Univerzita Pardubice
Fakulta ekonomicko-správní

Modelování ekonomických dat

Bc. Michal Bělský

Diplomová práce
2010

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: **2009/2010**

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Michal BĚLSKÝ**
Osobní číslo:	**E09987**
Studijní program:	**N6209 Systémové inženýrství a informatika**
Studijní obor:	**Regionální a informační management**
Název tématu:	**Modelování ekonomických dat**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

Predpokládaným výstupem práce bude:
Sběr a analýza dat z vybrané oblasti (ekonomické, environmentální nebo sociální).
Analýza tvorby modelu pro vybranou oblast (ekonomickou, environmentální, sociální).
Návrh modelu za využití např. rozhodovacích stromů, neuronových sítí, shlukové analýzy atd. v data-miningovém nástroji.

Rozsah grafických prací:

Rozsah pracovní zprávy:

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

Seznam odborné literatury:

BERKA, P., Dobývání znalostí z databází. Praha, Academia, 2003, 366 s., 2003.

BERRY M.J.A. Data mining techniques : for marketing sales and customer support. Indianapolis Wiley, 643 s., 2004.

BERRY M.J.A. Mastering data mining : the art and science of customer : relationship management. New York, John Wiley & Sons, 494 s., 2000.

KAHOUN, J. Ukazatele regionální konkurenceschopnosti v České republice. In Working Paper CES VŠEM, č. 5, 2007. s. 1-36.

MARTIN, R.L. A Study on the Factors of Regional Competitiveness. Cambridge: Cambridge Econometrics, 2003.

PROVAZNÍKOVÁ, R., KŘUPKA, J., KAŠPAROVÁ, M. Predictive Modelling on the Regional Level. In TILTAI. Social Sciences in Global Word: Possibilities, Changes and Perspectives., Klaipéda: Klaipéda University Press. 2009, roč. 39, s.150-158. ISSN 1648-3979.

Vedoucí diplomové práce: **doc. Ing. Jiří Křupka, Ph.D.**
Ústav systémového inženýrství a informatiky

Konzultant diplomové práce: **doc. Ing. Romana Provazníková, Ph.D.**
Ústav ekonomie

Datum zadání diplomové práce: **5. října 2009**

Termín odevzdání diplomové práce: **30. dubna 2010**

L.S.

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

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

V Pardubicích dne 5. října 2009

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 odst. 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 28. 4. 2010

Bc. Michal Bělský

Poděkování

Rád bych poděkoval svému vedoucímu práce doc. Ing. Jiřímu Křupkovi, Ph.D., za odborné vedení, náměty a připomínky, které mi poskytoval v průběhu celého období zpracovávání mé diplomové práce.

Také bych rád poděkoval doc. Ing. Romaně Provazníkové, Ph.D., za odborné ekonomické náměty, připomínky a cenné rady.

SOUHRN

Diplomová práce se zabývá oblastí data miningu. Je zaměřena na hledání smysluplných vlastností v ekonomickém datovém souboru pomocí shlukovacích metod. Analyzovanými atributy jsou míra nezaměstnanosti, míra růstu mezd, průměrná hrubá měsíční mzda, počet dokončených bytů a tržby z průmyslové činnosti. Všechny atributy jsou členěny podle jednotlivých krajů. Modelovacím nástrojem je statistický software SPSS Clementine.

KLÍČOVÁ SLOVA

Shlukovací metody, data mining, nezaměstnanost, tržby, mzdy, dokončené byty, modelování, Phillipsova křivka.

TITLE

Modelling of economical data

SUMMARY

My thesis deals with an area of data mining. It is specialized in searching for meaningful characteristics in the economical data set by the help of clustering methods. Analysed attributes mean unemployment rate, growing wages, average monthly gross wage, number of completed dwellings and revenues from industrial activity. All those attributes are divided in accordance with particular regions. Modelling tool is statistical software SPSS Clementine.

KEY WORDS

Clustering methods, data mining, unemployment, revenues, completed dwellings, modelling, Phillips curve.

Obsah

Úvod

Tato diplomová práce se zaměřuje na problematiku data miningu a na principy modelovacích metod shlukovací analýzy s aplikací na modelování ekonomických dat. Oblast data miningu určitě stojí za povšimnutí. Data mining lze přeložit jako dolování z dat, zabývá se hledáním zajímavých vlastností v rozsáhlých datových souborech.

Tento obor se v dnešní době rychle rozvíjí a je těžištěm pro modelovaní smysluplných vlastností na datech. Sklidil velké úspěchy v bankovnictví a pojišťovnictví, kde pomocí specifických metod odhalil více pojistných podvodů a rizikových zákazníků.

Obzvláště se věnuji metodám shlukovací analýzy, pomocí nichž modeluji závislosti na datech z ekonomické oblasti v softwaru Clementine.

Konkrétně se věnuji třem skupinám dat, které jsem vybral po konzultaci s odborníkem z Ústavu ekonomie. V první skupině modeluji závislost počtu dokončených bytů na průměrné hrubé měsíční mzdě. V druhé skupině dat řeším závislost mezi mírou registrované nezaměstnanosti a mírou růstu mezd v podobě Phillipsovy křivky. Do třetí skupiny patří průměrná hrubá měsíční mzda, míra registrované nezaměstnanosti a tržby z průmyslové činnosti.

Celý datový soubor je rozdělen podle 13 krajů v časové řadě od roku 2001 do roku 2009 po čtvrtletí. Praha byla vyloučena, jelikož má zvláštní postavení, hlavní město Praha se řadí mezi vyspělé metropole Evropské unie. Statistické údaje a další studie (např. Ekonomická situace českých krajů a měst - vypracované společností MasterCard a Vysokou školou ekonomickou) potvrzují výjimečné postavení hl. města Prahy, kde právě Praha zaujímá první místo ve všech srovnávacích ukazatelích, v socioekonomické úrovni a v investiční atraktivnosti.[20]

Z těchto důvodů nebyla tato lokalita zařazena do datového modelování, představovala by v dataminingové teorii odlehlé hodnoty „outliers“. S těmito hodnotami musím pracovat velmi opatrně. V tomto případě je nejlepší zbraní důkladná znalost zpracovávaného datového souboru.

Dle mého názoru má data mining velkou budoucnost. Nabízí širokou škálu metod a má všestranné uplatnění v mnoha oborech. Celý jeho proces má velmi dlouho cestu, začíná získáním dat od zákazníka přes modelování až po využití získaných výsledků v praxi. Uplatnění v tomto oboru najdou manažeři, databázoví administrátoři a specialisti na data mining.

Zabývám se konkrétně statistickou metodou shlukovací analýza, přímo Kohonenovou mapou, metodou TwoStep a K-Means. Všechny tyto metody se řadí do skupiny učení bez učitele, pro analýzu pozorování nepotřebují informaci od učitele. Shlukovací analýza se snaží najít zajímavé homogenní podskupiny (shluky) v datovém souboru tak, aby si členové uvnitř shluku byly co nejvíce podobní a mezi shluky byly co nejvíce rozdílní.

Chtěl bych hlavně poukázat na uplatnění oboru data mining v praxi. Pod tímto oborem se neskrývá jenom hledání zajínavých vlastností v datech, ale také využití širokého spectra metod ze statistiky a matematiky. Díky dokončeným analýzám v data miningu si většina firem upevnila postavení na dnešním silném konkurenčním trhu. Jsou zrealizovány rozsáhlé projekty z oblasti bankovnictví, pojišťovnictví, telekomunikací a marketingu.

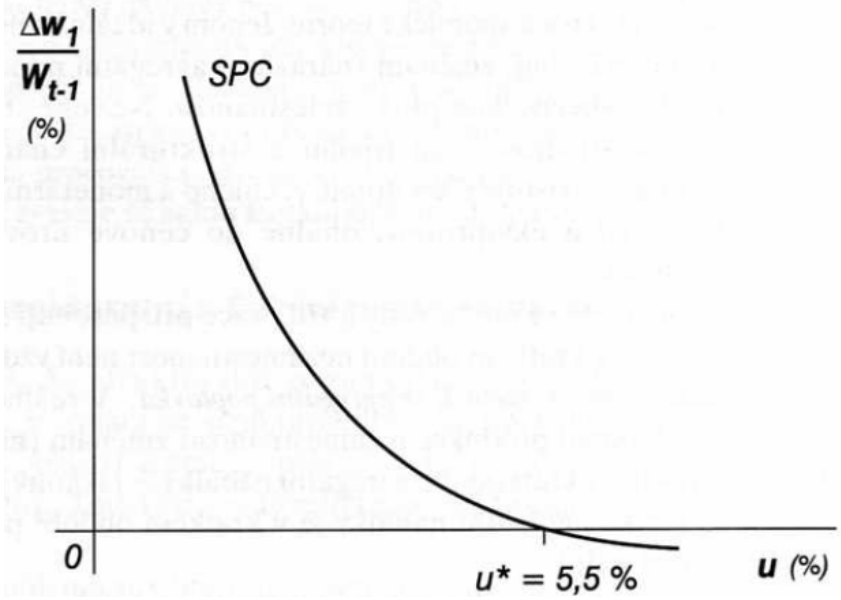
Cílem této práce je:

- Analyzovat možnosti využití dataminingových metod v ekonomické oblasti.
- Poukázat na závislosti mezi vybranými parametry (počet dokončených bytů a průměrná hrubá mzda, míra registrované nezaměstnanosti a míra růstu mezd atd.).

1 Původní mzdová Phillipsova křivka

Phillipsova křivka je popsána inverzním vzájemným vztahem mezi mírou nezaměstnanosti a mírou růstu peněžních (nominálních) mzdových sazeb.[13] Na tento důkaz přišel v roce 1958 novozélandský ekonom A. W. Phillips. Analyzoval chování mezd a míry nezaměstnanosti ve Velké Británii. *„Na základě empirického výzkumu vztahu změn peněžních mzdových sazeb a míry nezaměstnanosti v uvedeném období ve Velké Británii formuloval závěr o inverzním vzájemném vztahu mezi mírou nezaměstnanosti a mírou změny peněžních mzdových sazeb, jež je od té doby nazýván* ***Phillipsovou křivkou****.“*

„Původní mzdová Phillipsova křivka vyjadřuje vzájemný inverzní vztah mezi mírou nezaměstnanosti a mírou růstu peněžních (nominálních) mezd.“ [13]



Obrázek č. 1: Mzdová Phillipsova křivka [13]

Na obrázku č. 1 je zobrazen graf Phillipsovy křivky, na ose x měřím míru nezaměstnanosti (u) v procentech a na ose y míru změny peněžních mzdových sazeb v procentech, která je reprezentovaná vztahem $\Delta W_t / W_{t-1}$ a označovaná g_w. [21]

Z výše uvedeného grafu mohu vyvodit následující skutečnosti:

1. Phillipsova křivka má tvar hyperboly.
2. Má negativní sklon.
3. Protíná osu x.

Z uvedeného obrázku č. 1 je patrné, že čím vyšší je míra nezaměstnanosti, tím nižší je míra mzdové inflace. [13]

Phillipsova křivka protíná osu x v bodě $u^* = 5{,}5\%$, při míře nezaměstnanosti rovnající se přirozené míře.[13] V tomto bodě je míra růstu mezd nula procent. Když je nezaměstnanost pod 5,5 %, nominální mzdová sazba bude růst, a bude-li nad 5,5 %, nominální mzdová sazba klesne.

Nyní zformuluji mzdovou Phillipsovu křivku formálně, označím g_w jako tempo růstu nominálních mezd (míra mzdové inflace). Dále mohu pro míru mzdové inflace napsat:

$$g_w = \frac{W_t - W_{t-1}}{W_{t-1}},$$

Rovnice č. 1: Výpočet míry mzdové inflace [13]

kde W_t značí nominální mzdy v současném období a W_{t-1} značí nominální mzdy v minulém období. Mzdovou Phillipsovu křivku můžu zapsat jako $g_w = -e(u - u^*)$, kde ε je koeficient citlivosti změny míry nominálních mezd k procentní změně skutečné míry nezaměstanosti.[21]

„Z uvedené rovnice je patrné, že mzdy rostou, jestliže je skutečná míra nezaměstnanosti (u) nižší než přirozená míra nezaměstnanosti (u^*). Mzdy klesají tehdy, je-li skutečná míra nezaměstnanosti (u) větší než přirozená míra nezaměstnanosti (u^*).“ [13]

2 Ekonomický vývoj České republiky

„Makroekonomické ukazatele slouží k vytvoření celkového obrazu o dění v rámci jedné ekonomiky.“ Mezi nejvýznamnější indikátory patří inflace, nezaměstnanost, změna hrubého domácího produktu (HDP), vývoj platební bilance a státního rozpočtu. [14] [15]

Vývoj české ekonomiky v roce 1997 procházel transformační recesí, poté přišel hospodářský růst, který měl relativně krátké trvání a byl přerušen druhou recesí, která trvala až do roku 1999. Od 1999 došlo opět k pozitivnímu hospodářskému růstu. Změna hospodářské politiky podpořila růst domácí poptávky a spolu s pokrokem v realizaci strukturálních reforem přispěla k dosažení vyšší růstové dynamiky. Tyto pozitivní tendence jsou zřetelné zvláště po vstupu České republiky (ČR) do Evropské unie (EU) a po oživení evropské ekonomiky. Podařilo se dosáhnout většího vzájemného souladu mezi růstem HDP a změnami nezaměstnanosti. V roce 2009 pokračovala tendence meziročně nižšího výkonu. *„Hrubý domácí produkt poklesl proti stejnému období minulého roku o 5,5 %, což byl největší propad ekonomiky v novodobé historii samostatné ČR.“* [16] [19]

„Ve vývoji struktury produktu pokračovaly tendence odrážející slabou poptávku a zejména meziroční pokles obratu zahraničního obchodu a nižší investiční aktivita.“ [16]

Negativní vliv výdajových položek na HDP byl částečně kompenzován růstem konečné spotřeby. Ve srovnání s předchozím obdobím zpomalil svou dynamiku, a tím se zhoršila i situace

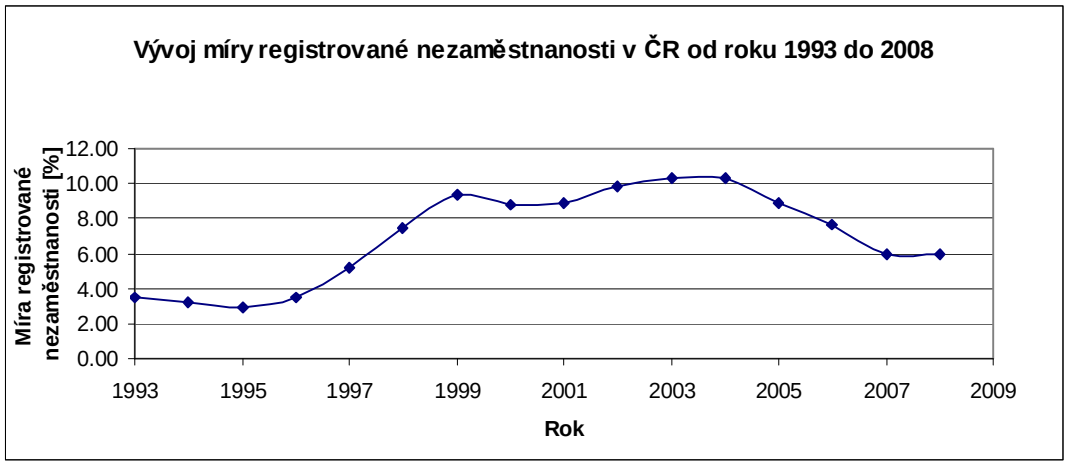
na trhu práce, poklesla zaměstnanost a vzrostla míra nezaměstnanosti. *„Reakcí na recesi bylo také prudké zpomalení růstu mezd.*“ [16]

Při cenovém vývoji došlo ke zpomalení meziročního růstu spotřebitelských cen a důsledkem toho došlo k poklesu tržních cen. [16]

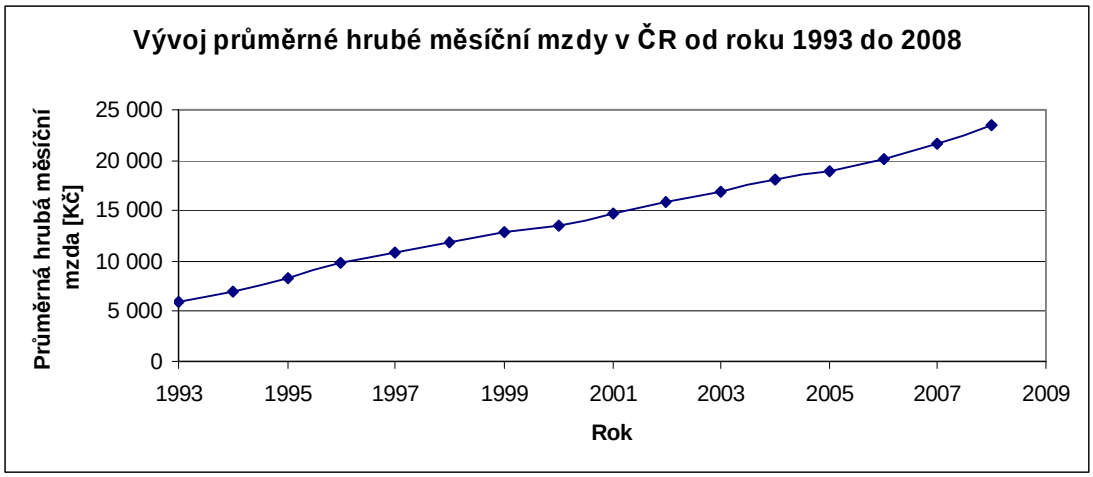
Průmyslová produkce pokračovala v poklesu, snížila se téměř ve všech odvětvích.

„Na vnitřním spotřebitelském trhu došlo k poklesu tržeb.“[16]Spotřebitelé odkládají především nákupy finančně náročnějšího průmyslového zboží, což souvisí se situací na trhu práce a i spotřebitelskou opatrností vyplývající z nejistých vyhlídek dalšího ekonomického vývoje.

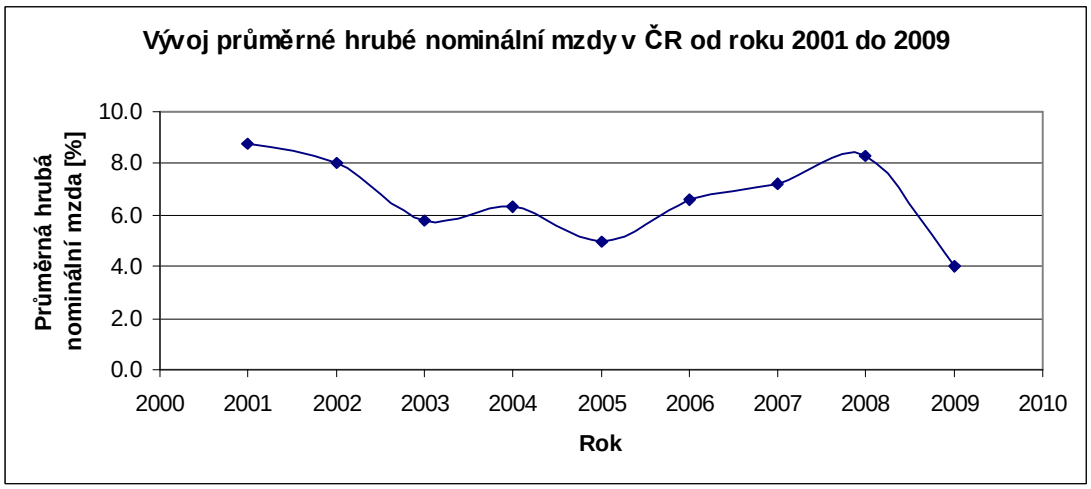
Na následujících grafech je vidět závislost jednotlivých ukazatelů míry registrované nezaměstnanosti, průměrné hrubé měsíční mzdy, průměrné nominální mzdy a míry růstu reálného HDP za celou ČR. Kolem roku 2008 je na všech grafech vidět pokles, pouze graf průměrné hrubé měsíční mzdy pořád roste. Od počátku roku 2008 je daný pokles způsobený ekonomickou krizí. Vývoj míry nezaměstnanosti odpovídá vývoji HDP. Uvedené grafy (graf č. 1 až graf č. 4) potvrdily závislosti v mnou získaných datech a ve výsledných analýzách v dalších částech mé diplomové práce.



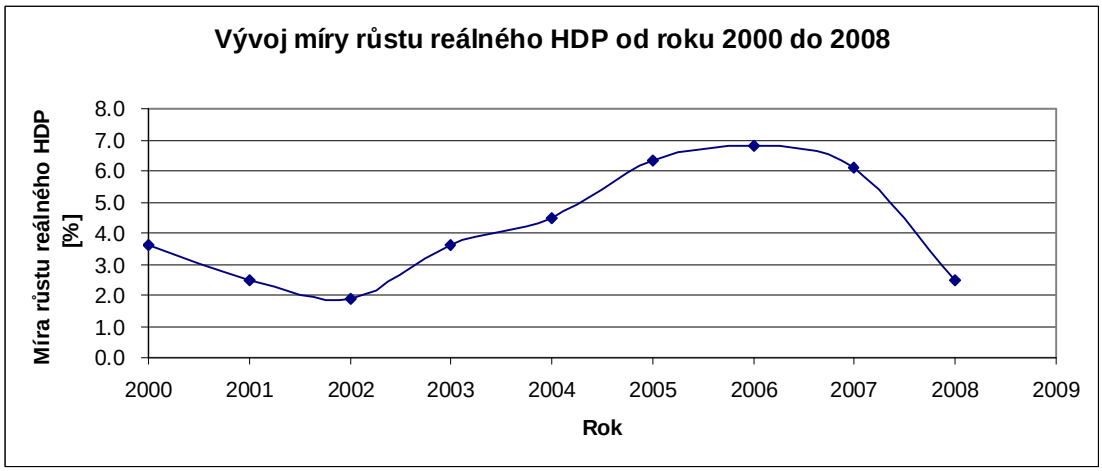
Graf č. 1: Vývoj míry registrované nezaměstnanosti v ČR [18]



Graf č. 2: Vývoj průměrné hrubé mesíční mzdy v ČR [18]



Graf č. 3: Vývoj průměrné nominální mzdy v ČR [18]



Graf č. 4: Vývoj míry růstu reálného HDP v ČR [18]

3 Data pro modelování

V této kapitole bubou popsána data pro modelování a jejich způsob získání. Dále se zaměřím na jednotlivé typy dat a popíšu jejich význam.

Data jsou dostupná na Českém statistickém úřadu (ČSÚ) Pardubického kraje. Na začátek bych podotknul, že tyto stránky jsou dobře strukturované. Když to srovnám s ostatními weby Českých krajských statistických úřadů, tak stránky ČSÚ Pardubického kraje mi byly nejbližší a nabízely nejširší škálu dat.

Hledal jsem čtvrtletní časovou řadu dat od roku 2001 do roku 2009. Tuto řadu má zveřejněnou pouze ČSÚ Pardubického kraje. Všechna data jsou dostupná z jejich webových stránek.

Kdybych tyto data chtěl sehnat ročně, tak bude v tabulce daleko méně záznamů, proto jsem je stahoval po čtvrtletí.

Celá tabulka záznamů spadá do ekonomické oblasti. Zkoušel jsem se dívat i po enviromentální oblasti záznamů, ale tato oblast se moc nevykazuje. Vykazuje se pouze za celou ČR a ročně. Škála ukazatelů z této oblasti není moc široká, proto jsem volil ekonomické zaměření datové tabulky.

Data lze najít ve statistickém bulletinu ČSÚ Pardubického kraje přímo pod sociálním a ekonomickým vývojem Pardubického kraje v rubrice Mezikrajské srovnání vybraných ukazatelů za každé čtvrtletí od roku 2001 do roku 2009.

