Univerzita Pardubice

Fakulta ekonomicko-správní

Nadstavbový modul v MS Excel pro metody hierarchického shlukování.

Jaroslav Lohynský

Diplomová práce

2011

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: **2010/2011**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Jaroslav LOHYNSKÝ**
Osobní číslo:	**E09807**
Studijní program:	**N6209 Systémové inženýrství a informatika**
Studijní obor:	**Informatika ve veřejné správě**
Název tématu:	**Nadstavbový modul v MS Excel pro metody hierarchického shlukování.**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

Základní pojmy související s tématem
Návrh modulů pro vybrané metody hierarchického shlukování
Tvorba modulů a jejich implementace do MS Excel
Metodický pokyn pro uživatele navržené aplikace

Rozsah grafických prací:

Rozsah pracovní zprávy: **cca 55 stran**

Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam odborné literatury:

BREDEN, M. SCHWIMMER, M. Excel 2007 VBA. 1. vyd. Brno: Computer Press, 2009 696 s. ISBN: 978-80-251-2698-1

KUBANOVÁ, J. Statistické metody pro ekonomickou a technickou praxi. 2. vyd. Bratislava: Statis, 2004 249 s. ISBN 80-85659-37-9

ŘEZANKOVÁ, H. HÚSEK, D. SNÁŠEL, V. Shluková analýza dat. 2. vyd. Praha: Professional Publishing, 2009 218 s. ISBN: 978-80-86946-81-8

WALKENBACH, J. Microsoft Office Excel 2007 : Programování ve VBA. Brno : Computer press, 2008. 912 s. ISBN 978-80-251-2011-8

Internetové zdroje

Vedoucí diplomové práce: **Ing. Hana Jonášová, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání diplomové práce: **4. října 2010**

Termín odevzdání diplomové práce: **6. května 2011**

L.S.

doc. Ing. Renáta Myšková, Ph.D.
děkanka

doc. Ing. Jiří Křupka, Ph.D.
vedoucí ústavu

V Pardubicích dne 4. října 2010

Prohlašuji:

Tuto práci jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odstavec 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně.

V Pardubicích dne 24. 4. 2011

Jaroslav Lohynský

Poděkování:

Touto cestou bych rád poděkoval Ing. Haně Jonášové. Ph.D, vedoucí mé práce, která věnovala svůj čas a poskytla mi důležité informace a rady při vytváření nadstavbového modulu a psaní této práce.

ANOTACE

Diplomová práce je zaměřena na shlukovou analýzu dat. Práce zahrnuje teoretické základy dané problematiky, vytvoření nadstavbového modulu pro metody hierarchického shlukování v MS Excel, popis procedur a algoritmů tohoto modulu a metodický pokyn pro používání modulu.

KLÍČOVÁ SLOVA

shluková analýza, hierarchické shlukování, nehierarchické shlukování, transformace dat, podobnost objektů, podobnost shluků, MS Excel, Visual Basic for Applications

TITLE

Upgrade module in MS Excel for execution of cluster analysis by hierarchical method.

ANNOTATION

This diploma work is focused on the cluster analysis. The work includes theoretical essentials of this issue, creation of advanced module for hierarchical clustering in MS Excel, description of procedures and algorithms used inside the module, and methodical instructions how to handle this module.

KEYWORDS

cluster analysis, hierarchical clustering, non-hierarchical clustering, data transformation, similarity of objects, similarity of clusters, MS Excel, Visual Basic for Applications

OBSAH

1 ÚVOD

Situací, kdy je potřeba setřídit objekty do skupin na základě jejich podobnosti, se v běžném životě vyskytuje mnoho a často není složité je řešit bez větších problémů na základě znalostí či intuice. Ale v současné době „informačního boomu“, s rozvíjející se vědou a technikou vzrůstá i počet informací, které je třeba nějakým způsobem roztřídit a zpracovat. V takovém případě je potřeba úlohy automatizovat tak, aby data bylo možné analyzovat pomocí vhodných nástrojů, zobrazovat strukturované výstupy nebo využívat jednotlivé postupy. Danou problematikou se v praxi zabývá shluková analýza.

Vznik shlukové analýzy spadá do čtyřicátých let dvacátého století, ale její rozvoj nastal až s masovým zavedením osobních počítačů. První monografie z oblasti shlukové analýzy byla napsána psychologem z kalifornské univerzity R. C. Tryonem roku 1939. Od té doby se shluková analýza stala nedílnou složkou zpracování informací, obsažených ve vícerozměrných pozorováních a je obsažena téměř ve všech běžně používaných statistických programech [5].

Hlavním cílem shlukové analýzy je shlukování objektů sobě podobných do shluků. Snahou je, aby ve shluku vždy byly objekty sobě podobné, podobnější než objekty v jiných skupinách, shlucích. Pro příklad je uvedena jedna z mnoha definic shlukové analýzy. Jak uvádí R.C.Tryon (1939): „Shluková analýza je obecný logický postup formulovaný jako procedura, pomocí níž seskupujeme objektivně jedince do skupin na základě jejich podobností a rozdílností.“ [13]

Pojem shluková analýza zahrnuje několik různých algoritmů a metod pro seskupování objektů podobného typu do příslušných kategorií. Jinými slovy je shluková analýza datový nástroj, který se zaměřuje na různé objekty a třídí je do skupin (shluků), a to způsobem, že podobnost mezi objekty je maximální, pokud patří do stejné skupiny a minimální, pokud nikoliv [2].

Dnes lze jen těžko najít vědní obor, v němž by tyto metody nenašly své uplatnění. Shluková analýza představuje nezbytnou součást každého automatizovaného systému analýzy vícerozměrných dat. V praxi se používá velké množství shlukovacích metod, které se dělí na dvě velké skupiny a to nehierarchické a hierarchické shlukovací metody. Cílem této práce je vytvořit nadstavbový modul pro zpracování dat vybranými metodami hierarchického shlukování. Tento modul bude vytvořen v MS Excel a bude určen pro podporu výuky předmětu Zpracování dat metodami shlukové analýzy, který je součástí studijního programu Systémové inženýrství a informatika, Univerzity Pardubice, Fakulty ekonomicko-správní.

2 ZÁKLADNÍ METODY SHLUKOVÁNÍ

Shluková analýza je vícerozměrnou statistickou metodou, která umožňuje roztřídění množiny objektů, obsahujících informace vícerozměrných pozorování, do několika co možná nejvíce stejnorodých tříd. Vzniklé třídy - shluky je dále nutné charakterizovat, tzn. najít vhodnou interpretaci vzniklého rozkladu [5].

Mezi další metody vícerozměrné statistické analýzy patří metoda hlavních komponent, faktorová analýza a diskriminační analýza. Pomocí metody hlavních komponent je možné řešit jeden z nejdůležitějších problémů statistiky – problém redukce dat a problém nalezení vztahu mezi pozorovanými proměnnými. Metoda umožňuje najít takovou množinu lineárních kombinací původních proměnných, která zachovává co nejvíce informací obsažených v údajích, ale počet jejich prvků je menší než počet původních proměnných [5]. Faktorová analýza je jednou z technik explorační[1] analýzy dat. Umožňuje popsat strukturu závislosti velkého množství veličin tak, že veličiny přiřadíme k určitému počtu hypotetických faktorů a na základě toho je roztřídíme do skupin. Veličiny přiřazené stejnému faktoru mají být silně závislé a zároveň se snažíme, aby faktorů bylo co nejméně [5]. Hlavním cílem diskriminační analýzy je nalézt nejvýhodnější způsob rozlišení skupin prvků a předpovědět, do které skupiny patří sledovaný prvek. U každého prvku se měří několik znaků, vyjadřujících jeho vlastnosti. Úkolem diskriminační analýzy je nalezení optimálního přiřazovacího pravidla, tzn. pravidla, které minimalizuje pravděpodobnost chybné klasifikace [5].

Protože pojem analýza dat je často chápán ve smyslu statistické analýzy, vznikly s rozvojem dalších matematicky a informaticky orientovaných metod nové zastřešující názvy typu dobývání znalostí z databází nebo data mining. Shlukování je pak označeno jako jeden ze základních typů získávání znalostí. V terminologii těchto metod se rozlišuje učení bez učitele a učení s učitelem.

Ke klasifikaci lze použít oba tyto přístupy. Předpokladem je, že jsou k dispozici údaje o určitých objektech, které se více či méně odlišují, takže může existovat několik skupin těchto objektů. Cílem shlukové analýzy je v dané množině objektů nalézt její podmnožiny – shluky objektů – tak, aby si objekty shluku byli navzájem podobni, ale nebyli si příliš podobni s objekty mimo tento shluk [11].

V případě učení s učitelem obsahuje vstupní datový soubor informace o příslušnosti objektů do známých skupin. Cílem je vytvořit model, na jehož základě by mohly být objekty bez známé příslušnosti zařazovány do daných skupin.

Při učení bez učitele není předem známa příslušnost žádného z objektů a obvykle není ani předem znám počet skupin. Shluková analýza patři mezi metody učení bez učitele [11].

---

[1]Explorační analýza dat (Exploratory data analysis, EDA) je ve statistice souhrn metod používaných pro průzkum dat a hledání hypotéz.

Ve vědních oblastech z oblasti informatiky se pojem shlukování nejčastěji spojuje se shlukováním objektů. Sleduje se podobnost vektorů X_i, které tvoří řádky matice. Ve skutečnosti existuje v praxi možností více. Při analýze dat mohou být také hledány shluky proměnných. Sleduje se podobnost vektorů, které tvoří sloupce matice. S touto variantou se můžeme setkat v oblasti vyhledávání informací (information retrieval), kde způsob shlukování s pojmem term-based clustering spočívá ve shlukování atributů (slov / termů) před shlukováním dokumentů. Pomocí shlukování proměnných se tedy snižuje rozměr vektoru charakterizující objekty. Možnou variantou je i současné shlukování objektů a proměnných [11].

Jiným hlediskem pro klasifikaci shluků je rozlišení, zda jde o shluky disjunktní, nebo překrývající se. V tomto smyslu rozlišujeme:

- disjunktní shlukování, ve kterém je objekt přiřazen jednoznačně do třídy, jde o tzv. hard clustering;
- překrývající se shlukování, někdy je označováno jako chumáčování – clumping.

Většina postupů ve shlukové analýze se zaměřuje na vytváření disjunktních shluků. [11]

Je známá řada metod shlukové analýzy, ale není snadné je jednoznačným způsobem utřídit. Existují tradiční metody, ale řada nových algoritmů byla navržena teprve v poslední době. Některé nové algoritmy modifikují tradiční metody shlukové analýzy, jiné se vydávají novými směry. Vzniká dokonce terminologický problém, co vše by se ještě mělo nazývat shlukovou analýzou a co by se už mělo označovat jinak [11].

Tradiční způsob třídění metod je nikoliv podle použitých matematických metod, ale podle systému použité klasifikace. Podle tohoto kritéria dělíme metody shlukové analýzy na dvě základní skupiny:

- hierarchické metody,
- nehierarchické metody.

Hierarchické metody lze charakterizovat tak, že každý shluk je současně podmnožinou jiného shluku s výjimkou samotné množiny objektů, která je považována za maximálně možný shluk. Z hierarchických metod jsou uvedeny dvě základní skupiny, lišící se způsobem shlukování:

- aglomerativní přístup, který je charakteristický tím, že vycházíme od jednotlivých objektů a jejich postupným seskupováním budujeme hierarchický systém podmnožin, až dospějeme ke konečnému spojení všech objektů do množiny objektů *O*,
- divizní přístup shlukování, který je založen na předpokladu, že vycházíme z množiny objektů určených ke klasifikaci jako celku a jejím postupným rozdělováním získáváme hierarchický systém podmnožin [5].

Při nehierarchickém shlukování se v prvním kroku určí optimální počet shluků, do kterých budou objekty rozděleny. Toto stanoveni tzv. optimálního počátečního rozkladu na *k* shluků může být následováno buď zachováním stejného počtu shluků, nebo změnou počtu shluků v průběhu vypočtu v závislosti na

řídících proměnných algoritmu. Zvláště v prvním typu algoritmů, kdy už nedochází ke změnám v počtu shluků je tedy nutno klást na počáteční rozklad značný důraz [4]. Z tohoto hlediska lze rozdělit nehierarchické metody shlukování na metody:

- neměnící počet shluků (Forgyova a Janceyova metoda, MacQueenova k-průměrová metoda),
- měnící počet shluků (Metody RELOC, ISODATA apod.).

Na klasifikaci metod nehierarchického shlukování ovšem existuje celá řada pohledů. Tyto metody se mohou dělit také na metody optimalizační a metody analýzy módů. Mezi metody optimalizační patří například metody založené na středech shluků, např. *k*-průměrová metoda. Metody analýzy módů jsou založené na pravděpodobnostním rozdělení *m*-rozměrného pozorování, tj. na hustotě pravděpodobnosti. Tyto metody jsou v literatuře také označovány jako metody založené na hustotě [11].

Dále lze metody nehierarchického shlukování rozdělit na metody pevného shlukování (objekt je ve shluku nebo není) a na fuzzy shlukovou analýzu, kdy je každému objektu nejprve přiřazena míra příslušnosti ke každému ze shluků, na základě níž je teprve provedeno binární přiřazení [11].

3 PŘEDZPRACOVÁNÍ VSTUPNÍCH DAT

3.1 Vstupní datová matice

Vstupem pro shlukování je datová matice a výstupem shlukování je identifikace shluků. Objekty jsou zařazeny do jednotlivých shluků na základě jejich podobnosti [11].

Shlukovou analýzu provádíme zpravidla na množině n objektů [O_1, O_2, …, O_n], z nichž je každý popsán prostřednictvím p ukazatelů [U_1, U_2, …, U_p], které má smysl na dané množině objektů sledovat. Výběr množiny sledovaných ukazatelů rozhoduje o úspěchu závěrů metody, proto je nutné věnovat mu patřičnou pozornost. Předpokládejme, že na objektu O_i, i = 1, 2, …, n bylo měřeno p znaků. Tím byl získán vektor $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. Výsledkem pozorování je matice $\boldsymbol{X}$ typu $n \times p$

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

kde x_{ij} značí hodnotu j-tého ukazatele u i-tého objektu. Tuto matici nazýváme datovou maticí. Jednotlivá měření, tzn. vektory X_i tvoří řádky matice. Sloupce datové matice $\boldsymbol{X}$ vyjadřují hodnoty j-tého znaku v množině objektů, j = 1,2, …, p. Předpoklad je zde, že byly měřeny pouze kvantitativní znaky, a tedy vektory X_i jsou číselné vektory [5]. Práce s ostatními typy dat bude popsána dále v textu.

3.2 Předzpracování dat

Předzpracování dat spočívá především v posouzení, zda mají být do analýzy zahrnuty všechny proměnné, nebo jen některé. Poté by měla být věnována pozornost typům jednotlivých proměnných a jejich významu pro analýzu. Často je vhodné provést transformaci dat, buď jejich standardizaci, nebo normalizaci. V neposlední řadě je součástí shlukové analýzy identifikace odlehlých hodnot a ošetření chybějících údajů [11].

3.2.1 Typy proměnných

Pro účely statistické analýzy, a teda i shlukové analýzy, je třeba u každé proměnné určit její typ. V této podkapitole bude popsán jeden z možných přístupů, který za hlavní kritérium považuje typy vztahů mezi hodnotami. Podle tohoto hlediska rozlišujeme proměnné nominální, ordinální, intervalové a poměrové.

Nominální proměnná je taková, o jejichž dvou hodnotách můžeme pouze říci, zda jsou stejné nebo různé. Hodnotami mohou být texty (písmena) i číslice. Lze u nich zjišťovat jen rozdělení četností, nelze provádět aritmetické operace [10].

Ordinální (pořadová) proměnná je taková, u jejichž dvou hodnot můžeme navíc určit pořadí. Jako hodnoty lze použít text, datum, číslo [10].

Intervalová (rozdílová) proměnná je taková, pro jejíž dvě hodnoty můžeme navíc (k možnostem ordinální proměnné) vypočítat, o kolik je jedna hodnota větší nebo menší než druhá. Hodnotami jsou tedy čísla [10].

Poměrová (podílová) proměnná je taková, pro jejíž dvě hodnoty můžeme navíc (k možnostem intervalové proměnné) vypočítat, kolikrát je jedna hodnota větší (resp. menší) než druhá, tj. jedná se pouze o kladné hodnoty.

Nominální a ordinální proměnné jsou souhrnně označovány jako kvalitativní, intervalové a poměrové proměnné jsou souhrnně označovány jako kvantitativní (numerické), v literatuře se můžeme setkat též s pojmem kardinální. Kvantitativní proměnné můžeme podle jiného hlediska dělit na diskrétní, které nabývají pouze celočíselných obměn a spojité (metrické), jež mohou nabývat libovolných hodnot z určitého intervalu [10].

Nominální, ordinální a kvantitativní diskrétní proměnné můžeme souhrnně označit jako kategoriální. Podle jiného hlediska je můžeme dělit na dichotomické (alternativní), které nabývají pouze dvou kategorií a vícekategoriální (množné), jež nabývají více než dvou kategorií.

U dichotomických proměnných se při výpočtech předpokládá, že jde o proměnné binární, které nabývají hodnot 0 a 1. U dichotomických proměnných můžeme dále rozlišit proměnné symetrické, které mají obě kategorie stejné důležitosti a asymetrické, jejichž jedna kategorie je důležitější [11].

3.2.2 Výběr proměnných

Ze statistického hlediska může být vhodné, aby v souboru dat pro shlukovou analýzu zůstaly pouze proměnné statisticky nezávislé. Základní způsoby testování závislosti pro dvojice proměnných závisí na jejich typu. Pro kvantitativní typ proměnné je možné pro určení míry lineární závislosti použít korelační koeficient, který se pro *j*-tou a *l*-tou proměnnou spočítá podle vzorce:

$$r_{jl} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}(x_{il} - \bar{x}_l)^2}}, \qquad (1)$$

kde

$\bar{x}_j$ je aritmetický průměr *j*-té hodnoty a

$\bar{x}_l$ je aritmetický průměr *l*-té hodnoty.

Tento korelační koeficient nabývá hodnot v intervalu od -1 do 1, přičemž hodnota 0 znamená lineární nezávislost [11].

Při práci s dichotomickými proměnnými jsou údaje o objektech vyjádřeny pomocí kontingenční tabulky četnosti. Symboliku dvourozměrné tabulky četností pro objekty $\mathbf{x}_i$ a $\mathbf{x}_k$ znázorňuje tabulka 1.

Tabulka 1 Kontingenční tabulka dichotomických dat. Zdroj [11]

Kat. objektu $\mathbf{x}_i$	Kategorie objektu $\mathbf{x}_k$	
	1	0
1	a	b
0	c	d

Uvedená kontingenční tabulka se nazývá čtyřpolní (matice četností obsahuje čtyři políčka). Někdy bývá označována jako asociační. Pro vyjádření závislosti dichotomických proměnných lze využít korelační koeficient, který lze spočítat podle vzorce:

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \tag{2}$$

Tato statistika se nazývá Pearsonův korelační koeficient a nabývá hodnot v intervalu od -1 do 1, přičemž hodnota -1 znamená nezávislost proměnných [11].

3.2.3 Transformace dat

Transformace dat spočívá v standardizaci anebo normalizaci dat[2]. Hodnoty jednotlivých znaků objektů jsou často v různých jednotkách. To může způsobovat, že se určité znaky jeví jako dominující a jiné znaky jen málo ovlivňuji průběh shlukování. Někdy je proto výhodné data upravit tak, aby byly všechny znaky souměřitelné. Jedním ze způsobů, jak toho docílit, je standardizace dat [4] . Standardizaci dat lze provést pomocí následujícího vztahu:

$$x_{ij}^{s} = \frac{x_{ij} - \bar{x}_j}{s_j}, \tag{3}$$

kde

x_{ij}^{s} je standardizovaná hodnota,

x_{ij} je původní hodnota,

$\bar{x}_j$ je střední hodnota *j*-tého znaku (sloupce),

s_j je směrodatná odchylka *j*- tého znaku (sloupce).

