

UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

BAKALÁŘSKÁ PRÁCE

Zpracování dat z oblasti energetiky pomocí
shlukové analýzy



Vedoucí bakalářské práce:
Mgr. Ondřej Vencálek Ph.D.
Rok odevzdání: 2014

Vypracoval:
Michal Polák
AST, III. ročník

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně pod vedením Mgr. Ondřeje Vencálka Ph.D. a výhradně s použitím uvedené literatury.

V Olomouci dne 9. dubna 2014

Poděkování

Na tomto místě bych velice rád poděkoval svému vedoucímu Mgr. Ondřeji Vencálkovi Ph.D. za čas, který mi věnoval a za všechny rady a připomínky, díky kterým se podařilo tuto práci dokončit. Poděkování patří i mé rodině za podporu během celého studia.

Obsah

Úvod

Cílem mé bakalářské práce je s využitím metod shlukové analýzy analyzovat systémovou odchylku trhu s elektřinou. Tato analýza by dále měla být využita při obchodování na trzích s elektrickou energií. Konečným výsledkem by mělo být určité pravidlo, které na základě minulých pozorování této veličiny bude předpovídat její hodnotu v budoucích časových okamžicích. Veškeré výpočty a výstupy budu provádět pomocí statistického softwaru *R*.

V první sekci popíšu, jakým způsobem je získávána sledovaná veličina, jakým způsobem se s ní manipuluje na trzích, a nakonec stručně představím analyzovaná data. V druhé kapitole budu na data aplikovat různé metody shlukové analýzy a ty mezi sebou porovnáám a popíšu určité výhody a nevýhody. V poslední kapitole vytvořím predikční model pro systémovou odchylku elektřiny a pokusím se navrhnout optimální obchodní strategii.

1. Systémová odchylka elektřiny

Analyzovaná data jsou záznamy o systémové odchylce elektřiny za rok 2012. Data jsem získal z webu společnosti **OTE, a.s.**, jejímž hlavním předmětem podnikání, jsou činnosti operátora trhu.

1.1. Trh s elektřinou

V této podkapitole popisuji analyzovanou veličinu a zde uvedené informace jsem získal z internetových stránek společnosti **OTE, a.s.**

Pro definici systémové odchylky a popisu závislosti ceny odchylky na velikosti systémové odchylky je nutné nejprve uvést, jak trh s elektřinou funguje.

Legislativa definuje účastníky trhu, pro které **OTE** vypořádává a zúčtovává odchylky, jako tzv. **subjekty zúčtování (SZ)**, a stanovuje postup výpočtu velikosti odchylek **SZ**. Základní pravidlo pro zpoplatnění všech odchylek určuje, že každý účastník trhu je buď sám odpovědný za odchylku, tedy je SZ, nebo předává zodpovědnost za odchylku jinému SZ. Jakýkoliv odběr elektřiny ze soustavy nebo dodávka elektřiny do soustavy musí být přiřazeny k některému z účastníků trhu.

Odchylkou subjektu zúčtování za každou obchodní hodinu je součet rozdílu mezi skutečně dodaným množstvím a sjednaným množstvím elektřiny pro dodávku a rozdílu mezi skutečně odebraným množstvím a sjednaným množstvím elektřiny k odběru.

Sjednaná množství dodávek elektřiny do soustavy a odběrů elektřiny ze soustavy jsou pro každou obchodní hodinu vypočtena operátorem trhu.

Skutečná množství dodané nebo odebrané elektřiny jsou evidována na základě obchodního měření.

1.1.1. Definice systémové odchylky

Systémovou odchylkou (SO) v každé obchodní hodině se rozumí saldo všech dodávek a odběrů SZ. Je rovna součtu odchylek SZ a je kryta regulační energií.

Regulační energii na krytí systémové odchylky opatřuje provozovatel přenosové soustavy aktivací podpůrných služeb, nákupem na vyrovnávacím trhu s regulační energií a nákupem regulační energie ze zahraničí.

1.1.2. Cena systémové odchylky

Cenou odchylky je nejvyšší nabídková cena regulační elektřiny dodaná v této obchodní hodině. Byla-li takto stanovená cena nižší než cena stanovená cenovým rozhodnutím ERÚ (Energetický regulační úřad), použila se cena stanovená ERÚ. V případě, že pro některou obchodní hodinu nebyla obstarána žádná regulační energie, použila se zúčtovací cena podle cenového rozhodnutí ERÚ.

