrání všech řádků s duplicitními objekty, které byly zjištěny a označeny během kontroly vstupních dat. Je mapována na tlačítko „Odstranit duplicity“.

Sub FillComboboxes() – naplnění prvků *ComboBox* názvy atributů

Tato procedura naplní dva prvky typu *ComboBox* na listu „Grafy“ názvy jednotlivých atributů. Tyto prvky slouží k výběru atributů, které budou porovnávány ve výsledném grafu.

Function Catenate(myRange As range, Optional delimiter As String) As String – sloučení všech řetězců v buňkách v daném rozsahu

Tato funce postupně prochází všechny buňky ve zvoleném rozsahu a spojí je do jednoho řetězce. Volitelným argumentem lze určit znak, který bude použit jako oddělovač. Funkce vrací řetězec spojených hodnot.

Function Warning(level As Integer) As Boolean – zobrazení varování o ztrátě dat

Funkce Warning je spouštěna jednotlivými rutinami po stisknutí některých tlačítek. Kontroluje úroveň řídící hodnoty *warning* a pokud je vyžadováno, zobrazí varování o možné ztrátě dat při provedení požadované akce. Pokud uživatel akci potvrdí, vrací *True*, pokud ne vrací *False*.

Soubor funcí pro výpočet SHA1

K odhalení duplicitních řádků je využita funkce k výpočtu SHA1 šifry. Soubor funkcí, který tento proces zajišťuje, tvoří jedinou část kódu, který nebyl vytvořen autorem práce, ale převzat z veřejně dostupných zdrojů[1].

1.3 Proces shlukování

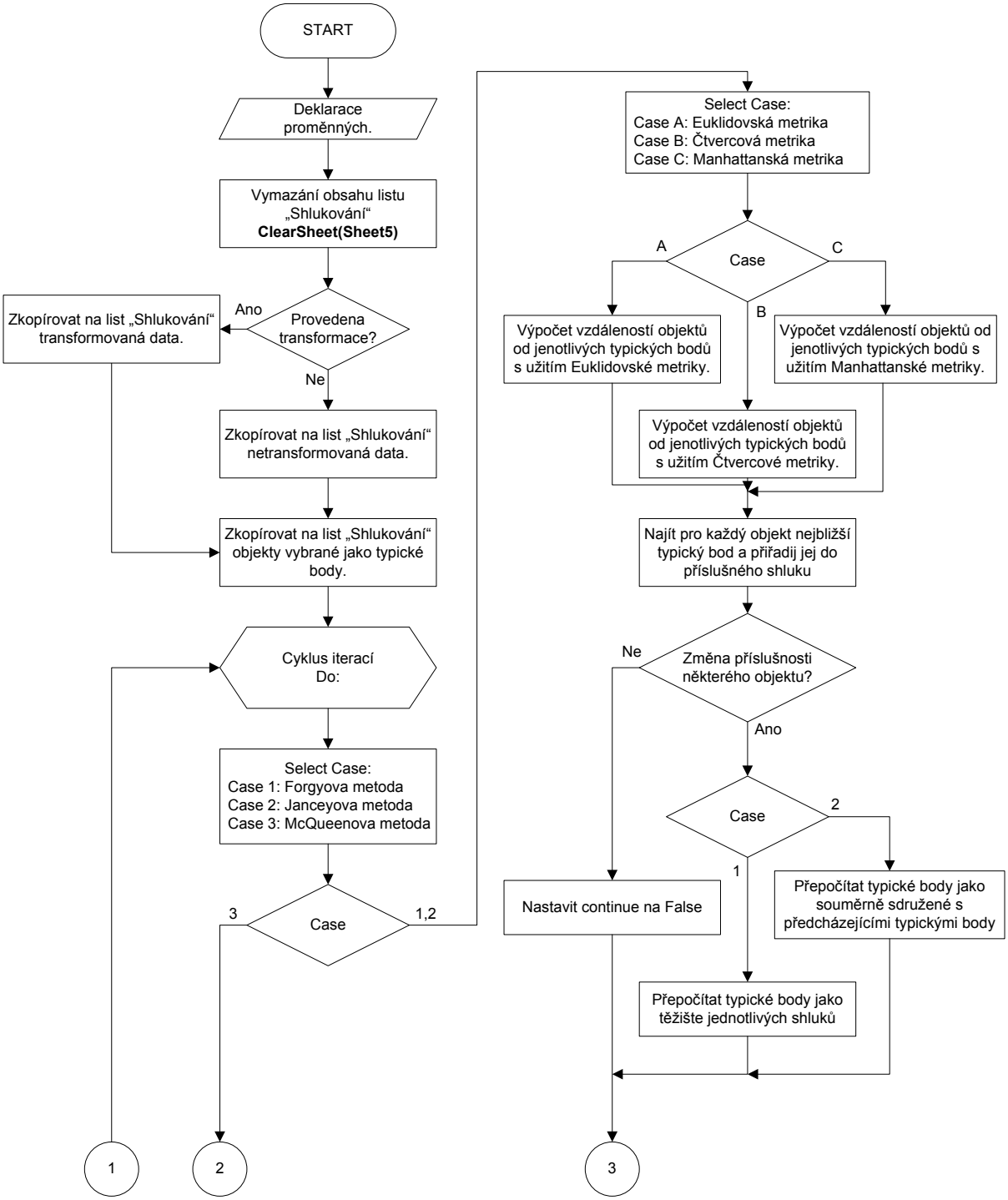
Samotný proces shlukování má několik voleb, které musí být před jejím spuštěním nastveny. Nejprve je uživatel vyzván zvolit prostřednictvím prvků *OptionButton* metodu shlukování a metriku která má být použita pro výpočet podobnosti objektů. Poté musí být vybrány objekty, které představují typické body pro shlukování a stanoví počet shluků. To může být provedeno ručně, výběrem objektů v prvku *ListBox* nebo náhodným výběrem. Při něm je uživatel po stisknutí tačítka „náhodně“ vyzván k zadání počtu shluků a poté jsou prvky vybrány automaticky. Po stisku tlačítka „Shlukovat“ dojde nejprve ke kontrole, zda byly vybrány typické body. Pokud není vybrán alespoň jeden, zobrazí se výzva uživateli, aby tak učinil. Pokud jsou body vybrány, proběhne kontrola, zda nejsou některé ze zvolených objektů duplicitní. Pokud ano, uživatel je opět vyzván k nápravě.