[2] V literatuře je možné se setkat i s odlišnou terminologií. Standardizace – normalizace sloupců, normalizace – normalizace řádků.

Výpočet střední hodnoty *j*-tého sloupce: $\bar{x}_j = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{ij}$.

Výpočet směrodatné odchylky *j*-tého sloupce: $s_j = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}$.

Objekty pro shlukovou analýzu, jsou určeny vektory o *p* znacích. Normy těchto vektorů mohou někdy nežádoucím způsobem ovlivňovat výsledky kvantitativního hodnoceni podobnosti objektů. V takových případech je vhodné normalizovat tyto vektory, aby měli stejnou normu (nejlépe jednotkovou) [4]. Normalizaci dat lze provést pomocí následujícího vztahu:

$$x_{ij}^n = \frac{x_{ij}}{n_i}, \tag{4}$$

kde

x_{ij}^n je normalizovaná hodnota,

x_{ij} je původní hodnota,

n_i je norma i-tého řádku.

Výpočet normy i-tého řádku: $n_i = \sqrt{\sum_{j=1}^{p} x_{ij}^2}$.

3.2.4 Identifikace odlehlých hodnot

Odlehlá hodnota je případ, kdy se hodnota proměnné vyskytuje jednou nebo při nízké frekvenci daleko od střední hodnoty a také od většiny ostatních hodnot této proměnné [10]. Analýza souboru dat bez vyloučení odlehlých hodnot může být zkreslená a nesprávná. Odlehlá hodnota může být zjištěna statistickým testem, lze použít například Dixonův test odlehlých hodnot [5].

3.2.5 Ošetření chybějících údajů

Při sbírání a kombinování dat se vyskytují chybějící hodnoty téměř v každé sadě dat. Jestliže programový systém záznamy s chybějícími hodnotami ignoruje, výsledek není ideální. Chybějící hodnoty je možné nahradit. Cíl při nahrazování chybějících hodnot je dvojí: zaplnit prázdná místa nejpravděpodobnějšími hodnotami a zachovat celkové rozdělení hodnot proměnné. Nahradit chybějící hodnotu je možné těmito způsoby:

- Substituce střední hodnotou – střední hodnota je založena na statistickém výpočtu nejmenší chyby čtverců. Tím se do rozdělení hodnot proměnné zavádí nejmenší možná variace. Je-li rozdělení velmi špičaté (nesouměrné), je lepší použít medián.
- Substituce střední hodnotou třídy – využívají se střední hodnoty podskupin jiných proměnných nebo kombinací proměnných. Tato metoda zachovává lépe původní rozdělení hodnot. Je vhodné, jsou-li vybrané proměnné značně korelované s proměnnou s chybějící hodnotou.

- Regresní substituce – podobně jako u substituce střední hodnotou třídy využívá regresní substituce střední hodnoty skupin jiných proměnných. Výhodou regrese je schopnost pracovat se spojitými proměnnými stejně jako hledat ve více proměnných přesnější míru. Výsledné hodnocení regrese slouží k dopočtení náhradních hodnot [10].

4 MĚŘENÍ PODOBNOSTI OBJEKTŮ

Základním problémem shlukové analýzy je kvantitativně vyjádřit podobnost či vzdálenost objektů. V jednotlivých krocích algoritmů je posuzována podobnost resp. vzdálenost dvou objektů, objektu a shluku nebo dvou shluků. V některých případech je způsob hodnocení podobnosti dán přímo shlukovací metodou, často jsou ale tyto kroky nezávislé, a je vybírána nejvhodnější míra podobnosti, jak z hlediska shlukovaných objektů, tak i z hlediska použité metody shlukování [5].

U míry podobnosti je zpravidla požadována, aby nabývala hodnoty 0 pro maximální rozdílnost a hodnoty 1 pro totožnost objektů. Často však praktické důvody vedou k použití různých měr vzdálenosti, kde je stejný jev měřen s opačnou interpretací, tj. hodnota 0 vyjadřuje totožnost objektů [5]. Míra podobnosti je definována následujícím způsobem:

Nezápornou reálnou funkci m, která každé dvojici vektorů X_i, X_k z E_p přiřazuje číslo m_{ik}, nazýváme mírou podobnosti těchto vektorů, jestliže pro všechny dvojice X_i, X_k platí:

a) $0 \leq m(X_i, X_k) < 1$ pro $X_i \neq X_k$,

b) $m(X_i, X_k) = 1$,

c) $m(X_i, X_k) = m(X_k, X_i)$.

Je-li definována míra podobnosti, lze její hodnoty vypočítat pro všechny dvojice objektů a sestavit je do symetrické matice podobnosti $\boldsymbol{M}$ typu $n \times n$. Na základě vlastnosti b) z definice míry podobnosti bude mít tato matice na hlavní diagonále jedničky [5].

Řada shlukovacích metod však vychází z duálního pojmu k míře podobnosti, a to z míry nepodobnosti, která se v mnoha případech jeví jako výhodnější. Míru nepodobnosti lze definovat analogicky k míře podobnosti, pouze vlastnost b) bude $m(X_i, X_i) = 0$. Jako míry nepodobnosti jsou typické funkce založené na vzdálenosti objektů. K určení míry podobnosti resp. nepodobnosti ve shlukové analýze může sloužit některá z funkcí vzdálenosti. Je nutné tedy definovat pojem vzdálenost.

Nezáporná reálná funkce $d(X_i, X_k)$ se nazývá vzdáleností bodů X_i a X_k z E_p jestliže platí:

a) $d(X_i, X_k) \geq 0$ pro všechny body X_i a X_k z E_p,

b) $d(X_i, X_k) = 0$ právě tehdy, když $X_i = X_k$,

c) $d(X_i, X_k) = d(X_k, X_i)$,

d) $d(X_i, X_k) + d(X_k, X_m) \geq d(X_i, X_m)$ pro každou trojici bodů X_i, X_k a X_m z E_p.

Takto definovanou vzdálenost bodů je možné interpretovat rovněž jako vzdálenost jim odpovídajících objektů O_i a O_k. K určení vzdálenosti objektů se ve shlukové analýze používají různé způsoby výpočtu vzdálenosti, většinou převzaté z matematické analýzy [5].

4.1 Míry vzdálenosti pro kvantitativní data

Euklidovská vzdálenost [5]

$$d_E(X_i, X_k) = \sqrt{\sum_{j=1}^{p}(x_{ij} - x_{kj})^2}\,, \tag{5}$$

kde x_{ij} je hodnota *j*-tého pozorování na *i*-tém prvku a x_{kj} je hodnota *j*-tého pozorování na *k*-tém prvku. Tato metoda je početně jednoduchá, ale předpokládá nekorelovanost proměnných [5]. Další způsoby výpočtu jsou:

Čtvercová Euklidovská vzdálenost [11]

$$d_{ES}(X_i, X_k) = \sum_{j=1}^{p}(x_{ij} - x_{kj})^2\,, \tag{6}$$

Vážená Euklidovská vzdálenost [11]

$$d_{EW}(X_i, X_k) = \sqrt{\sum_{j=1}^{p} w_j(x_{ij} - x_{kj})^2}\,, \tag{7}$$

Hammingova vzdálenost (manhattanská) [5]

$$d_B(X_i, X_k) = \sum_{j=1}^{p}\left|x_{ij} - x_{kj}\right|, \tag{8}$$

Čebyševova vzdálenost (sup-metrika) [5]

$$d_C(X_i, X_k) = \max\left|x_{ij} - x_{kj}\right|, \tag{9}$$

Minkovského vzdálenost [5]

$$d_M(X_i, X_k) = \sqrt[q]{\sum_{j=1}^{p}\left|x_{ij} - x_{kj}\right|^q} \quad \text{q = 1, 2, …, } \infty, \tag{10}$$

Sokalova vzdálenost [3]

$$d_S(X_i, X_k) = \sqrt{\frac{d_{ES}(X_i, X_k)}{p}}\,, \tag{11}$$

Hammingova, Euklidovská a Čebyševova vzdálenost jsou zvláštním případem Minkovského metriky. Položíme-li ve vztahu pro výpočet Minkovského vzdálenosti $q = 1$ dostaneme Hammingovu vzdálenost, pro $q = 2$ jde o Euklidovskou vzdálenost a pro $q \to \infty$ získáme Čebyševovu vzdálenost. Takto definované vzdálenosti mají i stejné vlastnosti. jejich společnou nevýhodou je to, že jsou závislé na jednotkách, ve kterých byly měřeny jednotlivé znaky. Řešením tohoto problému může být standardizace nebo normalizace údajů datové matice [5]. Viz kapitola 3.2.3 Transformace dat.

V literatuře je možné se setkat i s dalšími způsoby výpočtu vzdálenosti mezi objekty, které nemusí být vždy v souladu s výše uvedenou definicí vzdálenosti. Příkladem takovýchto vzdáleností jsou Jeffreys-Matusitu vzdálenost, Lanceyova-Williamsova vzdálenost, vzdálenost s názvem koeficient divergence atd. [5]

4.2 Míry vzdálenosti pro dichotomická data

V případě, že jsou objekty charakterizovány dichotomickými znaky, se míra podobnosti ve shlukové analýze nazývá koeficientem asociace. Při určování prvků v matici vzdáleností resp. podobností objektů O_i a O_k bude pozorována shoda či neshoda výsledků u p proměnných. Asociace dvou objektů je vyjádřena asociační čtyřpolní tabulkou.

Tabulka 2 Asociační čtyřpolní tabulka. Zdroj [5]

Objekt O_i	Objekt O_k	
	1	0
1	a	b
0	c	d

Počet pozitivních shod vektorů X_i , X_k je ve výše uvedené tabulce označen písmenem *a,* počet negativních shod je označen písmenem *d.* Počet neshod, pro případ že znaky ve vektoru X_i nabývají hodnoty 1 a znaky ve vektoru X_k nabývají hodnoty 0 je označen písmenem *b* a počet neshod, pro případ že znaky ve vektoru X_i nabývají hodnoty 0 a znaky ve vektoru X_k nabývají hodnoty 1 je označen písmenem *c*. Součet $a+d$ vyjadřuje celkový počet shod, součet $b+c$ je roven celkovému počtu neshod a součet $a+b+c+d$ je roven *p*, tj. celkovému počtu znaků kterými je každý objekt charakterizován. Přehled nejdůležitějších koeficientů asociace [5]:

Jaccardův koeficient asociace vyjadřuje podíl shodných výsledků při vyloučení shodně negativních výsledků. Je označován rovněž jako S-koeficient.

$$S_J = \frac{a}{a+b+c} \tag{12}$$

Sokalův-Michenerův koeficient asociace vyjadřuje podíl shodných výsledků, je označován rovněž jako M-koeficient.

$$S_{SM} = \frac{a+d}{a+b+c+d} \tag{13}$$

Russellův-Raoův koeficient asociace vyjadřuje podíl shodně pozitivních výsledků, je označován jako RR-koeficient. Jeho nevýhodou je, že různě hodnotí podobnost objektu se sebou samým.

$$S_{RR} = \frac{a}{a+b+c+d} \tag{14}$$

Diceův koeficient asociace.

$$S_D = \frac{2a}{2a+b+c} \tag{15}$$

Rogersův-Tanimonoův koeficient asociace.

$$S_{RT} = \frac{a+d}{a+2b+2c+d} \tag{16}$$

Nepojmenovaný N1 koeficient asociace.

$$S_{N1} = \frac{2(a+d)}{2(a+d)+b+c} \tag{17}$$

Nepojmenovaný N2 koeficient asociace.

$$S_{N2} = \frac{a}{a+2(b+c)} \tag{18}$$

Hamannův koeficient - jeho oborem hodnot je interval <−1, 1>. Hodnota -1 nastane v případě, kdy nedojde k žádné shodě a hodnota 0 v případě stejného počtu shod jako neshod [3].

$$S_H = \frac{a+d-(b+c)}{a+b+c+d} \tag{19}$$

Míru vzdálenosti lze z koeficientů asociace spočítat podle vztahu:

Míra vzdálenosti = 1 - koeficient asociace (tj. míra podobnosti).

Tento způsob lze použít u koeficientů asociace, jejichž oborem hodnot je interval <0,1>. Dále se jako míra vzdálenosti se používá například čtvercová euklidovská vzdálenost, což je součet četností *b* a *c*. Stejným součtem se dá vyjádřit rovněž manhattanská vzdálenost:

$$d_{ES} = d_B = b+c \tag{20}$$

Také je možné využít euklidovskou vzdálenost:

$$d_E = \sqrt{b+c} \tag{21}$$

Mezi koeficienty lze vyjádřit vzájemné vztahy, například mezi euklidovskou vzdáleností a Sokalův-Michenerův koeficientem asociace platí vztah:

$$d_E = \sqrt{(a+b+c+d)(1-S_{SM})} \tag{22}$$

Pro binární data se též používá Lanceyova-Williamsova nemetrická míra nepodobnosti, která se spočítá takto:

$$d_{LW} = \frac{b+c}{2a+b+c} \tag{23}$$

5 HIERARCHICKÉ SHLUKOVÁNÍ

Metoda hierarchického shlukování byla obecně popsána v kapitole 2 Základní metody shlukování. V této kapitole bude uvedený postup popsán podrobně.

Vstupem do hierarchického shlukování je tabulka vzdáleností mezi jednotlivými objekty. Od tohoto okamžiku je vhodnější uvažovat objekty v tabulce vzdáleností ne jako prosté objekty, ale jako shluky s jedním objektem. Dále je postup takový, že dva shluky, mezi nimiž je nejmenší vzdálenost (v případě aglomerativního přístupu) se sloučí v jeden nový shluk, který bude mít dva objekty. Vznikne nová tabulka vzdáleností mezi shluky, ve které již nebudou shluky, které se slučovaly, a místo nich se do tabulky doplní nově vzniklý shluk. Zároveň je nutné vypočítat vzdálenosti mezi tímto nově vzniklým shlukem a ostatními shluky v tabulce. Problematice měření vzdálenosti mezi shluky se věnuje následující podkapitola 5.1.

Celý postup se opakuje do té doby, dokud se nesloučí všechny objekty v jeden shluk. V průběhu shlukování se vytváří nová tabulka, do které se zapisují údaje z jednotlivých iteraci, tj. jaké dva shluky byly v daném kroku sloučeny, jaký shluk vzniknul, jaké obsahoval objekty a při jaké vzdálenosti ke sloučení došlo. Tato tabulka je výstupem hierarchického shlukování a zároveň je podkladem pro vytvoření dendrogramu. Dendrogram je stromový graf, který slouží jako grafický výstup hierarchického shlukování. Problematice tvorby dendrogramu se podrobně věnuje podkapitola 5.2.

5.1 Měření podobnosti shluků

Mezi nejčastěji používané metody výpočtu vzdálenosti mezi shluky patří:

- metoda nejbližšího souseda,
- metoda nejvzdálenějšího souseda,
- metoda průměrné vzdálenosti,
- centroidní metoda,
- mediánová metoda,
- Wardova-Wishartova metoda.

Jakýkoliv způsob výpočtu vzdálenosti shluků lze vyjádřit pomocí Lance-Williams formule [6]. Její obecný tvar je:

$$d(U,R) = \alpha_A \cdot d(U,P) + \alpha_B \cdot d(U,Q) + \beta \cdot d(P,Q) + \gamma \cdot \left| d(U,P) - d(U,Q) \right|, \qquad (24)$$

kde

$\alpha_A, \alpha_B, \beta, \gamma$ jsou koeficienty, které se mění podle použité metody výpočtu vzdálenosti shluků,

U je původní shluk,

R je nový shluk,

P a Q jsou původní shluky, jejichž spojením vzniká nový shluk R.

5.1.1 Metoda nejbližšího souseda (nearest neighbour)

Metodu poprvé uvedl P. Sneath pod názvem „single linkage“. Při aplikaci této metody se za vzdálenost dvou shluků považuje vzdálenost jejich nejbližších prvků, tzn.

$$d(S_g, S_h) = \min[d(X_i, X_k)];\ X_i \in S_g;\ X_j \in S_h \tag{25}$$

Při použití této metody se často mohou i značně odlehlé objekty dostat do stejného shluku, pokud je mezi nimi větší počet dalších objektů, které vytváří tzv. most. Toto řazení se objektů do tvaru mostu je charakteristické pro tuto metodu a je považováno za její nevýhodu, i když z ostatních hledisek je metoda nejbližšího souseda velmi příznivě hodnocena [5].

Lance-Williams formule pro metodu nejbližšího souseda má tento tvar:

$$d(U,R) = 0{,}5 \cdot d(U,P) + 0{,}5 \cdot d(U,Q) - 0{,}5 \cdot \left| d(U,P) - d(U,Q) \right|. \tag{26}$$

5.1.2 Metoda nejvzdálenějšího souseda (furthest neighbour)

Metoda je známá pod názvem „complete linkage“. Za vzdálenost dvou shluků je považována vzdálenost jejich nejodlehlejších prvků, tzn.

$$d(S_g, S_h) = \max[d(X_i, X_k)];\ X_i \in S_g;\ X_j \in S_h \tag{27}$$

Tato metoda má sklon k vytváření kompaktních shluků, které ale nejsou velké. Nepříznivý výskyt mostů, resp. řetězení shluků v této metodě odpadá [5].

Lance-Williams formule pro metodu nejvzdálenějšího souseda má tento tvar:

$$d(U,R) = 0{,}5 \cdot d(U,P) + 0{,}5 \cdot d(U,Q) + 0{,}5 \cdot \left| d(U,P) - d(U,Q) \right|. \tag{28}$$

5.1.3 Metoda průměrné vzdálenosti

U této metody je za vzdálenost dvou shluků považována průměrná vzdálenost mezi páry patřícími dvěma shlukům [5]. Výpočet se provede pomocí vztahu:

$$d(S_g, S_h) = \frac{1}{n_g n_h} \sum_{x_i \in S_g} \sum_{x_k \in S_h} d(X_i, X_k). \tag{29}$$

Lance-Williams formule pro metodu průměrné vzdálenosti má tento tvar:

$$d(U,R) = \frac{|P|}{|R|} \cdot d(U,P) + \frac{|Q|}{|R|} \cdot d(U,Q), \tag{30}$$

kde $|P|, |R| a |Q|$ jsou počty objektů shluků P, R a Q.

5.1.4 Centroidní metoda

Metoda byla poprvé použita R. Sokalem a C. Michenerem pod názvem „weighted group method". Autoři vyšli z geometrických představ a míru nepodobnosti dvou shluků vyjádřili jako euklidovskou vzdálenost jejich těžišť. Postupně se ukázalo, že je výhodnější vzít místo euklidovské vzdálenosti její druhou mocninu. Umístění shluků je reprezentováno centroidem všech bodů během shlukování [5]. Výpošet se provede na základě vztahu:

$$d(S_g, S_h) = d^2\left(\frac{1}{n_g}\sum_{x_i \in S_g} X_i; \frac{1}{n_h}\sum_{x_k \in S_h} X_k\right) = d^2(\bar{x}_g; \bar{x}_h). \tag{31}$$

Lance-Williams formule pro centroidní metodu má tento tvar:

$$d(U,R) = \frac{|P|}{|R|} \cdot d(U,P) + \frac{|Q|}{|R|} \cdot d(U,Q) - \frac{|P| \cdot |Q|}{|R|^2} \cdot d(P,Q), \tag{32}$$

kde $|P|, |R| a |Q|$ jsou počty objektů shluků P, R a Q.