Křivka závislosti ceny odchylky na velikosti systémové odchylky je vždy potvrzována cenovým rozhodnutím ERÚ pro příslušný rok. V roce 2011 a 2012 je křivka závislosti ceny odchylky pro každou obchodní hodinu stanovena následovně:

Pro zápornou a nulovou systémovou odchylku SO určená vzorcem:

$$C = 2350 + 5,5x|SO|[K/MWh; MWh]$$

Pro kladnou systémovou odchylku SO určená vzorcem:

$$C = 1 + 3,5x|SO|[K/MWh; MWh]$$

Od 1. ledna 2010 došlo k vytvoření těsnější závislosti ceny protiodchylky (odchylka SZ ve směru opačném systémové odchylce) na ceně regulační energie. Zúčtovací cenou protiodchylky je vážený průměr cen z aktivované regulační energie.

Pokud má odchylka SZ stejný směr jako systémová odchylka, pak je tento SZ povinen uhradit částku rovnou součinu odchylky SZ a stanovené zúčtovací ceny jakožto poplatek za jím způsobenou odchylku. V opačném případě SZ inkasuje částku rovnou násobku zúčtovací ceny protiodchylky a hodnoty protiodchylky SZ.

Tabulka 1: Pravidla pro zisk a ztrátu na trhu v závislosti na **směru** SO a SZ

SO	Odchylka SZ	Zúčtovací cena	Kdo komu platí (směr)
+	+	Odchylka“	SZ → OTE
+	-	Protiodchylka“	OTE → SZ
-	+	Protiodchylka“	OTE → SZ
-	-	Odchylka“	SZ → OTE

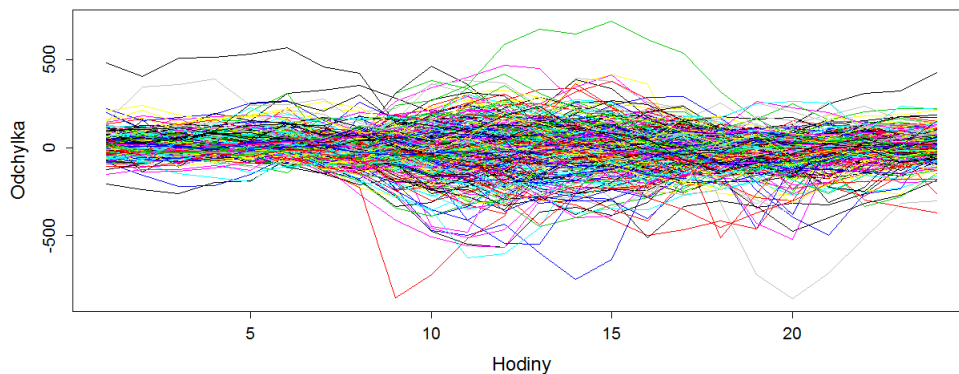
1.2. Presentace dat

Datovou množinu tvoří matice typu 8784×3 . Pracuji se třemi proměnnými a to dnem, hodinou a odchylkou.

Den	Hodina	odchylka
1.1.2012	1	199.110
1.1.2012	2	109.708
1.1.2012	3	136.202
1.1.2012	4	123.108
1.1.2012	5	156.914

Data předvedu na mnohorozměrnou úlohu, jednotlivá pozorování budou dny a každý den bude mít 24 proměnných (hodiny). Tím vznikne matice **M** typu 366×24 , se kterou budu dále pracovat ve shlukové analýze.

Obrázek 1: Denní průběhy systémové odchylky za rok 2012



Tento obrázek je důležitý pro další postup a především pro shlukovou analýzu. Pomocí shlukové analýzy rozdělím 366 dnů do určitého počtu skupin (shluků). Z grafu je vidět, že bude vhodné uvažovat 3 skupiny dnů a to podprůměrné, průměrné a nadprůměrné.

2. Shluková analýza

V této kapitole se budu zabývat otázkou, jak nejlépe roztrdit jednotlivé dny do tří skupin. K tomu využiji různé metody shlukové analýzy, porovnáám je mezi sebou a pokusím se vybrat tu nejoptimálnější. V první části se zaměřím na hierarchické shlukování, popíšu vybrané typy shlukovacích metod a zaměřím se na některé jejich výhody a nevýhody. To stejné provedu v části druhé, kde se zaměřím na některé nehierarchické metody shlukové analýzy.