Po splnění těchto podmínek lze tlačítkem „Shlukovat“ spustit vlastní shlukování. Průběh každé iterace je detailně vypsán v listu „Shlukování“ a na závěr jsou do tabulky vypsány příslušnosti jednotlivých objektů do shluků.

Procedura: ***Clustering(method As String, metrics As String)***

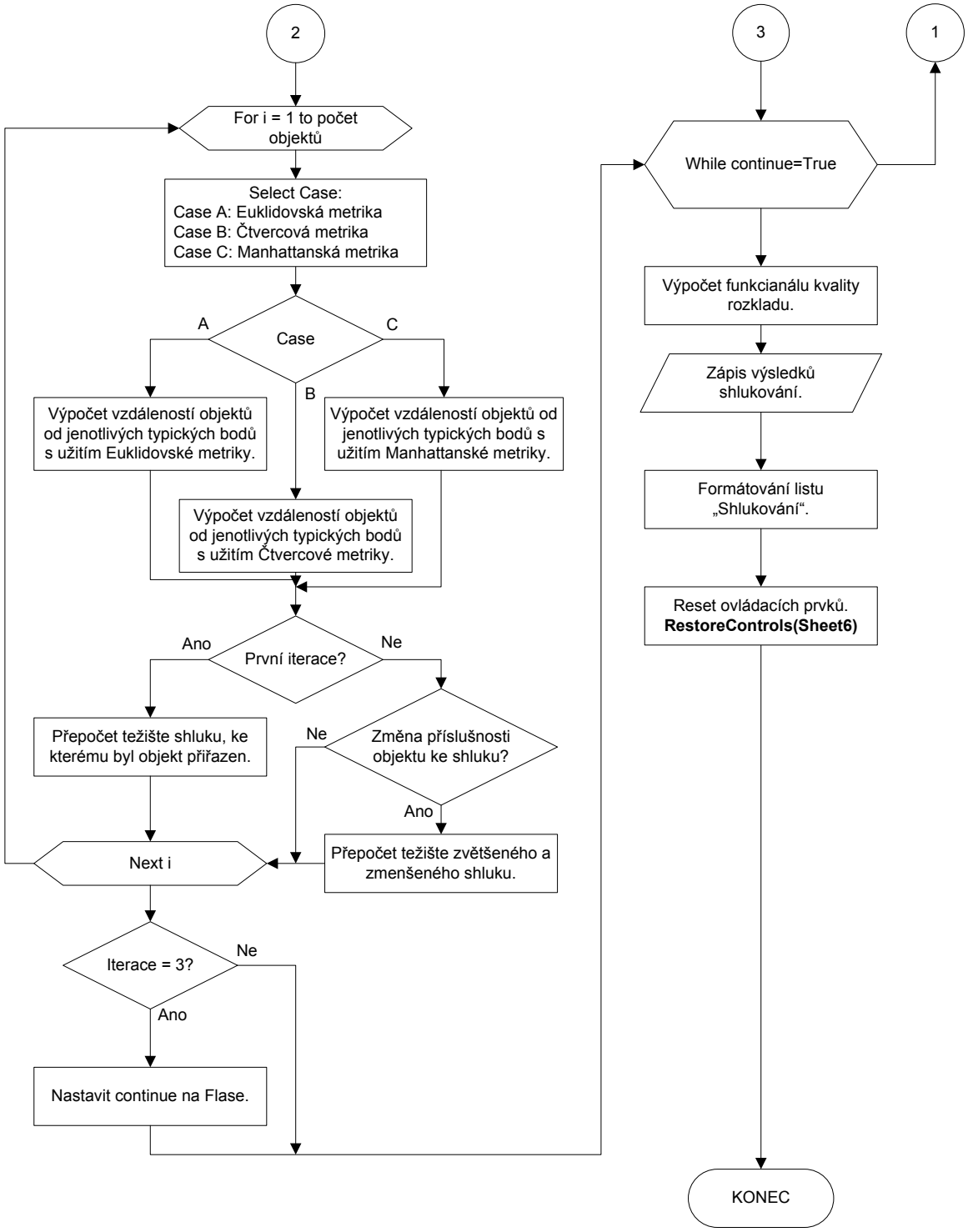
Vývojový diagram a popis procedury *Clustering(method As String, metrics As String)* je na obrázku 1 a 2.

[1] HULBERT, Chris. *Calculating an SHA1 hash in Excel* [online]. 2009 [cit. 2012-05-10]. Dostupný z WWW: <http://splinter.com.au/calculating-an-sha1-hash-in-excel>



Obrázek 1: Vývojový diagram shlukování - část 1.

zdroj: autor



Obrázek 2: Vývojový diagram shlukování - část 2.