5.1.5 Mediánová metoda

Metoda byla poprvé uvedena J. C. Gowerem pod názvem „unweighted group method". Cílem byla snaha odstranit nedostatky centroidní metody (rozdílné počty objektů shluků způsobí rozdílnou váhu složek v algoritmu centroidní metody). Mediánová metoda počítá vážený průměr vzdálenosti mezi páry náležícími dvěma shlukům [5].

Lance-Williams formule pro mediánovou metodu má tento tvar:

$$d(U,R) = 0{,}5 \cdot d(U,P) + 0{,}5 \cdot d(U,Q) - 0{,}25 \cdot d(P,Q). \tag{33}$$

5.1.6 Wardova-Wishartova metoda

Při použití této metody se vzdálenost mezi dvěma shluky počítá jako přírůstek vnitroskupinového součtu čtverců odchylek, který vznikne spojením těchto shluků. Výhodou této metody je, že vytváří shluky přibližně stejné velikosti [6].

Lance-Williams formule pro Wardovu-Wishartovu metodu má tento tvar:

$$d(U,R) = \frac{|P|+|U|}{|R|+|U|} \cdot d(U,P) + \frac{|Q|+|U|}{|R|+|U|} \cdot d(U,Q) - \frac{|U|}{|R|+|U|} \cdot d(P,Q), \tag{34}$$

kde $|P|,|R|,|Q|a|U|$ jsou počty objektů shluků P, R, Q a U.

5.2 Vytvoření dendrogramu

Výstupem z procesu shlukování je tabulka, která dává přehled o tom, jaké shluky resp. objekty byly v každém kroku spojeny, a také informaci při jaké vzdálenosti spojení proběhlo. Obecně lze tuto tabulku vyjádřit takto:

Tabulka 3 Tabulka po shlukování - obecná [zdroj: autor]

Krok	**Shluk 1**	**Shluk 2**	**Nový shluk**	**Vzdálenost**
1				
2				
…				
n - 1				

kde *n* je počet objektů. Tuto tabulku nelze v praxi explicitně použít pro vykreslení dendrogramu a musí být upravena.

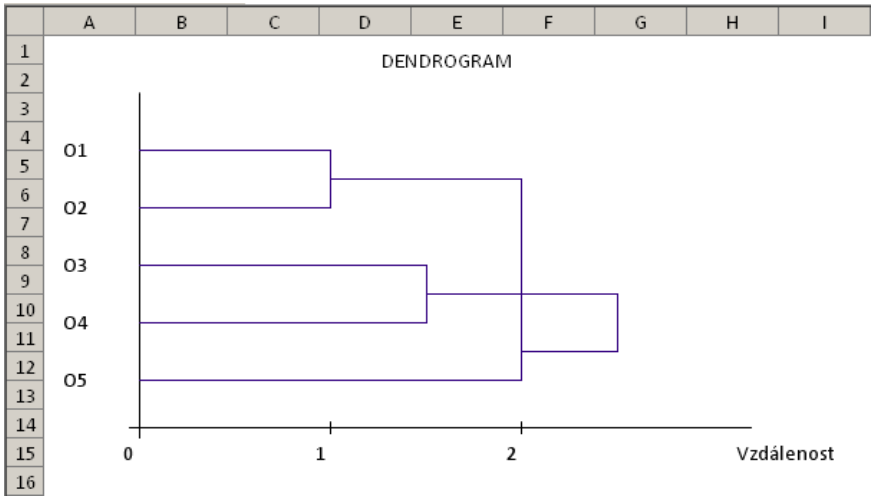
Úprava tabulky pro vykreslení dendrogramu

Důvod a princip úpravy tabulky je možné demonstrovat na příkladu s pěti objekty. Tj. *n* = 5. Nechť je výsledkem po shlukování tato tabulka:

Tabulka 4 Tabulka po shlukování - příklad [zdroj: autor]

Krok	**Shluk 1**	**Shluk 2**	**Nový shluk**	**Vzdálenost**
1	O1	O2	O1, O2	1
2	O3	O4	O3, O4	1,5
3	O1, O2	O5	O1, O2, O5	2
4	O1, O2, O5	O3, O4	O1, O2, O5,O3, O4	2,5

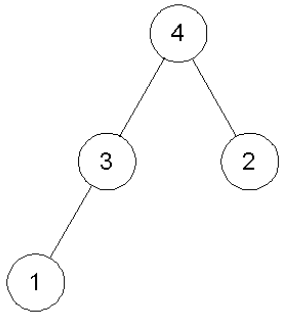
Při vykreslování dendrogramu krok po kroku od prvního řádku této tabulky dospějeme k tomuto výsledku:



Obrázek 1 Dendrogram z neupravené tabulky [zdroj: autor]

Nejprve bylo vykresleno spojení objektu 1 s objektem 2 (O1, O2), poté spojení objektu 3 s objektem 4 (O3, O4), dále spojení shluku objektů 1 a 2 s objektem 5 (O1, O2, O5) a nakonec spojení shluku objektů 1, 2 a 5 s shlukem objektů 3 a 4 (O1, O2, O5, O3, O4). Nedostatek tohoto postupu je zřejmý z výše uvedeného obrázku – došlo ke křížení vertikálních a horizontálních linek.

Při řešení tohoto problému programovým prostředkem se ukázalo být výhodné pohlížet na dendrogram jako na strom, tj. jako na typ grafu, ve kterém jsou kroky shlukování reprezentovány uzly a úlohu vhodného setřídění kroků řešit algoritmem procházení grafu (stromu) do hloubky. Strom analogický s daným příkladem znázorňuje obrázek 2. Uzly v grafu jsou řádky tabulky, hrany nejsou ohodnocené a vyjadřují vazby mezi řádky v tabulce.



Obrázek 2 Strom výsledku shlukování - příklad [zdroj: autor]

Pro procházení stromu do hloubky je potřeba vytvořit dva zásobníky. GENER pro ukládání generovaných uzlů a EXPAND pro ukládání expandovaných uzlů. Zásobník GENER pracuje v režimu LIFO – Last In First Out [8].

Algoritmus pracuje následovně:

1. Vlož kořen stromu (poslední řádek tabulky) do zásobníku GENER.
2. Je-li zásobník GENER prázdný – konec, jinak pokračuj krokem tři.
3. Přesuň poslední uzel z GENER do EXPAND.
4. Expanduj přesunutý uzel a následovníky vlož do zásobníku GENER v pořadí jak byly expandováni.
5. Pokračuj krokem dva.

Uzly v algoritmu odpovídají krokům v tabulce. Jestliže se expanduje uzel (řádek), tak se v podstatě hledají takové řádky v tabulce, jejichž „Nový shluk“ odpovídá „Shluku1“ a „Shluku2“ expandovaného uzlu (řádku). Cílem úlohy je projít celý graf, s tím že na konci úlohy jsou uzly v zásobníku EXPAND setříděny vhodným způsobem pro vykreslování dendrogramu. Pro bližší objasnění výše uvedeného postupu, je tento příklad podrobně vyřešen v příloze 11, která je uložena na CD.

6 NADSTAVBOVÝ MODUL V MS EXCEL – ALGORITMY

6.1 Úvod – programování v MS Excel

VBA (Visual Basic for Applications) je programovacím jazykem, jehož základy jsou společné hned několika aplikacím balíku MS Office. Kromě Excelu také Wordu, Accessu, Outlooku či PowerPoint (tento jazyk se dokonce uplatňuje i třeba v aplikacích Corel Draw, AutoCAD). VBA přitom není jediným jazykem, ve kterém lze programovat nadstavby aplikací MS Office. Vedle něj se nabízí ještě VBScript (Visual Basic Script) a případně JavaScript [12].

VBA je objektově orientovaným jazykem. MS Office má objekty. Každý objekt má své vlastnosti a na každý objekt můžeme aplikovat metody, lépe řečeno provést s objektem určitou činnost. Některé objekty navíc mohou rozpoznávat tzv. události, kdy kód proběhne na základě nějaké akce (klepnutí myší, apod.) [9].

6.1.1 Základní pojmy

Projekt - představuje souhrn všech modulů a formulářů patřících danému sešitu. Je explicitně součástí každého sešitu.

Objekt - je nejmenším stavebním prvkem objektového modelu a může být představován například buňkou, tlačítkem, listem, grafem, aplikací apod. Kolekce objektů, pokud existuje, je také objektem.

Modul - je skupina deklarací a procedur (obsahujících programový kód), které dohromady tvoří logickou část celku.

Formulář - tvoří uživatelské dialogové okno sloužící ke komunikaci s uživatelem za účelem získání vstupů.

Procedura - je pojmenovaná posloupnost příkazů, která se vykoná jako celek (v praxi je ztotožňována s pojmem makro).

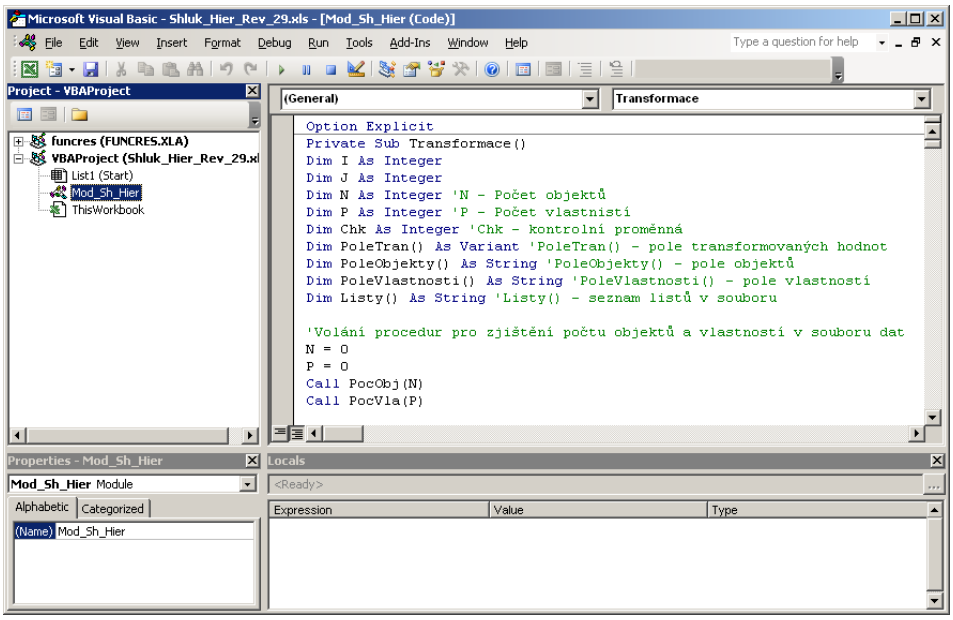
Programový kód - tvoří zápis instrukcí podle pravidel jazyka.

Deklarace - je definice typu objektu, například proměnné.

Kromě možnosti programování v VBA existuje s uvedením Office 2007 možnost programování v nastupující platformě Dotnet. Pro vytvoření nadstavbového modulu byl i přesto zvolen programovací jazyk VBA, především z důvodu jeho rozšířenosti a dobré znalosti mezi širokou skupinou programátorů, jeho dostupností – vývojové prostředí je součástí instalačního balíku MS Office, a také z důvodu zpětné kompatibility - vývojové prostředí v Excelu 2007 je naprosto stejné jako v dřívějších verzích Excelu [1].

6.2 Koncepce nadstavbového modulu

Nadstavbový modul pro hierarchické shlukování se jmenuje Mod_Sh_Hier je součástí souboru Shluk_Hier.xls. Tento soubor je uložen na CD, které je přílohou této diplomové práce. Přepínání mezi obrazovkou Excelu a vývojovým prostředím VBA se provádí pomocí kláves Alt-F11.



Obrázek 3 Vývojové prostředí VBA [zdroj: autor]

Nadstavbový modul Mod_Sh_Hier obsahuje procedury realizující hierarchické shlukování na základě událostí podle požadavků uživatele. Těmito událostmi jsou nejčastěji kliknutí na tlačítko v listu. Například kliknutím na tlačítko „Kvantitativní data – Transformace dat“ se spustí procedura *Transformace()*. VBA explicitně nevyžaduje deklaraci proměnných, ale jejich vyžadování je možné vynutit [7]. Protože, podle mého názoru, nedeklarování proměnných vnáší do programového kódu chyby a zmatek, je deklarování proměnných vynuceno příkazem Option Explicit (viz obrázek 3, řádek 1). Každé důležité místo nebo každý dílčí celek v programovém kódu a také každá proměnná, která má nějaký specifický význam, jsou v modulu popsány komentáři, tj. zkrácenými popisky zelené barvy, které stručně vysvětlují význam daného kódu nebo proměnné (viz obrázek 3).

Nadstavbový modul Mod_Sh_Hier je navržen tak, že zpracovává dva druhy dat – kvantitativní a dichotomická data.

V případě kvantitativních dat je postup následující:

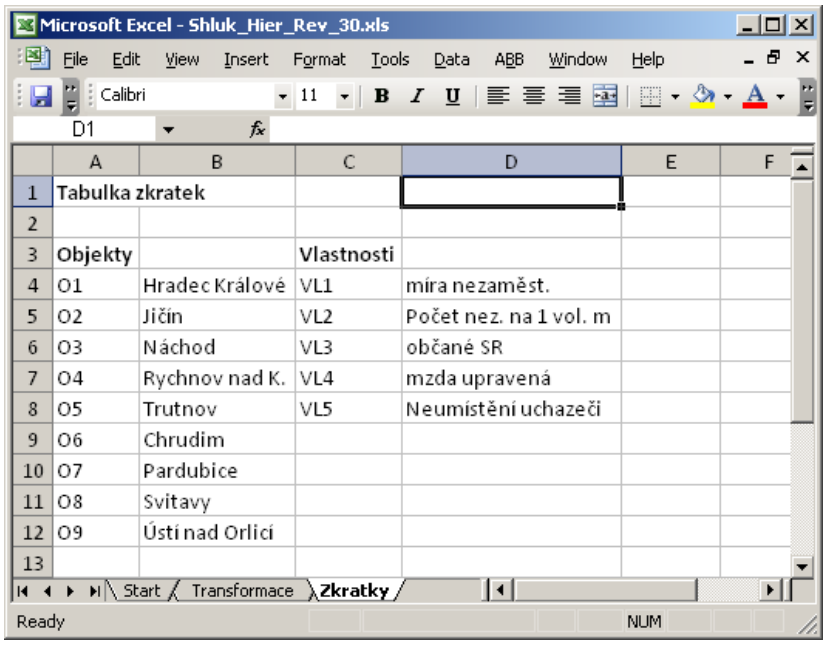
- transformace dat – vznikne list s názvem „Transformace“
- výpočet vzdáleností - vznikne list s názvem „Vzdálenosti“
- shlukování - vznikne list s názvem „Shlukování“
- vytvoření dendrogramu - vznikne list s názvem „Dendrogram“

V případě dichotomických dat je postup následující:

- vytvoření asociační tabulky – vznikne list s názvem „Asoc_Tab“
- vytvoření tabulky koeficientů asociace – vznikne list s názvem „Koef_asociace“
- výpočet vzdáleností - vznikne list s názvem „Vzdálenosti“
- shlukování - vznikne list s názvem „Shlukování“
- vytvoření dendrogramu - vznikne list s názvem „Dendrogram“

Při zpracování jak kvantitativních tak dichotomických dat je vždy na začátku prvního kroku prováděna kontrola dat. U kvantitativních dat je prováděna kontrola zda tabulka neobsahuje chybějící hodnoty a zda je obsah buněk číslo, u dichotomických dat je prováděna kontrola zda tabulka obsahuje pouze číselné hodnoty 0 a 1. V opačném případě program dál nepokračuje a uživatel je upozorněn na chybu v datech.

Protože názvy objektů a vlastností jsou často dlouhé textové řetězce, navíc různé délky, jsou v průběhu programu používány místo těchto názvů jejich zkratky. Objekty jsou značeny *O1, O2, ..., On*, kde *n* je počet objektů a vlastnosti jsou značeny *VL1, VL2, ..., VLp*, kde *p* je počet vlastností. Tabulka zkratek přiřazených jednotlivým objektům a vlastnostem je uložena na listu s názvem „Zkratky“, který je vygenerován vždy po prvním kroku zpracování dat. Viz obrázek 4.



	A	B	C	D	E	F
1	Tabulka zkratek					
2						
3	Objekty		Vlastnosti			
4	O1	Hradec Králové	VL1	míra nezaměst.		
5	O2	Jičín	VL2	Počet nez. na 1 vol. m		
6	O3	Náchod	VL3	občané SR		
7	O4	Rychnov nad K.	VL4	mzda upravená		
8	O5	Trutnov	VL5	Neumístění uchazeči		
9	O6	Chrudim				
10	O7	Pardubice				
11	O8	Svitavy				
12	O9	Ústí nad Orlicí				
13						

Obrázek 4 Tabulka zkratek [zdroj: autor]

6.3 Popis procedur a algoritmů nadstavbového modulu

V následujících podkapitolách jsou popsány procedury a algoritmy nadstavbového modulu pro hierarchické shlukování a to zvlášť pro zpracování kvantitativních dat a zvlášť pro zpracování dichotomických dat.

6.3.1 Zpracování kvantitativních dat

1. krok – transformace dat

O transformaci dat pojednává kapitola 3.2.3 této práce. V nadstavbovém modulu je tato činnost ošetřena procedurou s názvem *Transformace()*. Tato procedura je spuštěna kliknutím na tlačítko „Kvantitativní data – Transformace dat“. Po úspěšné transformaci dat je vytvořen list s názvem „Transformace“, kde jsou uložena transformovaná data. Před spuštěním procedury transformace dat musí uživatel zvolit způsob transformace pomocí roletového menu napravo od tlačítka pro spuštění transformace. Nadstavbový modul nabízí tyto možnosti transformace dat:

- kontrola dat,
- standardizace,
- normalizace,
- standardizace + normalizace,
- normalizace + standardizace.

V případě volby „kontroly dat“ jsou data pouze zkontrolována a do listu „Transformace“ nakopírována bez úpravy. Tato volba je zde proto, aby uživatel měl možnost provést hierarchické shlukování i takových dat, která má již předzpracovaná jiným způsobem, například jiným SW.

Popis procedury transformace dat:

Název procedury:

Transformace()

Výpočty v proceduře

Procedura volá proceduru *PocObj(N)*, která vrací počet objektů vstup. tabulky – proměnná N.

Procedura volá proceduru *PocVla(P)*, jenž vrací počet vlastností vstup. tabulky – proměnná P.

Procedura volá proceduru *Kontrola(N, P, Chk, PoleTran())*, které jsou předány proměnné *N* a *P*, a tato procedura vrací informaci o výsledku kontroly dat – proměnnou *Chk* a pole *PoleTran()* s načtenými a zkontrolovanými hodnotami ze vstupní tabulky.

Procedury *PocObj()*, *PocVla()* a *Kontrola()* jsou podrobněji popsány v příloze 1 uložené na CD.

Procedura dále podle volby uživatele volá proceduru *Standardizace(N, P, PoleTran())*, nebo *Normalizace (N, P, PoleTran())*, a tyto procedury vrací upravené pole *PoleTran()* se standardizovanými, nebo normalizovanými hodnotami.