Data jsou strukturovaná do 13 krajů. Vyřadil jsem z nich Prahu. Tato lokalita je ekonomicky nejvyspělejší. Je zde nejnižší nezaměstnanost. Z tohoto srovnání by tato lokalita představovala odlehlé hodnoty v datovém souboru oproti ostatním krajům, proto jsem Prahu nezařadil do datového souboru.

Pro každé čtvrtletí je zde na stránkách přehledně členěný excelovský soubor podle jednotlivých ukazatelů. Jsou zde ukazatele zaměstnanosti, nezaměstnanosti, bytové výstavby a zemědělství.

Všechna zvolená data byla vybrána po konzultaci s odborníkem z Ústavu ekonomie. Pro mou diplomovou práci jsou vybrány hlavně ukazatele týkající se trhu práce, migrace, průmyslu a bytové výstavby. Konkrétní ukazatele jsou zobrazené v nasledující tabulce č. 1.

Základní údaje	Trh práce	Obyvatelstvo a migrace	Bytová výstavba	Průmysl
Rozloha v km^2	Celkový počet zaměstnanců	Přistěhovalí	Dokončené byty	Tržby z průmyslové činnosti
	Průměrná hrubá měsíční mzda	Vystěhovalí		
	Míra registrované nezaměstnanosti	Počet obyvatel		

Tabulka č. 1: Vybrané ukazatele [zdroj: vlastní]

Nyní se zaměřím na jednotlivé ukazatele a uvedu ke každému jeho význam a popis.

1. **Rozloha v km^2**: Není moc měnící se údaj, zůstává skoro stejný v průběhu roku. Mění se ročně.
2. **Celkový počet zaměstnanců**: Do tohoto počtu zaměstnanců se zahrnují všichni stálí a dočasní zaměstnanci, kteří jsou v pracovním, služebním nebo členském poměru. Nejsou zde zahrnuty ženy na mateřské, osoby na rodičovské dovolené, učni a osoby vykonavající veřejnou funkci (senátoři, poslanci, soudci a členové zastupitelstev).[1]
3. **Průměrná hrubá měsíční mzda**: *„Do mezd se zahrnují základní mzdy a platy, příplatky ke mzdě nebo platu, prémie a odměny, náhrady mezd a platů, odměny za pracovní pohotovost a jiné složky mzdy nebo platu, které byly v daném období zaměstnancům zúčtovány k výplatě. Jedná se o hrubé mzdy, tj. před snížením o pojistné na všeobecné zdravotní pojištění a sociální zabezpečení, zálohové splátky daně z příjmů fyzických osob a další zákonné nebo se zaměstnancem dohodnuté srážky.*“[1]
4. **Míra registrované nezaměstnanosti**: *„Tento ukazatel se vypočte jako podíl vyjádřený v procentech, kde v čitateli je počet dosažitelných, neumístěných uchazečů o zaměstnání, občanů ČR a občanů EU, vedených úřady práce podle bydliště uchazeče ke konci sledovaného měsíce. Jedná se o evidované nezaměstnané, kteří nemají žádnou objektivní překážku pro přijetí do zaměstnání a při nabídce vhodného pracovního místa mohou do něj bezprostředně nastoupit. Jmenovatel tvoří pracovní sílu, tj. počet zaměstnaných v národním hospodářství s jediným nebo hlavním zaměstnáním podle výsledků výběrového šetření pracovních sil.*“[2]

5. **Přistěhovalí: „***Ukazatel vyjadřuje počet případů přistěhování na dané území. Přistěhováním se rozumí změna obce trvalého nebo dlouhodobého pobytu osoby na území ČR (vnitřní stěhování) nebo přes hranici ČR (zahraniční stěhování).“*[1]
6. **Vystěhovalí: „***Ukazatel vyjadřuje počet případů vystěhování z daného území. Vystěhováním se rozumí změna obce trvalého nebo dlouhodobého pobytu osoby na území ČR (vnitřní stěhování) nebo přes hranici ČR (zahraniční stěhování).“*[2]
7. **Počet obyvatel:** Jednou ze základních charakteristik, kterou sleduje demografická statistika je počet obyvatel k určitému okamžiku.
8. **Dokončené byty:** Jsou byty v dokončených budovách, které vyžadují stavební ohlášení nebo povolení.
9. **Tržby z průmyslové činnosti:** Oceňují se v základních běžných cenách, které fakturuje výrobce kupujícímu. Nezahrnují DPH, spotřební daň a clo. *„Tržby z průmyslové činnosti zahrnují tržby za prodej vlastních průmyslových výrobků a služeb průmyslové povahy, prodaných externím odběratelům.“*[1]

4 Aplikace metodiky CRISP-DM

Metodika CRISP-DM vznikla v rámci Evropského výzkumného projektu. Cílem projektu bylo navrhnout universální postup pro vytvoření modelu procesu dobývání znalostí z databází. Tato metodika umožní řešit úlohy pro dobývání znalostí rychleji, efektivněji a s nižšími náklady.[4]

Metodika CRISP-DM dělí životní cyklus projektu DM do šesti fází: porozumění problému, porozumění datům, příprava dat, modelování, hodnocení a využití v praxi.[5]

Úvodní fáze je zaměřena na pochopení cílů projektu z pohledu manažera a následné převedení na úlohy dobývání znalostí z databází. [5]

Trh práce patří k nejsledovanější části v ekonomické oblasti. Její ukazatele nám poskytnou informace o mzdách a o nezaměstnanosti.

Cílem modelování je najít:

- skryté závislosti, jak mzda ovlivňuje počet dokončených bytů
- jak výše nezaměstnanosti a mezd ovlivňuje výši tržeb z průmyslové činnosti
- mezi těmito skupinami dat zajímavé vlastnosti a chování

Všechny uvedené ukazatele v tabulce č. 1 jsem nakopíroval z jednotlivých souborů na webu do jednoho excelovského souboru typu xls.

Dále provedu prvotní náhled na data. Data jsou uspořádany do jednotlivých sloupců (proměnných) a řádků (záznamů). Nyní provedu první úpravy s datovým souborem. V každém

záznamu, ve kterém je použita desetinná čárka, je nutné tuto čárku nahradit desetinnou tečkou. Tuto operaci jsem provedl pomocí funkce DOSADIT v MS Excel. Jedná se konkrétně o sloupce míra registrované nezaměstnanosti a tržby z průmyslové činnosti.

Dále jsem odstranil interpunkci z názvů sloupců. Také jsem musel dávat pozor na délku řetězce v názvu sloupce. Software Clementine toleruje určitý počet znaků a překročením tohoto počtu daný sloupec při načítaní ignoruje.

Poslední úpravou v excelu je daný aktuální list uložit ve formátu CSV. CSV je jednoduchý formát pro výměnu tabulkových dat. Tento formát se skládá z řádků, ve kterém jsou všechny položky oddělené středníkem. [3]

Datová tabulka má celkem 12 sloupců a 442 řádků. Časová řada začíná od 1. čtvrtletí 2001 a končí v pololetí 2009.

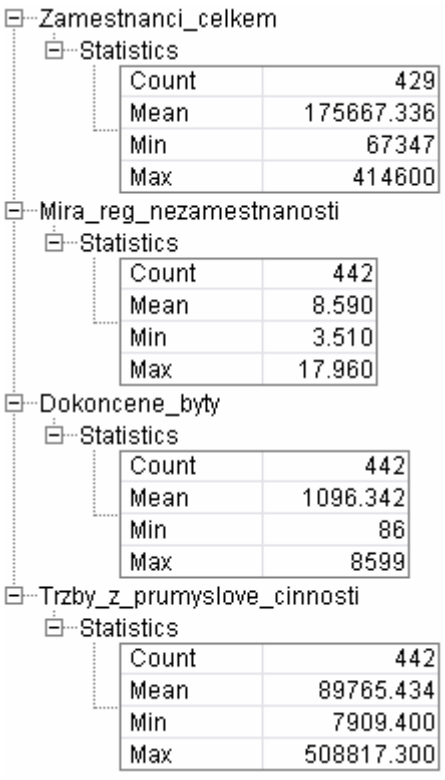
Dále pro seznámení s daty vytvořím datový slovník a provedu základní statistiku v programu Clementine.

Na obrázku č. 2 je uzel type v Clementine, který zobrazuje názvy atributů, jejich typy a rozsah.

Field	Type	Values
Rozloha_v_ha	Range	[316289,1101613]
Pristehovali	Range	[693,43053]
Vystehovali	Range	[682,18549]
Pocet_obyv	Range	[303051,1276384]
Zamestnanci_celkem	Range	[67347,414600]
Prum_hr_mes_mzda	Range	[11577,23735]
Mira_reg_nezamestnanosti	Range	[3.51,17.96]
Dokoncene_byty	Range	[86,8599]
Trzby_z_prumyslove_cinnosti	Range	[7909.4,508817.3]
Casova_rada	Set	"1.ctvrtleti","2.ctvrtleti","3.ctvrtleti","4.ctvrt...
Rok	Set	2001,2002,2003,2004,2005,2006,200...
Kraj	Set	Jihomoravský,Jihočeský,Karlovarský,K...

Obrázek č. 2: Uzel type a jednotlivé typy atributů [zdroj: vlastní]

Pro modelování nepoužívám atributy (časová řada "*Casova_rada*", rok "*Rok*" a kraj "*Kraj*".



Obrázek č. 3: Deskriptivní charakteristiky dat [zdroj: vlastní]

Obrázek č. 3 zobrazuje základní deskriptivní charakteristiky dat vybraných atributů: počet (*Count*), průměr (*Mean*), minimum (*Min*) a maximum (*Max*). Mezi další charakteristiky patří součet (*Sum*) a rozsah (*Range*), ale je třeba zvážit, které použiji, protože právě součet a rozsah se pro tato data nehodí. V případě míry registrované nezaměstnanosti dává součet (*Sum*) nesmyslnou hodnotu.

V tabulce č. 2 je zobrazen datový slovník. Jsou zde zobrazeny jednotlivé atributy a jejich vlastnosti. U každého atributu je definován jeho typ, rozsah, popis a výskyt. U atributů časová řada, rok a kraj je zvolen typ množina (*Set*), tento typ je využíván pro data s vícenásobnými odlišnými hodnotami v softwaru Clementine. Všem ostatním atributům je přiřazen typ range.

Atributy	Typ	Rozsah	Popis	Výskyt
Časová řada	Set	1. ctvrtleti, 2. ctvrtleti, 3.ctvrtleti, 4. ctvrtleti	Časová řada	2. ctvrtletí
Rok	Set	2001,…, 2009	Jednotlivé roky	2001
Kraj	Set	Středočeský, Jihočeský…..	Jednotlivé kraje	Středočeský
Rozloha v ha	Range	<316289;1101613>	Rozloha kraje	1 101 461
Přistěhovalí	Range	<693;43053>	Počet přistěhovalých	7 738
Vystěhovalí	Range	<682;18549>	Počet vystěhovalých	3 923
Počet obyvatel	Range	<303051; 1276384>	Počet obyvatel	630 353
Zaměstnanci celkem	Range	<67347;414600>	Celkový počet zaměstnanců	255 680
Průměrná hrubá měsíční mzda	Range	<11577;23735>	Průměrná měsíční mzda v jednotlivých krajích	13 689
Míra registrované nezaměstnanosti	Range	<3,5;17,9>	Vyjádření nezaměstnanosti v %	6,61
Dokončené byty	Range	<86;8599>	Počet dokončených bytů	580
Tržby z průmyslové činnosti	Range	<7909,4;508817,3>	Tržby za prodej vlastních průmyslových výrobků a služeb	77 712,3
Míra růstu mezd	Range	<-12,71;8,31>	Míra růstu mezd	1,35

Tabulka č. 2: Datový slovník [zdroj: vlastní]

4.1 Příprava dat pro modelování

Tato fáze je jedním nejdůležitějších kroků. Zároveň je to také nejpracnější a nejnáročnější část DM projektu. [5]

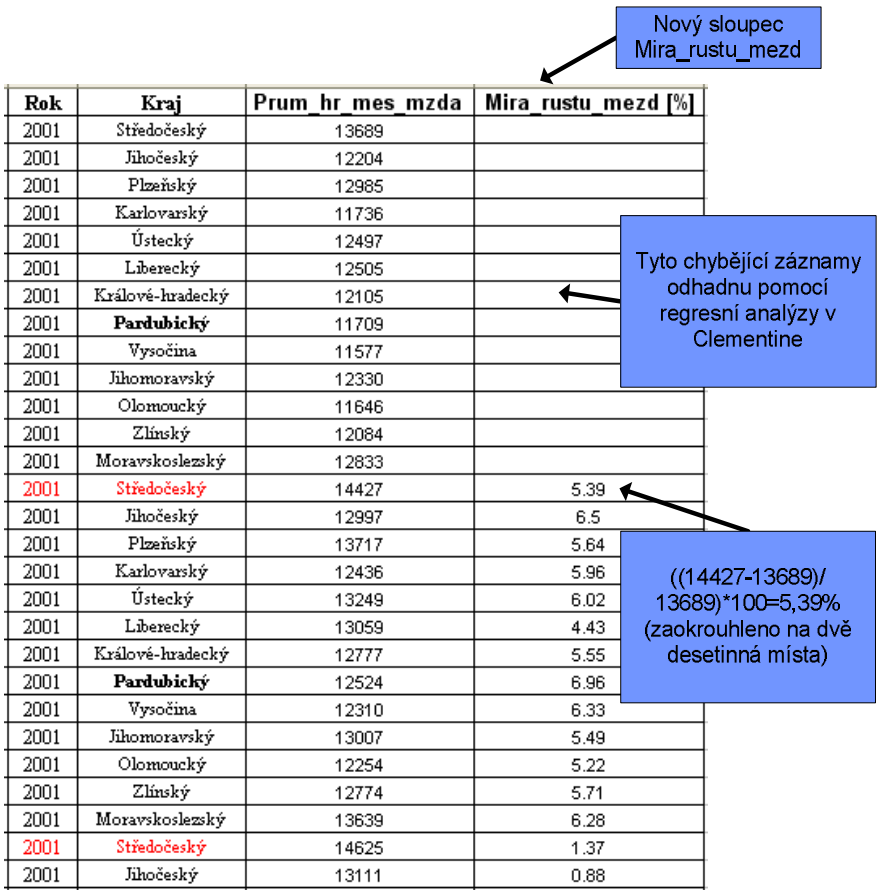
Příprava dat zahrnuje čistění dat, vytváření dat, formátování dat a selekci dat. Jednotlivé úkony jsou prováděny opakovaně.[4]

Celá tato fáze je modelována v programu Clementine od firmy SPSS. Pro úvodní posouzení kvality dat je možné využít uzel data audit.

4.1.1 Výběr vhodných atributů

Po konzultaci s odborníkem z Ústavu ekonomie byly vybrány následující tři skupiny dat pro modelování.

- Míra registrované nezaměstnanosti a průměrná hrubá měsíční mzda. Z průměrné hrubé měsíční mzdy vypočítám míru růstu mezd podle vzorce $(W_t - W_{t-1})/W_{t-1}$. Tuto hodnotu počítám pro každý kraj zvlášť. Na obrázku č. 4 vidím postup výpočtu nového sloupce míra růstu mezd (”*Mira_rustu_mezd*“) v programu MS Excel.



Rok	Kraj	Prum_hr_mes_mzda	Mira_rustu_mezd [%]
2001	Středočeský	13689	
2001	Jihočeský	12204	
2001	Plzeňský	12985	
2001	Karlovarský	11736	
2001	Ústecký	12497	
2001	Liberecký	12505	
2001	Královéhradecký	12105	
2001	**Pardubický**	11709	
2001	Vysočina	11577	
2001	Jihomoravský	12330	
2001	Olomoucký	11646	
2001	Zlínský	12084	
2001	Moravskoslezský	12833	
2001	Středočeský	14427	5.39
2001	Jihočeský	12997	6.5
2001	Plzeňský	13717	5.64
2001	Karlovarský	12436	5.96
2001	Ústecký	13249	6.02
2001	Liberecký	13059	4.43
2001	Královéhradecký	12777	5.55
2001	**Pardubický**	12524	6.96
2001	Vysočina	12310	6.33
2001	Jihomoravský	13007	5.49
2001	Olomoucký	12254	5.22
2001	Zlínský	12774	5.71
2001	Moravskoslezský	13639	6.28
2001	Středočeský	14625	1.37
2001	Jihočeský	13111	0.88

Obrázek č. 4: Výpočet atributu míra růstu mezd v MS Excel [zdroj: vlastní]

- Dokončené byty a průměrná hrubá měsíční mzda.
- Průměrná hrubá měsíční mzda, míra registrované nezaměstnanosti a tržby z průmyslové činnosti.

Dále budu pracovat pouze s těmito atributy, proto v obrázku č. 5 neuvádím celý datový soubor, ale pouze už 8 vybraných atributů. Přidal jsem k tomu ještě sloupec časová řada a rok.

4.1.2 Zjištění kvality dat

Pomocí uzlu data audit posoudím kvalitu dat. Podívám se na jednotlivé vybrané atributy. Na níže uvedeném obrázku č. 5 jsou zobrazené jednotlivé vybrané atributy. Když se podívám na poslední sloupec valid, tak vidím, že v atributech průměrná hrubá měsíční mzda a míra růstu mezd chybí nějaké hodnoty v datové matici. Celkem je v každém atributu 442 záznamů a v atributech průměrná hrubá měsíční mzda a míra růstu mezd je jich 429. Chybí zde 13 záznamů. Tyto chybějící záznamy sloupce průměrná hrubá měsíční mzda nebyly na webu Statistického úřadu zveřejněné. Ve sloupci míra růstu mezd chybějící hodnoty vznikly díky přepočtu z průměrné hrubé měsíční mzdy. Chybějící hodnoty je třeba pro další použití odhadnout. Dále přistoupím k metodám pro odhad chybějících hodnot.

Field	Graph	Type	Valid
Casova_rada		Set	442
Kraj		Set	442
Rok		Set	442
Prum_hr_mes_mzda		Range	429
Mira_reg_nezamestnanosti		Range	442
Dokoncene_byty		Range	442
Trzby_z_prumyslove_cinnosti		Range	442
Mira_rustu_mezd		Range	429

Obrázek č. 5: Analýza vybraných vstupních dat pomocí uzlu data audit [zdroj: vlastní]

Chybějící hodnoty jsou v softwaru Clementine reprezentovány hodnotou „$null$”. Na obrázku č. 6 jsou zobrazené chybějící hodnoty ve sloupci průměrná hrubá měsíční mzda, konkrétně pro druhé čtvrtletí roku 2002.

	Casova_rada	Rok	Kraj	Prum_hr_mes_mzda
61	1.ctvrtleti	2002	Vysočina	12516
62	1.ctvrtleti	2002	Jihomoravský	13108
63	1.ctvrtleti	2002	Olomoucký	12362
64	1.ctvrtleti	2002	Zlínský	12832
65	1.ctvrtleti	2002	Moravskoslezský	13792
66	2.ctvrtleti	2002	Středočeský	$null$
67	2.ctvrtleti	2002	Jihočeský	$null$
68	2.ctvrtleti	2002	Plzeňský	$null$
69	2.ctvrtleti	2002	Karlovarský	$null$
70	2.ctvrtleti	2002	Ústecký	$null$
71	2.ctvrtleti	2002	Liberecký	$null$
72	2.ctvrtleti	2002	Králové-hradecký	$null$
73	2.ctvrtleti	2002	Pardubický	$null$
74	2.ctvrtleti	2002	Vysočina	$null$
75	2.ctvrtleti	2002	Jihomoravský	$null$
76	2.ctvrtleti	2002	Olomoucký	$null$
77	2.ctvrtleti	2002	Zlínský	$null$
78	2.ctvrtleti	2002	Moravskoslezský	$null$
79	3.ctvrtleti	2002	Středočeský	15506

Obrázek č. 6: Chybějící hodnoty v atributu prům. hrubá měs. mzda [zdroj vlastní]

4.1.3 Ošetření chybějících hodnot

V data miningu se při sbírání a kombinování dat mohu dostat do situace, kdy mi budou některé záznamy chybět. Chybějící hodnoty se vyskytují snad v každé sadě dat. Řada softwarových nástrojů tyto hodnoty ignoruje a dělá z nich nesmyslné údaje.[4][6]

Cíl při nahrazování hodnot:

1. Zaplnit prázdná místa nejpravděpodobnější hodnotou.
2. Zachovat celkové rozdělení hodnot.

Na obrázku č. 5 je vidět, že datový soubor není úplný, ve sloupcích průměrná hrubá měsíční mzda a míra růstu mezd chybějí záznamy. S tímto datovým souborem nelze dále pracovat. Musím provést následující úpravy. Chybějící hodnoty musím ošetřit pomocí vhodné metody.

Nejjednodušší možností, jak doplnit chybějící hodnotu, je nahradit ji nejčastější hodnotou daného atributu. Existuje celá řada metod pro odhad chybějících hodnot. [4] [6]

1. Substituce jedné hodnoty
2. Substituce střední hodnoty třídy
3. Regresní substituce
4. Rozhodovací stromy

Zvolil jsem odhad pomocí regresní substituce. Tato metoda využívá střední hodnoty skupin jiných proměnných. *„Výhodou regrese je schopnost pracovat se spojitými proměnnými stejně jako hledat ve více proměnných přenější míru. Výsledné hodnocení regrese slouží k dopočtení náhradních hodnot.“* [6]

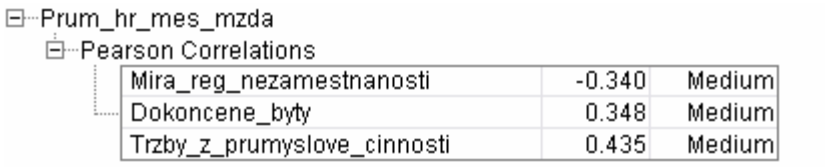
K odvození chybějících hodnot ve sloupcích průměrná hrubá měsíční mzda a míra růstu mezd je nutné použít atributy s nejvyšším stupněm korelace k danému atributu. Korelace vyjadřuje, do jaké míry jsou si dané atributy podobné. Korelační analýza zkoumá těsnost a sílu znaků. Počítá se pomocí korelačního koeficientu, který nabývá hodnot v intervalu <-1,1>.[8] Dále uvedu příklady těsnosti:

1. Do 0,2 je vztah zanedbatelný.
2. 0,2-0,4 je nepříliš těsný vztah.
3. 0,4-0,7 je středně těsný vztah.
4. 0,7-0,9 je velmi těsný vztah.
5. Více než 0,9 je extrémně těsný vztah.[9]

V této části se dále věnuji odhadu chybějících hodnot ve sloupci průměrná hrubá měsíční mzda a ve sloupci míra růstu mezd. Parametry nejsou odhadovány ze všech dat, ale pouze z dat, kreré používám pro modelování (obrázek č. 5).

Ke zjistění velikosti korelace použiji uzel statistics v softwaru Clementine. Tento uzel definuje Pearsonův korelační koeficient. Hodnoty od 0 do 0,333 vyjadřují slabý korelační vztah (*weak*), hodnoty od 0,333 do 0,666 vyjadřují střední korelační vztah (*medium*) a hodnoty od 0,666 do 1 vyjadřují silný korelační vztah (*strong*).[22]

Na obrázku č. 7 je vidět výstup z tohoto uzlu. Níže uvedený obrázek č. 7 zobrazuje velikost korelačního koeficientu mezi průměrnou hrubou měsíčnou mzdou a všemi ostatními atributy, které nemají žádnou chybějící hodnotu. Clementine posoudil u všech atributů střední závislost.