zdroj: autor

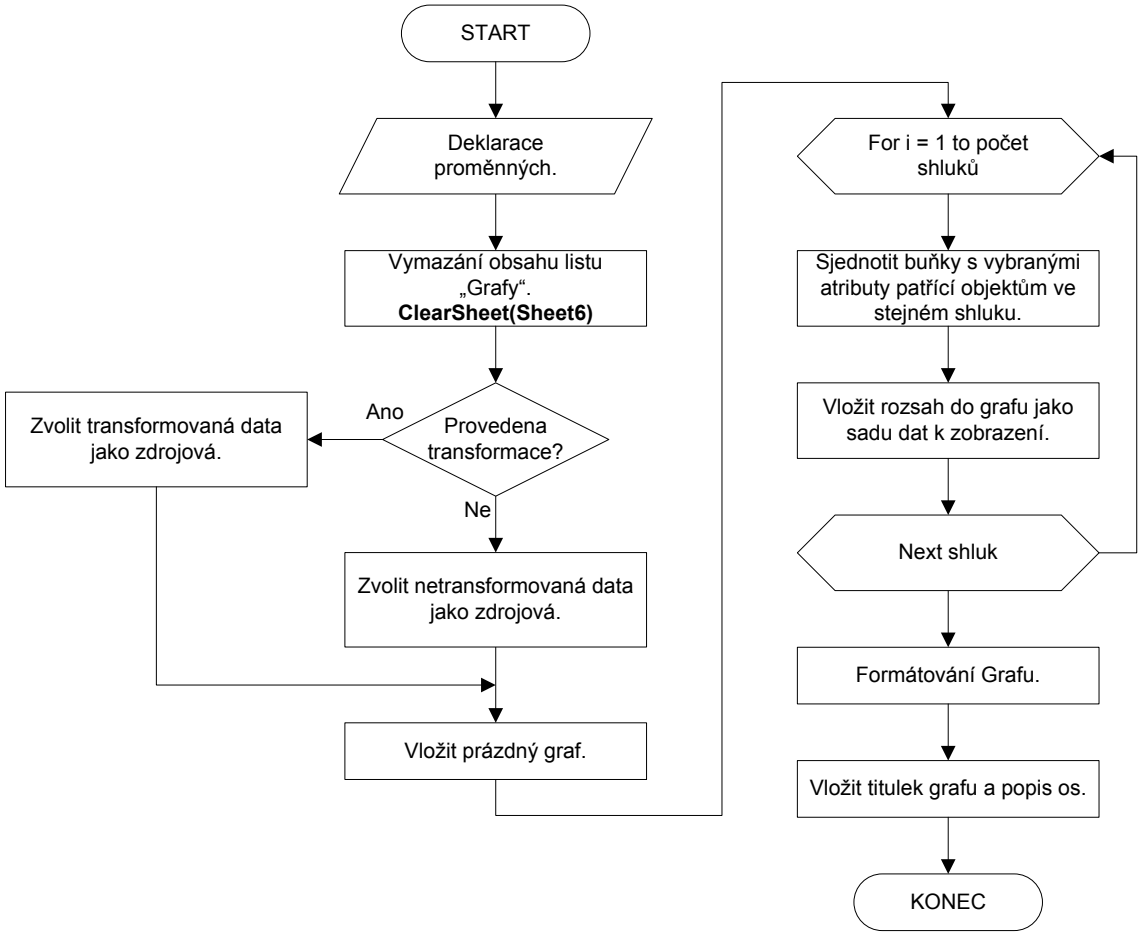
1.4 Grafický výstup

Grafickým výstupem shlukování jsou dvojrozměrné grafy zobrazující všechny objekty shlukování a porovnávající vždy dva atributy. Uživatel má možnost pomocí dvojice prvků ComboBox zvolit atributy pro osu X a Y, které chce porovnávat a poté tačítkem „Graf“

vygenerovat grafický výstup. Pokud bude zadána neexistující hodnota, tlačítko se deaktivuje, dokud nebude hodnota opravena.

Procedura: ***GenerateGraphs()***

Vývojový diagram a popis procedury *GenerateGraphs()* je na obrázku 3.



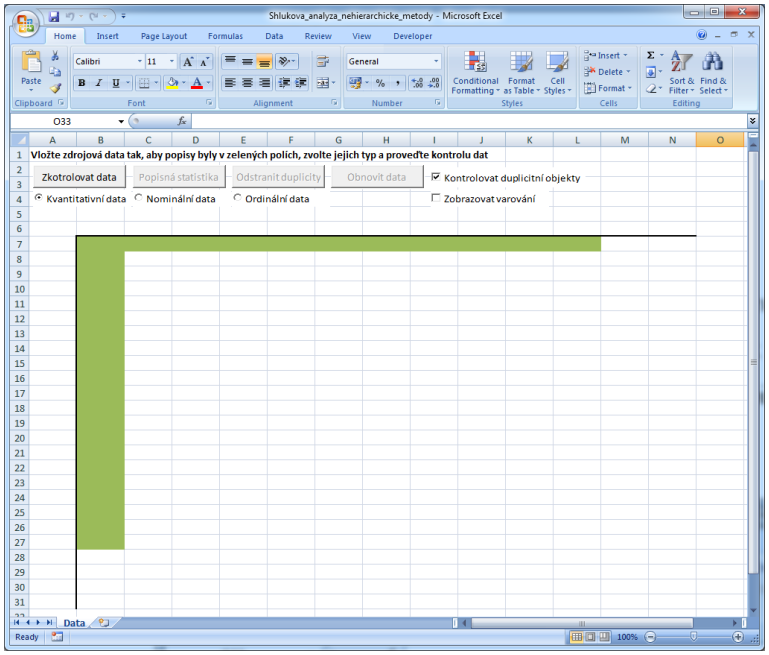
Obrázek 3: Vývojový diagram generování grafu

zdroj: autor

2 METODICKÉ POKYNY PRO UŽIVATELE

Tato část obsahuje metodické pokyny pro uživatele popisující použití nadstavbového modulu pro shlukovou analýzu nehierarchickými metodami. Počet dostupných metod a metrik je přizpůsoben rozsahu výuky předmětu Zpracování dat metodami shlukové analýzy, pro který byl primárně modul určen.

Po prvním otevření souboru v aplikaci Excel se zobrazí jeden list s názvem „Data“ viz Obrázek 4. Další obrázky již nebudou obsahovat celé okno aplikce Excelu ale pouze výseče s popisovanými oblastmi.