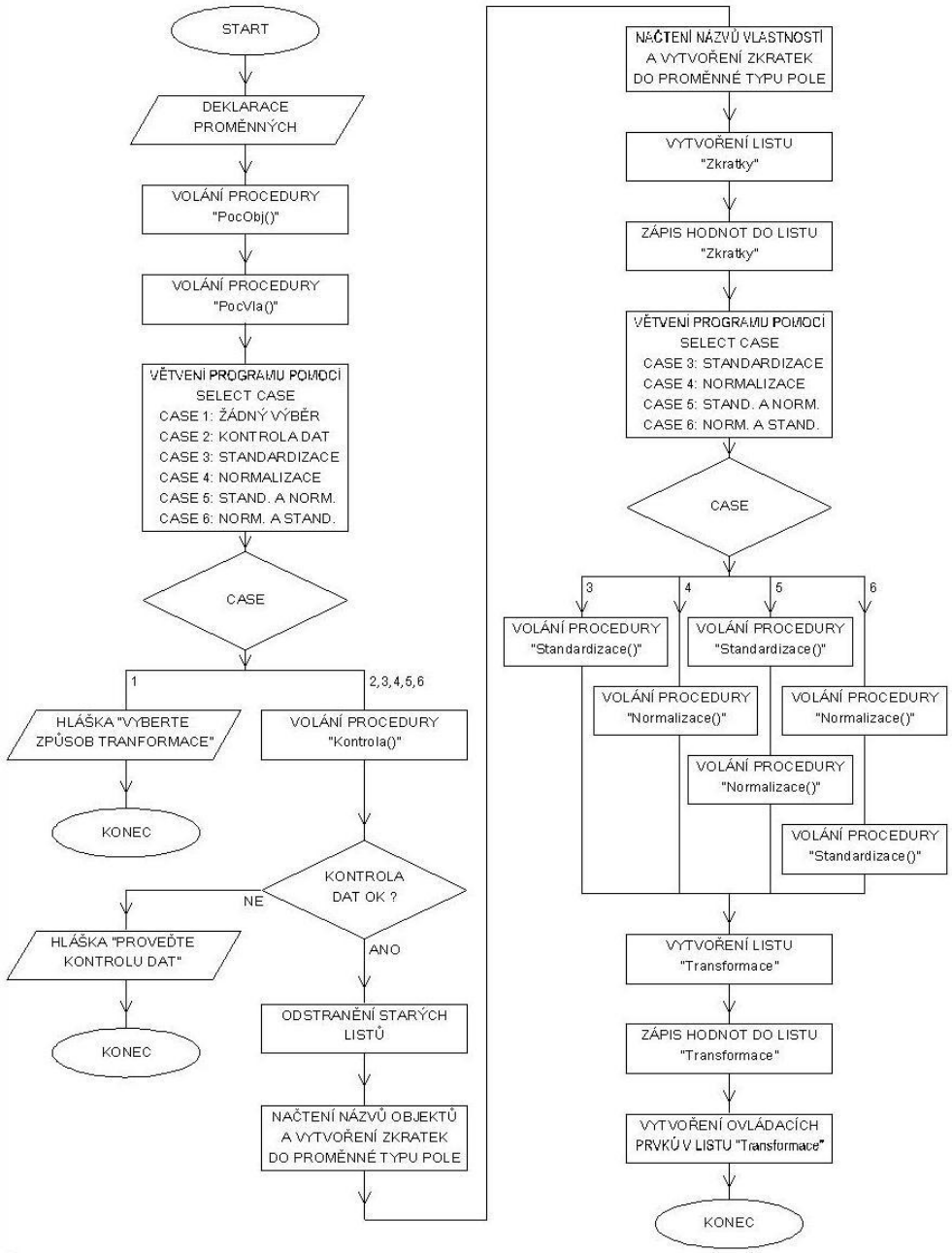
Vývojové diagramy a popis procedur *Standardizace()* s *Normalizace ()* jsou na CD v přílohách 2 a 3.

Výstup procedury:

Procedura vytvoří list s názvem „Zkratky“ s tabulkou zkratek objektů a vlastností.

Procedura vytvoří list s názvem „Transformace“ s tabulkou transformovaných hodnot.

Vývojový diagram procedury:



Obrázek 5 Vývojový diagram transformace dat [zdroj: autor]

2. krok – výpočet vzdáleností

O výpočtu vzdáleností mezi objekty, jejichž vlastnosti jsou determinovány kvantitativními daty, pojednává kapitola 4.1 této práce, s názvem Míry vzdálenosti pro kvantitativní data. V nadstavbovém modulu je tato činnost ošetřena procedurou s názvem *Vzdalenost()*. Tato procedura je spuštěna kliknutím na tlačítko „Vzdálenost“ které je umístěno na listu „Transformace“. Po výpočtu vzdálenosti je vytvořen list s názvem „Vzdálenosti“ s tabulkou vzdáleností mezi objekty. Před spuštěním procedury výpočtu vzdáleností musí uživatel zvolit způsob výpočtu vzdáleností mezi objekty pomocí roletového menu napravo od tlačítka „Vzdálenost“.

Nadstavbový modul nabízí tyto možnosti výpočtu vzdáleností:

- Euklidovská vzdálenost,
- Čtvercová Euklidovská vzdálenost,
- Hammingova vzdálenost,
- Čebyševova vzdálenost.

Popis procedury výpočtu vzdáleností:

Název procedury:

Vzdalenost ()

Vstupní parametry procedury:

N – počet objektů

P – počet vlastností

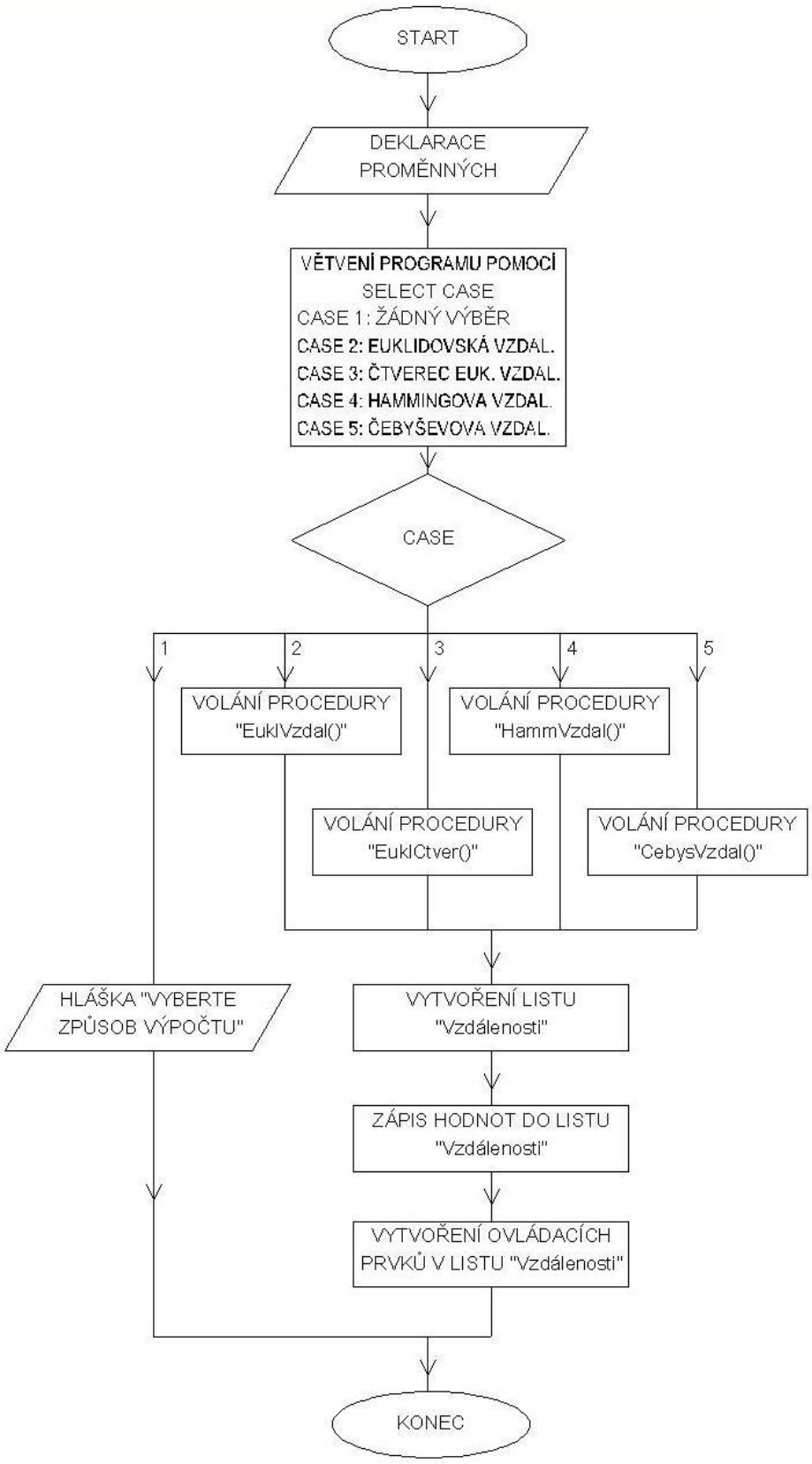
Výpočty v proceduře

Procedura podle volby uživatele volá proceduru *EuklVzdal(N, P, PoleVzdal()), EuklCtver(N, P, PoleVzdal()), HammVzdal(N, P, PoleVzdal())* nebo *CebysVzdal(N, P, PoleVzdal())*, kterým jsou předány proměnné *N* a *P* a pole *PoleVzdal()*, a tyto procedury vrací pole *PoleVzdal()* s hodnotami vzdáleností mezi objekty. Volané procedury jsou podrobněji popsány na CD v příloze 4.

Výstup procedury:

Procedura vytvoří list s názvem „Vzdálenosti“ s tabulkou vzdáleností mezi objekty.

Vývojový diagram procedury:

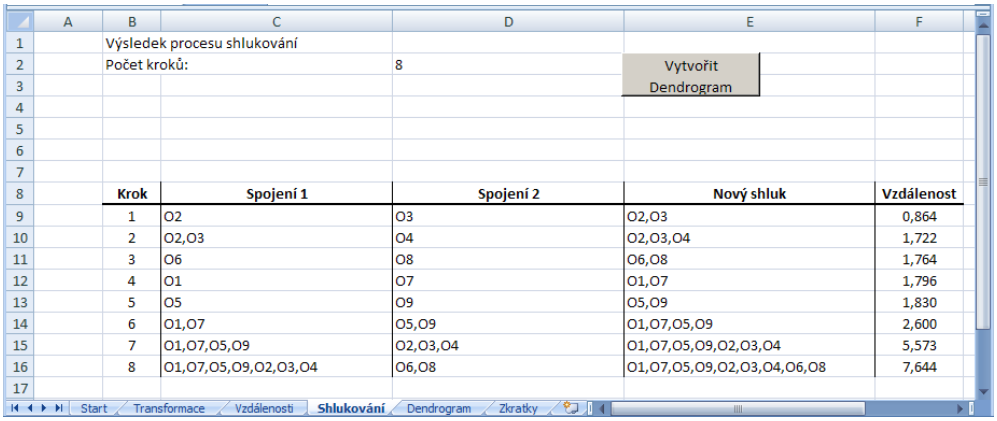


Obrázek 6 Vývojový diagram výpočtu vzdáleností [zdroj: autor]

3. krok – shlukování

Nadstavbový modul provádí hierarchické shlukování objektů aglomerativní metodou. Aglomerativní přístup je charakteristický tím, že se vychází od jednotlivých objektů, tj. na začátku algoritmu jsou jednotlivé objekty brány jako shluk s jedním objektem. Jejich postupným slučováním se buduje hierarchický systém, až se dospěje ke konečnému sloučení všech objektů do jednoho shluku. Toto slučování se provádí tak, že se v jednotlivých iteracích spojují dva shluky s nejmenší vzájemnou vzdáleností. Vzájemnou vzdálenost mezi shluky lze počítat různými způsoby a této problematice se věnuje kapitola 5.1 Měření podobnosti shluků. V nadstavbovém modulu je tato činnost shlukování ošetřena procedurou s názvem *Shluky()*.

Proces shlukování se odstartuje tlačítkem „Shlukování“ v listu „Vzdálenosti“ a výstupem z tohoto procesu je tabulka, ze které je zřejmé jaké shluky byly spojeny v jednotlivých krocích a při jaké vzdálenosti ke sloučení došlo. Viz obrázek 7.



Výsledek procesu shlukování

Počet kroků: 8

Vytvořit Dendrogram

Krok	Spojení 1	Spojení 2	Nový shluk	Vzdálenost
1	O2	O3	O2,O3	0,864
2	O2,O3	O4	O2,O3,O4	1,722
3	O6	O8	O6,O8	1,764
4	O1	O7	O1,O7	1,796
5	O5	O9	O5,O9	1,830
6	O1,O7	O5,O9	O1,O7,O5,O9	2,600
7	O1,O7,O5,O9	O2,O3,O4	O1,O7,O5,O9,O2,O3,O4	5,573
8	O1,O7,O5,O9,O2,O3,O4	O6,O8	O1,O7,O5,O9,O2,O3,O4,O6,O8	7,644

Start | Transformace | Vzdálenosti | Shlukování | Dendrogram | Zkratky

Obrázek 7 Tabulka po shlukování [zdroj: autor]

Před spuštěním procedury shlukování musí uživatel zvolit metodu, která bude použita pro výpočet vzdáleností mezi shluky. Tato volba se provádí pomocí roletového menu napravo od tlačítka „Shlukování“. Nadstavbový modul nabízí tyto možnosti výpočtu vzdáleností mezi shluky:

- metoda nejbližšího souseda,
- metoda nejvzdálenějšího souseda,
- metoda průměrné vzdálenosti,
- centroidní metoda,
- mediánová metoda,
- Wardova-Wishartova metoda.

Postup shlukování aglomerativní metodou se skládá z několika úkonů, které jsou sami o sobě natolik složité, že kdyby se prováděli všechny v jedné proceduře, tak by tato procedura byla velmi nepřehledná. Proto samotná procedura *Shluky()* nevykonává žádné výpočty, ale „pouze“ zajišťuje volání dalších procedur

a výpis hodnot do listů. Procedury, které jsou procedurou *Shluky()* volány jsou *PocObjShl(), MinHodnota(), LegTab()* a *ShVypocet()*. Postup shlukování lze zjednodušeně popsat takto:

- Určení shluků, které budou v daném kroku sloučeny na základě minimální hodnoty. Toto zajišťuje procedura *MinHodnota()*.
- Sloučení shluků a přepočítání tabulky vzdáleností. Toto zajišťuje procedura *ShVypocet()*.
- Pro výpočet vzdálenosti mezi nově vzniklým shlukem a ostatními shluky je použita Lance-Williams formule, ve které se podle zvolené metody přepočítávají koeficienty $\alpha_A, \alpha_B, \beta, \gamma$. Pro výpočet těchto koeficientů je potřeba mít informaci o počtu objektů v jednotlivých shlucích v daném kroku. Tuto informaci zajišťuje procedura *PocObjShl()*.
- Přepočítané tabulky vzdáleností v jednotlivých krocích se vkládají do listu „Vzdálenosti“ pod sebe. Aby byli dobře čitelné, je potřeba vytvořit legendu pro jednotlivé řádky a sloupce. Tuto legendu zajišťuje procedura *LegTab()*.

Popis procedury shlukování:

Název procedury:

Shluky ()

Vstupní parametry procedury:

N – počet objektů

Tabulka vzdáleností mezi objekty

Výpočty v proceduře

Procedura nevykonává žádné výpočty, ale zajišťuje volání dalších procedur, které potřebné výpočty provádějí. Volané procedury jsou *PocObjShl(), MinHodnota(), LegTab()* a *ShVypocet()*. Vývojové diagramy a popisy těchto procedur jsou na CD v přílohách 5 - 8.

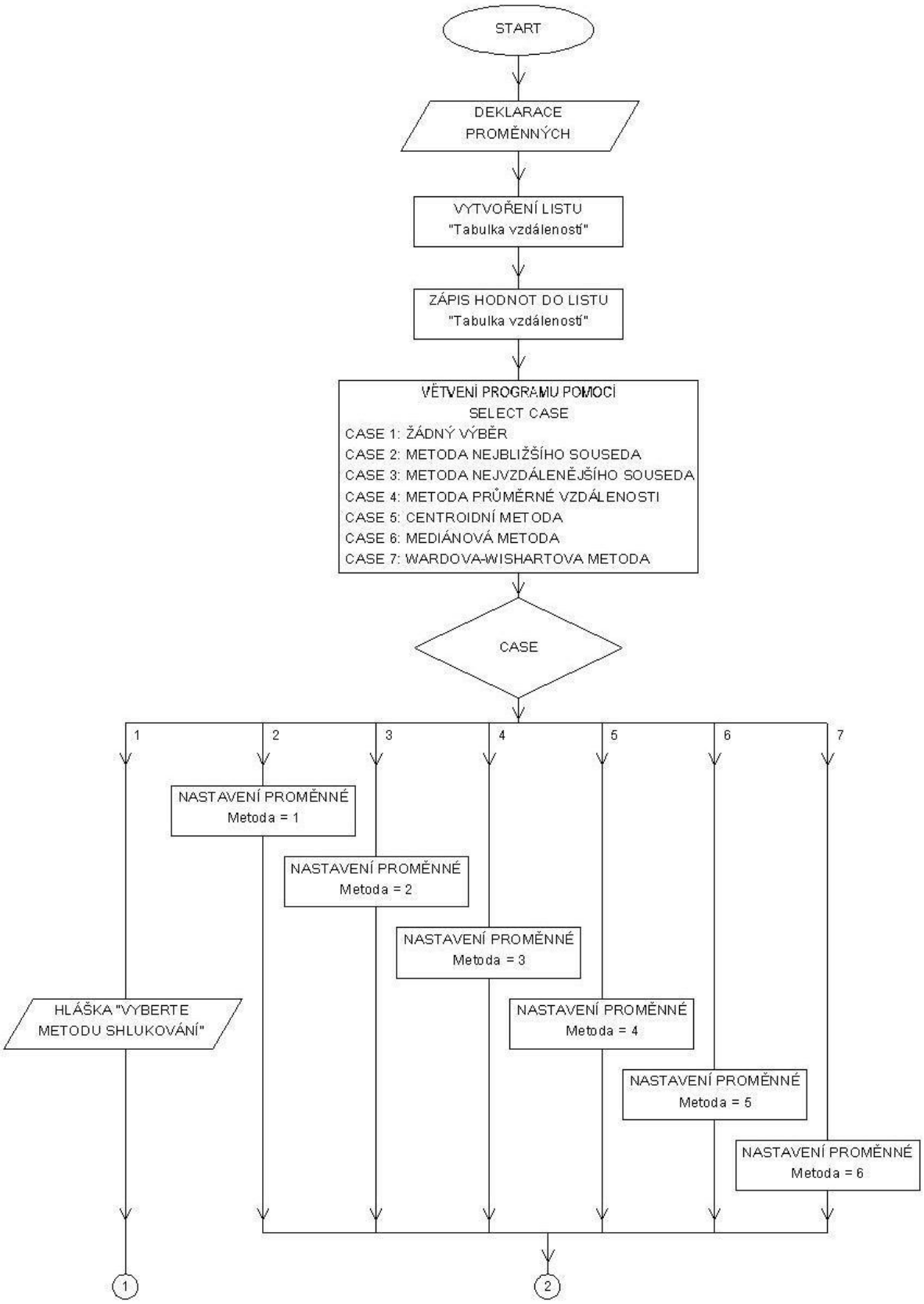
Výstup procedury:

Procedura vytvoří list s názvem „Shlukování“ s výstupní tabulkou po shlukování.