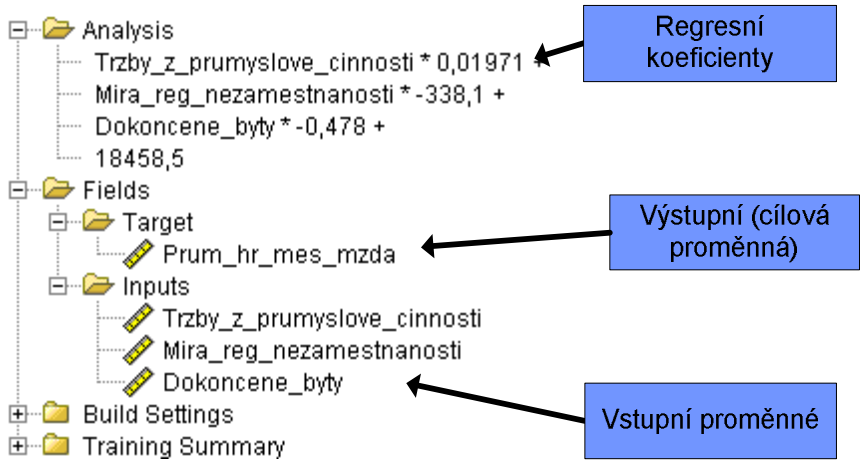


Obrázek č. 7: Korelační koeficient mezi průměrnou hrubou měsíční mzdou a dalšími atributy [zdroj: vlastní]

Dále musím podotknout, že daný atribut lze odhadovat pouze z atributů, které mají úplnou sadu záznamů (v daném sloupci nesmí chybět žádný záznam). Na obrázku č. 5 vidím, že záznamy chybí ve sloupcích průměrná hrubá měsíční mzda a míra růstu mezd.

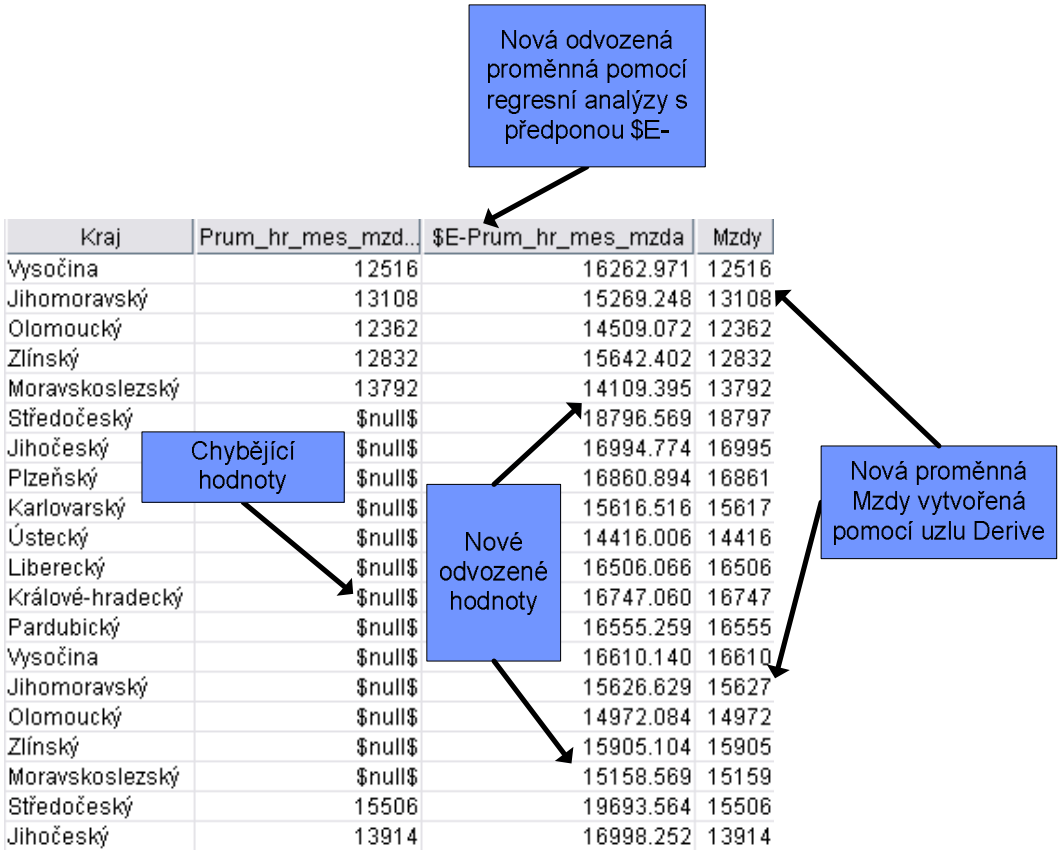
Podle obrázku č. 7 vyberu atributy s největším korelačním koeficientem k odhadu

chybějících hodnot ve sloupci průměrná hrubá měsíční mzda. Vybral jsem atributy míra registrované nezaměstnanosti, dokončené byty a tržby z průmyslové činnosti (vstupní proměnné na obrázku č. 8).



Obrázek č. 8: Parametry modelu Lineární regrese pro prům. hrubou měs. mzdu [zdroj: vlastní]

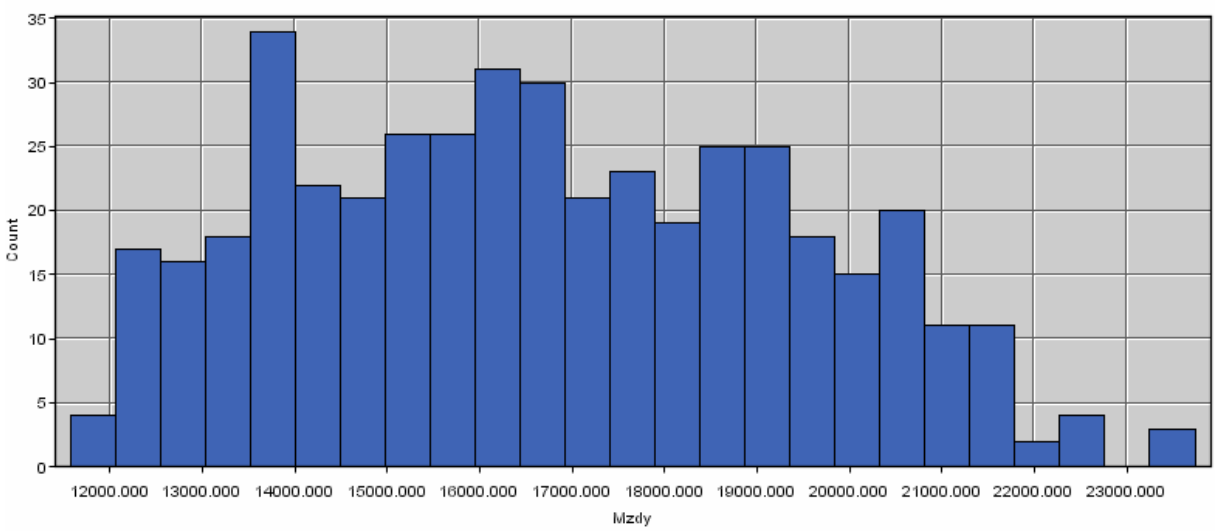
Na obrázku č. 8 jsou zobrazeny parametry uzlu regression. Udává tři vstupní proměnné a jednu výstupní proměnnou. Dále jsou vidět regresní koeficienty jednotlivých vstupních proměnných. Chybějící záznamy ve sloupci průměrná hrubá měsíční mzda jsem odhadnul pomocí tří vstupních proměnných (míra registrované nezaměstnanosti, dokončené byty a tržby z průmyslové činnosti).



Kraj	Prum_hr_mes_mzd..	$E-Prum_hr_mes_mzda	Mzdy
Vysočina	12516	16262.971	12516
Jihomoravský	13108	15269.248	13108
Olomoucký	12362	14509.072	12362
Zlínský	12832	15642.402	12832
Moravskoslezský	13792	14109.395	13792
Středočeský	$null$	18796.569	18797
Jihočeský	$null$	16994.774	16995
Plzeňský	$null$	16860.894	16861
Karlovarský	$null$	15616.516	15617
Ústecký	$null$	14416.006	14416
Liberecký	$null$	16506.066	16506
Králové-hradecký	$null$	16747.060	16747
Pardubický	$null$	16555.259	16555
Vysočina	$null$	16610.140	16610
Jihomoravský	$null$	15626.629	15627
Olomoucký	$null$	14972.084	14972
Zlínský	$null$	15905.104	15905
Moravskoslezský	$null$	15158.569	15159
Středočeský	15506	19693.564	15506
Jihočeský	13914	16998.252	13914

Obrázek č. 9: Tabulka s novými odvozenými hodnotami v atributu mzdy [zdroj: vlastní]

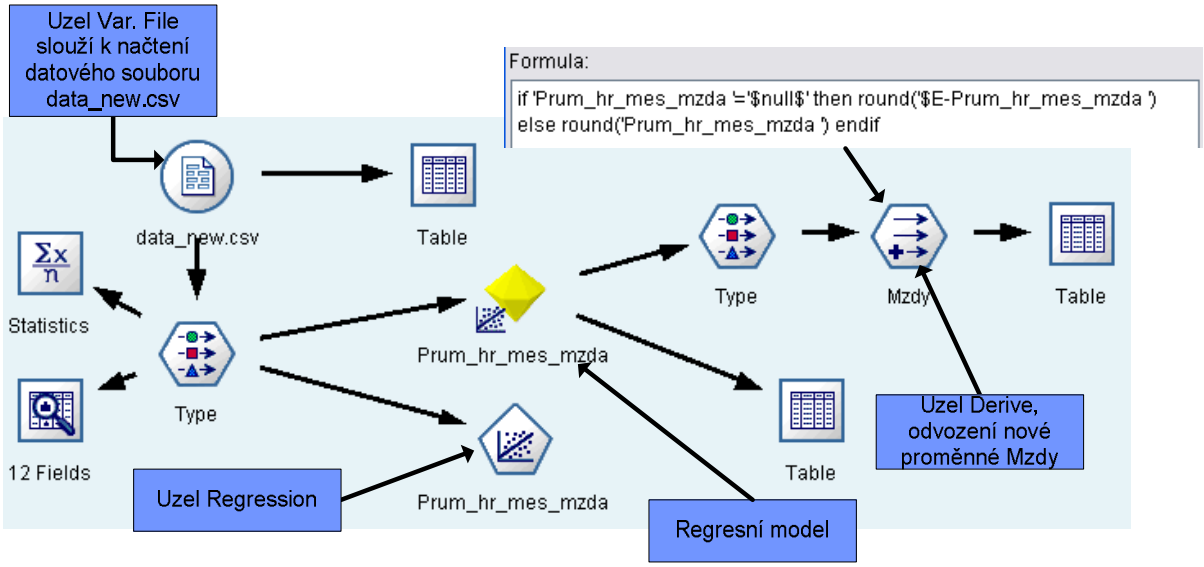
Nová odhadnutá předpovězená proměnná pomocí regresní substituce se jmenuje (“*$E-Prum_hr_mes_mzda*“) (zobrazuje obrázek č. 9). V tomto sloupci jsou pro hodnoty $null$ ze sloupce (“*prum_hr_mes_mzda”)* odvozené nové hodnoty. Atribut mzdy je nový sloupec, se kterým budu dále pracovat místo sloupce (“*prum_hr_mes_mzda*“).



Obrázek č. 10: Histogram průměrné hrubé měsíční mzdy [zdroj: vlastní]

Na obrázku č. 10 je zobrazen histogram sloupce průměrná hrubá měsíční mzda (“*Mzdy*“) s novými odhadnutými hodnotami. Mzda se pohybuje přibližně v rozsahu od 12 000 Kč do 23 000 Kč. Největší hodnoty dosahuje Středočeský a Jihomoravský kraj.

Na obrázku č. 11 je vidět celý model z programu Clementine pro odhad chybějících hodnot ve sloupci (“*prum_hr_mes_mzda*“). Uzel derive v sobě skrývá nadefinovanou funkci, která je zobrazená na obrázku č. 11 v rámečku nad uzlem derive s názvem mzdy. Funkce se skládá ze základní podmínky if-then-endif. Další použitá funkce round zaokrouhlí hodnoty na celá čísla. Do čísla pět zaokrouhlí dolů a od čísla šest nahoru (zobrazené na obrázku č. 9 výše). Uzlem derive vytvořím nový sloupec mzdy, ve kterém už budou nahrazeny chybějící hodnoty. Sloupec mzdy můžu považovat za uplný a bude použit k další analýze.



Obrázek č. 11: Model odhadu chybějících hodnot [zdroj: vlastní]

Na obrázku č. 12 je zobrazen korelační koeficient mezi vypočtenou mírou růstu mezd a ostatními atributy. Slabá závislost se objevila v atributech dokončené byty a tržby z průmyslové činnosti. Ve zbylých dvou atributech je nižší závislost (míra registrované nezaměstnanosti a mzdy).

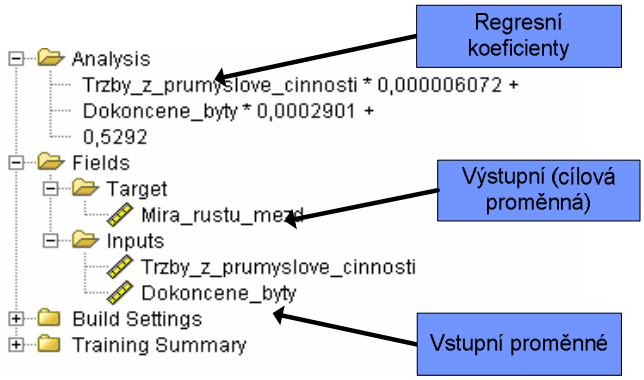
Mira_rustu_mezd

Pearson Correlations

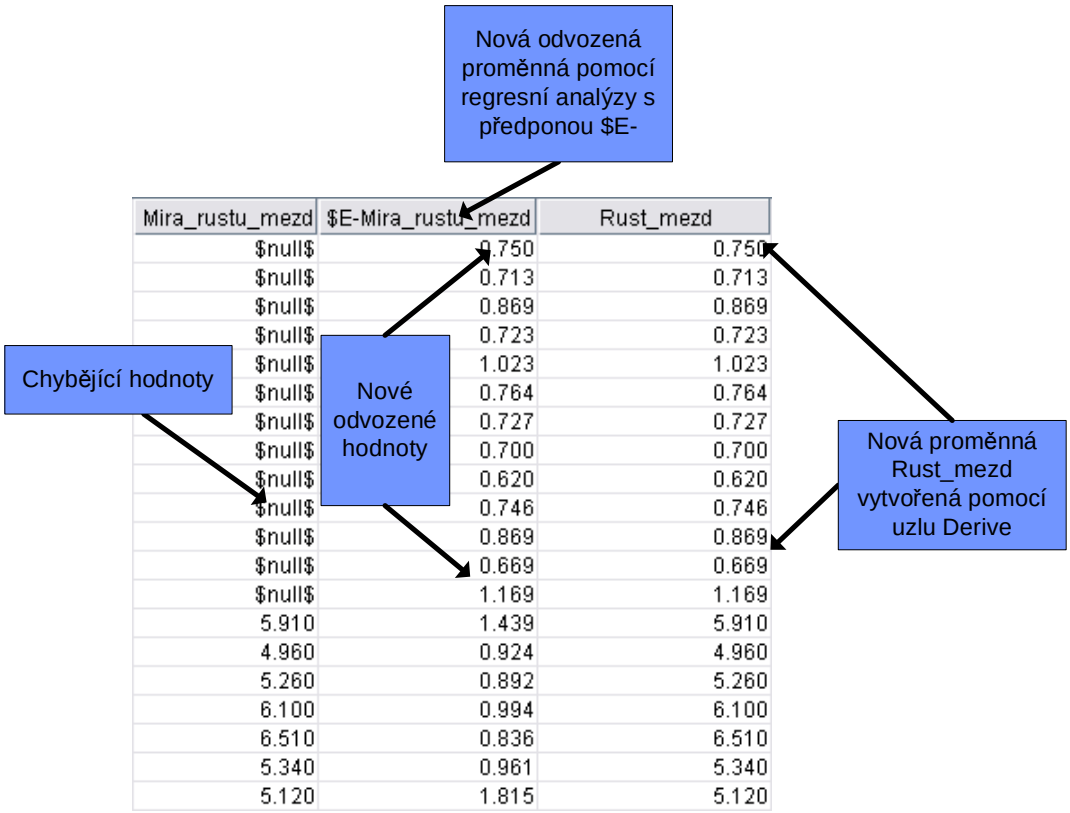
Mira_reg_nezamestnanosti	-0.054	Weak
Dokoncene_byty	0.218	Weak
Trzby_z_prumyslove_cinnosti	0.228	Weak
Mzdy	-0.036	Weak

Obrázek č. 12: Korelační koeficient mezi mírou růstu mezd a dalšími atributy [zdroj: vlastní]

Pro odhad chybějících záznamů pomocí regresní analýzy použiji atributy dokončené byty a tžby z průmyslové činnosti.



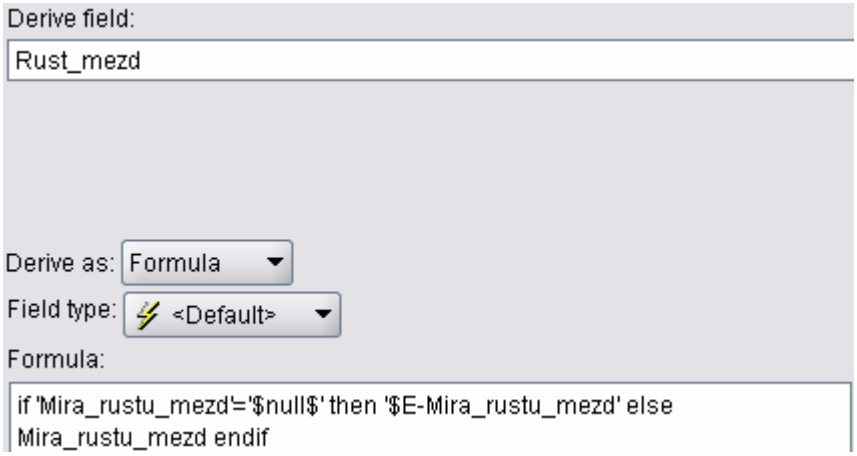
Obrázek č. 13: Parametry modelu Lineární regrese pro míru růstu mezd [zdroj: vlastní]



Mira_rustu_mezd	$E-Mira_rustu_mezd	Rust_mezd
$null$	0.750	0.750
$null$	0.713	0.713
$null$	0.869	0.869
$null$	0.723	0.723
$null$	1.023	1.023
$null$	0.764	0.764
$null$	0.727	0.727
$null$	0.700	0.700
$null$	0.620	0.620
$null$	0.746	0.746
$null$	0.869	0.869
$null$	0.669	0.669
$null$	1.169	1.169
5.910	1.439	5.910
4.960	0.924	4.960
5.260	0.892	5.260
6.100	0.994	6.100
6.510	0.836	6.510
5.340	0.961	5.340
5.120	1.815	5.120

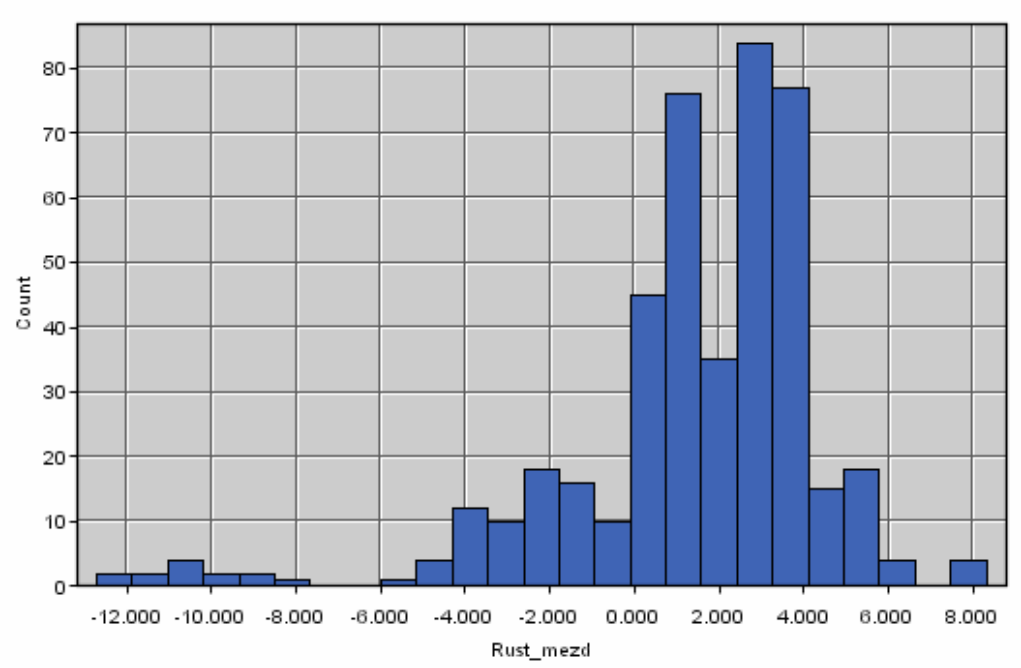
Obrázek č. 14: Tabulka s novými odvozenými hodnotami atributu růst mezd [zdroj: vlastní]

Na obrázku č. 14 je sloupec (“*$E-Mira_rustu_mezd*“), který vznikl odhadem pomocí regresní analýzy. Sloupec (“*rust_mezd*“) je nový sloupec, vytvořený pomocí uzlu derive. Tento sloupec je už kompletní a mohu s ním dále pracovat a použít ho v modelování.



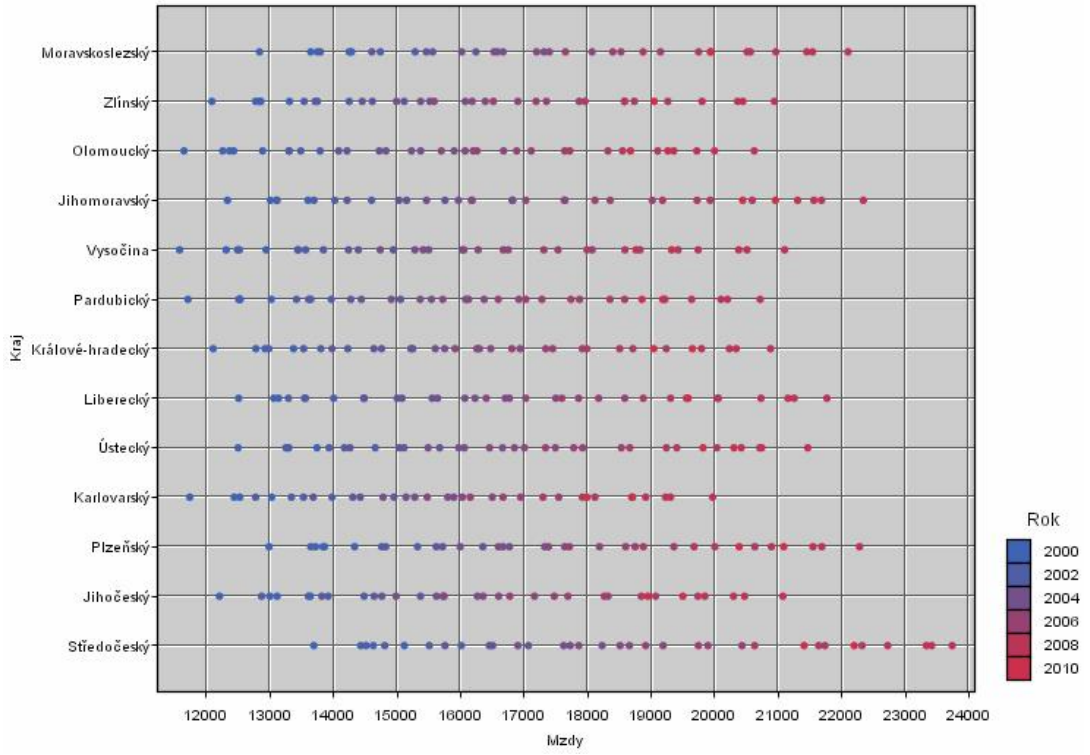
Obrázek č. 15: Nový atribut růst mezd [zdroj: vlastní]

Na obrázku č. 15 je nadefinovaná základní podmínka uzlu derive, pomocí níž jsem vytvořil nový sloupec (“*rust_mezd*“). Podmínka říká, vezmi všechny hodnoty z původního sloupce (“*mira_rustu_mezd*“), kromě těch, ve kterých je hodnota $null$ a zkopíruj je do nového sloupce (“*rust_mezd*“). Když je hodnota rovná $null$, tak kopíruj hodnoty z odhadnutého sloupce (“*$E-Mira_rustu_mezd*“).