Obrázek 4: Úvodní obrazovka aplikace

zdroj: autor

V závislosti na nastavené úrovni zabezpečení Excelu může být spouštění maker zakázáno. V takovém případě je nutné v menu „Macro security“ snížit úroveň zabezpečení. Při typickém nastavení „Disable all macros with notification“ se po spuštění souboru zobrazí výzva k povolení maker. Pokud nebudou makra povolena, nadstavbový modu pro shlukování nebude možné použít.

2.1 Zdrojová data Ukázkového příkladu

Jako zdrojová data jsou použity statistické údaje ze serveru www.policie.cz popisující počty zjištěných trestných činů v jednotlivých krajích za rok 2011 a jejich celkovou objasněnost, uvedné v Tabulka 1: Zdrojová data pro vzorový příklad Vzhledem k rozdílnému

počtu obyvatel v jednotlivých krajích jsou v zájmu jejich porovnatelnosti přepočteny absolutní součty všech trestných činů a udávány jako počet trestných činů na 100 000 obyvatel.

Tabulka 1: Zdrojová data pro vzorový příklad

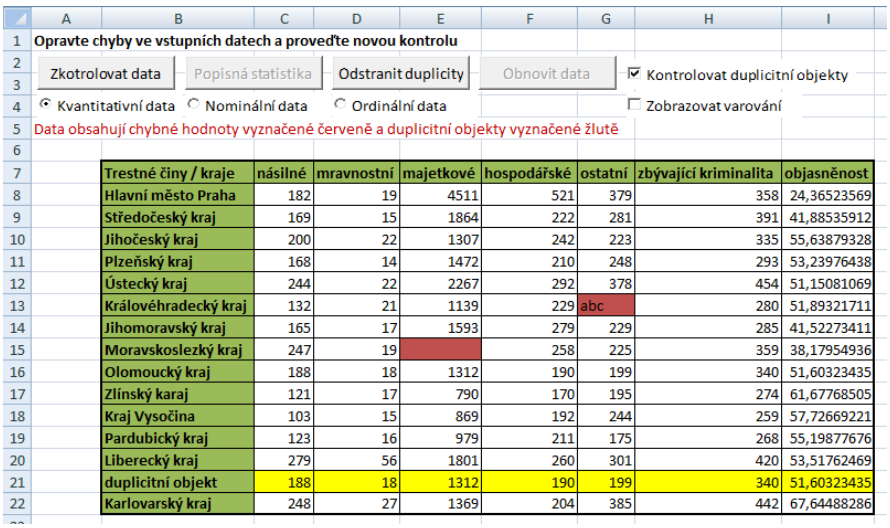
Trestné činy / kraje	násilné	mravnostní	majetkové	hospodářské	ostatní	zbývající kriminalita	objasněnost
Hlavní město Praha	182	19	4 511	521	379	358	24
Středočeský kraj	169	15	1 864	222	281	391	42
Jihočeský kraj	200	22	1 307	242	223	335	56
Plzeňský kraj	168	14	1 472	210	248	293	53
Ústecký kraj	244	22	2 267	292	378	454	51
Královéhradecký kraj	132	21	1 139	229	186	280	52
Jihomoravský kraj	165	17	1 593	279	229	285	42
Moravskoslezký kraj	247	19	2 343	258	225	359	38
Olomoucký kraj	188	18	1 312	190	199	340	52
Zlínský karaj	121	17	790	170	195	274	62
Kraj Vysočina	103	15	869	192	244	259	58
Pardubický kraj	123	16	979	211	175	268	55
Liberecký kraj	279	56	1 801	260	301	420	54
Karlovarský kraj	248	27	1 369	204	385	442	68

Zdroj: Upraveno podle www.policie.cz

Jednotlivé kraje tvoří objekty ke shlukování a počty trestných činů a celková objasněnost jsou jejich atributy.

2.2 Analýza dat

Zdrojová data se vkládají do úvodního listu s názem „Data“ do vyznačené výseče (počínaje buňkou B7). Data je před zpracováním nutné nejdříve zkontrolovat. Kontrola se spouští stiskem tlačítka „Zkontrolovat data“. Předtím je možné zaškrtnout volbu „Kontrolovat duplicitní objekty“ která umožňuje odhalovat řádky se shodnými atributy. Jak je vidět na Obrázek 5, pokud jsou při kontrole odhaleny nesprávné nebo chybějící hodnoty, příslušné buňky se zvýrazní červeně, případné odhalené duplicitní objekty se zvýrazní žlutě.



Opravte chyby ve vstupních datech a proveďte novou kontrolu

Zkotrolovat data | Popisná statistika | Odstranit duplicity | Obnovit data | Kontrolovat duplicitní objekty

Kvantitativní data | Nominální data | Ordinální data | Zobrazovat varování

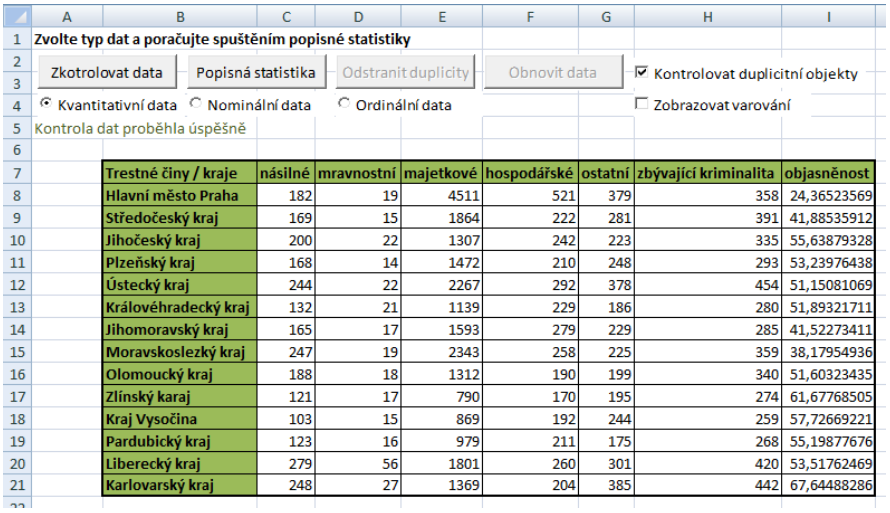
Data obsahují chybné hodnoty vyznačené červeně a duplicitní objekty vyznačené žlutě