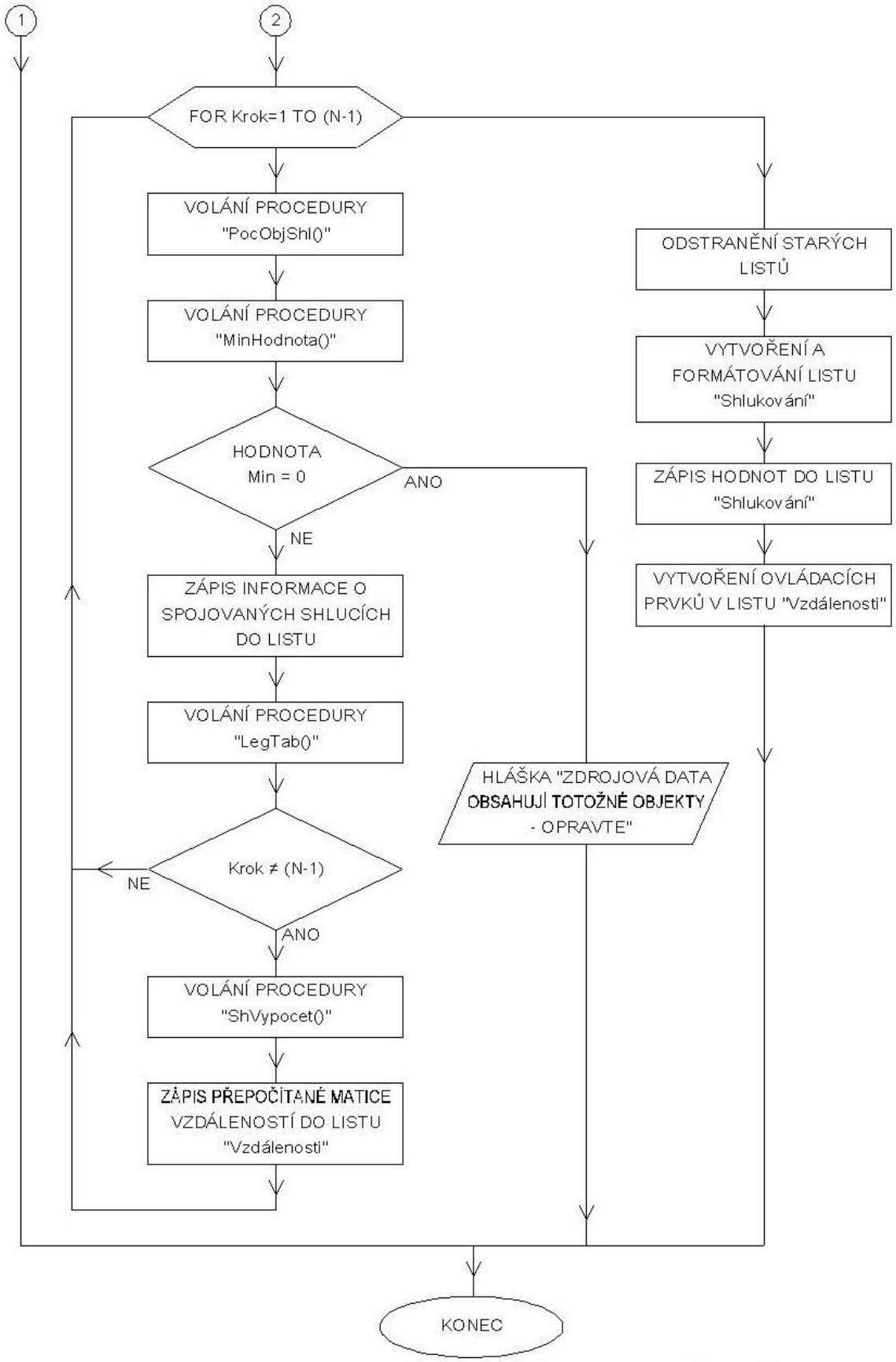
Procedura vytvoří list s názvem „Tabulka vzdáleností“ s výstupní tabulkou po shlukování.

Vývojový diagram procedury:

Viz obrázek 8 a obrázek 9.



Obrázek 8 Vývojový diagram procedury shlukování – část 1 [zdroj: autor]



Obrázek 9 Vývojový diagram procedury shlukování – část 2 [zdroj: autor]

4. krok – vytvoření dendrogramu

Vytvoření dendrogramu se odstartuje tlačítkem „Vytvořit Dendrogram“ v listu „Shlukování“. Po kliknutí na toto tlačítko je vytvořen list s názvem „Dendrogram“ který obsahuje vytvořený dendrogram.

Výstupem z procesu shlukování je tabulka, která dává přehled o tom, jaké shluky resp. objekty byly v každém kroku spojeny, a také informaci při jaké vzdálenosti spojení proběhlo. Jak bylo popsáno v kapitole 5.2, tuto tabulku nelze použít pro vykreslení dendrogramu a musí být upravena. Úpravu tabulky pro vykreslení dendrogramu realizuje procedura s názvem *UpTabDend ()*. Tato procedura je popsána v příloze 9, která je uložena na CD.

Vykreslení dendrogramu

Vykreslení dendrogramu realizuje procedura Dendrogram() a je založeno na vkládání objektů typu Shape do listu „Dendrogram“. Příkaz

Worksheets("Dendrogram").Shapes.AddLine()

vloží linku a příkaz

Worksheets("Dendrogram").Shapes.AddLabel()

vloží text do listu Dendrogram. AddLine a AddLabel jsou metody objektu Shapes a v závorkách se uvádějí parametry - souřadnice vložení objektu a text, jenž má být vložen. Podrobný příklad pro vysvětlení principu kreslení dendrogramu je uložen na CD v příloze 10.

Popis procedury vykreslení dendrogramu:

Název procedury:

Dendrogram ()

Vstupní parametry procedury:

M – počet kroků shlukování

Procedura volá proceduru *UpTabDend()*, která vrací upravenou výstupní tabulku z procesu shlukování.

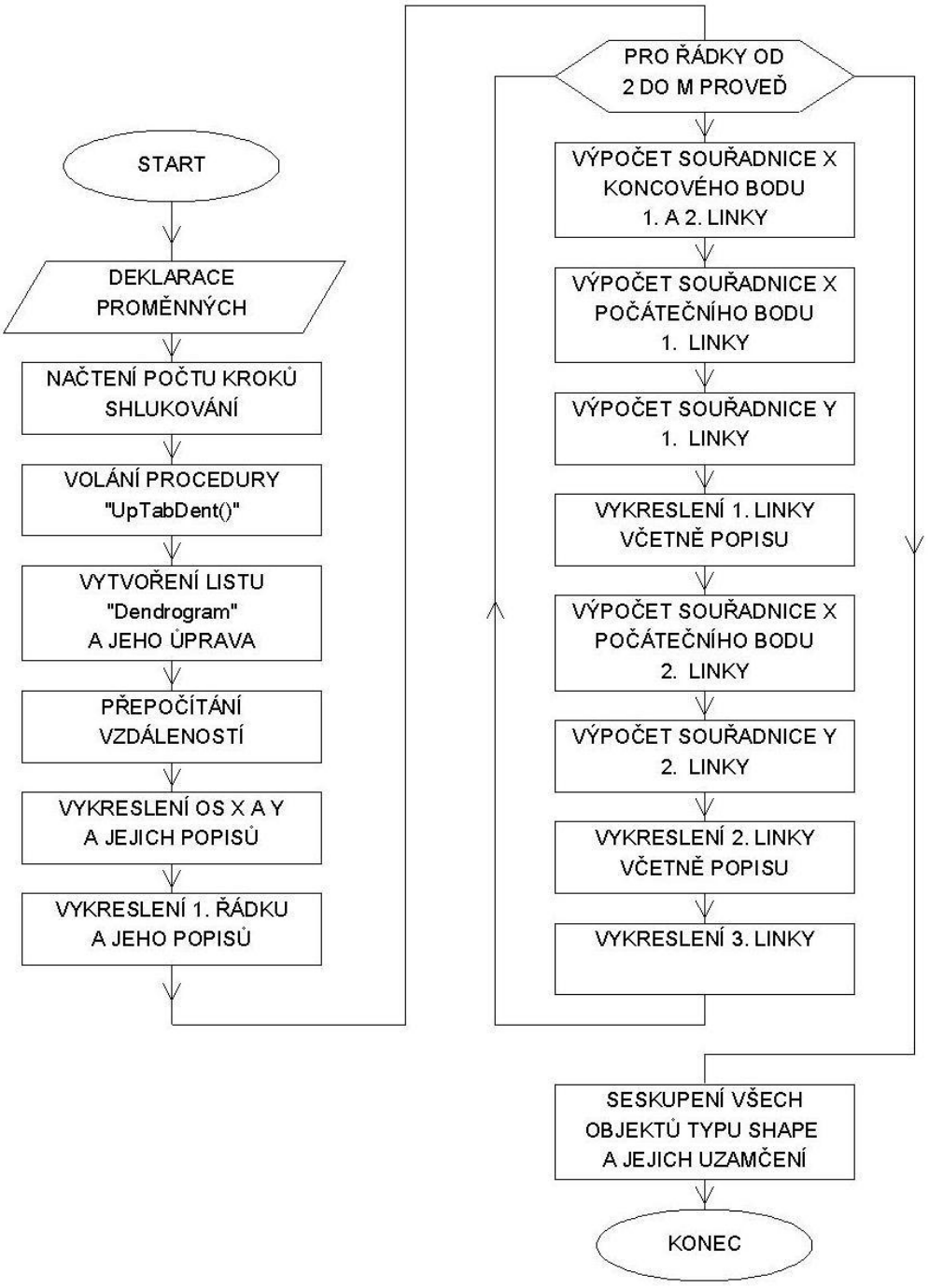
Výpočty v proceduře

Výpočty počátečních a koncových souřadnic objektů Shape.

Výstup procedury:

Vykreslený dendrogram na list s názvem „Dendrogram“.

Vývojový diagram procedury:



Obrázek 10 Vývojový diagram procedury Dendrogram() [zdroj: autor]

6.3.2 Zpracování dichotomických dat

1. krok – vytvoření asociační tabulky

V případě, že jsou objekty charakterizovány dichotomickými znaky, se ve shlukové analýze míra podobnosti nazývá koeficientem asociace. Při určování prvků v matici podobností objektů O_i a O_j bude pozorována shoda či neshoda výsledků u p proměnných. Asociace dvou objektů je vyjádřena asociační tabulkou, viz kapitola 4.2.

V nadstavbovém modulu je pro zahájení práce s dichotomickými daty určena procedura s názvem *StartDich()*. Tato procedura je spuštěna kliknutím na tlačítko „Dichotomická data – Asociační tabulka“. Po spuštění této procedury je provedena kontrola, zda jsou data skutečně dichotomické, tj. zda matice čísel obsahuje pouze 0 a 1. V případě, že jsou data pořádku, je vytvořen list s názvem „Asoc_Tab“, kde je uložena asociační tabulka.

Popis procedury pro zahájení práce s dichotomickými daty:

Název procedury:

StartDich()

Vstupní parametry procedury:

Vstupní tabulka s dichotomickými daty.

Procedura volá proceduru *PocObj(N)*, která vrací počet objektů vstup. tabulky – proměnná *N*.

Procedura volá proceduru *PocVla(P)*, jenž vrací počet vlastností vstup. tabulky – proměnná *P*.

Procedury *PocObj(N)* a *PocVla(P)* jsou již popsány v kapitole 6.3.1.

Výpočty v proceduře

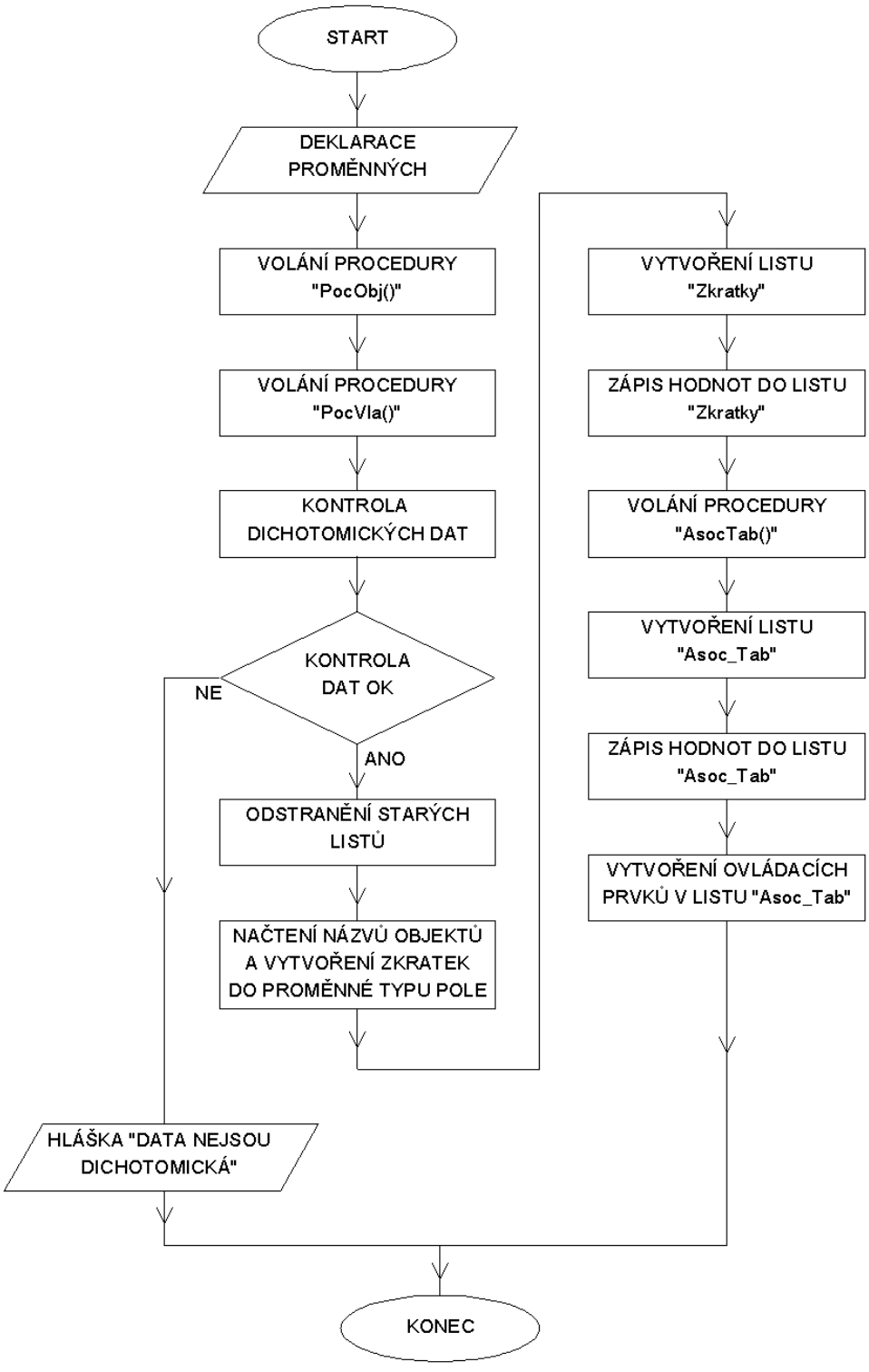
Procedura volá dále proceduru *AsocTab(N, P, AsocMat())*, která provádí vlastní výpočet asociační tabulky a vrací upravené třírozměrné pole *AsocMat()* s touto asociační tabulkou. Vývojový diagram a popis procedury *AsocTab()* je na CD v příloze 12.

Výstup procedury:

Procedura vytvoří list s názvem „Zkratky“ s tabulkou zkratek objektů.

Procedura vytvoří list s názvem „Asoc_Tab“ s asociační tabulkou.

Vývojový diagram procedury:



Obrázek 11 Vývojový diagram procedury StartDich() [zdroj: autor]

2. krok – výpočet vzdáleností

O výpočtu vzdáleností mezi objekty, jejichž vlastnosti jsou charakterizovány dichotomickými daty, pojednává kapitola 4.2 Míry vzdálenosti pro dichotomická data. Výpočet vzdáleností je možné provést více způsoby. Tento nadstavbový modul realizuje výpočet tak, že provádí výpočet koeficientů asociace z asociační tabulky. Protože koeficienty asociace vyjadřují míru podobnosti, je pro proces shlukování potřeba míru podobnosti převést na míru nepodobnosti (funkce založené na vzdálenosti objektů jsou prakticky míry nepodobnosti).

V nadstavbovém modulu je výpočet koeficientů asociace ošetřen procedurou s názvem *KoefAsoc()*. Tato procedura je spuštěna kliknutím na tlačítko „Výpočet koeficientů asociace“ které je umístěno na listu „Asoc_tab“. Po výpočtu koeficientů je vytvořen list s názvem „Koef_asociace“ s tabulkou koeficientů asociace. Před spuštěním procedury výpočtu koeficientů asociace musí uživatel zvolit způsob jejich výpočtu pomocí roletového menu napravo od tlačítka „Výpočet koeficientů asociace“.

Nadstavbový modul nabízí tyto možnosti výpočtu koeficientů asociace:

- Sokalův-Michenerův koeficient asociace,
- Rogersův-Tanimonoův koeficient asociace,
- Nepojmenovaný 1 koeficient asociace.

Protože tyto koeficienty nabývají hodnot v intervalu <0,1>, je možný jejich převod na míru nepodobnosti takto:

míra nepodobnosti = 1 - míra podobnosti.

Tento převod je v nadstavbovém modulu ošetřen procedurou s názvem *VzdalDich()*. Tato procedura je spuštěna kliknutím na tlačítko „Vzdálenost“, které je umístěno na listu „Koef_asociace“. Po výpočtu je vytvořen list s názvem „Vzdálenosti“ s tabulkou vzdáleností (měr nepodobností) mezi objekty.

Popis procedury výpočtu koeficientů asociace:

Název procedury:

KoefAsoc()

Vstupní parametry procedury:

Vstupní asociační tabulka

N – počet objektů

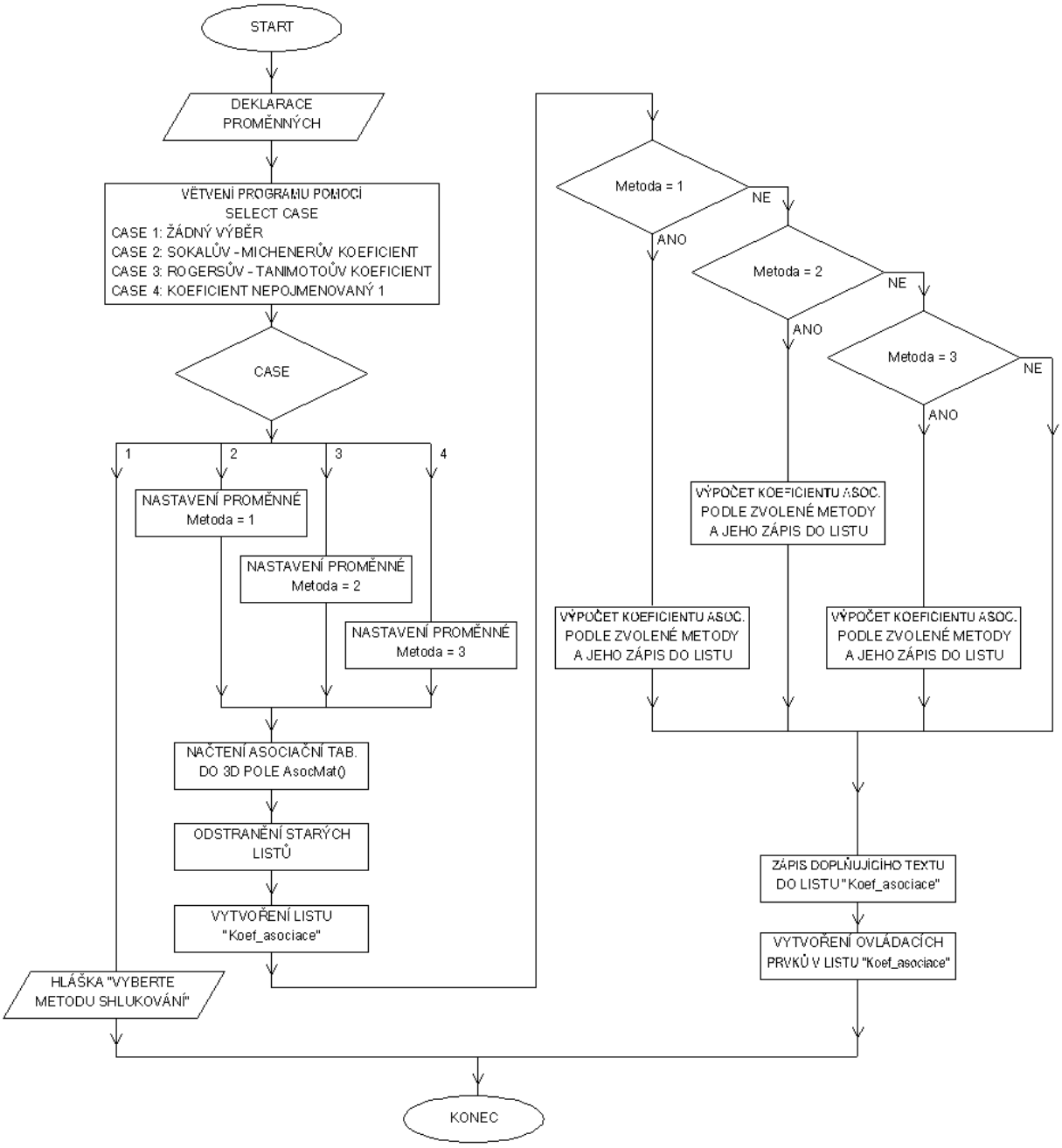
Výpočty v proceduře

Podle zvolené metody provádí procedura výpočty koeficientů asociace.