Obrázek č. 16: Histogram růstu mezd [zdroj: vlastní]

Výše uvedený histogram (obrázek č. 16) zobrazuje sloupec růst mezd. Hodnota má docela velké rozpětí a často se mění. Je to způsobeno tím, že v jednotlivých krajích průměrná hrubá měsíční mzda hodně kolísá během každého roku a čtvrtletí (zobrazeno na obrázku č. 17). Určité zkreslení způsobil i odhad 13 chybějících hodnot pomocí regresní analýzy. Na obrázku č. 17 je vidět vývoj průměrné hrubé měsíční mzdy v jednotlivých krajích, stoupá od roku 2000 do roku 2010.



Obrázek č. 17: Vývoj mezd v jednolivých krajích [zdroj: vlastní]

Nyní mám nové odhadnuté chybějící hodnoty, data jsou kompletní a připravená pro modelování.

4.2 Vybrané metody shlukovací analýzy

„*V této fázi jsou nasazeny analytické metody (algoritmy pro dobývání znalostí).*“ Existuje celá řada metod pro řešení dané úlohy. „*Je třeba vybrat tu nejvhodnější. Doporučuje se použít více různých metod a jejich výsledky kombinovat.*“ [4]

Pro mou práci jsem zvolil shlukovací analýzu. Shlukování je dataminingová metoda, která se snaží v dané datové množině nalézt skupiny (shluky) objektů tak, aby si členové shluku byli navzájem podobní, ale na druhou stranu si nebyli podobní s objekty mimo tento shluk. [4][7][10]

Hledáme takové skupiny záznamů, které jsou si podobné (stejné zákazníky nebo dodavatele) a chovají se podobným způsobem. [7]

1. Metoda K-Means: Tato metoda je nejběžněji užívaná v praxi. Jedna z nejjednodušších metod učení bez učitele.[7][11]
2. Metoda TwoStep: Z názvu je možné odvodit, že se jedná o dvoufázovou shlukovací metodu. Je vhodná pro velké soubory dat.
3. Kohonenova mapa: Je známá jako self-organizing map (SOM), v překladu samoorganizující mapa. SOM je speciální druh neuronové sítě, který je použit k nalezení shluků.[7]

V tabulce č. 3 uvádím rozdíly jednotlivých shlukovacích metod. Tak jak jsou tyto algoritmy implementovány v Clementine 10.1.

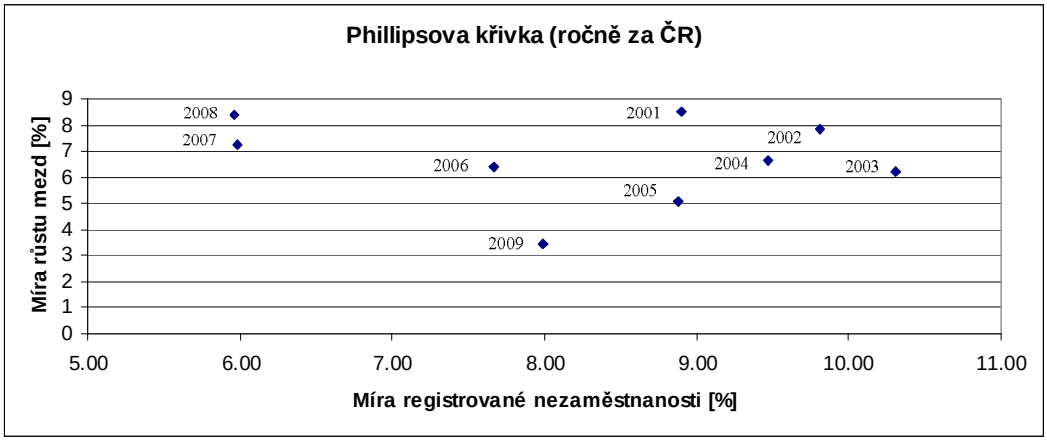
	Silné stránky	*Slabé stránky*
Kohonenova mapa	• pro spojité i kategorizované proměnné • nahradí chybějící hodnoty • jednoduchá	• výsledek závisí na subjektivní představě (navrhnout rozměry mapy)
K-Means	• jednoduchá a efektivní • nejrychlejší způsob shlukování pro velké datové soubory • chybějící hodnoty jsou nahrazeny hodnotou 0,5	• potřeba specifikovat počet shluků předem • citlivá na odlehlé hodnoty („outliers”)
TwoStep	• efektivně zvládá velké datové soubory • pro spojité i kategorizované proměnné • automaticky najde optimální počet shluků	• nepodporuje prázdná místa, chybějící hodnoty (vyloučí je)

Tabulka č. 3: Srovnání shlukovacích metod [22] [23]

4.3 Phillipsova křivka s reálnými daty

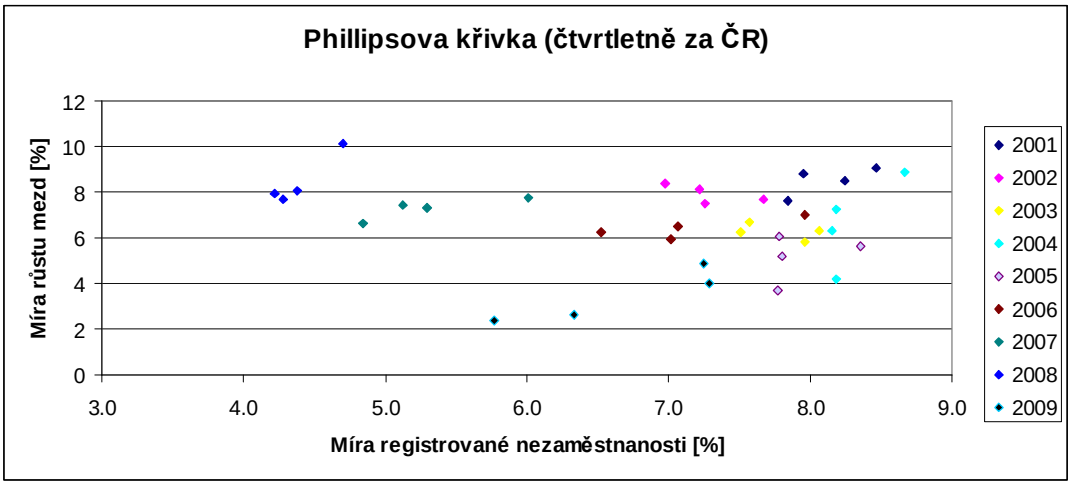
Uvedené tři grafy jsou sestrojeny podle Phillipsovy křivky na obrázku č. 1 na straně 9.

1. Roční data za Českou republiku. Graf č. 5 zobrazuje Phillipsovu křivku s ročními daty od roku 2001 do roku 2009 za ČR.



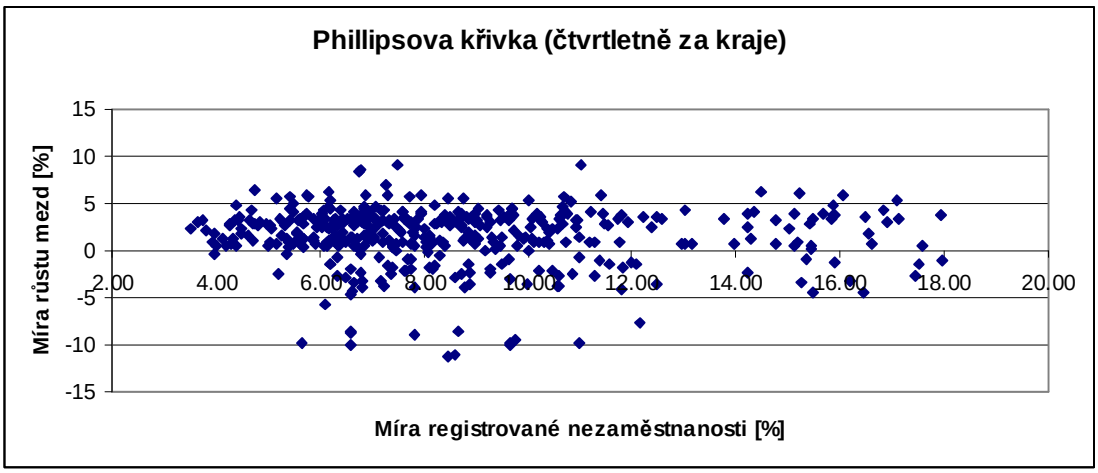
Graf č. 5: Phillipsova křivka (ročně za ČR) [zdroj: vlastní]

2. Čtvrtletní data za Českou republiku. Graf č. 6 zobrazuje Phillipsovu křivku se čtvrtletními daty od roku 2001 do roku 2009.



Graf č. 6: Phillipsova křivka (čtvrtletně za ČR) [zdroj: vlastní]

3. Čtvrtletní data za kraje. Graf č. 7 zobrazuje Phillipsovu křivku se čtvrtletními daty od roku 2001 do pololetí roku 2009.

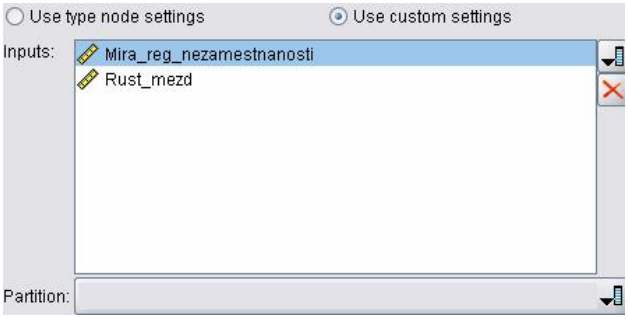


Graf č. 7: Phillipsova křivka (čtvrtletně za kraje) [zdroj: vlastní]

4.4 Modelování Phillipsovy křivky

V této kapitole se zaměřím na první vybranou skupinu atributů, na průměrnou hrubou měsíční mzdu a míru růstu mezd. Cílem je najít ve výsledných shlucích podobu Phillipsovy křivky.

Na níže uvedeném obrázku č. 18 vidím vstupní parametry Kohonenovy mapy (míra registrované nezaměstnanosti a růst mezd). Nedefinuji žádné cílové pole, poněvadž se tato metoda řadí mezi učení bez učitele.

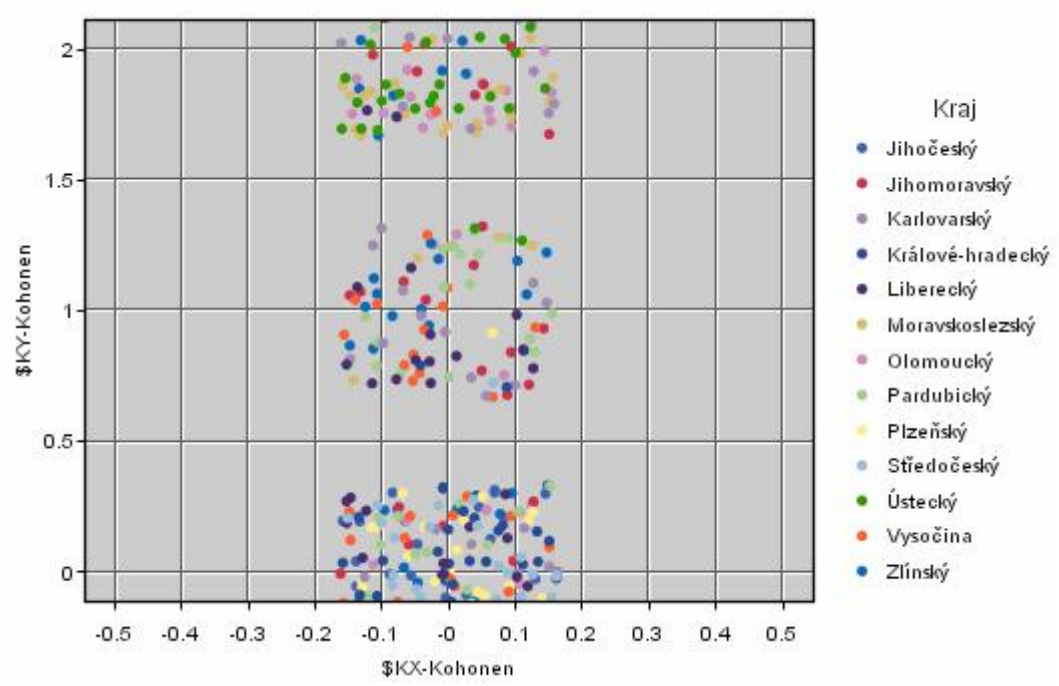


Obrázek č. 18: Nastavení Kohonenovy mapy [zdroj: vlastní]

Obrázek č. 19: Shluky Kohonenovy mapy pro nezaměstnanost a růst mezd [zdroj: vlastní]

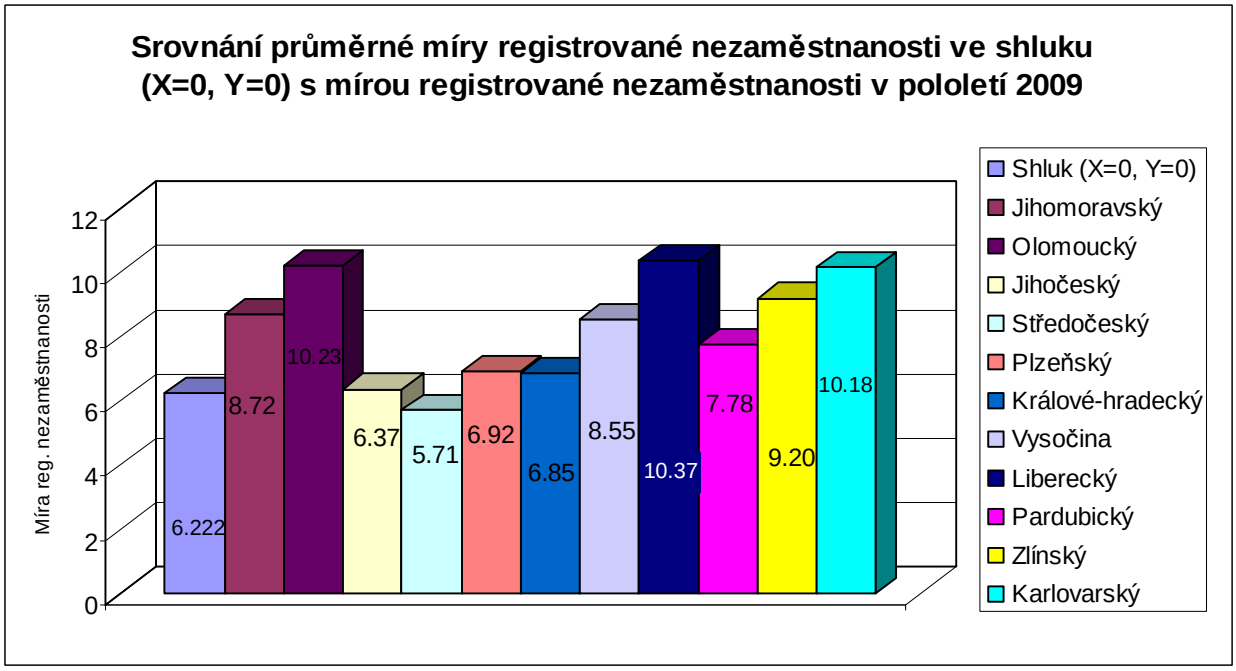
Obrázek č. 19 ukazuje jednotlivé shluky vytvořené pomocí Kohonenovy mapy. Tato metoda rozdělila data defaultně do tří shluků. Každý shluk je reprezentován x a y souřadnicemi. Nejvíce záznamů má první shluk o souřadnicích (X=0 Y=0). Z tohoto obrázku je možné dále vyčíst průměrnou hodnotu každého atributu, která je uvedena vždy v závorce. U shluku o souřadnicích X=0 a Y=0 je průměrná hodnota míry registrované nezaměstnanosti 6,222.

Na grafu v obrázku č. 20 vidím koncentraci jednotlivých shluků, které jsou určeny x a y souřadnicemi. Nejvíce záznamů je ve shlucích se souřadnicemi (X=0 a Y=0, X=0 a Y=2).



Obrázek č. 20: Shluky Kohonenovy mapy [zdroj: vlastní]

Shluk o souřadnicích X=0 a Y=0 má 219 záznamů, je zde vybráno 11 krajů (graf č. 8). Tento shluk obsahuje záznamy s průměrnou hodnotou míry registrované nezaměstnanosti 6,222 a průměrnou hodnotou růstu mezd 1,934 (zobrazeno na obrázku č. 19). Hodnota průměrné míry nezaměstnanosti je velmi nízká. Graf č. 8 srovnává průměrnou hodnotu míru registrované nezaměstnanosti ve shluku X=0 a Y=0 (6,222) s reálnými hodnotami míry nezaměstnanosti, které jsou uvedené v datové matici v MS Excel za pololetí 2009 pro uvedené kraje v grafu č. 8.



Graf č. 8: Srovnání nezaměstnanoti ve shluku (X=0 a Y=0) a v pololetí 2009 [zdroj: vlastní]

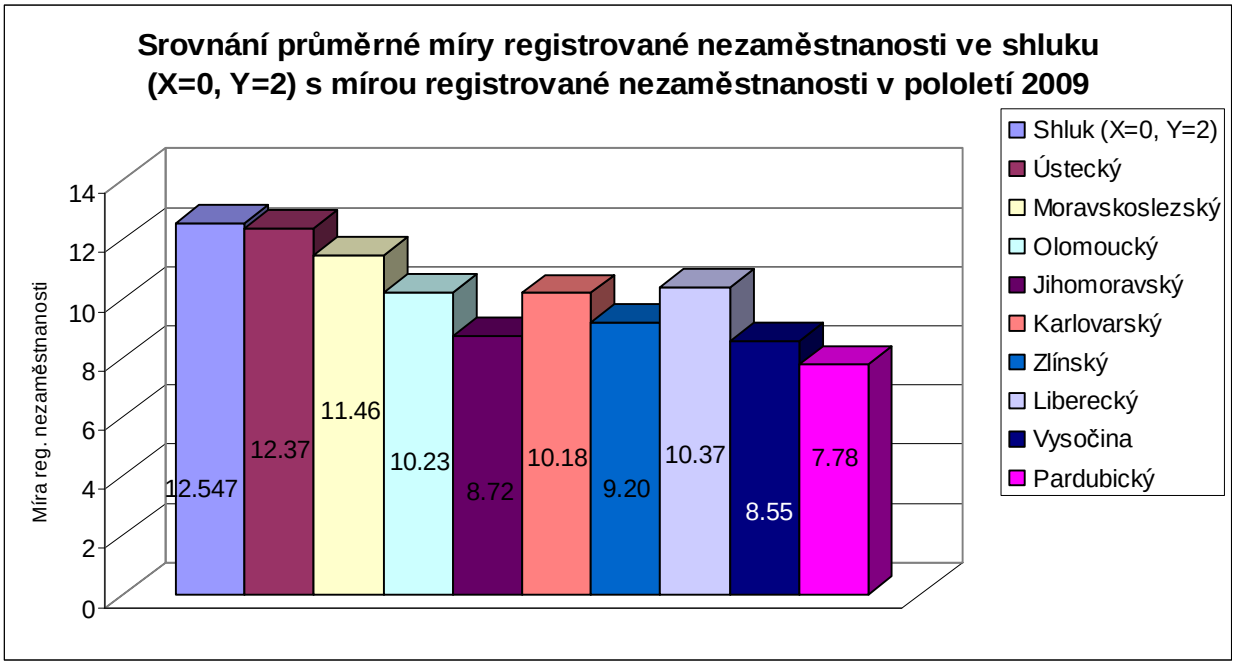
Obrázek č. 21 zobrazuje počet záznamů v jednotlivých krajích, které vybral skluk o souřadnicích X=0 a Y=0. Kraje Královéhradecký, Plzeňský, Středočeský a Jihočeský mají nejvyšší četnost záznamů (obrázek č. 21) a také se řadí mezi kraje s nejnižší nezaměstnaností podle grafu č. 8.

Kraj	Record_Count
Jihomoravský	7
Olomoucký	7
Karlovarský	8
Zlínský	11
Pardubický	17
Liberecký	18
Vysočina	19
Královéhradecký	32
Plzeňský	33
Středočeský	33
Jihočeský	34

Obrázek č. 21: Počet záznamů v krajích ve shluku (X=0 Y=0) [zdroj: vlastní]

V posledním shluku o souřadnicích X=0 a Y=2 je průměrná míra registrované nezaměstnanosti vysoká (12,547). Průměrný růst mezd je 0,82 (obrázek č. 19). Je zde vybráno 9 krajů (obrázek č. 22).

Graf č. 9 srovnává průměrnou hodnotu míru registrované nezaměstnanosti ve shluku X=0 a Y=2 (12,547) s reálnými hodnotami míry nezaměstnanosti, které jsou uvedené v datové matici v MS Excel za pololetí 2009 pro uvedené kraje v grafu č. 9.



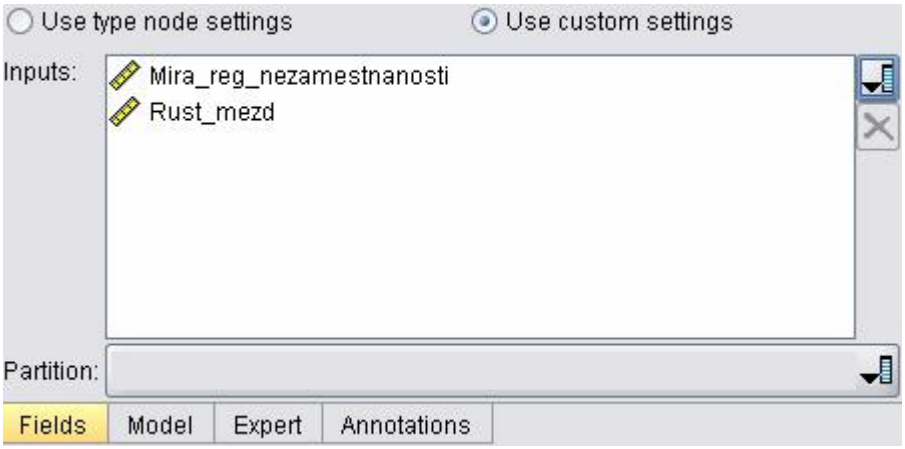
Graf č. 9: Srovnání nezaměstnanoti ve shluku (X=0 a Y=2) a v pololetí 2009 [zdroj: vlastní]

Obrázek č. 22 zobrazuje počet záznamů v jednotlivých krajích, které vybral skluk o souřadnicích X=0 a Y=2. Kraje Ústecký, Moravskoslezský a Olomouský mají nejvyšší četnost záznamů a také se řadí mezi kraje s nejvyšší nezaměstnaností podle grafu č. 9.