Pro zde uvedené výpočty a teorii jsem čerpal z knihy Vícerozměrné statistické metody [3] (Hebák, P. a kol.), Applied multivariate statistical analysis (Johnson, R.A. a Wichern, D.V.) a balíčku programu *R* - The Cluster Package.

2.1. Hierarchické shlukování

Tyto metody vytvářejí hierarchické posloupnosti rozkladů. Důležitou skutečností je, že při výpočtu není předem daný počet shluků. Já se zaměřím pouze na postupy aglomerativní, které lze obecně popsat takto:

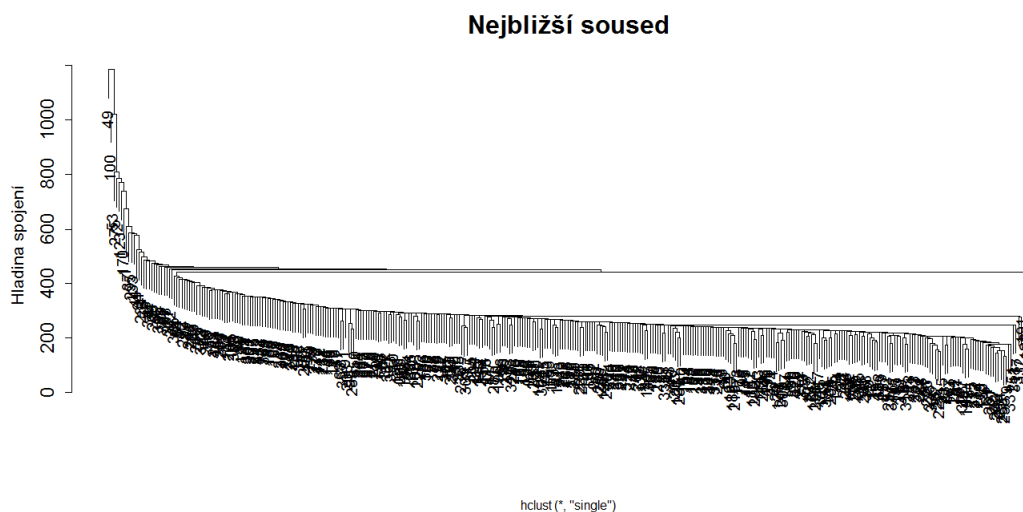
1. Vypočteme matici vzdáleností $\mathbf{D}$ mezi jednotlivými objekty.
2. Začneme proces rozkladu $S^{(n)}$, tj. od n shluků, z nichž každý obsahuje jeden objekt.
3. Prohlédáme matici $\mathbf{D}$ (vzhledem k symetrii jen její dolní nebo horní trojúhelníkovou část) a najdeme dva shluky (h -tý a h' -tý), jejichž vzdálenost $D_{hh'}$ je minimální.
4. Spojíme h -tý a h' -tý shluk do nového g -tého shluku. V matici $\mathbf{D}$ odstraníme h -tý a h' -tý řádek i sloupec a nahradíme je řádkem i sloupcem pro nový shluk; dimenze matice $\mathbf{D}$ se snížil o jedničku.
5. Poznamenanáme pořadí cyklu $v=1,2,\dots,n-1$, identifikaci spojených objektů h a h' a hladinu pro spojení $d_1 = D_{hh'}$.
6. Pokud proces vytváření rozkladů již neskončil spojením všech objektů do jediného shluku $S^{(1)}$, pokračujeme krokem 3.

Výsledky hierarchických shlukovacích postupů lze výhodně zachytit graficky v podobě stromového diagramu, který se nazývá dendrogram. Osa y znázorňuje hladinu spojení shluků.