Výstup procedury:

Procedura vytvoří list s názvem „Koef_asociace“ s tabulkou koeficientů asociace.

Vývojový diagram procedury:



Obrázek 12 Vývojový diagram procedury KoefAsoc() [zdroj: autor]

Popis procedury *VzdalDich()* pro převodu koeficientů asociace na míru nepodobnosti je uveden na CD v příloze 13.

3. krok – shlukování / 4. krok – vytvoření dendrogramu

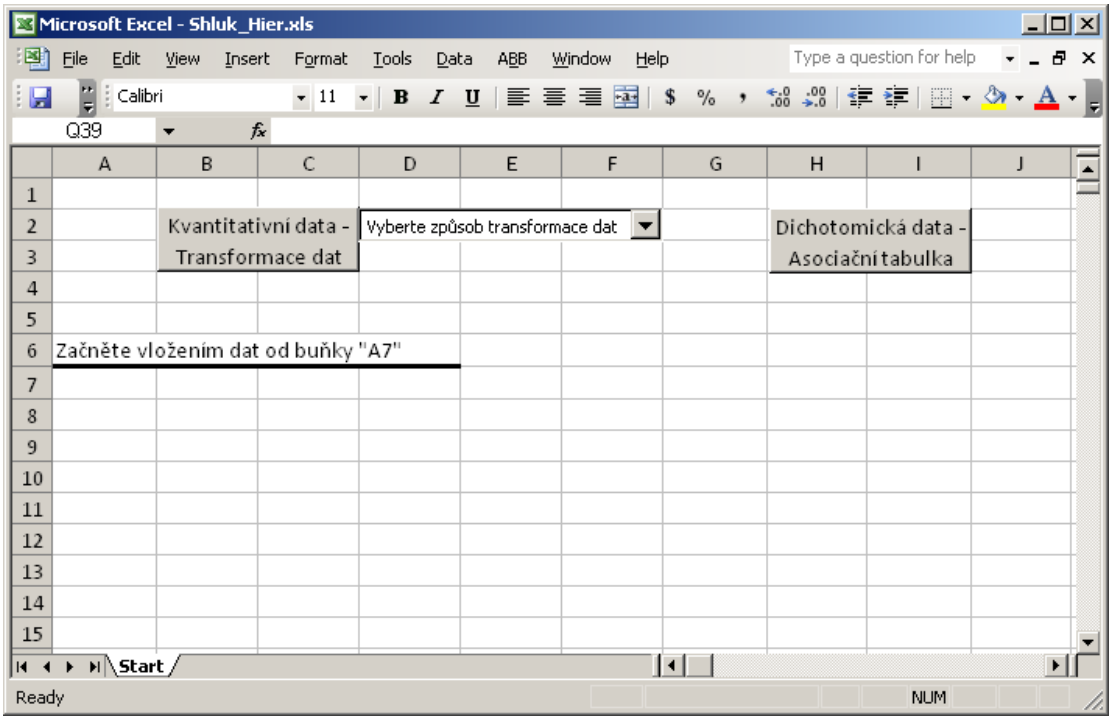
Poté co je vytvořen list vzdáleností mezi objekty zadanými dichotomickými daty se pokračuje ve shlukové analýze stejným způsobem jako v případě práce s kvantitativními daty. Po stisknutí tlačítka „Shlukování“ na listu „Vzdálenost“ dojde k provedení hierarchického shlukování objektů podle zvolené metody. Dále je možné vygenerovat dendrogram kliknutím na tlačítko „Vytvořit Dendrogram“ na listu „Shlukování“.

Shlukování a vytvoření dendrogramu v případě práce s dichotomickými daty je prováděno stejnými procedurami jako v případě práce s kvantitativními daty.

7 METODICKÉ POKYNY K POUŽITÍ MODULU

V této kapitole je popsaný způsob práce s nadstavbovým modulem pro provádění shlukové analýzy hierarchickými (aglomerativními) shlukovacími metodami. Tento modul je součástí souboru Shluk_Hier.xls, který je součástí této diplomové práce a je k dispozici na přiloženém CD. Nadstavbový modul je určený jako učební pomůcka pro předmět Zpracování dat metodami shlukové analýzy a jeho funkční možnosti jsou přizpůsobeny rozsahu výuky tohoto předmětu. Soubor Shluk_Hier.xls a nadstavbový modul je možné používat v MS Office Excel 2003, MS Office Excel 2007 a MS Office Excel 2010.

Po spuštění souboru Shluk_Hier.xls se otevře aplikace Excel s jedním listem s názvem Start, viz obrázek 13. Dále je možné pokračovat zpracováním buď kvantitativních dat, nebo dichotomických dat.



Obrázek 13 Úvodní list souboru Shluk_Hier.xls [zdroj: autor]

7.1 Práce s kvantitativními daty

Návod jak postupovat při práci s kvantitativními daty bude ukázán pomocí následujícího příkladu.

Příklad 1:

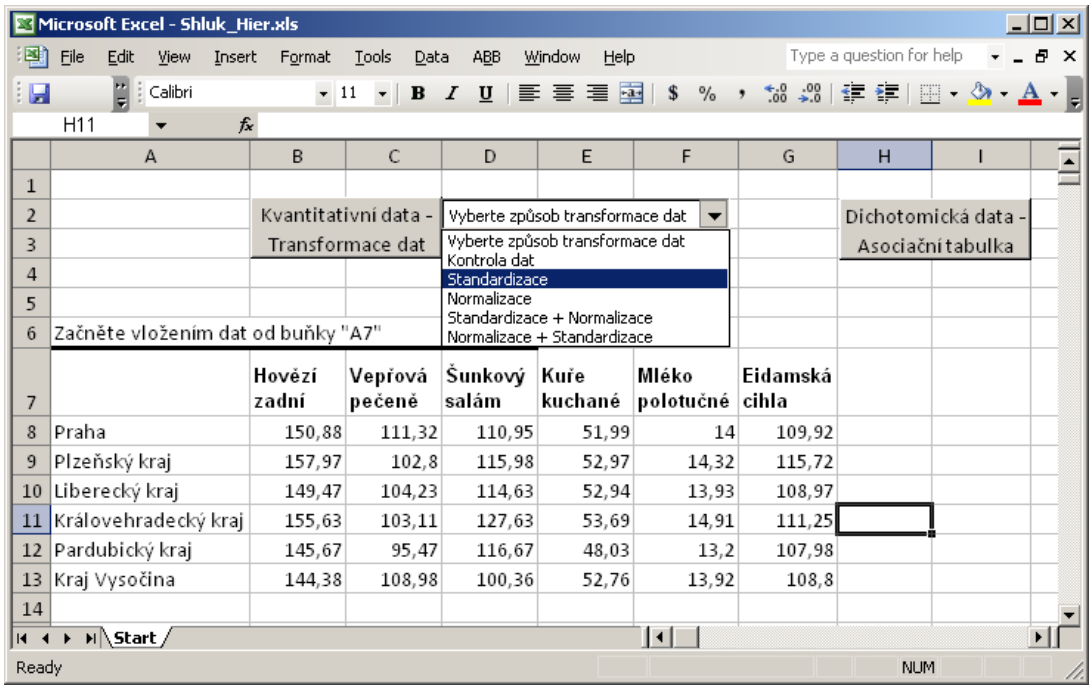
Tabulka 5 vyjadřuje ceny vybraných potravin v obchodech v jednotlivých krajích. Ve kterých krajích jsou ceny potravin nejvíce podobné? Použijte shlukovou analýzu, hierarchickou metodu, aglomerativní přístup. Volte Euklidovskou vzdálenost a metodu nejbližšího souseda.

Tabulka 5 Příklad pro kvantitativní data [zdroj: autor]

	Hovězí zadní	Vepřová pečeně	Šunkový salám	Kuře kuchané	Mléko polotučné	Eidamská cihla
Praha	150,88	111,32	110,95	51,99	14	109,92
Plzeňský kraj	157,97	102,8	115,98	52,97	14,32	115,72
Liberecký kraj	149,47	104,23	114,63	52,94	13,93	108,97
Královehradecký kraj	155,63	103,11	127,63	53,69	14,91	111,25
Pardubický kraj	145,67	95,47	116,67	48,03	13,2	107,98
Kraj Vysočina	144,38	108,98	100,36	52,76	13,92	108,8

Řešení:

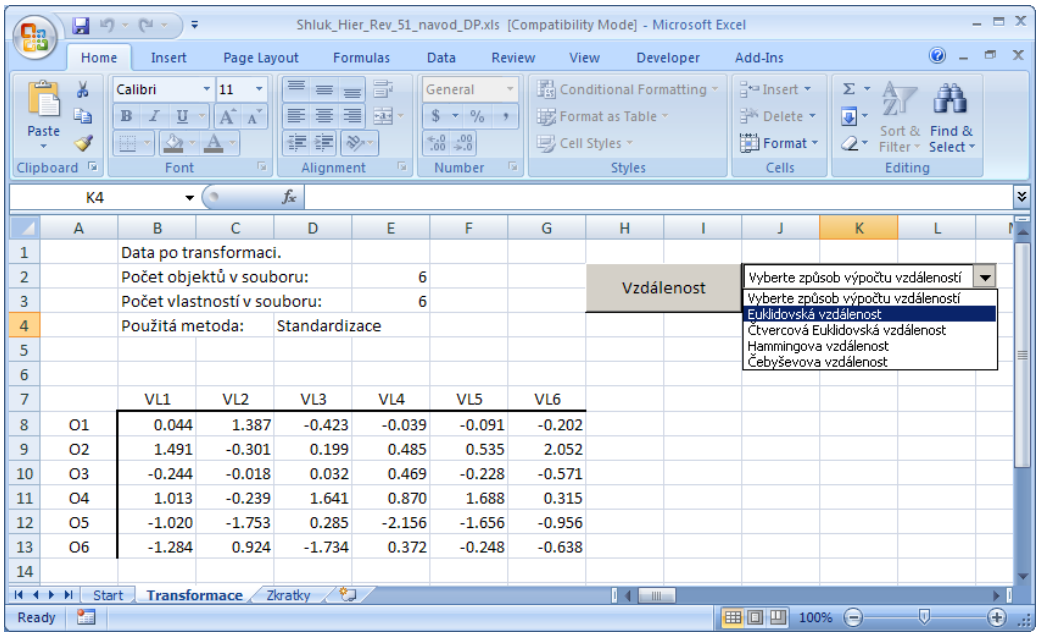
Tabulku je potřeba zkopírovat do listu „Start“ tak, že levý horní roh tabulky bude na buňce A7, sloupec A obsahuje názvy objektů shlukování a řádek 7 obsahuje názvy vlastností těchto objektů. Data začínají na buňce B8.



Obrázek 14 Příklad 1 – zkopírování dat do souboru Shluk_Hier.xls [zdroj: autor]

Protože se jedná o kvantitativní data, prvním krokem při jejich zpracování je transformace dat, viz kapitola 3.2.3. Volba způsobu transformace se provede pomocí roletového menu vedle tlačítka Kvantitativní data – transformace dat. Nadstavbový modul umožňuje provést standardizaci, normalizaci, standardizaci a následně normalizaci dat nebo obráceně normalizaci a následně standardizaci dat. V případě, že není požadována žádná transformace dat, provede se volba Kontrola dat, viz obrázek 14.

V příkladu je zvolena standardizace dat, aby rozdílná cena jednotlivých produktů neovlivňovala jejich vliv na rozdíl cen v krajích. Kliknutím na tlačítko Kvantitativní data – transformace dat dojde k provedení transformace dat. Vzniknou nové listy s názvem „Transformace“ a „Zkratky“. Na listu Transformace jsou transformovaná data, na listu Zkratky je legenda k vlastnostem VL1 … VL6 a k objektům O1 … O6. Viz obrázek 15.



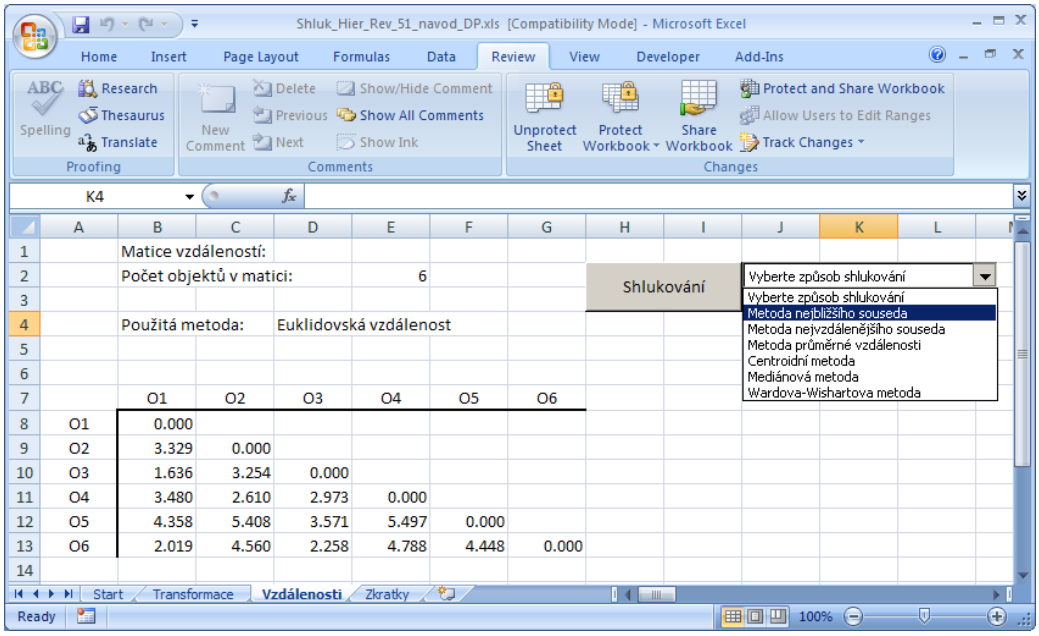
		VL1	VL2	VL3	VL4	VL5	VL6
8	O1	0.044	1.387	-0.423	-0.039	-0.091	-0.202
9	O2	1.491	-0.301	0.199	0.485	0.535	2.052
10	O3	-0.244	-0.018	0.032	0.469	-0.228	-0.571
11	O4	1.013	-0.239	1.641	0.870	1.688	0.315
12	O5	-1.020	-1.753	0.285	-2.156	-1.656	-0.956
13	O6	-1.284	0.924	-1.734	0.372	-0.248	-0.638

Obrázek 15 Příklad 1 – výsledek transformace dat [zdroj: autor]

Dalším krokem je výpočet vzdálenosti mezi objekty. O výpočtu vzdáleností mezi objekty pojednává kapitola 4.1 této práce. Volba způsobu výpočtu vzdálenosti se provede pomocí roletového menu vedle tlačítka Vzdálenost. Nadstavbový modul umožňuje výpočet euklidovské vzdálenosti, čtvercové euklidovské vzdálenosti, hammingovy vzdálenosti a čebyševovy vzdálenosti.

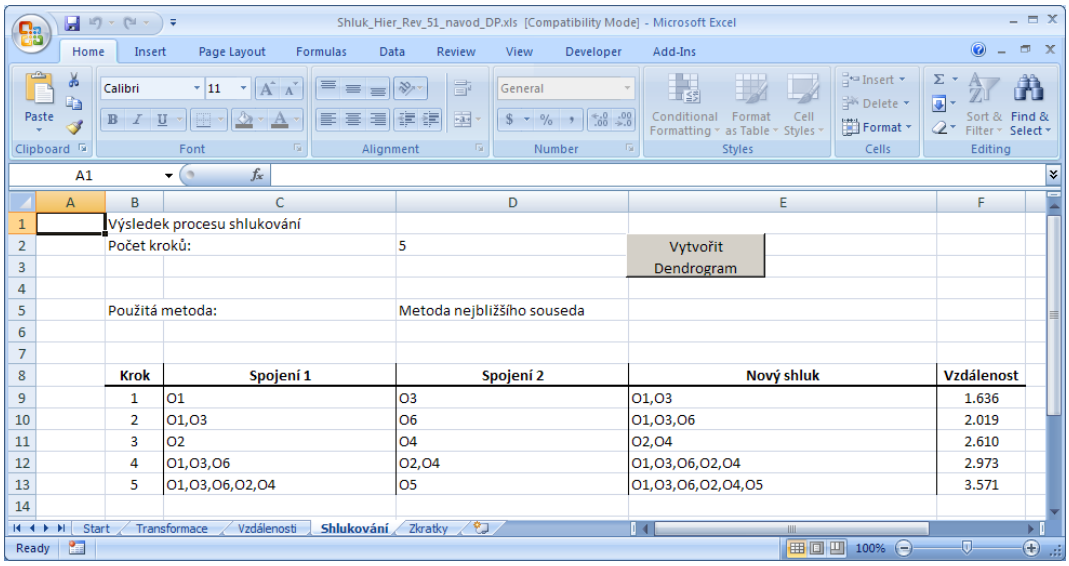
V příkladu je zvolena Euklidovská vzdálenost. Kliknutím na tlačítko Vzdálenost dojde k provedení výpočtu vzdálenosti. Vznikne nový list s názvem „Vzdálenosti“. Na tomto listu je tabulka vzdáleností mezi objekty. Viz obrázek 16.

Po výpočtu vzdálenosti následuje proces shlukování. Nadstavbový modul provádí shlukování hierarchickou metodou, aglomerativním přístupem. Proto je nyní potřeba zadat způsob výpočtu vzdálenosti mezi shluky. O výpočtu vzdáleností mezi shluky pojednává kapitola 5.1 Měření podobnosti shluků. Volba způsobu výpočtu vzdálenosti se provede pomocí roletového menu vedle tlačítka Shlukování. Nadstavbový modul umožňuje výpočet vzdálenosti metodou nejbližšího souseda, nejvzdálenějšího souseda, průměrné vzdálenosti, centroidní metodou, mediánovou metodou a Ward-Wishartovou metodou.