Kraj	Record_Count
Pardubický	1
Vysočina	2
Liberecký	3
Zlínský	9
Karlovarský	14
Jihomoravský	16
Olomoucký	23
Moravskoslezský	29
Ústecký	32

Obrázek č. 22: Počet záznamů v krajích ve shluku (X=0 Y=2) [zdroj: vlastní]

Metodu K-Means použiji také pro první vybranou skupinu atributů. Jedná se o atributy míra registrované nezaměstnanosti a míra růstu mezd. K-Means se řadí mezi metodu učení bez učitele, tudíž definuji pouze vstupní pole, není zde žádná cílová proměnná. Jsou zvoleny dva atributy (”*Mira_reg_nezamestnanosti*” a ”*Rust_mezd*”) na obrázku č. 23.



Obrázek č. 23: Nastavení metody K-Means [zdroj: vlastní]

Na záložce model v nastavení uzlu K-Means je možné nastavit, do kolika shluků budou data rozdělena. Před spuštěním této metody bych měl nastavit počet shluků, defaultně je nastaveno pět. Nastavil jsem 3 shluky, jelikož Kohonenova mapa rozdělila data také do třech shluků. K-Means je metoda s pevným počtem shluků.

Obrázek č. 24: Shluky metody K-Means [zdroj: vlastní]

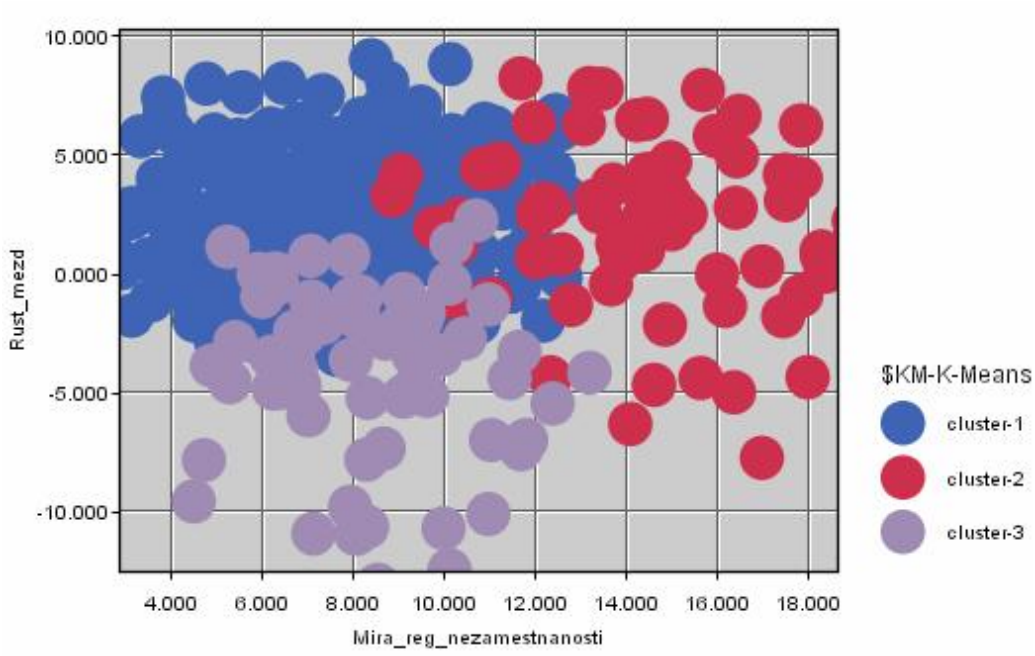
Metoda K-Means rozdělila atributy do třech shluků. Nejvíce záznamů je ve shluku 1 (cluster-1), celkem 306 záznamů (obrázek č. 24), ale tento shluk vybral všechny kraje stejně tak jako shluk 3 (cluster-3). Zajímavějším shlukem je shluk 2 (cluster-2), který vybral 5 krajů zobrazených na obrázku č. 25.

Kraj	Record_Count
Karlovarský	1
Jihomoravský	4
Olomoucký	12
Moravskoslezský	26
Ústecký	29

Obrázek č. 25: Počet záznamů v krajích ve shluku 2 (cluster-2) [zdroj: vlastní]

Obrázek č. 25 zobrazuje počet záznamů v jednotlivých krajích ve shluku 2 (cluster-2).

V tomto shluku jsou vybrány kraje s vysokou průměrnou mírou registrované nezaměstnanosti (pro srovnání graf č. 9). Průměrná hodnota míry nezaměstnanosti je 14,256 % a průměrná hodnota míry růstu mezd je 1,896 % (zobrazeno na obrázku č. 24).



Obrázek č. 26: Shluky K-Means v závislosti míry nezaměstnanosti na růstu mezd [zdroj: vlastní]

Na obrázku č. 26 jsou vidět jednotlivé shluky K-Means v závislosti míry registrované nezaměstnanosti na růstu mezd.

Metoda TwoStep stejně jako předchozí metody nepoužívá cílové pole. TwoStep je dvoufázová shlukovací metoda.

Obrázek č. 27: Shluky metody TwoStep [zdroj: vlastní]

Kraj	Record_Count
Olomoucký	5
Moravskoslezský	24
Ústecký	27

Obrázek č. 28: Počet záznamů v krajích ve shluku 3 (cluster-3) [zdroj: vlastní]

V nastavení TwoStep jsem nadefinoval také rozdělení do tří shluků (obrázek č. 27). Tato metoda zařadila do třetího shluku kraje s vysokou mírou nezaměstnaností (Olomoucký, Moravskoslezský a Ústecký) – obrázek č. 28. To odpovídá průměrné míře nezaměstnanosti v tomto shluku 15,086.

4.4.1 Zhodnocení všech tří metod

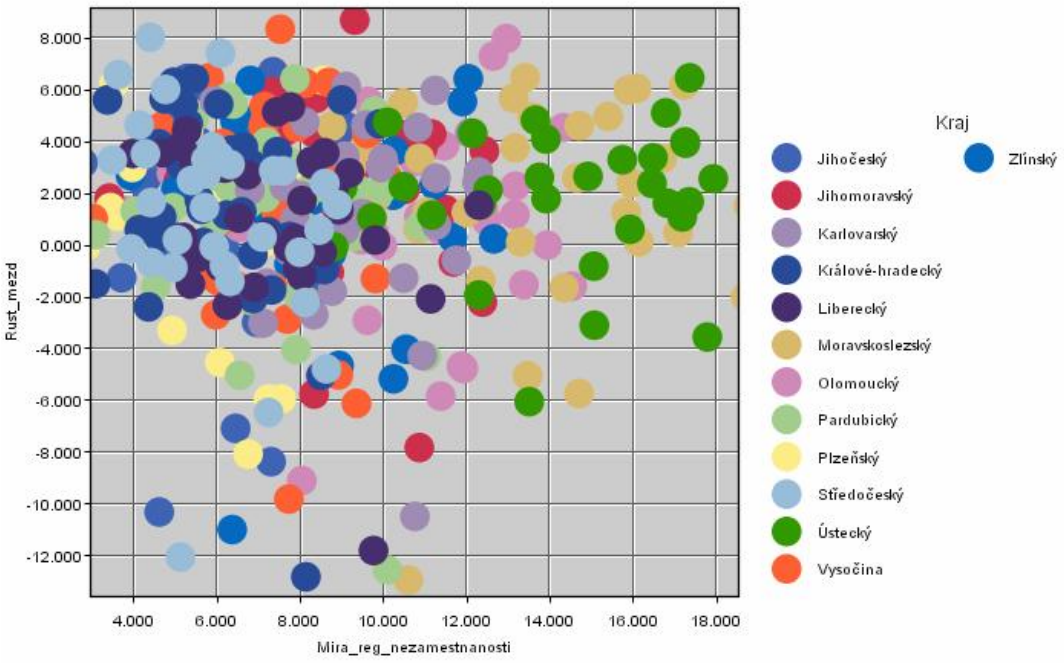
Phillipsovu křivku nepotvrdila ani jedna z metod shlukovací analýzy. Důvodem může být odlišnost zkonstruování této křivky a princip shlukovací metody. Grafická závislost míry registrované nezaměstnanosti a míry růstu mezd u K-Means je zobrazena na obrázku č. 26. Uvedený graf neodpovídá grafu Phillipsovy křivky.

4.5 Experimentování s modelem Phillipsovy křivky

V této kapitole se budu snažit najít podobu Phillipsovy křivky. Budu zkoumat závislost mezi mírou nezaměstnanosti a mírou růstu mezd. Uvedenou závislost budu hledat:

1. Globálně pro celou pro celou Českou republiku bez Prahy.
2. Pro Kohonenovu mapu s nastavením pěti shluků.
3. Pro Kohonenovu mapu s nastavením sedmi shluků.

Závislost mezi mírou růstu mezd a mírou registrované nezaměstnanosti za celou ČR ilustruje obrázek č. 29 uvedený níže.

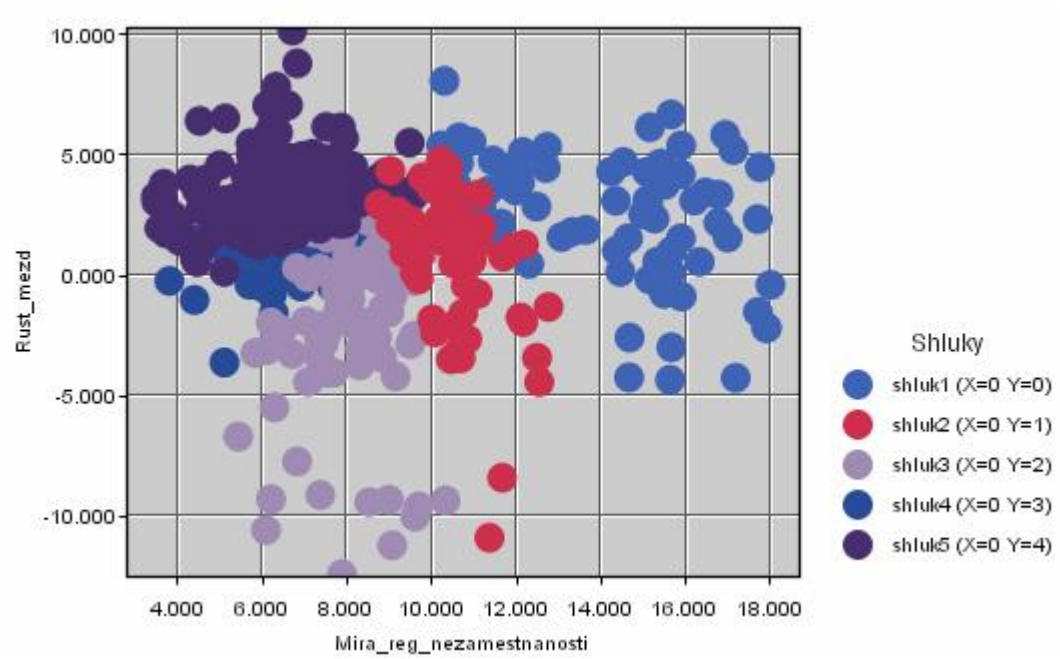


Obrázek č. 29: Závislost míry růstu mezd a míry nezaměstnanosti za ČR [zdroj: vlastní]

Na obrázku č. 29 jsem podobu Phillipsovy křivky nenašel. Jsou zde vidět jednotlivé kraje s závislostí mezi mírou růstu mezd a mírou registrované nezaměstnanosti. Z výše uvedeného grafu je možné vyčíst, že Ústecký, Moravzkoslezský a Olomoucký kraj má nejvyšší nezaměstnanost.

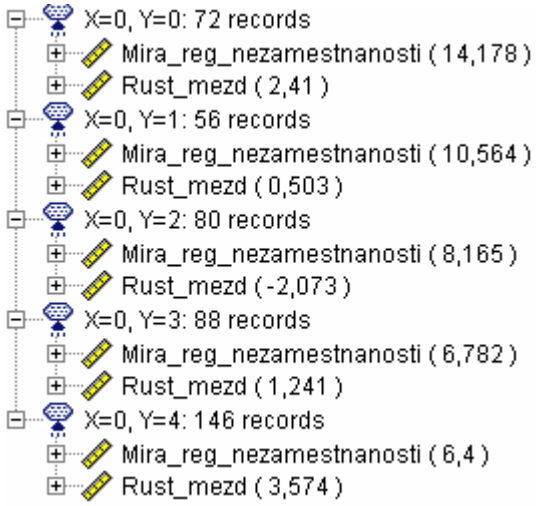
4.5.1 Kohonenova mapa s pěti shluky

Dále jsme použil Kohonenovu mapu a v nastavení metody jsem v záložce expert nastavil délku sítě na 5 a šířku na jedna. Uvedená metoda rozdělí data do pěti shluků. Na obrázku č. 30 je vidět grafický výsledek vztahu míry růstu mezd a míry registrované nezaměstnanosti pro pět shluků. Podoba Phillipsovy křivky nebyla nalezena. Nejvíce záznamů je ve shluku 5 (X=0 a Y=4), zobrazeno na obrázku č. 31.



Obrázek č. 30: Závislost míry růstu mezd a míry nezaměstnanosti pro tři shluky [zdroj: vlastní]

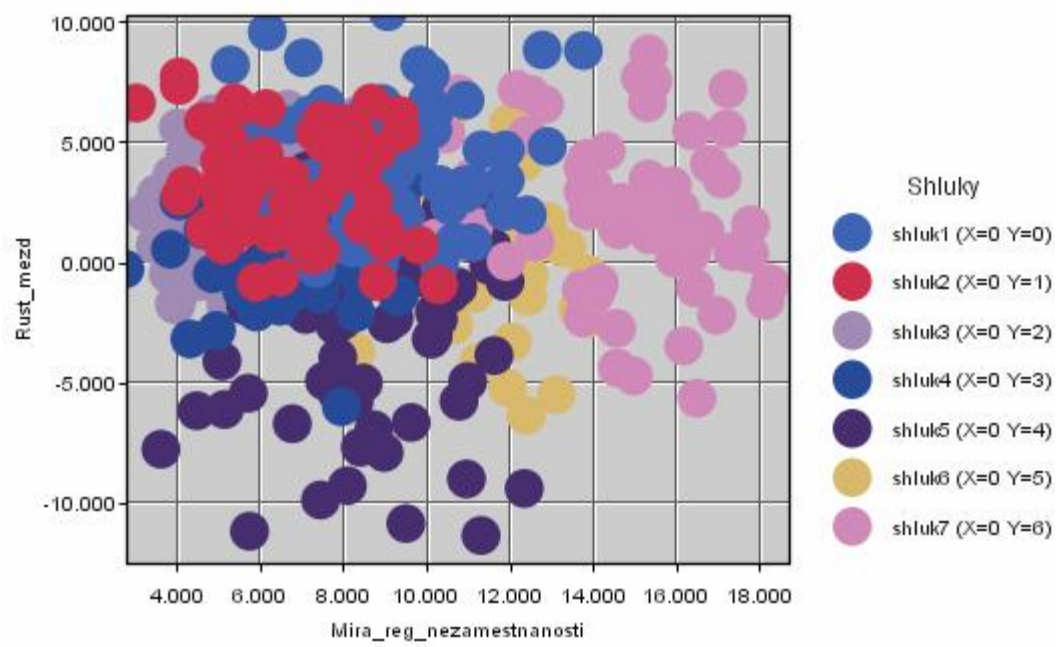
Na obrázku č. 31 vidím pět shluků Kohonenovy mapy. V prvním shluku (X=0 a Y=0) je 72 záznamů s vysokou průměrnou hodnotou míry registrované nezaměstnanosti 14,178 a průměrnou hodnotou míry růstu mezd 2,41. Ve shluku pět (X=0 a Y=4) je vybráno 146 záznamů a disponuje nízkou průměrnou mírou nezaměstnanosti 6,4.



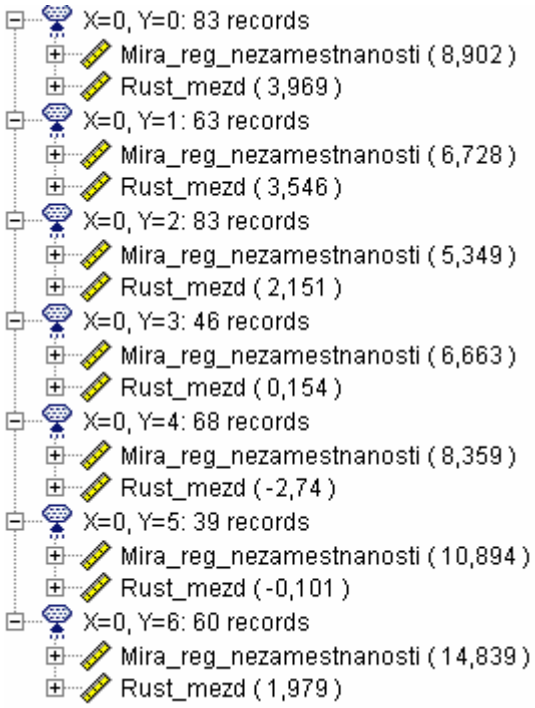
Obrázek č. 31: Shluky Kohonenovy mapy pro míru nezaměstnanosti a růst mezd [zdroj: vlastní]

4.5.2 Kohonenova mapa se sedmi shluky

Phillipsova křivka se ani v tomto případě nepotvrdila. Na níže uvedém obrázku č. 32 vidím závislost míry registrované nezaměstnanosti a růstu mezd, která je ilustrována sedmi shluky (obrázek č. 33). Mezi shluky s nejvyšší průměrnou mírou nezaměstnaností patří shluk 6 (X=0 a Y=5) a shluk 7 (X=0 a Y=6), zobrazené na obrázku č. 33.



Obrázek č. 32: Závislost míry růstu mezd a míry nezaměstnanosti pro 7 shluků [zdroj: vlastní]



Obrázek č. 33: Shluky Kohonenovy mapy pro míru nezaměstnanosti a růst mezd [zdroj: vlastní]

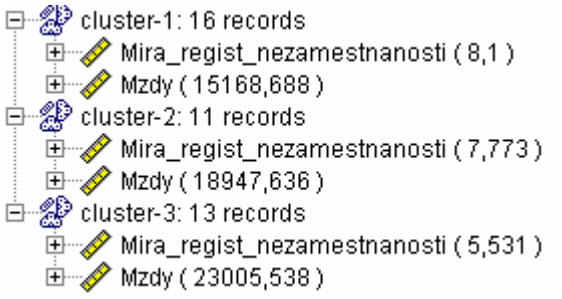
4.6 Modelování závislosti průměrné hrubé měsíční mzdy a míry registrované nezaměstnanosti

Původní mzdová Phillipsova křivka ilustruje závislost mezi mírou nezaměstnanosti a mírou růstu peněžních mzdových sazeb pro časovou řadu od roku 2000 do roku 2009. Parametr míra růstu mezd jsem spočítal z průměrné hrubé měsíční mzdy podle vzorce $(W_t - W_{t-1})/W_{t-1}$.[13]

Po prozkoumání dané časové řady, jsem dospěl k názoru, že můj vytvořený graf se nepodobá Phillipsově křivce. Zkoušel jsem na osu Y zobrazit ukazatel průměrná hrubá měsíční mzda místo uvedené míry růstu mezd a na ose X zůstala původní míra registrované nezaměstnanosti, ale rovněž jsem podobu Phillipsovy křivky nenašel.

Dále se zaměřím na vytvoření závislosti mezi průměrnou hrubou měsíční mzdou a mírou registrované nezaměstnanosti na naměřených datech v časové řadě po čtvrtletí od roku 2000 do roku 2009.

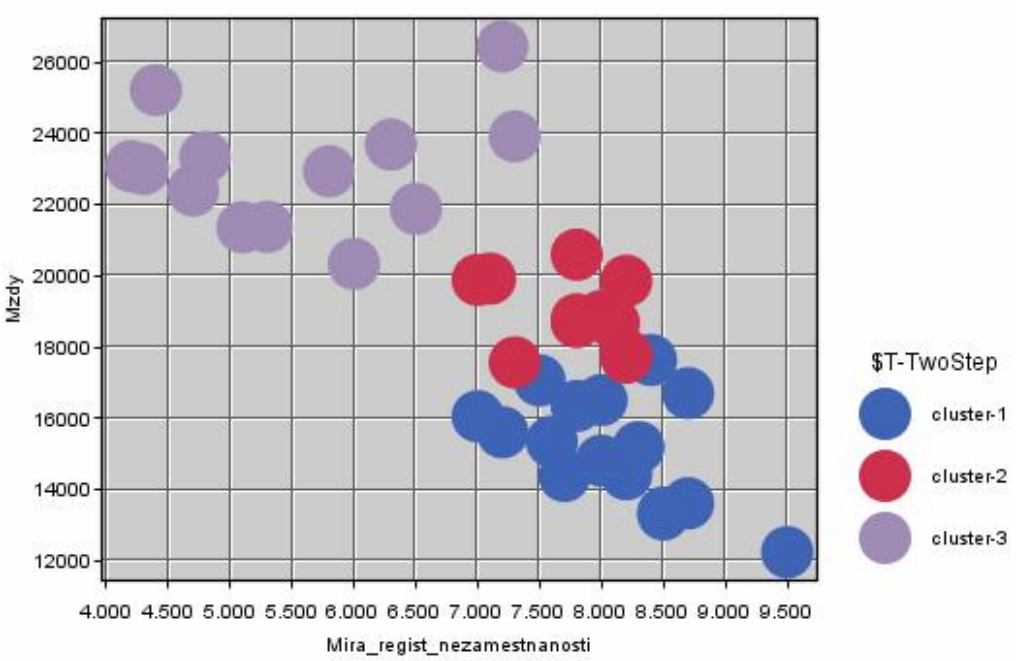
Metoda TwoStep dala nejlépe interpretovatelné výsledky ze všech tří metod shlukové analýzy. Na obrázku č. 34 jsou zobrazeny jednotlivé shluky.



Obrázek č. 34: Shluky metody TwoStep pro míru nezaměstnanosti a mzdy [zdroj: vlastní]

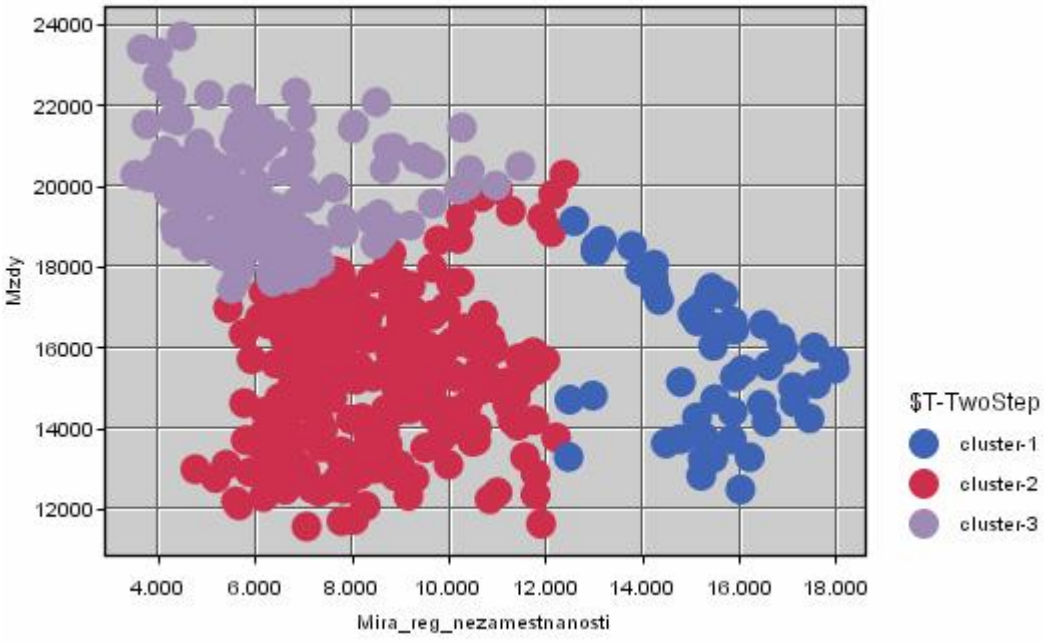
Dále si pro přehlednost a lepší orientaci zobrazím tyto shluky do grafické podoby v závislosti míry registrované nezaměstnanosti a průměrné hrubé měsíční mzdy, která je interpretována atributem mzdy.