Trestné činy / kraje	násilné	mravnostní	majetkové	hospodářské	ostatní	zbývající kriminalita	objasněnost
Hlavní město Praha	182	19	4511	521	379	358	24,36523569
Středočeský kraj	169	15	1864	222	281	391	41,88535912
Jihočeský kraj	200	22	1307	242	223	335	55,63879328
Plzeňský kraj	168	14	1472	210	248	293	53,23976438
Ústecký kraj	244	22	2267	292	378	454	51,15081069
Královéhradecký kraj	132	21	1139	229	abc	280	51,89321711
Jihomoravský kraj	165	17	1593	279	229	285	41,52273411
Moravskoslezký kraj	247	19		258	225	359	38,17954936
Olomoucký kraj	188	18	1312	190	199	340	51,60323435
Zlínský karaj	121	17	790	170	195	274	61,67768505
Kraj Vysočina	103	15	869	192	244	259	57,72669221
Pardubický kraj	123	16	979	211	175	268	55,19877676
Liberecký kraj	279	56	1801	260	301	420	53,51762469
duplicitní objekt	188	18	1312	190	199	340	51,60323435
Karlovarský kraj	248	27	1369	204	385	442	67,64488286

Obrázek 5: Vzorový příklad – neúspěšná kontrola dat

zdroj: autor

Pokud jsou odhaleny duplicitní řádky, zpřístupní se tlačítko „Odstranit duplicity“ pomocí kterého lze všechny takové řádky vymazat. Ostatní buňky s chybnými hodnotami opraví uživatel. Jakmile jsou chyby odstraněny, musí proběhnout nová kontrola dat. Takto se postupuje do chvíle, než data projdou kontrolou úspěšně viz Obrázek 6.



Zvolte typ dat a poračujte spuštěním popisné statistiky

Zkotrolovat data | Popisná statistika | Odstranit duplicity | Obnovit data | Kontrolovat duplicitní objekty

Kvantitativní data | Nominální data | Ordinální data | Zobrazovat varování

Kontrola dat proběhla úspěšně

Trestné činy / kraje	násilné	mravnostní	majetkové	hospodářské	ostatní	zbývající kriminalita	objasněnost
Hlavní město Praha	182	19	4511	521	379	358	24,36523569
Středočeský kraj	169	15	1864	222	281	391	41,88535912
Jihočeský kraj	200	22	1307	242	223	335	55,63879328
Plzeňský kraj	168	14	1472	210	248	293	53,23976438
Ústecký kraj	244	22	2267	292	378	454	51,15081069
Královéhradecký kraj	132	21	1139	229	186	280	51,89321711
Jihomoravský kraj	165	17	1593	279	229	285	41,52273411
Moravskoslezký kraj	247	19	2343	258	225	359	38,17954936
Olomoucký kraj	188	18	1312	190	199	340	51,60323435
Zlínský karaj	121	17	790	170	195	274	61,67768505
Kraj Vysočina	103	15	869	192	244	259	57,72669221
Pardubický kraj	123	16	979	211	175	268	55,19877676
Liberecký kraj	279	56	1801	260	301	420	53,51762469
Karlovarský kraj	248	27	1369	204	385	442	67,64488286

Obrázek 6: Vzorový příklad – úspěšná kontrola dat

zdroj: autor

Jakákoliv změna ve vstupních datech provedená po jejich úspěšné kontrole znamená zásah do již započatého procesu shlukování. Z toho důvodu je oblast dat monitorována a jejich případná změna vyvolá deaktivaci všech ovládacích prvků a naopak zpřístupnění tlačítka „Obnovit data“. Uživatel má tak dvě možnosti. Buď může pomocí tlačítka data obnovit, čímž se opět aktivují ovládací prvky a je mu tak umožněno pokračovat v práci nebo provést novou kontrolu dat, která způsobí odstranění veškerého dosavadního průběhu shlukování a zahájení

práce s novými daty. Při každé činnosti, která je z nějakého důvodu opakována a způsobí tak přepsání již existujících hodnot se zobrazuje varovné hlášení, upozorňující uživatele na možnou ztrátu dat. Pokud si uživatel nepřeje tato hlášení zobrazovat, lze je zakázat odznačením možnosti „Zobrazovat varování“.

2.3 Popisná statistika

Po úspěšné kontrole dat dojde k aktivaci talčítka „Popisná statistika“. Uživatel je vyzván ke zvolení typu dat, na němž závisí typ statistických ukazatelů, které budou generovány. Po spuštění se aktivuje druhý list s názvem „Popisná statistika“ který obsahuje výpis příslušných hodnot viz Obrázek 7.

	A	B	C	D	E	F	G	H	I
1	**Pokračujte transformací dat, nebo korelací s použitím původních hodnot.**								
2	Transformace	Korelace							
3									
4									
5	Popisná statistika pro kvantitativní data:								
6									
7			násilné	mravnostní	majetkové	hospodářské	ostatní	zbývající kriminalita	objasněnost
8		Počet hodnot	14	14	14	14	14	14	14
9		Minimum	103	14	790	170	175	259	24,36523569
10		Maximum	279	56	4511	521	385	454	67,64488286
11		Součet	2569	297	23616	3480	3648	4758	705,2443597
12		Stř. Hodnota	183,5	21,21428571	1686,857143	248,5714286	260,5714286	339,8571429	50,37459712
13		Chyba stř. Hodnoty	14,59329592	2,819045453	251,397716	22,97531267	19,67167716	17,85868125	2,898649857
14		Useknutý průměr	183,5	21,21428571	1686,857143	248,5714286	260,5714286	339,8571429	50,37459712
15		Median	175,5	18,5	1420,5	225,5	236,5	337,5	52,56649074
16		Modus	#N/A	19	#N/A	#N/A	#N/A	#N/A	#N/A
17		Dolní kvartil	140,25	16,25	1181	205,5	205	281,25	44,20172201
18		Horní kvartil	233	21,75	1848,25	259,5	296	383	55,52878915
19		Směr. Odchylka	54,60311346	10,54790224	940,6441211	85,96574835	73,60467615	66,82106663	10,84575465
20		Variační rozpětí	176	42	3721	351	210	195	43,27964717
21		Variační koeficient	0,297564651	0,497207513	0,557631169	0,345839218	0,282474086	0,19661516	0,215302062
22		Rozptyl výběru	2981,5	111,2582418	884811,3626	7390,10989	5417,648352	4465,054945	117,6303939
23		Špičatost	-0,975753064	10,71138578	6,38231396	8,666198265	-0,707555046	-1,128622356	1,472497726
24		Šikmost	0,267055773	3,135452648	2,272089915	2,717719938	0,80764346	0,465160359	-0,942807351