Obrázek 16 Příklad 1 – list se vzdálenostmi mezi objekty [zdroj: autor]

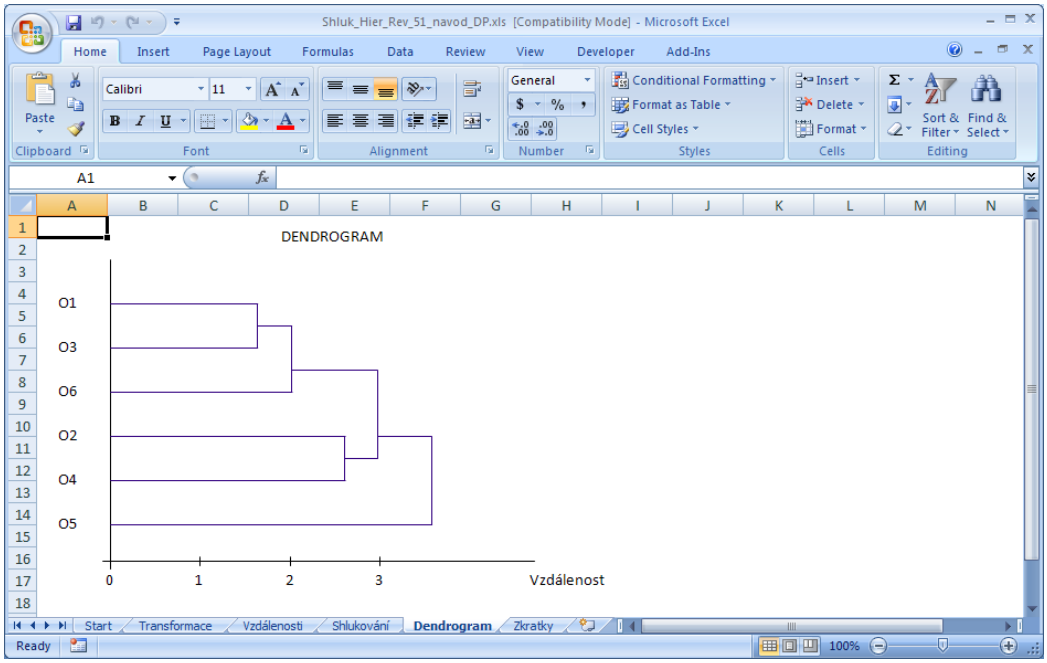
V příkladu je zvolena metoda nejbližšího souseda. Kliknutím na tlačítko Shlukování dojde ke spuštění procesu shlukování. Vznikne nový list s názvem „Shlukování“. Na tomto listu je tabulka, která vystihuje, jaké shluky byly spojeny v jednotlivých krocích a při jaké vzdálenosti ke spojení došlo. Viz obrázek 17.



Krok	Spojení 1	Spojení 2	Nový shluk	Vzdálenost
1	O1	O3	O1,O3	1.636
2	O1,O3	O6	O1,O3,O6	2.019
3	O2	O4	O2,O4	2.610
4	O1,O3,O6	O2,O4	O1,O3,O6,O2,O4	2.973
5	O1,O3,O6,O2,O4	O5	O1,O3,O6,O2,O4,O5	3.571

Obrázek 17 Příklad 1 – list s výsledkem procesu shlukování [zdroj: autor]

Pro grafickou reprezentaci výsledku shlukování se používá stromový graf, který se nazývá dendrogram. Ten je možné vygenerovat kliknutím na tlačítko Vytvořit Dendrogram. Vznikne nový list s názvem „Dendrogram“. Dendrogram pro tento příklad zobrazuje obrázek 18.



Obrázek 18 Příklad 1 – vytvoření dendrogramu [zdroj: autor]

7.2 Práce s dichotomickými daty

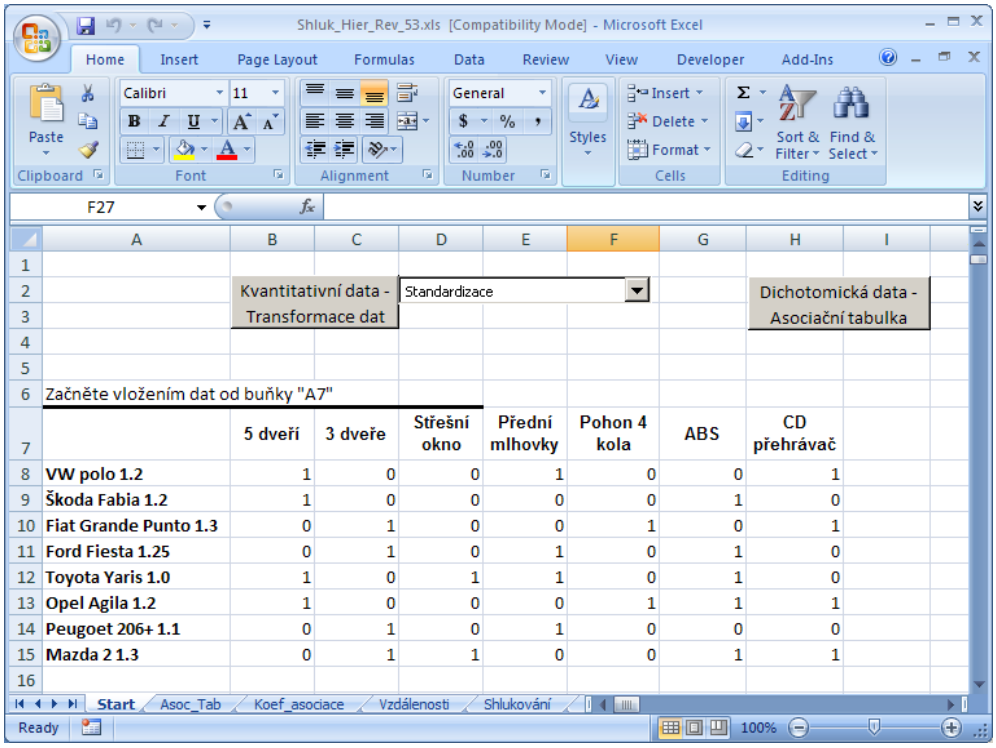
Návod jak postupovat při práci s dichtomickými daty bude demonstrován pomocí příkladu 2:

Tabulka 6 uvádí některé vlastnosti v základní výbavě vybraných automobilů s obsahem motoru 1,0 až 1.3. Které automobily jsou si z hlediska těchto vlastností nejvíce podobné, a které lze zařadit do stejných skupin? Použijte shlukovou analýzu, hierarchickou metodu, aglomerativní přístup. Volte Sokalův a Michenerův koeficient asociace a Wardova-Wishartova metodu pro výpočet vzdálenosti mezi shluky.

Tabulka 6 Příklad pro dichotomická data [zdroj: autor]

	5 dveří	**3 dveře**	**Střešní okno**	**Přední mlhovky**	**Pohon 4 kola**	**ABS**	**CD přehrávač**
VW polo 1.2	1	0	0	1	0	0	1
Škoda Fabia 1.2	1	0	0	0	0	1	0
Fiat Grande Punto 1.3	0	1	0	0	1	0	1
Ford Fiesta 1.25	0	1	0	1	0	1	0
Toyota Yaris 1.0	1	0	1	1	0	1	0
Opel Agila 1.2	1	0	0	0	1	1	1
Peugoet 206+ 1.1	0	1	0	1	0	0	0
Mazda 2 1.3	0	1	1	0	0	1	1

Řešení: Tabulku zkopírujeme do listu „Start“ analogicky jako v případě práce s kvantitativními daty. Viz obrázek 19.

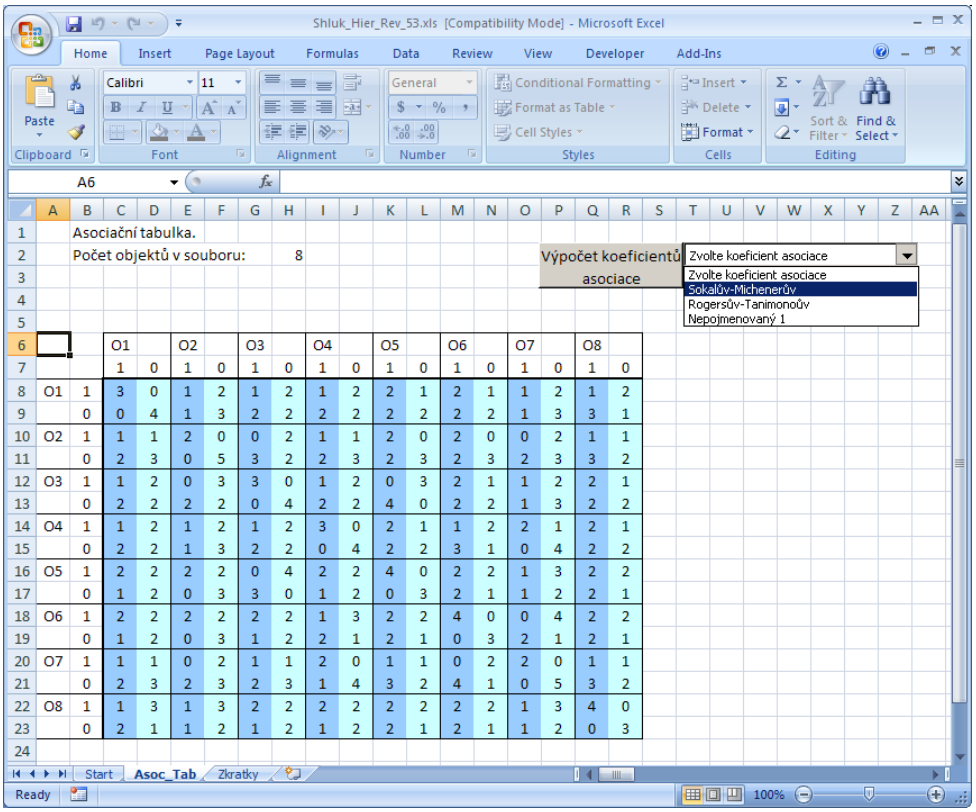


Obrázek 19 Příklad 2 – zkopírování dat do souboru Shluk_Hier.xls [zdroj: autor]

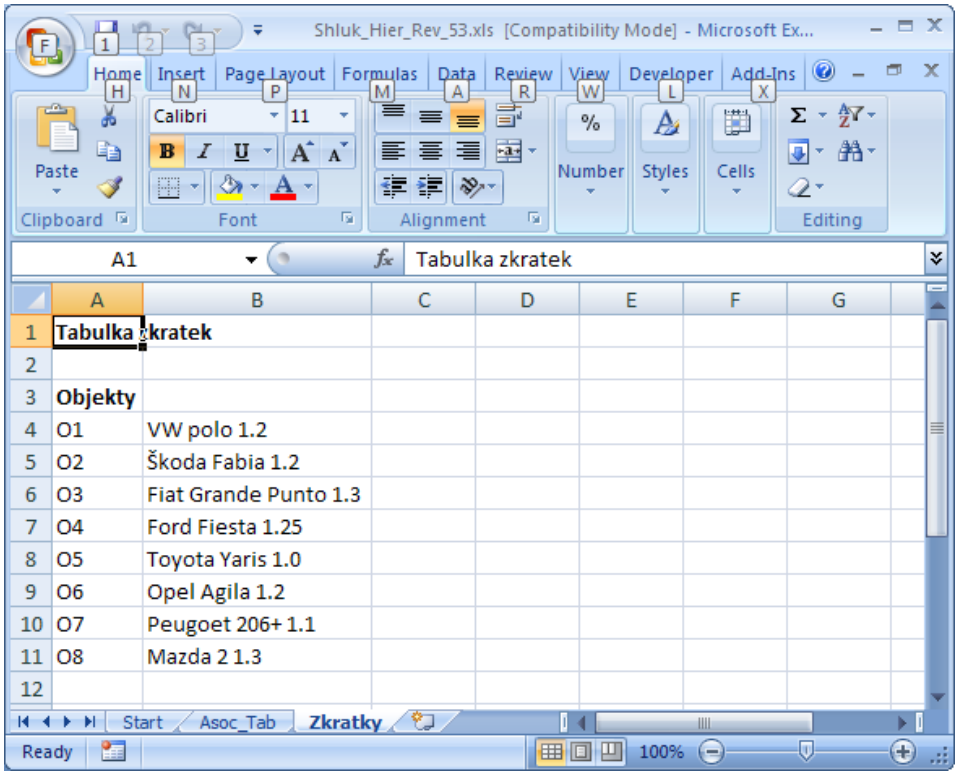
Při práci s dichotomickými daty, stejně tak jako při práci s kvantitativními daty, je nutné určit vzdálenost mezi objekty. Nadstavbový modul realizuje výpočet této vzdálenosti tak, že nejprve vytvoří asociační tabulku. Po té se z asociační tabulky vypočítají koeficienty asociace. Protože koeficienty asociace vyjadřují míru podobnosti, je potřeba pro proces shlukování hierarchickou metodou tuto míru podobnosti převést na míru nepodobnosti, tj. vzdálenost mezi objekty.

Kliknutím na tlačítko Dichotomická data – Asociační tabulka, dojde nejprve ke kontrole dat a v případě, že jsou dichotomická data v pořádku (soubor hodnot nabývá pouze 0 a 1), dojde k vytvoření nového listu s názvem „Asoc_Tab“, viz obrázek 20. Zároveň dojde k vytvoření listu s názvem „Zkratky“ s tabulkou zkratek - obrázek 21.

Poté následuje výpočet koeficientů asociace. Volba typu koeficientu asociace se provede pomocí roletového menu vedle tlačítka Výpočet koeficientů asociace. Nadstavbový modul umožňuje výpočet těchto koeficientů asociace: Sokalův-Michenerův koeficient asociace, Rogersův-Tanimonoův koeficient asociace a Nepojmenovaný 1 koeficient asociace.

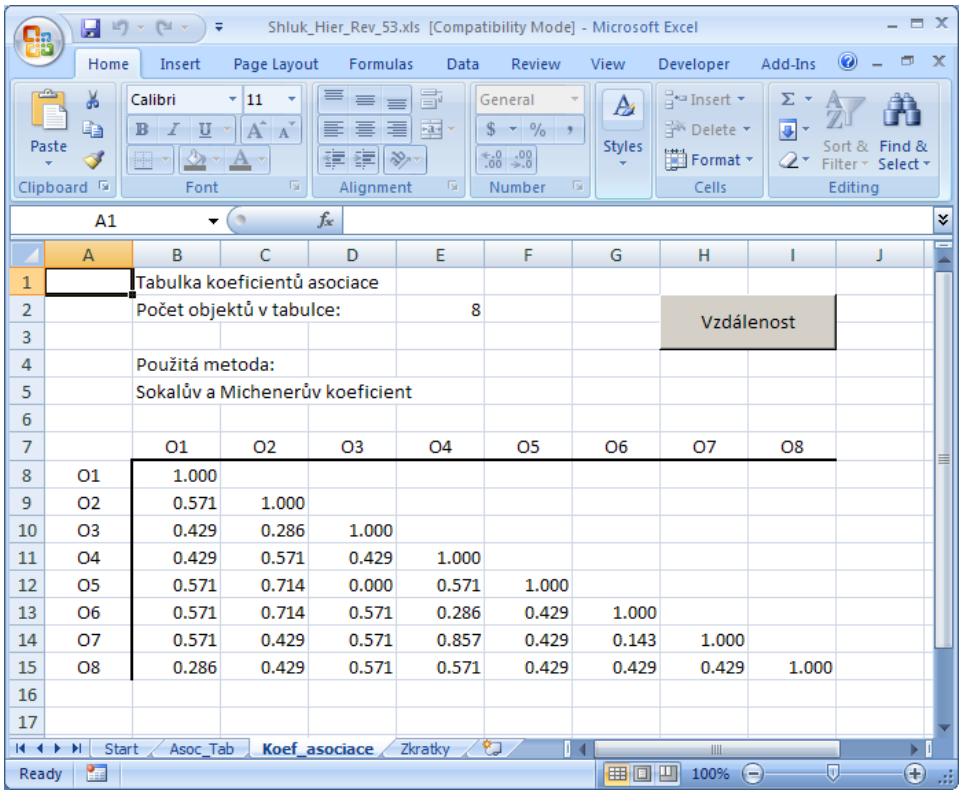


Obrázek 20 Příklad 2 – vytvoření listu s asociační tabulkou [zdroj: autor]



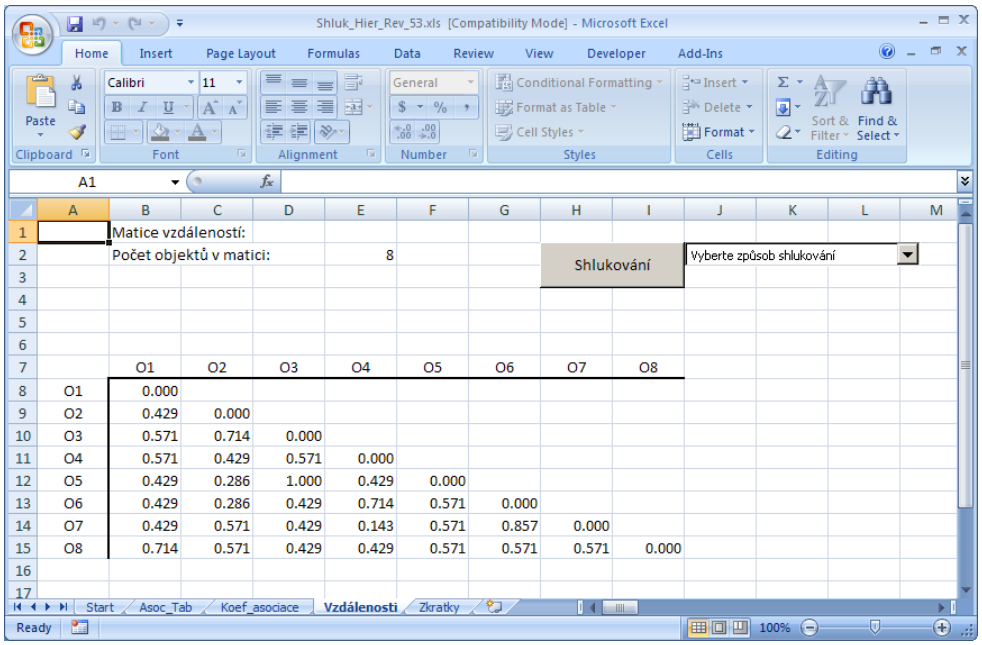
Obrázek 21 Příklad 2 – ukázka listu Zkratky [zdroj: autor]

V příkladu je zvolen Sokalův-Michenerův koeficient asociace. Kliknutím na tlačítko Výpočet koeficientů asociace dojde k provedení výpočtu koeficientů asociace. Vznikne nový list s názvem „Koef_asociace“. Na tomto listu je tabulka koeficientů asociace. Viz obrázek 22.



Obrázek 22 Příklad 2 – list s tabulkou koeficientů asociace [zdroj: autor]

Dalším krokem je přepočet koeficientů asociace na vzdálenost mezi objekty. Způsob výpočtu je uveden v kapitole 6.3.2 Zpracování dichotomických dat. Kliknutím na tlačítko Vzdálenost dojde k provedení výpočtu vzdálenosti. Vznikne nový list s názvem „Vzdálenosti“. Na tomto listu je tabulka vzdáleností mezi objekty. Viz obrázek 23.