2.1.1. Metoda nejbližšího souseda

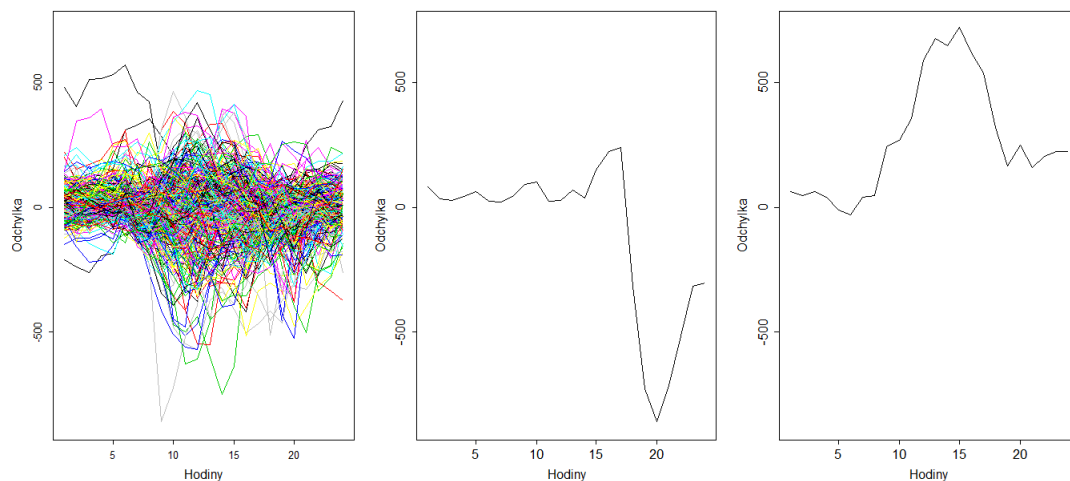
Jako první jsem vybral metodu nejbližšího souseda, jejímž kritériem pro spojování shluků je minimální hodnota z mezishlukových vzdáleností objektů. Při úpravě matice vzdálenosti postupujeme podle vztahu: $D_{gg'} = \min(D_{g'h}, D_{g'h'})$. Tedy vzdálenost nového shluku g vzniklého spojením shluků h a h' od libovolného stávajícího shluku g' je minimum z těchto dvou vzdáleností.

Obrázek 2: Dendrogram - Nejbližší soused



Z tohoto stromového grafu je možné vyčíst největší nedostatek této metody. Při velkém počtu proměnných se často i značně vzdálené objekty mohou sejít ve stejném shluku, pokud větší počet dalších objektů mezi nimi vytvoří jakýsi most a dochází k tzv. řetězení.

Obrázek 3: Denní průběhy systémové odchylky pro 3 shluky vytvořené metodou jednoduchého spojení

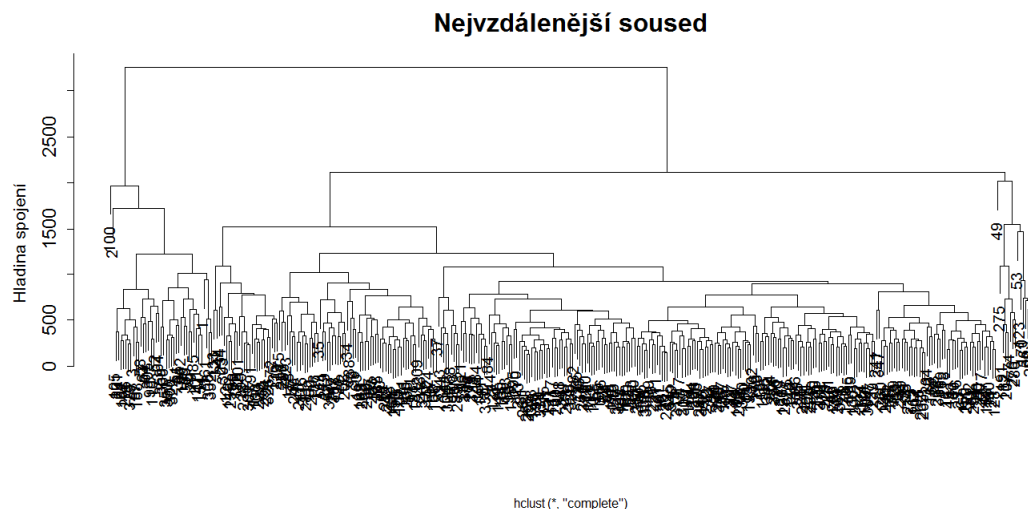


Z obrázku 3 je patrné, že do jednoho shluku bylo přiřazeno 364 dnů do zbylých dvou po jednom dni. Dny s maximální a minimální hodnotou SO, tvoří samostatné shluky o jednom dni. Metoda nejbližšího souseda se ukázala jako nevhodná, kvůli řetězení objektů. Dobrou vlastností této metody by mohlo být určení odlehlých pozorování.