Pokud bych proložil jednotlivé body v rámci grafu (obrázek č. 35) logaritmickou křivkou bylo by možné při zjednodušení ilustrovat uvedenou křivku závislosti. [17]



Obrázek č. 35: Závislost míry nezaměstnanosti a průměrné hrubé mzdy u metody TwoStep [zdroj: vlastní]

Dále se zaměřím na zkonstruování křivky závislosti mezi průměrnou hrubou měsíční mzdou a mírou registrované nezaměstnanosti podle jednotlivých krajů. Průměrná hrubá měsíční mzda je v Clementine pod názvem mzdy. Nejlépe ze shlukovací analýzy vyšla metoda TwoStep, do které vstupují míra registrované nezaměstnanosti a mzdy (průměrná hrubá měsíční mzda). Data na obrázku č. 36 zobrazují uvedenou křivku závislosti.



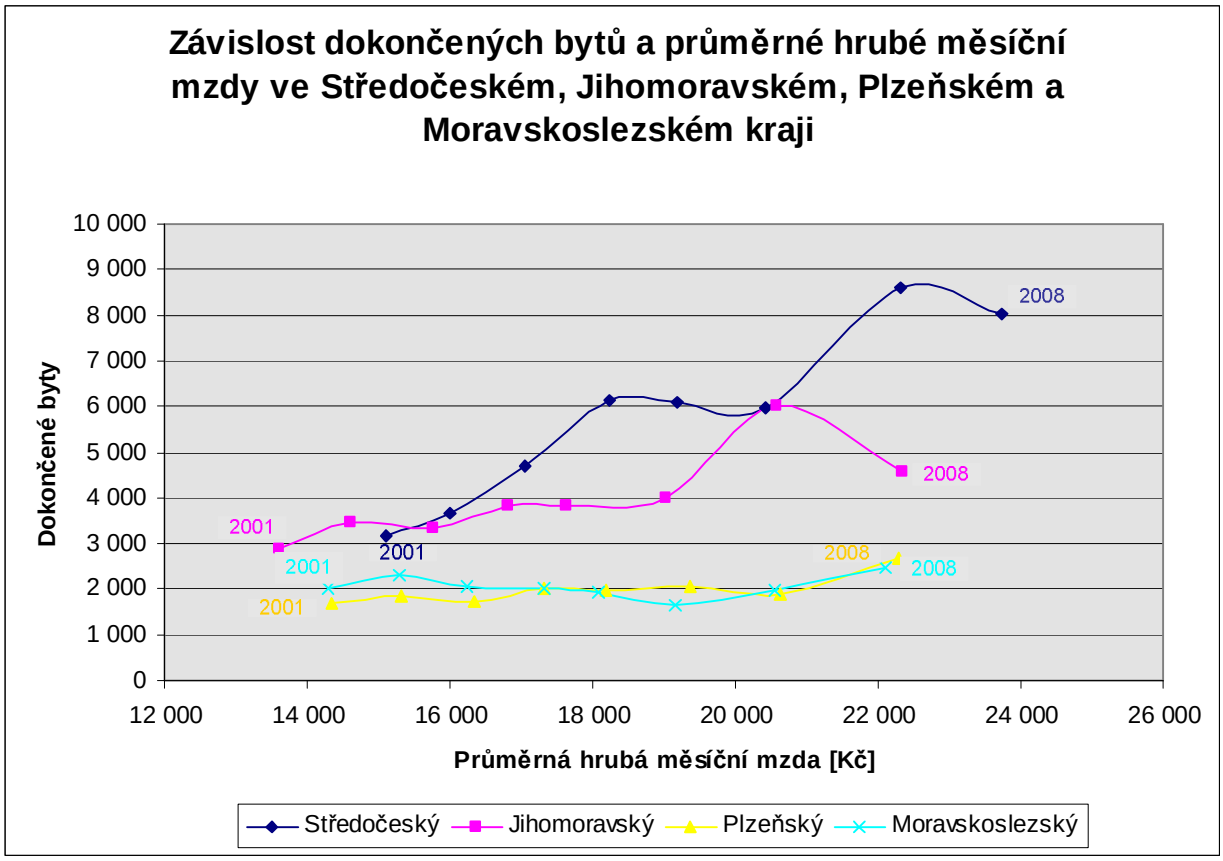
Obrázek č. 36: Závislost míry nezaměstnanosti a mzdy u metody TwoStep [zdroj: vlastní]

4.7 Analýza průměrné hrubé měsíční mzdy a dokončených bytů

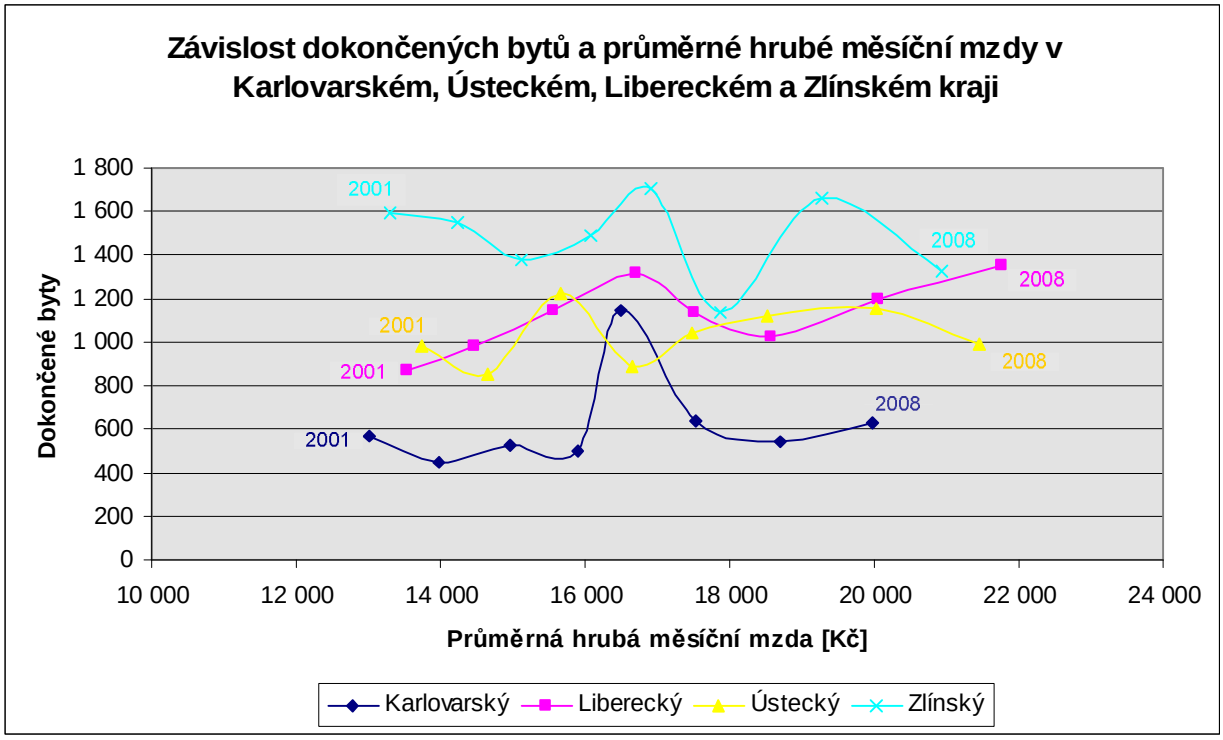
V této kapitole budu testovat další skupinu záznamů uvedenou v kapitole 4.1.1. Budu hledat závislost mezi atributy průměrná hrubá měsíční mzda a počet dokončených bytů v jednotlivých krajích. Použiji shlukovací analýzu.

Poptávka po nemovitostech a po dokončených bytech závisí a je determinována příjmem disponibilního důchodu, jehož hlavní složkou je výše mzdy jednotlivých domácností. Tato složka ovlivňuje jak bohatství domácností, tak dostupnost jednotlivých úvěrů a jejich možnost splácení. [12]

Na grafu č. 10 jsou zobrazeny kraje s nejvyššími hodnotami průměrné hrubé měsíční mzdy a počtu dokončených bytů. Patří sem Středočeský, Jihomoravský, Plzeňský a Moravskoslezský kraj. Mezi maximální hodnoty se řadí Středočeský kraj následovaný Jihomoravským. Uvedené grafy odpovídají časové řadě od roku 2001 do roku 2008 (vždycky je bráno 4. čtvrtletí).



Graf č. 10: Závislost počtu dokončených bytů a průměrné hrubé mzdy v uvedených krajích [zdroj vlastní]



Graf č. 11: Závislost počtu dokončených bytů a průměrné hrubé mzdy v uvedených krajích [zdroj vlastní]

Na výše uvedém grafu č. 11 vidíme opak předcházejícího grafu. Jsou zde zobrazeny kraje s nejnižšími hodnotami počtu dokončených bytů a průměrné hrubé měsíční mzdy. Řadí se sem Karlovarský, Ústecký, Liberecký a Zlínský kraj. Zajímavostí v tomto grafu je vztah Libereckého a Karlovarského kraje. V počtu dokončených bytů mají srovnatelné hodnoty, ale průměrná hrubá měsíční mzda je v Libereckém kraji vyšší zhruba o 1 700 Kč (zobrazeno na obrázku č. 38).

Field	Graph	Type	Min	Max
Dokoncene_byty		Range	86	8599
Mzdy		Range	11577	23735

Obrázek č. 37: Atributy rok, dokončené byty a mzdy [zdroj vlastní]

Na obrázku č. 37 je možné vidět maximální a minimální hodnoty vybraných atributů a jejich typ. Tyto charakteristiky použiji pro srovnání s hodnotami záznamů v jednotlivých shlucích.

Kraj	Rozloha_v_ha_Min	Rozloha_v_ha_Max	Dokoncene_byty_Min	Dokoncene_byty_Max	Mzdy_Min	Mzdy_Max
Středočeský	1101442	1101613	580	8599	13689	23735
Liberecký	316289	316312	100	1355	12505	21763
Ústecký	533425	533503	155	1226	12497	21462
Královéhradecký	475817	475853	184	1919	12105	20877
Karlovarský	331433	331457	86	1149	11736	19967
Pardubický	451846	451867	240	1866	11709	20713
Vysočina	679547	692541	256	1729	11577	21098
Plzeňský	756089	756123	263	2663	12985	22277
Moravskoslezský	542645	555441	341	2453	12833	22096
Zlínský	396349	396405	173	1706	12084	20937
Jihomoravský	706545	719650	559	6013	12330	22337
Olomoucký	513943	526690	220	1839	11646	20619
Jihočeský	1005634	1005731	273	2708	12204	21070

Obrázek č. 38: Maximální a minimální hodnoty vybraných atributů [zdroj vlastní]

Výše uvedená tabulka na obrázku č. 38 ukazuje maximální a minimální hodnoty atributů rozloha, dokončené byty a mzdy, která je rozčleněná podle jednotlivých krajů. Přidal jsem zde sloupec rozloha, protože si myslím, že na počtu dokončených bytů závisí také velikost rozlohy kraje. Není tomu tak v každém případě, když se kouknu na uvedenou tabulku, ale většina krajů tomu odpovídá. Příkladem je Středočeský kraj, který má největší rozlohu a zároveň má největší počet dokončených bytů. Jihočeský kraj má také velkou rozlohu, ale počet dokončených bytů tomu neodpovídá. Podle mě na osídlení má vliv řada jiných faktorů, jako jsou přírodní vlivy, uspořádání krajiny nebo podmínky pro život. Jihočeský kraj je těmito faktory hodně ovlivněn a je znám nízkou hustotou osídlení.

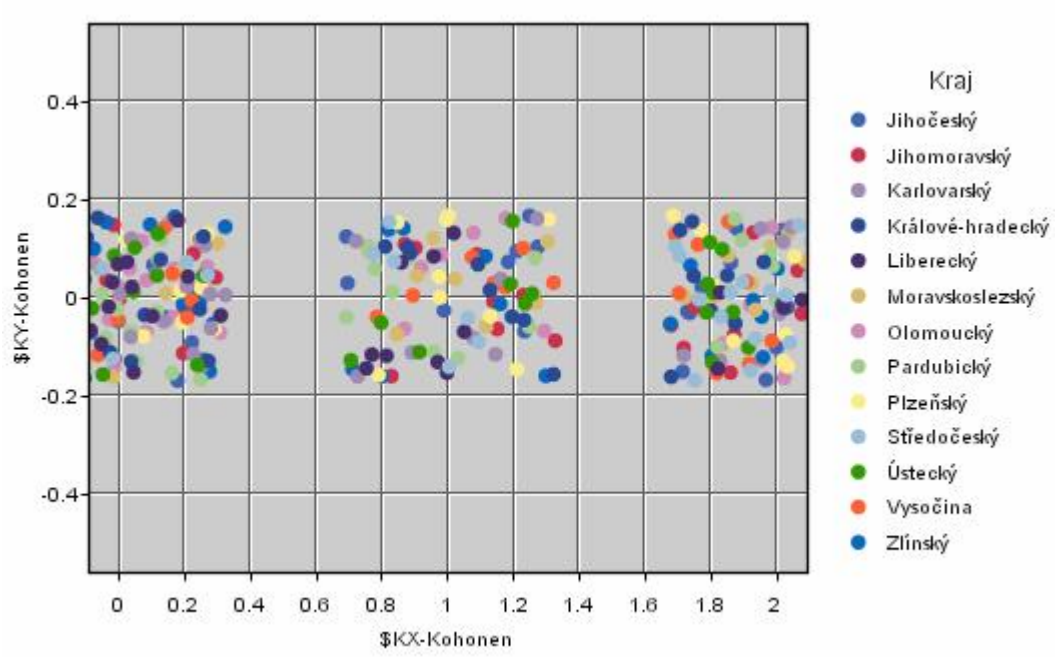
K nastavení počátečních atributů pro metody shlukovací analýzy použiji uzel type. Vstupní atributy mají ve sloupci direction nastavenou hodnotu „*In*".

Casova_rada	Set	"1.ctvrtleti","2.ctvrtleti","3.ctvrtleti","4.ctvrtleti"	None	None
Kraj	Set	Jihomoravský,Jihočeský,Karlovarský,Královéhra...	None	None
Rok	Set	2001,2002,2003,2004,2005,2006,2007,2008,2009	None	None
Rozloha_v_ha	Range	[316289,1101613]	None	None
Dokoncene_byty	Range	[86,8599]	None	In
Mzdy	Range	[11577,23735]	None	In

Obrázek č. 39: Nastavení parametrů shlukovací analýzy [zdroj vlastní]

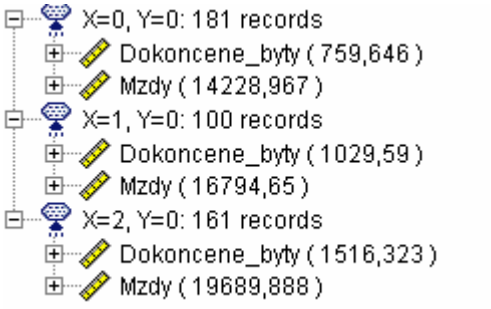
Všem metodám jsem nastavil pevný počet shluků. Zvolil jsem celkem tři shluky. Tento počet se mi zdá optimální, jelikož když nastavím více shluků, tak ve výsledku jsou mzdy docela různorodé a je velký rozdíl mezi maximální a minimální částkou mzdy.

Na obrázku č. 40 jsou zobrazeny jednotlivé shluky Kohonenovy mapy.

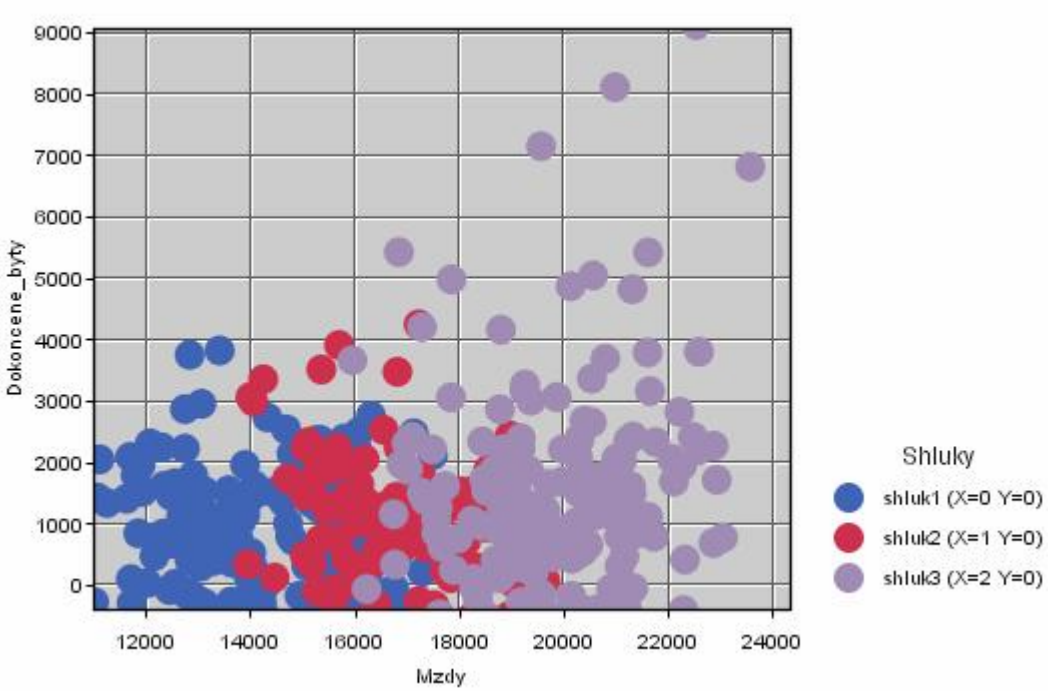


Obrázek č. 40: Souřadnice shluků Kohonenovy mapy pro mzdu a dokončené byty [zdroj: vlastní]

Na obrázku č. 41 jsou nejzajímavějšími první a třetí shluk. První shluk se souřadnicemi X=0 a Y=0 obsahuje záznamy s nízkými hodnotami průměrné mzdy, což odpovídá nižšímu počtu dokončených bytů. Třetí shluk se souřadnicemi X=2 a Y=0 je úplný opak prvního. Vyšší průměrná mzda odpovídá vyššímu počtu dokončených bytů. Dále se podívám na jednotlivé záznamy v těchto shlucích.



Obrázek č. 41: Shluky Kohonenovy mapy pro mzdu a dokončené byty [zdroj vlastní]



Obrázek č. 42: Závislost mzdy a dokončených bytů Kohonenovy mapy [zdroj: vlastní]

Závislost průměrné hrubé měsíční mzdy na počtu dokončených bytů zobrazuje obrázek č. 42 pro tři shluky Kohonenovy mapy. První shluk obsahuje záznamy s nízkými hodnotami obou atributů. Ve třetím shluku se zvýšily hodnoty průměrné měsíční mzdy oproti prvnímu.

Obrázek č. 43 ukazuje maximální a minimální hodnoty atributů v prvním shluku Kohonenovy mapy. Časová řada je od roku 2001 do roku 2006, kdy byla mzda nižší, než v následujících letech. Nejmenší počet záznamů má Středočeský kraj. Tento výsledek byl docela očekáván, jelikož v tomto kraji je nejvyšší průměrná mzda podle tabulky na obrázku č. 38, která je důsledkem nejvyššího počtu dokončených bytů.

Kraj	Rok_Min	Rok_Max	Dokoncene_byty_Mean	Dokoncene_byty_Min	Dokoncene_byty_Max	Mzdy_Mean	Mzdy_Min	Mzdy_Max	Record_Count
Středočeský	2001	2003	1102.800	580	1991	14600.800	13689	15753	5
Plzeňský	2001	2004	848.636	263	1845	14631.455	12985	15993	11
Jihomoravský	2001	2004	1514.417	559	3437	14103.167	12330	15627	12
Moravskoslezský	2001	2004	1028.667	341	2297	14592.167	12833	16016	12
Liberecký	2001	2004	487.769	100	1143	14293.846	12505	16067	13
Jihočeský	2001	2005	931.800	273	1900	14256.667	12204	15739	15
Pardubický	2001	2005	687.067	245	1479	13904.400	11709	15718	15
Královéhradecký	2001	2005	730.933	184	1671	14217.467	12105	15912	15
Ústecký	2001	2004	510.000	155	1226	14458.867	12497	16060	15
Zlínský	2001	2005	765.500	233	1590	14192.812	12084	15905	16
Vysočina	2001	2005	732.438	256	1558	13943.062	11577	16051	16
Olomoucký	2001	2005	797.882	220	1753	13948.471	11646	16069	17
Karlovarský	2001	2006	282.158	98	579	14287.895	11736	16150	19

Obrázek č. 43: Atributy v prvním shluku Kohonenovy mapy a jejich charakteristiky [zdroj vlastní]

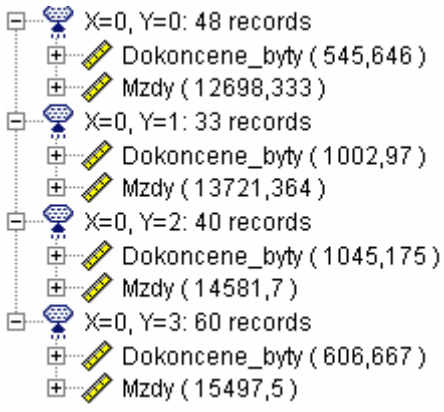
V tabulce na obrázku č. 44 jsou jednotlivé charakteristky atributů ve třetím shluku se souřadnicemi X=2 a Y=0.

Kraj	Rok_Min	Rok_Max	Dokoncene_byty_Mean	Dokoncene_byty_Min	Dokoncene_byty_Max	Mzdy_Mean	Mzdy_Min	Mzdy_Max	Record_Count
Karlovarský	2007	2009	357.250	138	628	18863.250	17988	19967	8
Pardubický	2006	2009	972.000	251	1866	19269.100	17871	20713	10
Olomoucký	2006	2009	1059.800	421	1839	19122.300	17639	20619	10
Zlínský	2006	2009	804.000	269	1661	19232.364	17866	20937	11
Jihočeský	2006	2009	1304.273	288	2708	19485.182	18259	21070	11
Vysočina	2006	2009	896.545	285	1729	19335.182	17989	21098	11
Králové-hradecký	2006	2009	929.909	338	1919	19296.727	17913	20877	11
Liberecký	2006	2009	676.167	256	1355	19920.500	18172	21763	12
Ústecký	2006	2009	624.833	183	1153	19767.667	17918	21462	12
Moravskoslezský	2005	2009	1159.929	394	2453	19979.643	18067	22096	14
Plzeňský	2005	2009	1249.643	354	2663	20137.143	18184	22277	14
Jihomoravský	2005	2009	2746.571	979	6013	20057.500	17641	22337	14
Středočeský	2002	2009	3907.478	1304	8599	20248.261	17065	23735	23

Obrázek č. 44: Atributy ve třetím shluku Kohonenovy mapy a jejich charakteristiky [zdroj vlastní]

Ve výše uvedené tabulce třetího shluku jsou vybrány záznamy s vyšší průměrnou mzdou (19 689 Kč) od roku 2005 do 2009. Nejmenší počet záznamů má Karlovarský kraj, protože se vyznačuje nejnižšími hodnotami průměrné hrubé mzdy a průměrného počtu dokončených bytů, proto také časová řada je od roku 2007 do roku 2009. Středočeský kraj má nejvíce záznamů v tomto shluku, protože se vyznačuje nejvyššími hodnotami průměrné hrubé mzdy a počtu dokončených bytů (tabulka na obrázku č. 38), proto je časová řada už od roku 2002 do roku 2009.