Obrázek 7: Vzorový příklad – popisná statistika

zdroj: autor

2.4 Transformace dat

V této chvíli se uživatel musí rozhodnout, zda bude potřebné na vstupních datech provést transformaci. Pokud jsou jednotlivé atributy v různých měrných jednotkách a hrozilo by, že by byl některý z nich dominantní a mohl tak výrazně ovlivnit shlukování, je vhodné použít standardizaci, která přiřadí jednotlivým znakům příslušnou váhu a zajistí tak jejich souměřitelnost. Pokud jsou data souměřitelná, ale normy jednotlivých vektorů, kterými jsou objekty tvořeny jsou výrazně rozdílné, je vhodné data normalizovat a převést tak všechny vektory na stejnou normu.

K přechodu na list umožňující transformaci dat slouží tlačítko „Transformace“. Zde je nutné zvolit, zda provést standardizaci, normalizaci nebo obojí a provést ji stiskem tlačítka

„Transformovat“. Pokud se uživatel rozhodne ji neprovádět, pokračuje stiskem tlačítka „Korelace“.

Ve vzorovém případě jsou všechny atributy dány počtem zjištěných činů na 100 000 obytel, kromě posledního, který je udáván v procentech. Bude tedy použita standardizace, jejíž výsledné hodnoty jsou na Obrázek 8.

Pokračujte spuštěním korelace s použitím transformovaných hodnot.

Transformovat | Korelace

☑ Standardizovat ☐ Normalizovat

Transformovaná data:

Trestné činy / kraje	násilné	mravnostní	majetkové	hospodářské	ostatní	zbývající kriminalita	objasněnost
Hlavní město Praha	-0,028507963	-0,217851173	3,115685818	3,28866432	1,669719142	0,281763426	-2,488641022
Středočeský kraj	-0,27557698	-0,611388777	0,195429734	-0,320761176	0,288021517	0,794262255	-0,812271612
Jihočeský kraj	0,313587598	0,077302029	-0,419070696	-0,079328033	-0,529717894	-0,075432728	0,503691509
Plzeňský kraj	-0,294582289	-0,709773178	-0,237037355	-0,465621063	-0,17724401	-0,727703966	0,274146399
Ústecký kraj	1,149821194	0,077302029	0,640032379	0,524254826	1,655620186	1,772669111	0,074270064
Královéhradecký kraj	-0,978773413	-0,021082372	-0,604413734	-0,236259576	-1,051379243	-0,929597444	0,145305374
Jihomoravský kraj	-0,351598216	-0,414619975	-0,103546238	0,367323283	-0,445124162	-0,851946106	-0,846968484
Moravskoslezký kraj	1,206837121	-0,217851173	0,72387804	0,113818482	-0,501519983	0,297293693	-1,166852797
Olomoucký kraj	0,08552389	-0,316235574	-0,413554534	-0,707054206	-0,868092823	0,00221861	0,117559096
Zlínský karaj	-1,187831812	-0,414619975	-0,989441831	-0,94848735	-0,924488644	-1,022779049	1,081507841
Kraj Vysočina	-1,529927374	-0,611388777	-0,902286474	-0,682910892	-0,233639832	-1,255733063	0,703466923
Pardubický kraj	-1,149821194	-0,513004376	-0,780930914	-0,453549405	-1,206467751	-1,115960655	0,461589624
Liberecký kraj	1,815007009	3,422371661	0,125926095	0,137961796	0,570000624	1,244640014	0,300732772
Karlovarský kraj	1,22584243	0,470839633	-0,350670289	-0,538051006	1,754312874	1,586305901	1,652464313

Obrázek 8: Vzorový příklad – transformace

zdroj: autor

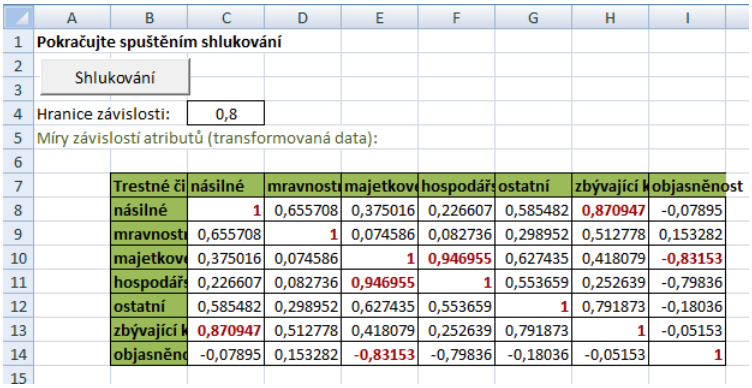
Pokud dojde k tranformaci dat, budou k veškerým dalším výpočtům, počínaje korelací spuštěnou z tohoto listu, použita transformovaná data. Pokud se uživatel rozhodne i po tomto kroku použít data původní, může buď začít práci od počátku novou kontrolou dat, nebo spuštěním korelace původních dat z listu „Popisná statistika“. V obou případech budou všechna data získaná výpočty s předchozím typem dat ztracena.