2.1.2. Metoda nejvzdálenějšího souseda

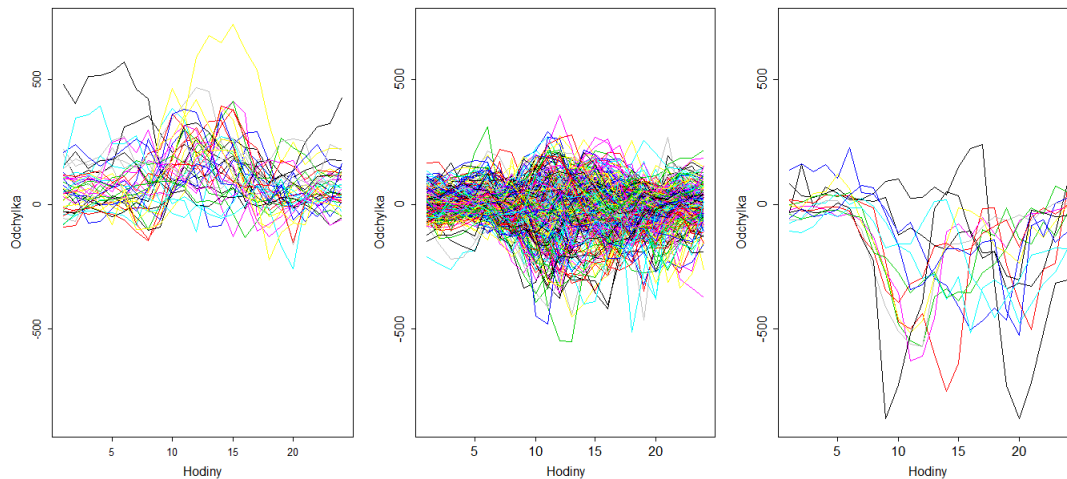
Druhou metodou je metoda nejvzdálenějšího souseda, jejímž kritériem pro spojování shluků je maximální hodnota z mezishlukových vzdáleností objektů. Při úpravě matice vzdálenosti postupujeme podle: $D_{gg'} = \max(D_{g'h}, D_{g'h'})$. Tedy vzdálenost nového shluku g vzniklého spojením shluků h a h' od libovolného stávajícího shluku g' je maximum z těchto dvou vzdáleností.

Obrázek 4: Dendrogram - Nejvzdálenější soused



Dendrogram pro tuto metodu ukazuje, že nežádoucí efekt řetězení použitím metody nejvzdálenějšího souseda odpadá.

Obrázek 5: Denní průběhy systémové odchylky pro 3 shluky vytvořené metodou úplného spojení



Tato metoda má tendenci tvořit kompaktní, ale nerovnoměrně velké shluky. To je viditelné na obrázku 5. Metodou nejvzdálenějšího souseda jsem vytvořil shluk

nadprůměrných dní o rozsahu 41 dní, shluk průměrných dní s 311 dny a shluk podprůměrných dní o rozsahu 14 dní. To není úplně optimální výsledek protože shluk průměrných dní, přestože se dá očekávat, že bude největší, je oproti dvěma zbylým shlukům příliš velký.

2.1.3. Wardova metoda

Wardova metoda se od předešlých dvou metod liší tím, že ke shlukování objektů nedochází na základě vzdálenosti mezi objekty, ale používá určité kritérium. Kritériem G pro spojování shluků je přírůstek celkového vnitroskupinového součtu čtverců odchylek pozorování od shlukového průměru. Vzniká-li tedy spojením shluků h a h' nový shluk g přepočítá se kritérium následovně:

$$\Delta G = \sum_i \sum_{j=1}^p (x_{gij} - \bar{x}_{gj})^2 - \sum_i \sum_{j=1}^p (x_{hij} - \bar{x}_{hj})^2 - \sum_i \sum_{j=1}^p (x_{h'ij} - \bar{x}_{h'j})^2.$$

Kde i je index pozorování, j určuje složku p -rozměrného pozorování.