V další kapitole se budu zabývat podrobnější analýzou jednotlivých shluků Kohonenovy mapy, které byly vytvořeny v předcházející kapitole, konkrétně na obrázku č. 41. Podrobně se podívám na první a třetí shluk. Pomocí uzlu select v programu Clementine vyberu pouze záznamy z prvního shluku a použiji znovu tuto metodu. V tomto shluku je celkem 181 záznamů (obrázek č. 41).



Obrázek č. 45: Shluky Kohonenovy mapy pro mzdy, dokončené byty [zdroj vlastní]

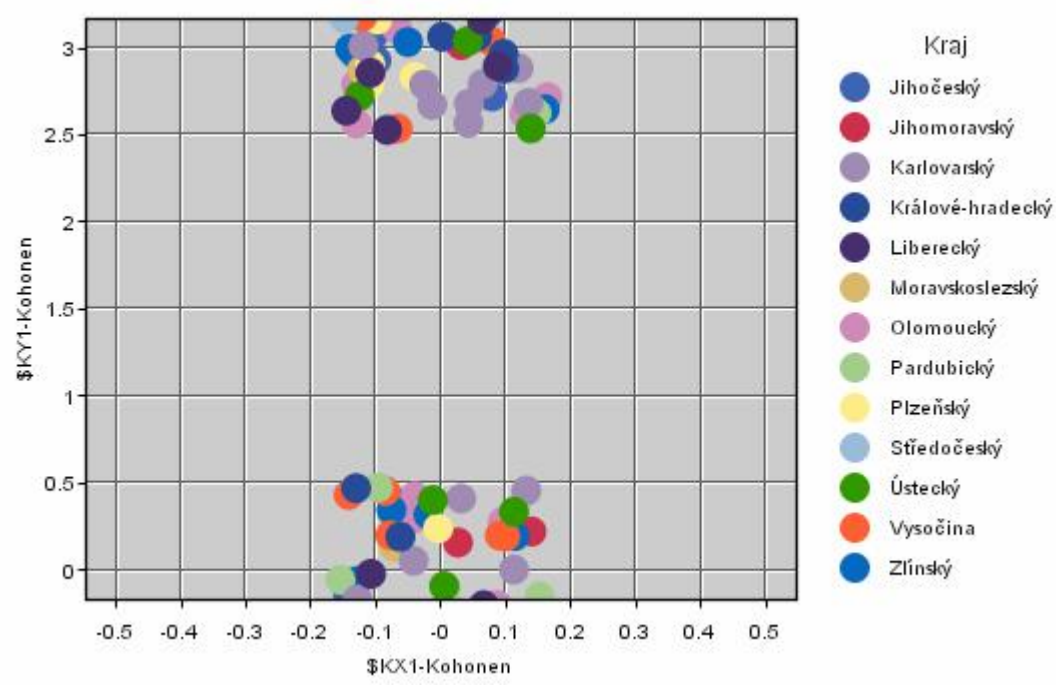
Na obrázku č. 45 jsou vidět čtyři shluky Kohonenovy mapy, které tato metoda vytvořila z původního prvního shluku na obrázku č. 41.

Obrázek č. 46 ukazuje záznamy zařazené do jednotlivých shluků o různých souřadnicích podle Kohonenovy mapy. Novou proměnnou shluky jsem vytvořil pomocí uzlu derive pro větší přehlednost v datovém souboru.

Kraj	Dokoncene_byty	Mzdy	$KX1-Kohonen	$KY1-Kohonen	Shluky
Jihočeský	287	12204	0	0	shluk1 (X=0 Y=0)
Olomoucký	344	11646	0	0	shluk1 (X=0 Y=0)
Jihomoravský	565	12330	0	0	shluk1 (X=0 Y=0)
Zlínský	314	12084	0	0	shluk1 (X=0 Y=0)
Moravskoslezský	468	12833	0	0	shluk1 (X=0 Y=0)
Plzeňský	309	12985	0	0	shluk1 (X=0 Y=0)
Vysočina	321	11577	0	0	shluk1 (X=0 Y=0)
Pardubický	307	11709	0	0	shluk1 (X=0 Y=0)
Karlovarský	137	11736	0	0	shluk1 (X=0 Y=0)
Královéhradecký	287	12105	0	0	shluk1 (X=0 Y=0)
Ústecký	155	12497	0	0	shluk1 (X=0 Y=0)
Liberecký	129	12505	0	0	shluk1 (X=0 Y=0)
Středočeský	580	13689	0	1	shluk2 (X=0 Y=1)
Moravskoslezský	841	13639	0	1	shluk2 (X=0 Y=1)
Olomoucký	762	12254	0	0	shluk1 (X=0 Y=0)
Královéhradecký	460	12777	0	0	shluk1 (X=0 Y=0)
Jihočeský	664	12997	0	0	shluk1 (X=0 Y=0)
Pardubický	468	12524	0	0	shluk1 (X=0 Y=0)
Plzeňský	537	13717	0	1	shluk2 (X=0 Y=1)
Středočeský	1114	14427	0	2	shluk3 (X=0 Y=2)

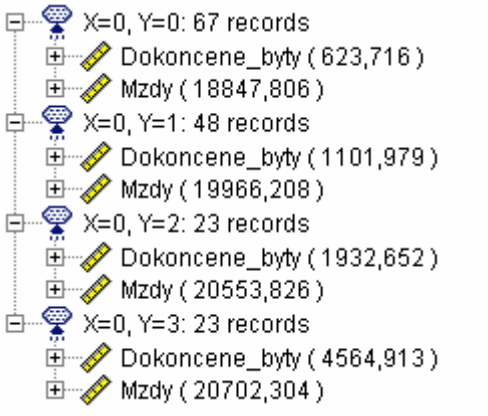
Obrázek č. 46: Jednotlivé záznamy ve shlucích [zdroj vlastní]

Pro lepší orientaci jsem si graficky znázornil dva shluky s nejvyšším počtem záznamů (X=0 a Y=0, X=0 a Y=3) na obrázku č. 47. Na tomto obrázku je vidět zastoupení jednotlivých krajů v těchto dvou shlucích. Do shluku o souřadnicích X=0 a Y=0 patří např. Kalovarský, Ústecký kraj a kraj Vysočina.



Obrázek č. 47: Shluky (X=0 a Y=0, X=0 a Y=3) pro dokončené byty a mzdy [zdroj vlastní]

Dále se podrobně podívám na třetí shluk (X=2 a Y=0) z obrázku č. 41, který se vyznačuje vysokou průměrnou hrubou měsíční mzdou a tím i vysokým počtem dokončených bytů.



Obrázek č. 48: Shluky Kohonenovy mapy pro mzdy, dokončené byty [zdroj vlastní]

Obrázek č. 48 zobrazuje podrobně původní třetí shluk z obrázku č. 41, na který jsem použil znovu Kohonenovu mapu. Shluk o souřadnicích X=0 a Y=3 na obrázku č. 48 představuje nejvyšší průměrnou hrubou měsíční mzdu (20 702 Kč) a tím i nejvyšší hodnotu počtu dokončených bytů (4 564).

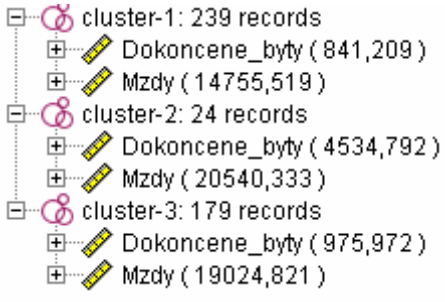
Kraj	Record_Count
Jihočeský	1
Plzeňský	1
Moravskoslezský	1
Jihomoravský	6
Středočeský	14

Obrázek č. 49: Shluk (X=0 a Y=3) Kohonenovy mapy pro dokončené byty a mzdy [zdroj vlastní]

Na výše uvedeném obrázku č. 49 jsou zobrazeny jednotlivé kraje ve shluku (X=0 a Y=3). Nejvyšší zastoupení záznamů má Jihomoravský a Středočeský kraj. Oba tyto kraje disponují nejvyšší průměrnou hrubou měsíční mzdou podle grafu č. 10 na straně 41.

Dále použiji metody K-Means a TwoStep pro analýzu průměrné hrubé měsíční mzdy s počtem dokončených bytů. Pro obě metody jsem nastavil stejný počet shluků (tři shluky) jako pro předchozí Kohonenovu mapu.

Na obrázku č. 50 jsou vidět jednotlivé shluky metody K-Means. Nejvíce záznamů je ve shluku 1 (cluster-1).



Obrázek č. 50: Shluky metody K-Means pro mzdu a dokončené byty [zdroj vlastní]

Kraj	Rok_Min	Rok_Max	Dokoncene_byty_Mean	Dokoncene_byty_Min	Dokoncene_byty_Max	Mzdy_Mean	Mzdy_Min	Mzdy_Max	Record_Count
Středočeský	2001	2003	1941.800	580	3652	15258.700	13689	16503	10
Moravskoslezský	2001	2005	1046.188	341	2297	15069.438	12833	16666	16
Plzeňský	2001	2005	874.375	263	1845	15261.375	12985	16861	16
Jihomoravský	2001	2005	1600.059	559	3437	14712.941	12330	16814	17
Ústecký	2001	2005	508.889	155	1226	14824.444	12497	16844	18
Liberecký	2001	2005	525.833	100	1316	14912.389	12505	16767	18
Jihočeský	2001	2006	969.368	273	2193	14728.211	12204	16772	19
Zlínský	2001	2006	784.000	173	1590	14612.000	12084	16514	20
Pardubický	2001	2006	706.700	240	1580	14514.250	11709	16591	20
Královéhradecký	2001	2006	731.550	184	1671	14791.800	12105	16804	20
Vysočina	2001	2006	804.048	256	1581	14543.714	11577	16748	21
Olomoucký	2001	2006	797.182	220	1753	14501.500	11646	16881	22
Karlovarský	2001	2006	354.091	98	1149	14575.364	11736	16666	22

Obrázek č. 51: Atributy prvního shluku K-Means a jednotlivé charakteristiky [zdroj vlastní]

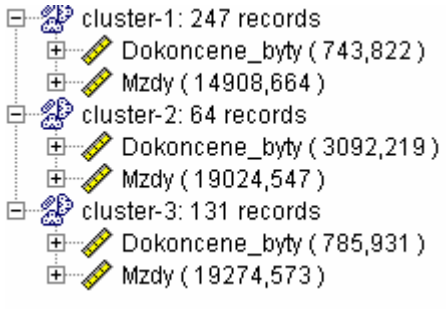
Na obrázku č. 51 jsou vidět jednotlivé charakteristiky atributů prvního shluku metody K-Means. Tento shluk má nízkou průměrnou hrubou měsíční mzdu a nízký počet dokončených bytů. Časová řada je vybrána zhruba od roku 2001 do 2006. Pouze Středočeský kraj má časovou řadu od roku 2001 do roku 2003 a má také nejnižší počet záznamů. Důvodem může být vyšší průměrná hrubá měsíční mzda během celé časové řady oproti ostatním krajům.

Kraj	Rok_Min	Rok_Max	Dokoncene_byty_Mean	Dokoncene_byty_Min	Dokoncene_byty_Max	Mzdy_Mean	Mzdy_Min	Mzdy_Max	Record_Count
Moravskoslezský	2008	2008	2453.000	2453	2453	22096.000	22096	22096	1
Plzeňský	2008	2008	2663.000	2663	2663	22277.000	22277	22277	1
Jihočeský	2008	2008	2708.000	2708	2708	21070.000	21070	21070	1
Jihomoravský	2004	2008	4108.571	3251	6013	19714.571	16815	22337	7
Středočeský	2003	2009	5160.786	3249	8599	20680.214	17065	23735	14

Obrázek č. 52: Atributy druhého shluku K-Means a jednotlivé charakteristiky [zdroj vlastní]

V tabulce na obrázku č. 52 jsou zobrazeny záznamy metody K-Means druhého shluku (cluster-2). Obsahuje záznamy s nejvyšší průměrnou hrubou měsíční mzdou a nejvyšším průměrným počtem dokončených bytů (je vidět na obrázku č. 50). Nejvíce záznamů ma Středočeský kraj. Tento kraj má záznamy v tomto shluku už od roku 2003, protože patří mezi kraje s nejvyšší průměrnou mzdou.

Poslední metodou shlukovací analýzy, kterou použiji, je TwoStep. Jsou zde nastaveny také tři shluky jako v předcházejících dvou metodách (obrázek č. 53). Nejvyšší počet záznamů je ve shluku 1 (cluster-1).



Obrázek č. 53: Tři shluky metody TwoStep pro mzdu a dokončené byty [zdroj vlastní]

Kraj	Rok_Min	Rok_Max	Dokoncene_byty_Mean	Dokoncene_byty_Min	Dokoncene_byty_Max	Mzdy_Mean	Mzdy_Min	Mzdy_Max	Record_Count
Středočeský	2001	2004	1099.833	580	1991	14984.000	13689	16900	6
Jihomoravský	2001	2005	1182.154	559	2072	14615.538	12330	16814	13
Moravskoslezský	2001	2005	895.188	341	1997	15261.188	12833	17397	16
Plzeňský	2001	2005	862.000	263	1845	15386.000	12985	17380	17
Jihočeský	2001	2006	905.421	273	1900	14766.789	12204	16995	19
Liberecký	2001	2005	533.842	100	1316	15023.474	12505	17023	19
Ústecký	2001	2006	515.650	155	1226	15058.950	12497	17336	20
Zlínský	2001	2006	766.364	173	1590	14853.409	12084	17349	22
Vysočina	2001	2006	795.864	256	1581	14669.364	11577	17308	22
Olomoucký	2001	2006	794.957	220	1753	14614.913	11646	17110	23
Pardubický	2001	2006	735.043	240	1580	14847.957	11709	17279	23
Králové-hradecký	2001	2006	748.391	184	1671	15111.130	12105	17447	23
Karlovarský	2001	2007	347.875	86	1149	14787.333	11736	17295	24

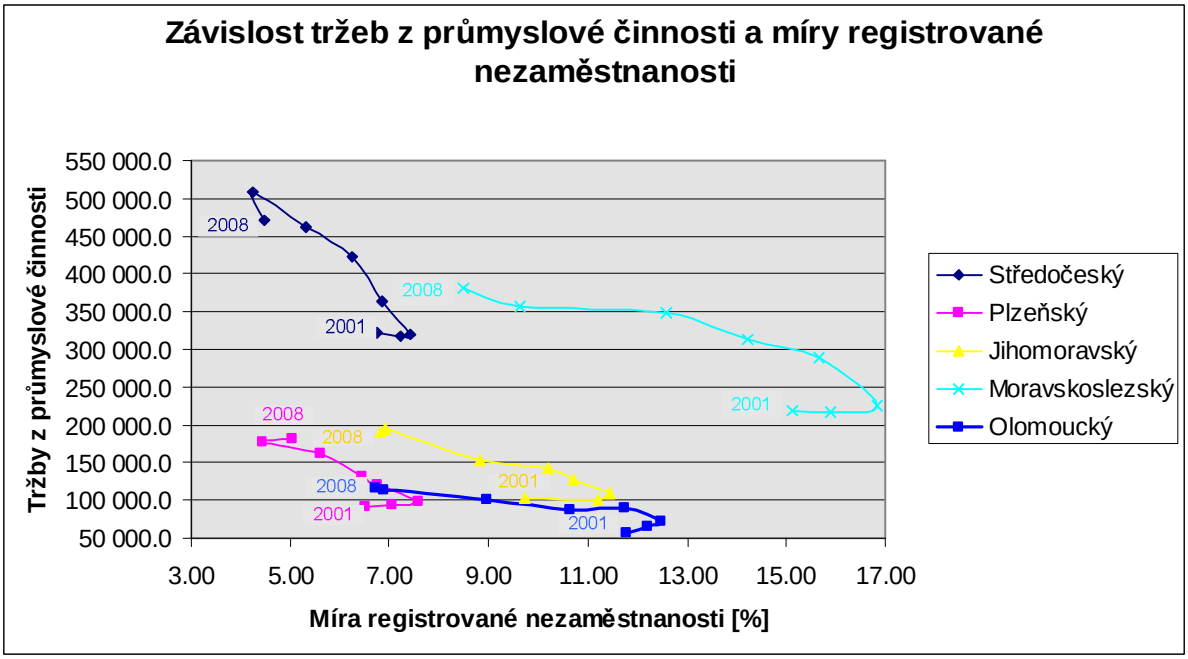
Obrázek č. 54: Atributy prvního shluku TwoStep a jednotlivé charakteristiky [zdroj vlastní]

V tabulce na obrázku č. 54 jsou vidět jednotlivé charakteristiky atributů prvního shluku metody TwoStep. Nejvíce záznamů má Karlovarský kraj, který má nejnižší průměrnou hrubou měsíční mzdu podle obrázku č. 38.

4.8 Analýza průměrné hrubé měsíční mzdy, míry registrované nezaměstnanosti a tržeb z průmyslové činnosti

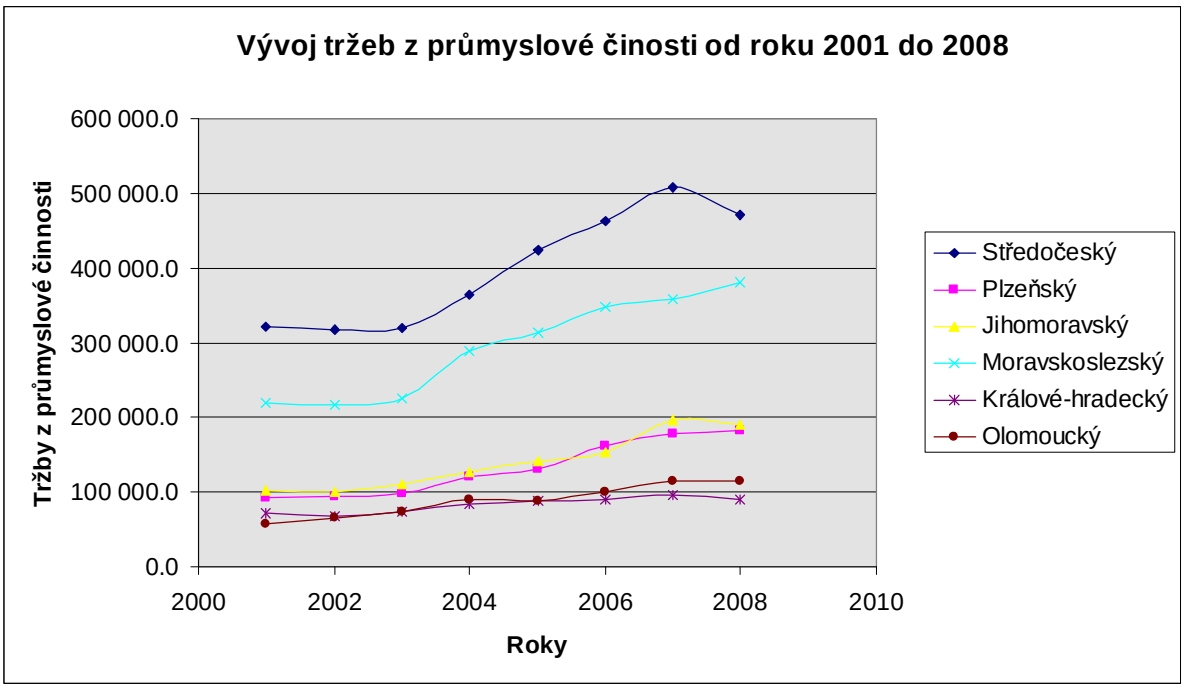
Poslední vybranou skupinou pro analýzu je vztah průměrné hrubé měsíční mzdy, míry registrované nezaměstnanosti a tržeb z průmyslové činnosti.

Na grafu č. 12 je zobrazena závislost tržeb z průmyslové činnosti a míry registrované nezaměstnanosti v časové řadě od roku 2001 do 2008 vždy pro čtvrté čtvrtletí. Všechny křivky na grafu mají podobný charakter, postupem času se míra registrované nezaměstnanosti snižovala a tím se zvyšovaly tržby z průmyslové činnosti. V roce 2008 se míra nezaměstnanosti ve většině krajích zvýšila (Středočeský,Plzeňský, Olomoucký).



Graf č. 12: Závislost tržeb z prům. činnosti a míry nezaměstnanosti ve vybraných krajích [zdroj vlastní]

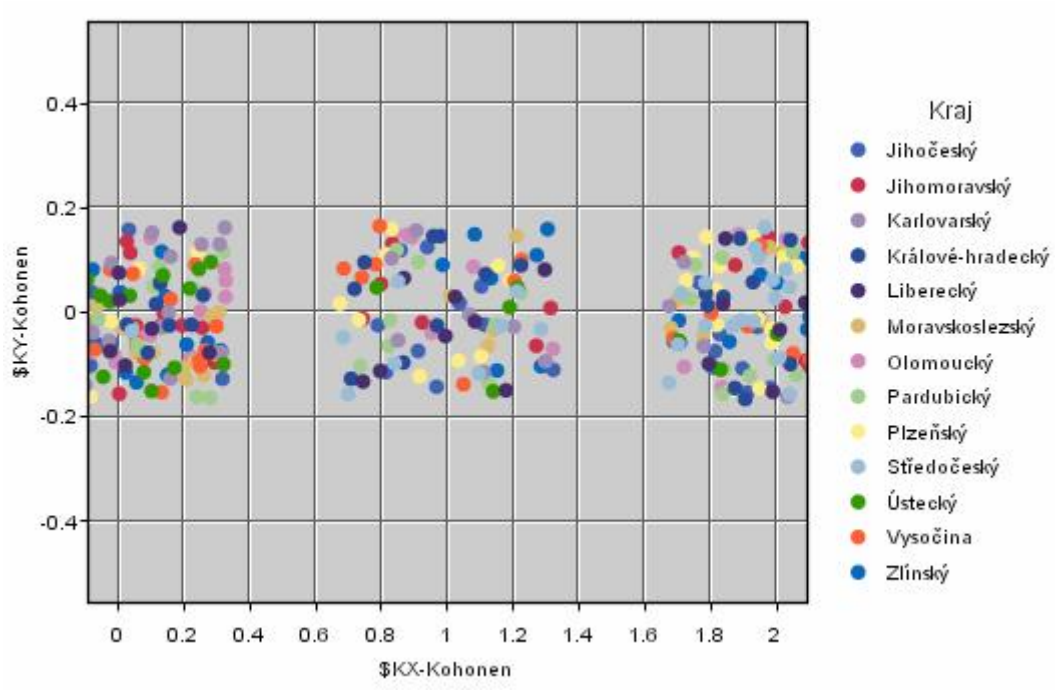
Graf č. 13 ilustruje závislost tržeb z průmyslové činnosti na časové řadě. Z grafu je možné vyčíst, že od roku 2001 tržby stoupaly a v roce 2007 nastal zlom, tržby se snižují. Hlavním důvodem snižování tržeb je převládající dnešní ekonomická krize.



Graf č. 13: Vývoj tržeb z průmyslové činnosti od roku 2000 do 2008 [zdroj vlastní]

Na následujícím obrázku č. 55 je vidět obsah jednotlivých shluků. V prvním shluku o souřadnicích X=0 a Y=0 je 189 záznamů. Průměrná míra registrované nezaměstnanosti je vysoká a činí 10,491 %. Z toho vyplývá poměrně nízká průměrná hodnota tržeb z průmyslové činnosti a nízká hodnota průměrné hrubé měsíční mzdy. Ve třetím shluku (X=2 Y=0) je pravý opak prvního shluku, nižší míra nezaměstnanosti a tím vyšší průměrná hodnota tržeb a mezd.