2.5 Korelace

Koeficienty korelace vyjadřují vzájemnou závislost jednotlivých atributů. Pokud je vzájemná korelace dvou atributů velmi vysoká, je doporučeno jeden z nich odstranit nebo oba vhodným způsobem sloučit. Korelace objektů, která by odhalila případné shody ve všech atributech, je nahrazena kontrolou duplicit při úvodní analýze dat.

Jak vidět na Obrázek 9, uživatel má možnost zadat hranici závislosti, která udává od jaké hodnoty budou atributy považovány za závislé a budou vyznačeny červeně. V případě vzorového příkladu je patrná velmi vysoká závislost atributů majetková a hospodářská trestná činnost. Oba atributy budou tedy sloučeny do jednoho a jednotlivé nové hodnoty budou tvořeny střední hodnotou každé z dvojic. Úpravu je nutné provést ve vstupních datech na prvním listu a zopakovat všechny dosavadní kroky, počínaje kontrolou dat, s novými

hodnotami. Na jakékoliv změny provedené přímo v listu transformace, korelace a následujících není brán zřetel.



Pokračujte spuštěním shlukování

Shlukování

Hranice závislosti: 0,8

Míry závislostí atributů (transformovaná data):

Trestné či	násilné	mravnostn	majetkové	hospodářs	ostatní	zbývající k	objasněnost
násilné	1	0,655708	0,375016	0,226607	0,585482	0,870947	-0,07895
mravnostn	0,655708	1	0,074586	0,082736	0,298952	0,512778	0,153282
majetkové	0,375016	0,074586	1	0,946955	0,627435	0,418079	-0,83153
hospodářs	0,226607	0,082736	0,946955	1	0,553659	0,252639	-0,79836
ostatní	0,585482	0,298952	0,627435	0,553659	1	0,791873	-0,18036
zbývající k	0,870947	0,512778	0,418079	0,252639	0,791873	1	-0,05153
objasněnc	-0,07895	0,153282	-0,83153	-0,79836	-0,18036	-0,05153	1

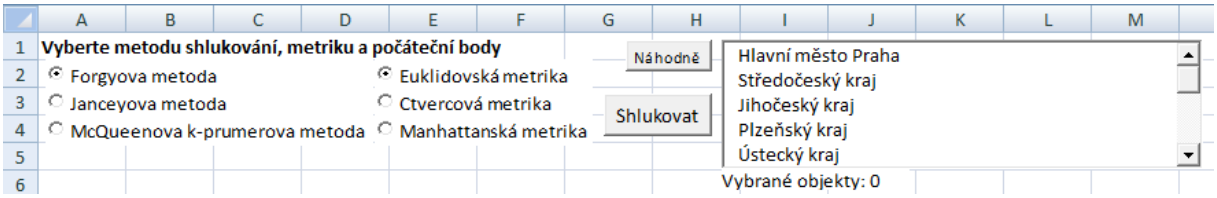
Obrázek 9: Vzorový příklad – korelace

zdroj: autor

Stiskem tlačítka „Shlukování“ je možné přejít na další list.

2.6 Shlukování

Na listu „Shlukování“ již dochází k vlastnímu měření podobností objektů a jejich přiřazování do shluků. Nejprve je nutné zvolit metodu a metriku, které budou ke shlukování použity. Následně je nutné zvolit některé objekty jako typické body. Jejich počet zároveň udá počet shluků, ke kterým se budou objekty přiřazovat. Pro výběr typických bodů existuje více metod, jako např. využití hierarchického shlukování, náhodný výběr nebo volba prvních x objektů. K výběru slouží list objektů zobrazený na Obrázek 10 vpravo. Výběr tedy může být proveden ručně uživatelem nebo náhodně pomocí příslušného tlačítka.



Vyberte metodu shlukování, metriku a počáteční body

Forgyova metoda
Janceyova metoda
McQueenova k-prumerova metoda
Euklidovská metrika
Ctvercová metrika
Manhattanská metrika

Náhodně
Shlukovat

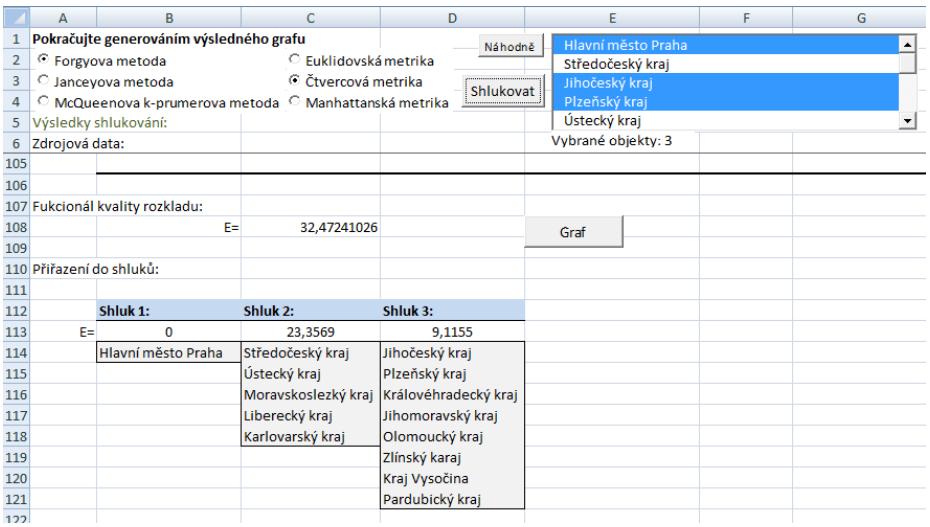
Hlavní město Praha
Středočeský kraj
Jihočeský kraj
Plzeňský kraj
Ústecký kraj

Vybrané objekty: 0

Obrázek 10: Shlukování - ovládací prvky

zdroj: autor

Typické body by ideálně měly představovat co nejvíce vzájemně nepodobné objekty. Pokud by byly jako typické body zvoleny duplicitní objekty, je to špatná volba a došlo by při shlukování k výskytu prázdných shluků. Tomu je zabráněno kontrolou vybraných objektů a při výskytu shodných typických bodů je uživatel vyzván k opravě. Náhodný výběr vrací vždy pouze unikátní typické body. Shlukování je zahájeno stiskem tlačítka „Shlukovat“.