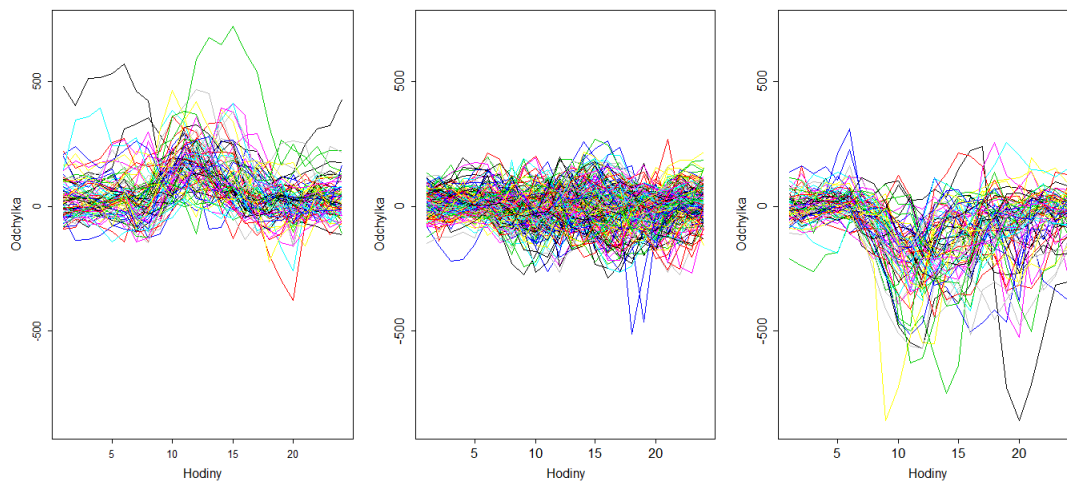
Přírůstek je vyjádřen jako součet čtverců v nově vznikajícím shluku, zmenšený o součty čtverců v obou zanikajících shlucích. Protože provádíme spojení zajišťující minimalizaci kritéria ΔG , má Wardova metoda tendenci odstraňovat malé shluky. Vznikají shluky zhruba shodné velikosti, což je často vítaná vlastnost.

Obrázek 6: Dendrogram - Wardova metoda



Z grafu je vidět to, co bylo zmíněno výše. Touto metodou získám shluky, jejichž velikost má nižší variabilitu než předchozí metody. Tímto postupem vytvořím shluk nadprůměrných dní s rozsahem 89 dní, průměrných 196 dní a shluk podprůměrných s 81 dny

Obrázek 7: Denní průběhy systémové odchylky pro 3 shluky vytvořené Wardovou metodou



Důležitým parametrem pro aglomerativní hierarchické shlukování je míra vzdálenosti použitá pro výpočet vzdálenosti mezi objekty (Nejbližší a Nejvzdálenější soused) nebo pro výpočet kritéria (Wardova metoda). Ve všech dosud uvedených výpočtech v této kapitole jsem použil euklidovskou metriku.

2.1.4. Míry vzdálenosti

Zde se zaměřím na to, jaký vliv mají různé míry vzdálenosti na konečný výsledek shlukování. Různé metriky použiji ve třech dosud uvedených hierarchických metodách a budu sledovat, jak se mění rozsahy shluků při různých metrikách. Budu uvažovat tyto míry vzdálenosti:

Manhatanská vzdálenost: $D_H(x_i, x_{i'}) = \sum_{j=1}^n |x_{ij} - x_{i'j}|$

Euklidovská vzdálenost: $D_E(x_i, x_{i'}) = \sum_{j=1}^n \sqrt{(x_{ij} - x_{i'j})^2}$

Čebyševova vzdálenost: $D_C(x_i, x_{i'}) = \max_j |x_{ij} - x_{i'j}|$

Tabulka 2: Metoda nejbližšího souseda

	Manhatanská	Euklidovská	Čebyševova
Shluk 1	364	364	364
Shluk 2	1	1	1
Shluk 3	1	1	1

Tabulka 3: Metoda nejvzdálenějšího souseda

	Manhatanská	Euklidovská	Čebyševova
Shluk 1	3	41	95
Shluk 2	326	311	256
Shluk 3	37	14	15

Tabulka 4: Wardova metode

	Manhatanská	Euklidovská	Čebyševova
Shluk 1	74	89	110
Shluk 2	222	196	152
Shluk 3	70	81	104

Z tabulky 3 a 4 lze vidět, že Manhatanská vzdálenost má tendenci tvořit velký shluk průměrných dní a menší shluk dní podprůměrných a nadprůměrných. Čebyševova vzdálenost naopak tyto dva shluky svou velikostí ke shluku průměrných dní přiblíží nejvíce.