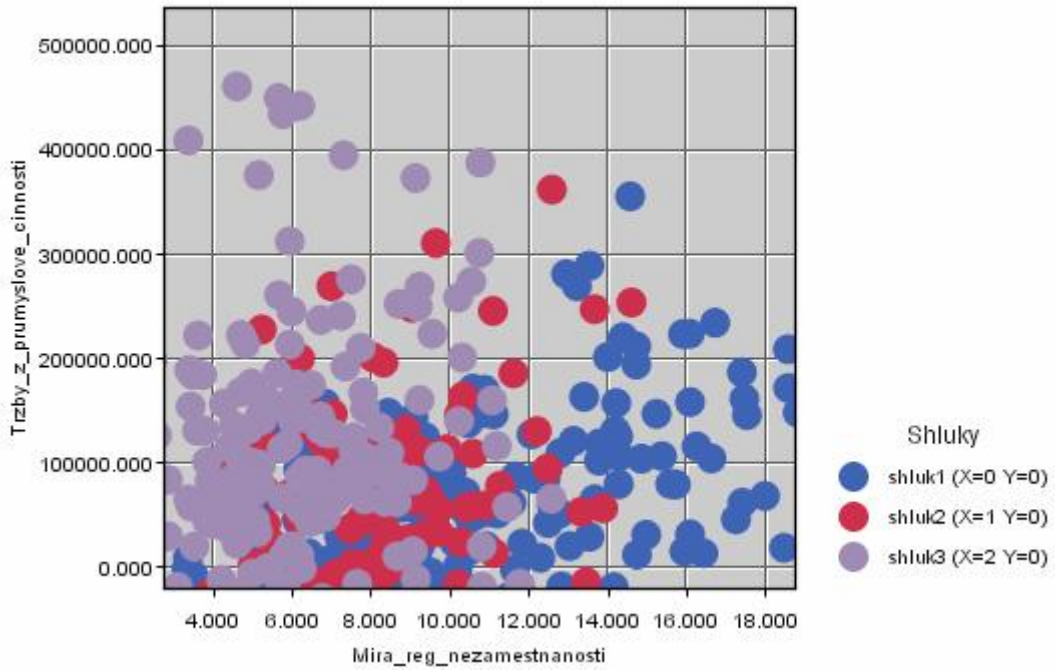
Obrázek č. 55: Tři shluky Kohonenovy mapy pro nezaměstnanost, mzdy a tržby [zdroj vlastní]



Obrázek č. 56: Shluky Kohonenovy mapy pro nezaměstnanost, mzdy a tržby [zdroj: vlastní]

Na obrázku č. 56 jsou vidět jednotlivé souřadnice Kohonenovy mapy. Ve shluku o souřadnicích X=1 a Y=0 je nejnižší počet záznamů (uprostřed na obrázku č. 56).

Obrázek č. 57 ilustruje grafickou závislost mezi mírou registrované nezaměstnanosti a tržeb z průmyslové činnosti všech shluků u Kohonenovy mapy.



Obrázek č. 57: Závislost míry nezaměstnanosti a tržeb z prům. činnosti Kohonenovy mapy [zdroj: vlastní]

Na obrázku č. 58 je zobrazena tabulka průměrných hodnot míry registrované nezaměstnanosti, tržeb z průmyslové činnosti a mezd v prvním shluku o souřadnicích X=0 a Y=0 v časové řadě od roku 2001.

Kraj	Rok_Min	Rok_Max	Mira_reg_nezamestnanosti_Mean	Trzby_z_prumyslove_cinnosti_Mean	Mzdy_Mean	Record_Count
Středočeský	2001	2002	6.720	77882.450	14099.500	2
Plzeňský	2001	2003	6.487	48892.729	14042.857	7
Jihočeský	2001	2004	6.074	48178.689	13556.222	9
Královéhradecký	2001	2004	6.792	43687.850	13833.500	12
Liberecký	2001	2004	8.269	44255.754	14293.846	13
Pardubický	2001	2005	8.392	56088.600	13787.643	14
Vysočina	2001	2005	7.958	45498.700	13802.533	15
Jihomoravský	2001	2005	10.364	62510.781	14581.625	16
Zlínský	2001	2005	9.492	46778.200	14192.812	16
Moravskoslezský	2001	2006	15.501	145134.515	15533.050	20
Olomoucký	2001	2006	11.498	44605.214	14388.190	21
Karlovarský	2001	2006	9.749	20233.829	14475.810	21
Ústecký	2001	2006	16.063	111593.587	15407.174	23

Obrázek č. 58: Průměrné hodnoty atributů prvního shluku (X=0, Y=0) Kohonenovy mapy [zdroj vlastní]

Obrázek č. 59 ilustruje jednotlivé shluky metody K-Means. Nejvíce záznamů je v prvním shluku (cluster-1). Ve shluku 2 (cluster-2) je nejnižší míra registrované nezaměstnanosti (6,587) a tím jsou vyšší tržby z průmyslové činnosti a zároveň vyšší průměrná hrubá měsíční mzda (“*Mzdy*“).

Obrázek č. 59: Tři shluky K-Means pro nezaměstnanost, mzdy a tržby [zdroj vlastní]

V níže uvedené tabulce na obrázku č. 60 jsou zobrazeny průměrné hodnoty atributů nezaměstnanost, tržeb z průmyslové činnosti a mezd třetího shluku (cluster-3) metody K-Means. Tento shluk obsahuje 55 záznamů. Jsou zde vybrány tři kraje (Olomoucký, Ústecký a Moravskoslezský). Tyto kraje disponují vysokou mírou nezaměstnanosti. Průměrná míra registrované nezaměstnanosti je vysoká (15,149).

Kraj	Rok_Min	Rok_Max	Mira_reg_nezamestnanosti_Mean	Trzby_z_prumyslove_cinnosti_Mean	Mzdy_Mean	Record_Count
Olomoucký	2003	2004	12.720	46690.050	14772.000	2
Ústecký	2001	2009	15.457	113112.471	16103.750	28
Moravskoslezský	2001	2007	14.999	163234.936	16146.400	25

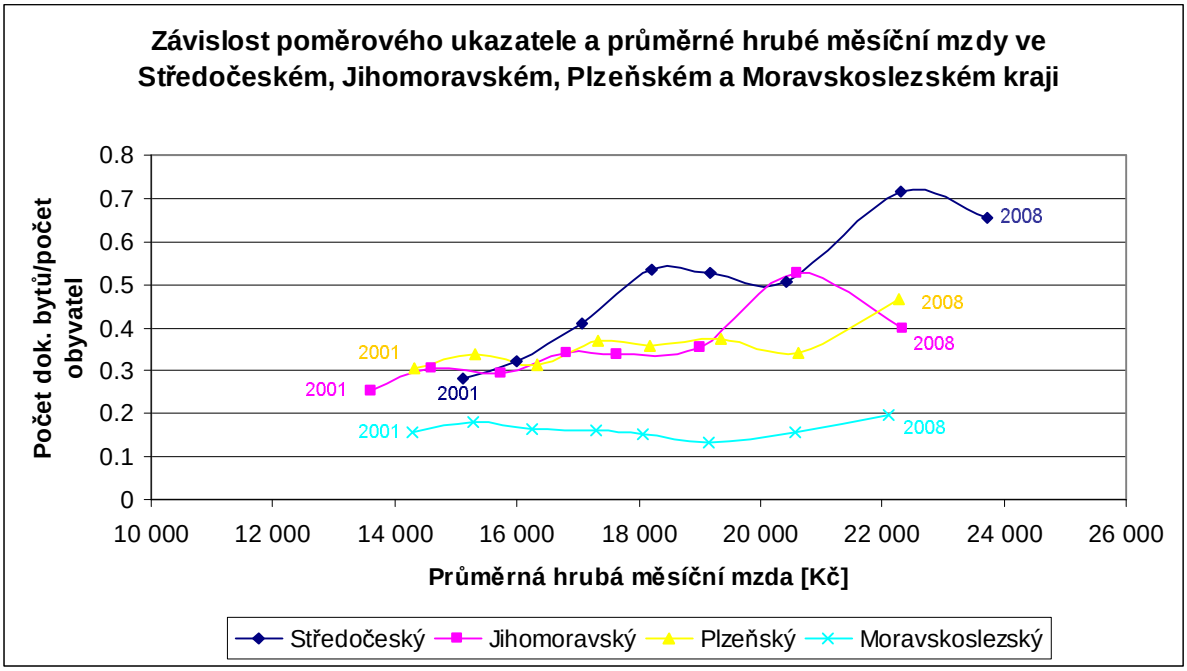
Obrázek č. 60: Průměrné hodnoty atributů třetího shluku metody K-Means [zdroj vlastní]

K analýze dat je možné přistoupit i z pohledu poměrových ukazatelů. Vypočítal jsem poměrový ukazatel (PU) jako počet dokončených bytů / počet obyvatel (rovnice č. 2). Daný poměrový ukazatel je vypočítán jako počet dokončených bytů / celkový počet obyvatel v uvedených krajích na grafu č. 14 za dané čtvrtletí v roce (je bráno vždy 4. čtvrtletí od roku 2001 do roku 2008). Český statistický úřad (ČSÚ) vykazuje tento poměrový ukazatel (PU) jako počet dokončených bytů na 1000 obyvatel. [24]

$$PU = \frac{\text{pocet dokoncenych bytu}}{\text{pocet obyvatel}} * 100\%$$

Rovnice č. 2: Výpočet poměrového ukazatele (PU) [zdroj vlastní]

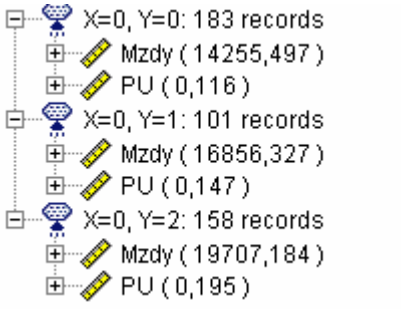
Na grafu č. 14 je vidět závislost tohoto poměrového ukazatele (PU) a průměrné hrubé měsíční mzdy.



Graf č. 14: Závislost pom. ukazatele a průměrné hrubé mzdy od roku 2001 do 2008 [zdroj vlastní]

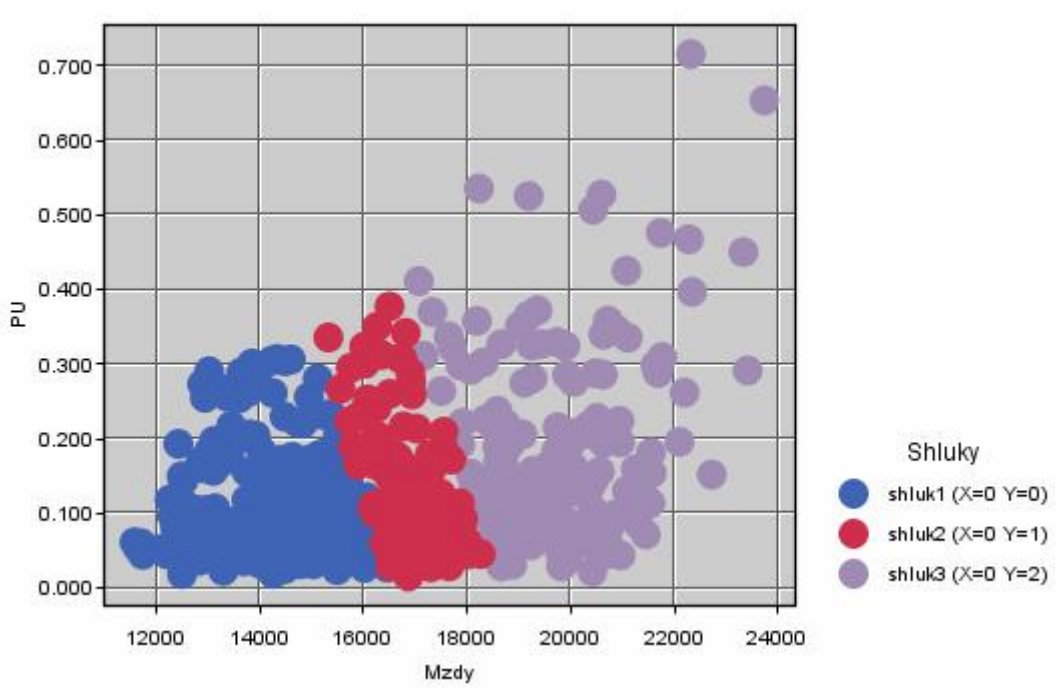
Graf č. 14 mohu srovnat s grafem č. 10 na staně 41. Když porovnám všechny křivky uvedených krajů, tak na grafu č. 14 křivka Plzeňského kraje stoupla proti grafu č. 10. Domnívám se, že tato skutečnost je dána tím, že tento kraj má poloviční počet obyvatel než tři zbývající kraje (Středočeský, Jihomoravský a Moravskoslezský).

Na obrázku č. 61 jsou zobrazené tři shluky Kohonenovy mapy. Nejvíce záznamů (183) je v prvním shluku (X=0 a Y=0). Tento shluk vykazuje nízkou průměrnou hrubou měsíční mzdu (“*Mzdy*“) a poměrový ukazatel (“*PU*“) 0,116. Ve třetím shluku je vyšší průměrná hrubá měsíční mzda (“*Mzdy*“) a poměrový ukazatel (“*PU*“) je také vyšší (0,195).



Obrázek č. 61: Tři shluky Kohonenovy mapy pro mzdy a poměrový ukazatel [zdroj vlastní]

Na obrázku č. 62 je vidět závislost poměrového ukazatele a průměrné hrubé měsíční mzdy ("*Mzdy*"). Čím vyšší poměrový ukazatel ("*PU*"), tím vyšší průměrná hrubá měsíční mzda.



Obrázek č. 62: Závislost pom. ukazatele (PU) a průměrné hrubé mzdy Kohonenovy mapy [zdroj: vlastní]

Závěr

V této části svou diplomovou práci zhodnotím jako celek a vyzdvihnu její význam a přínos pro společnost. Téma *Modelování ekonomických dat* jsem si zvolil záměrně, jelikož obor data mining sleduji v praktickém životě. Zajímám se o vývoj dataminingového softwaru od firmy SPSS a mapuji odborné články od předních společností, jež se tímto oborem zabývají.

Chci zdůraznit, že všechny cíle mé diplomové práce byly naplněny. Je zde popsána teorie Phillipsovy křivky a ekonomický vývoj ČR, který se potvrdil v dalších analýzách mé práce.

Věnuji se shlukovacím metodám a jejich aplikaci na tři vybrané skupiny dat z ekonomické oblasti. Použil jsem Kohonenovu mapu, metodu K-Means a TwoStep. Na základě zkušeností nejlepší výsledky poskytla Kohonenova mapa, která si nejlépe poradila s nestrukturovanými daty.

První testovanou skupinou dat byla závislost míry registrované nezaměstnanosti a míry růstu mezd za kraje po čtvrtletí od roku 2001 do 2009. Cílem této skupiny bylo nalézt Phillipsovu křivku. Ani jedna z metod shlukovací analýzy nedospěla k tomuto výsledku. Další experiment byl proveden s daty za celou ČR v časové řadě po čtvrtletí, od roku 2000 do 2009 pro křivku závislosti mezi průměrnou hrubou měsíční mzdou a mírou registrované nezaměstnanosti. V tomto případě se jako nejlepší jevila metoda TwoStep (obrázek č. 35). Pokud bych proložil jednotlivé body v rámci grafu logaritmickou křivkou, bylo by možné při zjednodušení ilustrovat uvedenou křivku závislosti.

Domnívám se, že podoba Phillipsovy křivky nebyla nelezena z důvodu odlišnosti zkonstruování této křivky a principu shlukovací metody.

Druhou testovanou skupinou dat byla závislost průměrné hrubé měsíční mzdy a počtu dokončených bytů. Potvrdilo se pravidlo, že čím vyšší průměrná hrubá měsíční mzda, tím vyšší počet dokončených bytů. Nejzajímavějšími shluky Kohonenovy mapy na obrázku č. 41 jsou první a třetí. V prvním shluku jsou nižší hodnoty průměrné měsíční mzdy a počtu dokončených bytů. Ve třetím shluku jsou zase vyšší hodnoty. Tyto skutečnosti jsou dané vybranou časovou řadou (obrázek č. 43 a 44).

Třetí testovanou skupinou dat jsou atributy míra registrované nezaměstnanosti, průměrná hrubá měsíční mzda a tržby z průmyslové činnosti. Čím nižší míra registrované nezaměstnanosti tím vyšší tržby z průmyslové činnosti. Kohonenova mapa na obrázku č. 55 tuto závislost potvrdila. V prvním shluku je vyšší průměrná míra nezaměstnanosti a tím nižší tržby z průmyslové činnosti. Třetí shluk obsahuje nižší míru nezaměstnanosti a tím vyšší tržby z průmyslové činnosti.

K analýze jsem přistoupil i pomocí poměrových ukazatelů (rovnice č. 2). Pro srovnání uvádím graf č. 10 a graf č. 14. Na obou grafech jsou zobrazeny čtyři kraje (Středočeský, Jihomoravský, Plzeňský a Moravskoslezský) v závislosti průměrné hrubé měsíční mzdy a počtu dokončených bytů na grafu č. 10 a v závislosti průměrné hrubé měsíční mzdy a poměrového ukazatele (PU) na grafu č. 14. Na grafu č. 14 nastala změna v křivce Plzeňského kraje, která stoupla proti grafu č. 10. Domnívám se, že tato změna je způsobená polovičním počtem obyvatel v tomto kraji proti třem zbývajícím krajům.

Shlukovací analýza je mocným nástrojem pro modelování rozlehlých datových souborů s velkým množstvím proměnných. Tato metoda je většinou aplikovaná jako první v dataminingovém projektu. Její shluky jsou často použity pro následné analýzy.[7] [22]

5 Použité zdroje

1. *Metodické vysvětlivky* [online]. 2010 [cit. 2010-02-12]. Dostupný z WWW: <http://www.pardubice.czso.cz/xe/edicniplan.nsf/o/531302-09-za_1__3_ ctvrtleti_2009-metodicke_vysvetlivky>.
2. *Ekonomicko-statistický slovník L až P* [online]. 2010 [cit. 2010-02-12]. Dostupný z WWW: <http://www.businessinfo.cz/cz/clanek/analyzy-statistiky/ekonomicko-statisticky-slovnik-l-p/1000431/39670/>.
3. *CSV* [online]. 2010 [cit. 2010-02-12]. Dostupný z WWW: <http://cs.wikipedia.org/wiki/CSV>.
4. BERKA, P. *Dobývání znalostí z databází*. Academia: Praha, 2003. ISBN 80-200-1062-9.
5. PAVEL, P. *DATA MINING: Díl I*. Pardubice: Univerzita Pardubice, 2006. 144 s. ISBN 80-7194-886-1.
6. RUD, O. *Data Mining - Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM)*. Computer Press: Praha, 2001. ISBN 80-7226-577-6.
7. LINOFF, G., BERRY, M. *Data Mining Techniques – For Marketing, Sales and Customer Support*. John Wiley & Sons:New York,1997. ISBN 0-471-17980-9.
8. KUBANOVÁ, J. *Statistické metody pro ekonomickou a technickou praxi*. Statis: Bratislava, 2004. ISBN 80-85659-37-9.
9. ABC (Slovník cizích slov). *Pearsonův korelační koeficient* [online]. 2005-2006 [cit. 2010-02-21]. Dostupný z WWW: < http://slovnik-cizich-slov.abz.cz/web.php/slovo/pearsonuv-korelacni-koeficient>.
10. *Shluková analýza* [online] 2010 [cit. 2010-02-21]. Dostupný z WWW: <http://staff.utia.cas.cz/nagy/skola/Projekty/Classification/ShlukovaAnalyza.pdf>.
11. *A Tutorial on Clustering Algorithms* [online] 2010 [cit. 2010-02-21]. Dostupný z WWW: <http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/ kmeans.html>.
12. *Determinanty cen nemovitostí pro jednotlivé regiony ČR* [online] 2010 [cit. 2010-04-04]. Dostupný z WWW: <http://www.cnb.cz/m2export/sites/ www.cnb.cz/cs/financni_stabilita/zpravy_fs/fs_2008-2009/FS_2008-2009_clanek_2.pdf>.

13. MACH, M. *Makroekonomie II. pro magisterské studium*. Melandrium: Slaný, 2001. ISBN 80-86175-18-9.
14. *Finance.cz* [online] 2010 [cit. 2010-04-19]. Dostupný z WWW: <http://www.finance.cz/ekonomika/prace/mzda/>.
15. *Vývoj HDP v ČR* [online] 2010 [cit. 2010-04-19]. Dostupný z WWW: <http://www.kurzy.cz/makroekonomika/hdp/>.
16. *Usporim.cz* [online] 2010 [cit. 2010-04-19]. Dostupný z WWW: <http://www.usporim.cz/products/analyza-vyvoje-ekonomiky-cr-a-odvetvi-v-pusobnosti-mpo-za-rok-2009/>.
17. PROVAZNÍKOVÁ, R., KŘUPKA, J., KAŠPAROVÁ, M. *Modelování konkurenceschopnosti regionů v podmínkách globalizace*. In Scientific Papers of the University of Pardubice, Special Edition, Series D6, Pardubice: Upa, 2009, s.113-124, ISSN 1211-555X.
18. *Český statistický úřad: Makroekonomické údaje* [online]. 2010 [cit. 2010-04-24]. Dostupné z WWW: <http://www.czso.cz/csu/redakce.nsf/i/cr:_makroekonomicke_udaje/$File/HLMAKRO.xls>.
19. *Makroekonomický vývoj České republiky v období 1997 – 2006* [online]. 2010 [cit. 2010-04-24]. Dostupné z WWW: <http://www.cers.tuke.sk/cers2007/PDF/Burianova.pdf>.
20. *Ekonomická situace českých krajů a měst* [online]. 2010 [cit. 2010-04-24]. Dostupné z WWW: <http://www.mmr.cz/Pro-media/Media-o-ministerstvu/Ekonomicka-situace-ceskych-kraju-a-mest>.
21. DORNBUSCH, R. FISCHER, S. *Makroekonomie*, 6th ed. Praha: SPN a Nadace Economics, 1996. ISBN 80-04-25556-6.
22. SPSS. *SPSS Inc. Clementine® 10.1 Desktop User´s Guide.* 2006.
23. *Web Data Mining* [online]. 2010 [cit. 2010-04-24]. Dostupné z WWW: <http://books.google.cz/books?id=6Mh50Uaq6AIC&pg=PA124&lpg=PA124&dq=strengths+K-means&source=bl&ots=NuvTCNIono&sig=PVfM_hy9_LX8kO5wSSgSZ5R8pDE&hl=cs&ei=RsHVS9mbMseLOMa3vJAO&sa=X&oi=book_result&ct=result&resnum=3&ved=0CA4Q6AEwAg#v=onepage&q=strengths%20K-means&f=false>.
24. *Český statistický úřad Vysočina* [online]. 2010 [cit. 2010-04-24]. Dostupné z WWW: < http://www.brno.czso.cz/xj/redakce.nsf/i/pocet_dokoncenych_bytu_na_vysocine_se_zvysil>.

Seznam obrázků

Seznam tabulek

Seznam grafů

Seznam rovnic

Seznam příloh

Použité zkratky

ČR	Česká republika
DM	Data mining
ČSÚ	Český statistický úřad
VŠE	Vysoká škola ekonomická
EU	Evropská unie
HDP	Hrubý domácí produkt
CRISP-DM	CRoss-Industry Standard Proces for Data Mining (souhrnná dataminingová metodologie)
MS	Microsoft
CSV	Comma-separated values (hodnoty oddělené čárkami)
SOM	Self-organizing map (samoorganizující se mapy)
DPH	Daň z přidané hodnoty
PU	Poměrový ukazatel

Příloha č. 1: Můj stream v SPSS Clementine 10.1