Obrázek 11: Vzorový příklad – shlukování

zdroj: autor

Ve vzorovém příkladě je použit náhodný výběr tří typických bodů. Na Obrázek 11 je výsledek shlukování zobrazující příslušnost jednotlivých objektů do shluků. Dále je zde uveden celkový funkcionál kvality rozkladu a stejné funkcionály pro každý ze shluků. Po odscrollování výše je možné sledovat celý průběh shlukování po jednotlivých iteracích.

Pro zjištění optimálního rozkladu je vhodné vyzkoušet více kombinací metod a metrik a porovnat funkcionály kvality rozkladu. Výsledný rozklad by měl mít tuto hodnotu minimální.

V některých případech může dojít k tomu, že některý ze shluků bude prázdný. To je způsobeno nevhodně zvolenými počátečními body. V takovém případě je nutné provést nový výběr typických bodů a shlukování opakovat. Po několika pokusech s různými typickými body, metodami a metrikami bylo shlukování ukončeno s následujícím výsledkem a celkovým funcionálem rozkladu 29,9 viz Tabulka 2.

Tabulka 2: Výsledek shlukování

	Shluk 1:	Shluk 2:	Shluk 3:
E=	0	20,2284	9,7063
	Hlavní město Praha	Středočeský kraj	Ústecký kraj
		Jihočeský kraj	Liberecký kraj
		Plzeňský kraj	Karlovarský kraj
		Královéhradecký kraj	
		Jihomoravský kraj	
		Moravskoslezký kraj	
		Olomoucký kraj	
		Zlínský karaj	
		Kraj Vysočina	
		Pardubický kraj	

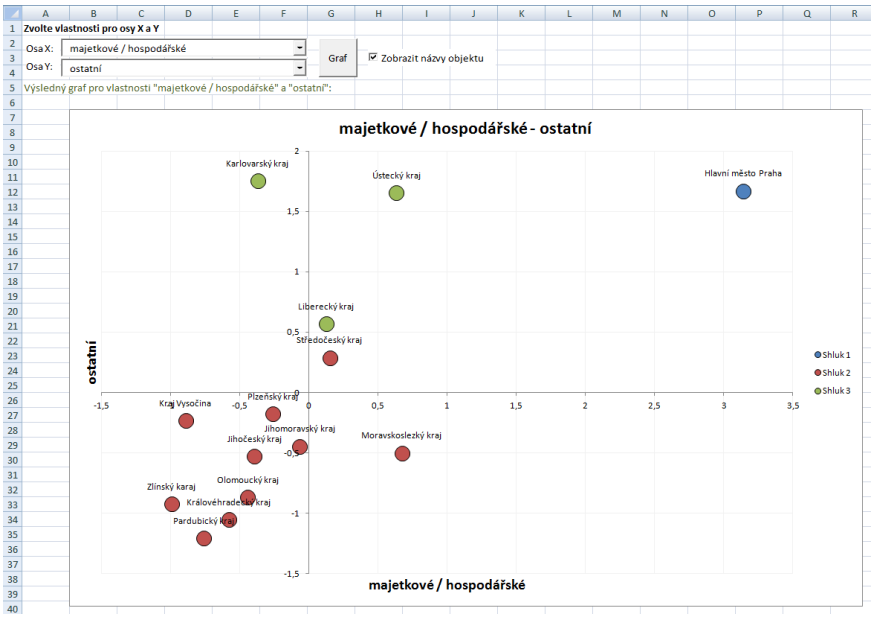
zdroj: autor

Stiskem tlačítka „Graf“ dojde k aktivaci posledního listu, na kterém je možné generovat grafy.

2.7 Grafický výstup

Na listu „Grafy“ jsou dvě okna s rozbalovací lištou, kde je možné vybrat atributy, pomocí kterých chce uživatel objekty porovnávat. Stiskem tlačítka „Graf“ potom dojde k vygenerování příslušného grafu. Grafický výstup je velmi užitečný nástroj k odhalování různých vztahů jednotlivými shluky. Pro snažší identifikaci jednotlivých objektů je každý z nich opatřen svým názvem. V případě většího počtu podobných objektů by ovšem mohlo dojít k vzájemnému překrývání popisů a tím k degradaci čitelnosti grafu, proto je možné zobrazení popisů vypnout. Stejným způsobem lze opakovaně generovat grafy pro libovolné kombinace atributů.

Graf na Obrázek 12 zobrazuje porovnání atributů „majetkové/hospodářské“ a „ostatní trestné činy“. Je zde patrná výrazná odlišnost těchto hodnot v jednotlivých shlucích. Hlavní město Praha je jako samostatný shluk odloučeno od zbytku objektů. Karlovarský, Ústecký a Liberecký kraj, tvoří třetí shluk a nachází se v levé horní části, zbytek objektů tvoří shluk č.2 a jsou koncentrovány převážně ve spodním levém kvadrantu.



Obrázek 12: Vzorový příklad – graf závislostí

zdroj: autor

V této chvíli je proces shlukování u konce. Uživatel může kdykoliv přepnout na některý z předchozích listů a kontrolovat jeho průběh, provést shlukování od kteréhokoliv kroku znovu, případně celý proces opakovat od počátku např. s novými vstupními daty.

2.8 Testování

Počet objektů a atributů vstupních dat není nijak limitován a je teoreticky omezen pouze dostupným počtem řádků a sloupců v MS Excel nutných pro dokončení procesu shlukování. Shlukování bylo úspěšně testováno na datech čítajících 1000 objektů a 50 atributů.