2.2. Nehierarchické shlukování

Nehierarchické metody se od těch hierarchických liší především v tom, že počet shluků je předem známý. Jejich výhodou proti hierarchickým metodám je, že je lze aplikovat na mnohem větší datové množiny. Algoritmus nehierarchických metod je založen na funkcionálu kvality rozkladu. Chceme tedy pro daný počet shluků k dosáhnout extrému některého funkcionálu, obvykle Wardova kritéria. Nevýhodou toho přístupu ke shlukování může být to, že při nevhodně zvoleném počtu shluků můžeme obdržet nesmyslné výsledky.

2.2.1. Metoda K-průměrů

Algoritmus je založen na přesunování objektů mezi shluky. Můžeme ho popsat následujícími kroky.

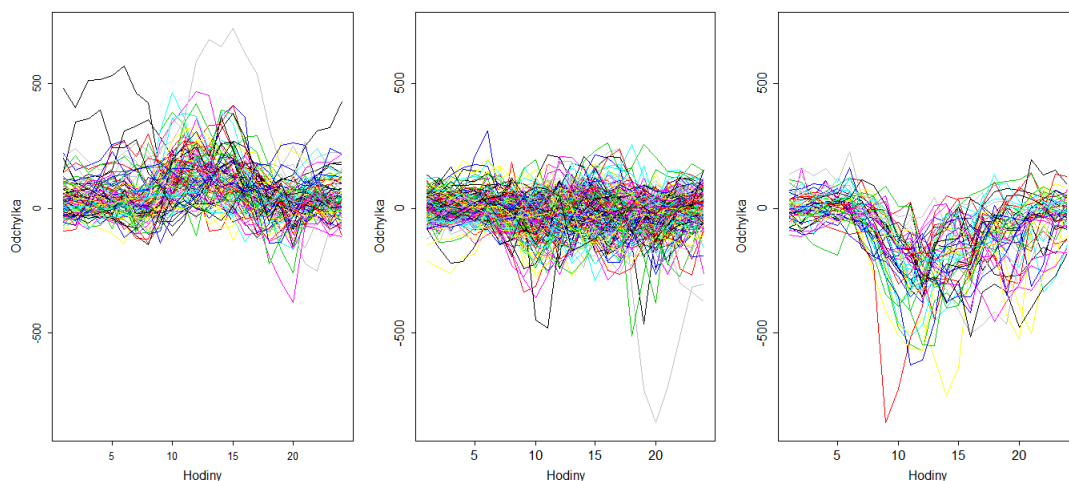
1. Zvolíme počáteční rozklad do k shluků, nejčastěji náhodně, podkladem však může být nějaká vnější informace, někdy také výsledky již provedeného shlukování.

2. Určíme centroidy pro všechny shluky v aktuálním rozkladu. Máme-li p -rozměrná pozorování, pak centroid je p -rozměrný vektor jehož j -tá složka je průměr j -té proměnné.

3. Probereme po řadě všechny objekty. Pokud má právě zkoumaný objekt nejbližší k centroidu shluku do něhož náleží, ponecháme jej na místě, jinak jej přesuneme do shluku, k jehož centroidu má nejbližší. Nedošlo-li v tomto kroku k žádným přesunům, považujeme aktuální rozklad za definitivní (suboptimální) řešení úlohy. Jinak se vracíme k 2. kroku (stačí přepočítat centroidy jen pro shluk, ze kterého byl objekt odebrán, a pro shluk, do kterého byl tento objekt zařazen).

Při výpočtu dochází k optimalizaci Wardova kritéria. Cílem metody K-průměrů je tedy minimalizovat vnitroshlukovou variabilitu. Přestože prvotní rozčlenění do k shluků je náhodné, k optimalizaci kritéria dojde většinou již po malém počtu iterací. Pro výpočet se narozdíl od hierarchických metod nevyužívá matice vzdáleností mezi objekty a není ji nutné v každé iteraci přepočítávat, což má velký vliv na rychlost výpočtu. To je důvod, proč jsou nehierarchické metody při velkém počtu objektů vhodnějším nástrojem.