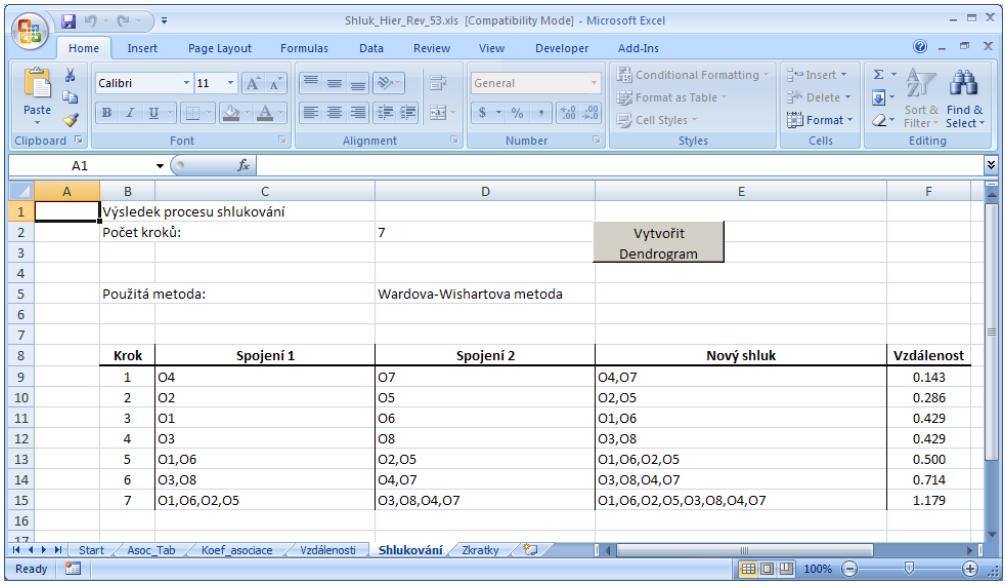


	A	B	C	D	E	F	G	H	I
7		O1	O2	O3	O4	O5	O6	O7	O8
8	O1	0.000							
9	O2	0.429	0.000						
10	O3	0.571	0.714	0.000					
11	O4	0.571	0.429	0.571	0.000				
12	O5	0.429	0.286	1.000	0.429	0.000			
13	O6	0.429	0.286	0.429	0.714	0.571	0.000		
14	O7	0.429	0.571	0.429	0.143	0.571	0.857	0.000	
15	O8	0.714	0.571	0.429	0.429	0.571	0.571	0.571	0.000

Obrázek 23 Příklad 2 – list se vzdálenostmi mezi objekty [zdroj: autor]

Nyní se již postupuje stejným způsobem jako při zpracovávání kvantitativních dat. Po výpočtu vzdálenosti následuje proces shlukování. Je potřeba zadat způsob výpočtu vzdálenosti mezi shluky. Ta se provede pomocí roletového menu vedle tlačítka Shlukování. V příkladu je zvolena Wardova-Wishartova metoda. Kliknutím na tlačítko Shlukování dojde ke spuštění procesu shlukování. Vznikne nový list s názvem „Shlukování“. Na tomto listu je tabulka, která vystihuje, jaké shluky byly spojeny v jednotlivých krocích a při jaké vzdálenosti ke spojení došlo. Viz obrázek 24.

Pro grafickou reprezentaci výsledku shlukování slouží stromový graf, který se nazývá dendrogram. Ten je možné vygenerovat kliknutím na tlačítko Vytvořit Dendrogram. Vznikne nový list s názvem „Dendrogram“. Dendrogram pro tento příklad zobrazuje obrázek 25.



Výsledek procesu shlukování

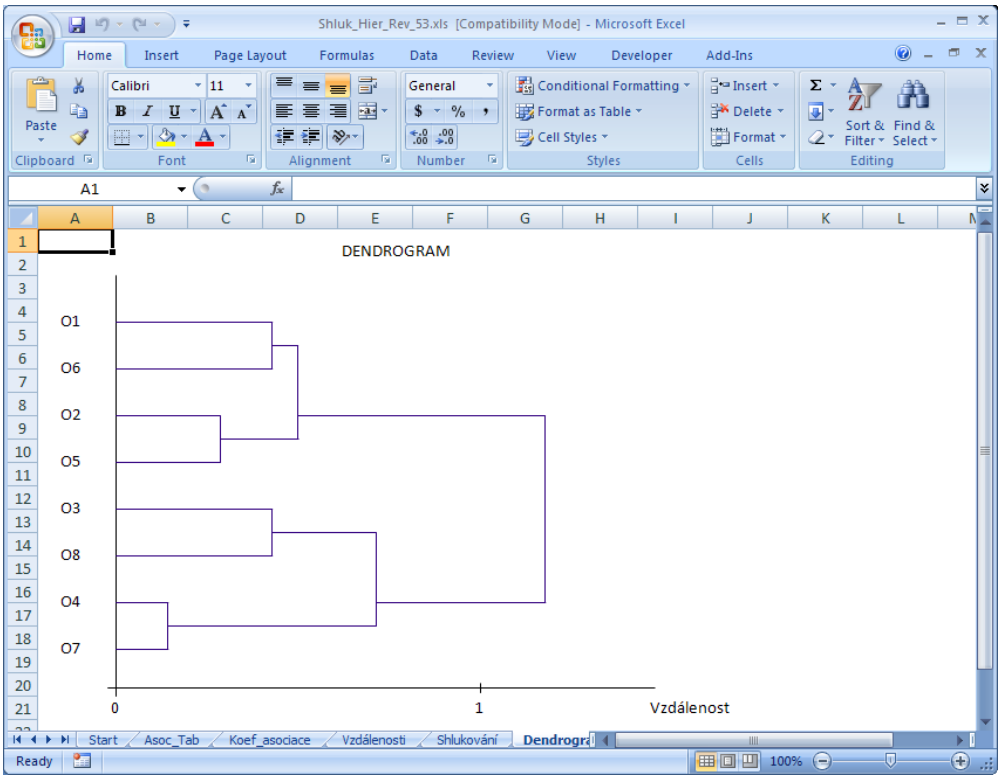
Počet kroků: 7

Vytvořit Dendrogram

Použitá metoda: Wardova-Wishartova metoda

Krok	Spojení 1	Spojení 2	Nový shluk	Vzdálenost
1	O4	O7	O4,O7	0.143
2	O2	O5	O2,O5	0.286
3	O1	O6	O1,O6	0.429
4	O3	O8	O3,O8	0.429
5	O1,O6	O2,O5	O1,O6,O2,O5	0.500
6	O3,O8	O4,O7	O3,O8,O4,O7	0.714
7	O1,O6,O2,O5	O3,O8,O4,O7	O1,O6,O2,O5,O3,O8,O4,O7	1.179

Obrázek 24 Příklad 2 – list s výsledkem procesu shlukování [zdroj: autor]



Obrázek 25 Příklad 2 – vytvoření dendrogramu [zdroj: autor]

7.3 Řešení známých problémů s nadstavbovým modulem

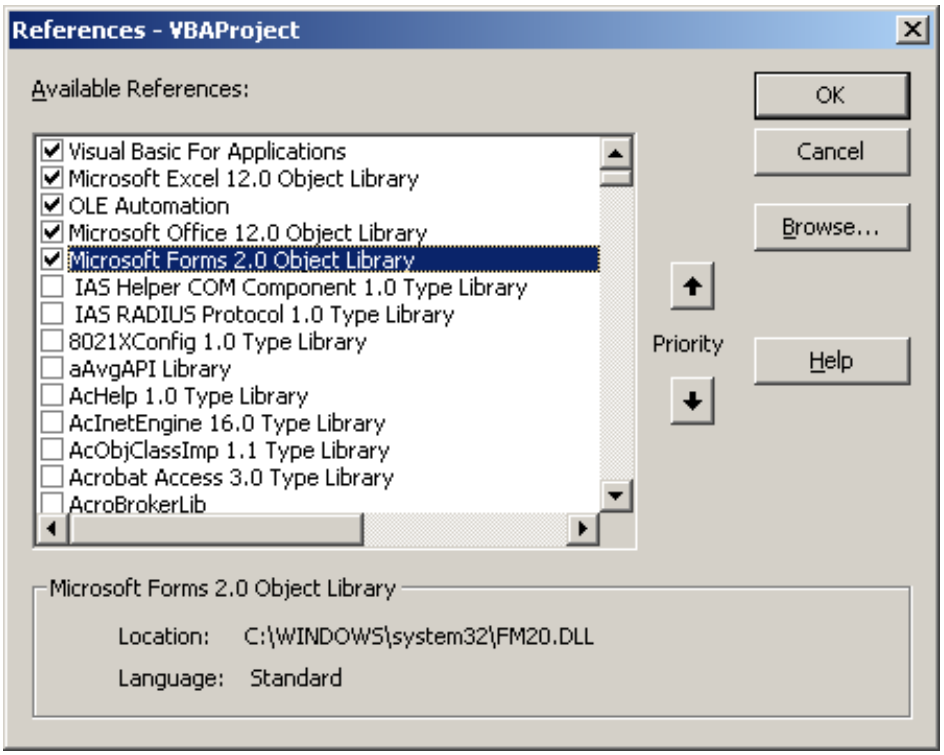
7.3.1 Povolení maker

Pro správnou funkci nadstavbového modulu je zapotřebí mít v aplikaci Excel povolená makra. Úprava nastavení spouštění maker se provádí volbou tlačítka Office → Možnosti aplikace Excel → Centrum zabezpečení → Nastavení Centra zabezpečení →Nastavení maker.

7.3.2 Error in loading DLL

Nadstavbový modul je testován v těchto verzích aplikace Excel: Excel 97-2003, Excel 2007, Excel 2010. Může se stát, že při různých pokusech s modulem, zejména je-li modul používán v různých verzích aplikace Excel a zároveň je spouštěn MS Visual Basic editor, dojde k vyvolání této chybové hlášky: „Error in loading DLL“.

V tomto případě je potřeba při otevřeném nadstavbovém modulu spustit MS Visual Basic editor, vybrat menu Tools / References a v otevřeném okně upravit nastavení DLL knihoven. Viz obrázek 26.



Obrázek 26 Nastavení knihoven DLL [zdroj: autor]

7.3.3 Vykreslování dendrogramu

Vykreslování dendrogramu může trvat dlouhou dobu (v závislosti na rychlosti počítače), nebo může skončit neúspěchem, tj. Visual Basic skript nahlásí chybu a bude ukončen. Tento problém je nejčastěji způsoben chybnou, nebo dokonce žádnou transformací vstupních dat.

Hodnoty na ose X se nemění dynamicky. Osa X je popsána celými čísly od nuly do vzdálenosti, při které došlo k sloučení všech objektů do jednoho shluku[3]. Jestliže je tato vzdálenost příliš vysoká, vypisování všech hodnot do dendrogramu je časově náročné.

Dalším důvodem neúspěšného vykreslení dendrogramu může být jeho velikost. Nadstavbový modul vykresluje dendrogram takovým způsobem, že z důvodu jeho čitelnosti je nejmenší vzdálenost sloučení shluků realizována spojem délky 100 pixelů. Jestliže je poměr mezi nejmenší vzdáleností sloučení shluků a největší vzdáleností sloučení shluků příliš vysoký, může dojít k situaci, že program nemůže vykreslit dendrogram, protože se nevejde na daný list.

Řešením těchto situací je volba vhodného způsobu transformace dat, případně vhodná volba výpočtu vzdáleností mezi objekty a shluky.

[3] Má-li tato hodnota desetinná místa, pak je popis osy zaokrouhlen dolu na celé číslo.

8 ZÁVĚR

Tato diplomová práce se věnuje jedné z několika metod vícerozměrné statistické analýzy - shlukové analýze. Součástí diplomové práce je nadstavbový modul pro zpracování dat metodou hierarchického shlukování. Tímto jsou cíle práce, které jsou uvedené v úvodu, splněny.

Úvodní kapitoly diplomové práce se zabývají teoretickými aspekty souvisejícími se shlukovou analýzou. Jsou zde uvedeny základní pojmy z oblasti zpracování vícerozměrných dat, různé metody vícerozměrné statistické analýzy a základní metody shlukové analýzy. Dále jsou v teoretické části popsány způsoby předzpracování a transformace dat, metody pro měření vzdáleností mezi objekty a shluky, a postup při hierarchickém shlukování.

Následující kapitoly se věnují nadstavbovému modulu. Je zde uvedena základní koncepce nadstavbového modulu a popsány jednotlivé procedury programu včetně výpočtů. Kapitola s názvem „Metodické pokyny k použití modulu“ je prakticky návod, jak modul používat. Součástí této podkapitoly je také doporučení, jak řešit problémy, které mohou při práci s programem nastat.

Nadstavbový modul pro zpracování dat metodou hierarchického shlukování, který je součástí této diplomové práce, je určen pro podporu výuky předmětu Zpracování dat metodami shlukové analýzy. Tento modul je naprogramován v programovém prostředí Microsoft Visual Basic for Aplication, který je součástí MS Excel. Ačkoli postupy a matematický aparát pro hierarchické shlukování jsou již dobře známy, algoritmy použité v tomto nadstavbovém modulu jsou originální. Nadstavbový modul umožňuje zpracovávat jak kvantitativní data, tak i dichotomická data. V případě kvantitativních dat umožňuje modul transformaci dat, výpočet vzdáleností mezi objekty a provedení procesu shlukování. U dichotomických dat je ze vstupních dat nejprve vytvořena asociační tabulka, v dalším kroku jsou spočítány koeficienty asociace, poté následuje výpočet vzdáleností mezi objekty a proces shlukování. Výsledky jednotlivých kroků jsou postupně vkládány do listů souboru Excel. Tato vlastnost nadstavbového modulu je výhodou pro uživatele, kterými budou zejména studenti, protože je možné výsledky v každém kroku posupně sledovat a analyzovat.

Vstupem do procesu shlukování je tabulka, kde řádky vyjadřují jednotlivé objekty a sloupce jsou vlastnosti těchto objektů. Tuto tabulku je možné do modulu nakopírovat běžným způsobem v MS Office (například klávesovými zkratkami Ctrl+C / Ctrl+V). I přesto, že diplomová práce obsahuje metodické pokyny pro práci s modulem, je jeho ovládání intuitivní a snadno pochopitelné. Modul je ovládán pouze pomocí rozbalovacích menu a tlačítky pro potvrzení volby.

Výstupem shlukování je tabulka, která uvádí které objekty (shluky) byli sloučeny v každém kroku shlukování a při jaké vzdálenosti ke sloučení došlo. Grafický výstupem shlukování je dendrogram.

Dendrogram je typ grafu, který není v MS Excel běžně k dispozici. Proto bylo nutné tento graf manuálně naprogramovat. Podkladem pro vykreslení dendrogramu je výstupní tabulka po shlukování. V průběhu vývoje této části modulu došlo k zjištění, že tuto výstupní tabulku nelze explicitně použít pro

vytvoření dendrogramu, protože by docházelo k nežádoucímu křížení linek. Aby k tomuto křížení nedocházelo, bylo potřeba nalézt algoritmus pro vhodné uspořádání řádků ve výstupní tabulce. Tento problém byl velmi obtížně řešitelný, protože u rozsáhlých souborů dat nebylo snadné definovat pravidla pro uspořádání řádků. Programový kód vytvořený běžnými programovacími postupy (větvení, různé typy cyklů) byl velmi složitý a právě u rozsáhlých souborů dat vykazoval chyby a jeho odladění by vyžadovalo zdlouhavé testování a úpravy.

Nakonec se podařilo problém hledání pravidel pro uspořádání řádků obejít, protože se ukázalo jako velmi vhodné využít teorie z předmětu Umělá a výpočetní inteligence a pohlížet na dendrogram jako na strom, tj. jako na typ grafu, ve kterém jsou řádky v tabulce reprezentovány uzly a úlohu vhodného setřídění řádků řešit algoritmem procházení grafu (stromu) do hloubky. Tento postup je velmi efektivní a rychlý. Je jedinečný a byl vyvinut v rámci této diplomové práce.

POUŽITÁ LITERATURA

[1] BREDEN, M.; SCHWIMMER, M. *Excel 2007 VBA.* 1. vyd. Brno: Computer Press, 2009 696 s. ISBN: 978-80-251-2698-1.

[2] *Cluster Analysis. Elektronická učebnice StatSoft* [online]. StatSoft, 2008. [cit. 2009-02-14]. Dostupný z WWW: <http://www.statsoft.com/textbook/stcluan.html >.

[3] HYNAR, M. *Metody shlukování* [online]. 2003 [cit. 2010-07-20]. Dostupný z WWW: <http://www.fit.vutbr.cz/study/courses/ZZD/public/seminar0304/Shlukovani1-text.pdf>.

[4] KELBEL, J.; ŠILHÁN, D. *Shluková analýza* [online]. [cit. 2010-07-16]. Dostupný z WWW: <http://staff.utia.cas.cz/nagy/skola/Projekty/Classification/ShlukovaAnalyza.pdf>.

[5] KUBANOVÁ, J. *Statistické metody pro ekonomickou a technickou praxi.* 2. vyd. Bratislava: Statis, 2004 249 s. ISBN 80-85659-37-9.

[6] KUMAR, V.; STIENBACH, M.; TAN, P. *Introduction to Data Mining. Chapter 8 - Cluster Analysis: Basic Concepts and Algorithms* [online]. 2005 [cit. 2010-07-22]. Dostupný z WWW: <http://www-users.cs.umn.edu/~kumar/dmbook/index.php>.

[7] Nápověda programu MS Excel a VBA.

[8] OLEJ, V.; PETR, P. *Expertní a znalostní systémy v managementu. Část, Expertní systémy - distanční opora.* 1. vyd. Pardubice: Univerzita Pardubice, 2004. ISBN 80-7194-688-5.

[9] PECHÁČEK, P. *Excelentně v Excelu.* [cit. 2010-07-26]. Dostupný z WWW: <http://excelplus.net/news.php>.

[10] PETR, P. *Data Mining. Díl 1.* 1. vyd. Pardubice: Univerzita Pardubice, 2006. ISBN 80-7194-886-1.

[11] ŘEZANOVÁ, H.; HÚSEK, D.; SNÁŠEL, V. *Shluková analýza dat.* 2. vyd. Praha: Professional Publishing, 2009 218 s. ISBN: 978-80-86946-81-8.

[12] WALKENBACH, J. *Microsoft Office Excel 2007: Programování ve VBA.* 1. vyd. Brno: Computer Press, 2008 912 s. ISBN: 978-80-251-2698-1.

[13] ŽÁK, L. *Shluková analýza.* Automatizace, Vol.2004, (2004), No.3, pp.184-189, ISSN 0005-125.

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

PŘÍLOHY

Seznam souborů uložených na CD:

Heslo.txt – Heslo pro přístup k nadstavbovému modulu v prostředí VBA.

Příloha 1.doc - Popis procedury pro zjištění počtu objektů, vlastností a kontroly dat.

Příloha 2.doc - Popis procedury standardizace dat.

Příloha 3.doc - Popis procedury normalizace dat.

Příloha 4.doc - Popis procedur výpočtu vzdáleností.

Příloha 5.doc - Popis procedury pro určení počtu objektů ve shlucích.

Příloha 6.doc - Popis procedury pro určení shluků, které budou sloučeny.

Příloha 7.doc - Popis procedury pro vytvoření legendy řádků a sloupců přepočítané tabulky vzdáleností.

Příloha 8.doc - Popis procedury pro přepočítání tabulky vzdáleností.

Příloha 9.doc – Popis procedury úpravy tabulky pro vykreslení dendrogramu.

Příloha 10.doc - Podrobný popis postupu vykreslení dendrogramu.

Příloha 11.doc - Podrobný popis postupu úpravy tabulky pro vykreslení dendrogramu.

Příloha 12.doc - Popis procedury vytvoření asociační tabulky.

Příloha 13.doc - Popis procedury převodu koeficientů asociace na míru nepodobnosti.

Shluk_Hier.xls – Soubor MS Excel s nadstavbovým modulem.