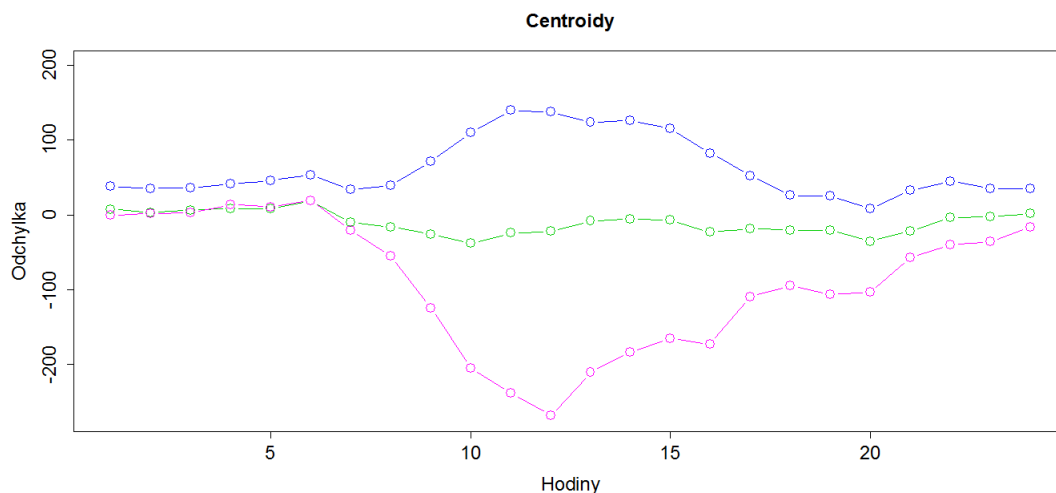
Obrázek 8: Denní průběhy systémové odchylky pro 3 shluky vytvořené metodou K-průměrů



K výsledku 109 nadprůměrných, 202 průměrných a 55 podprůměrných dní lze stabilně dojít zhruba po 50 iteracích.

Při tomto přístupu ke shlukování nedochází k postupnému zařazování objektů do shluků a není proto možné vykreslit dendrogram. Proto jsem se kvalitu rozkladu rozhodl posuzovat na základě centroidů, které jsou v programu *R* jedním z výstupů nehierarchických metod.

Obrázek 9: Průběhy centoridů v jednotlivých shlucích vytvořených metodou K-průměrů

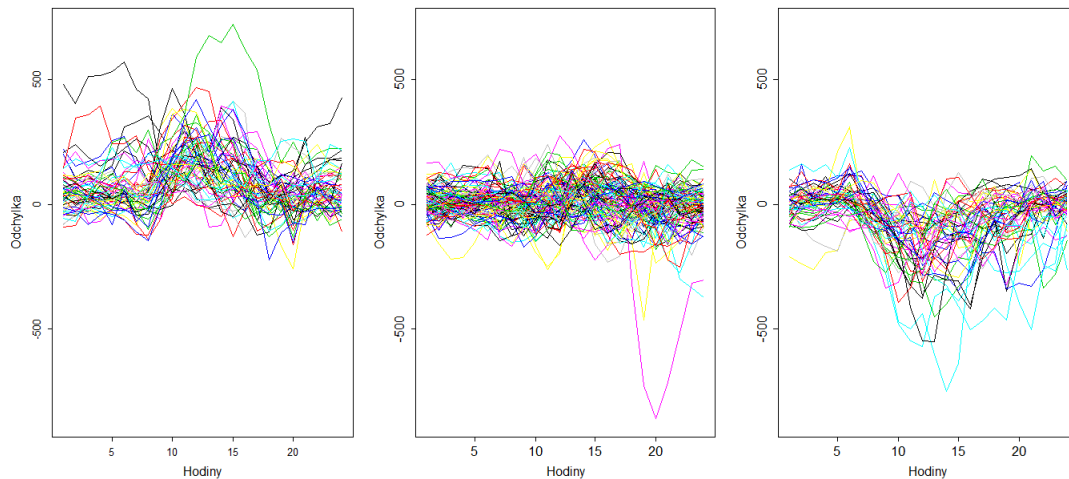


Z průběhů centoridů je vidět, že shluky se začínají výrazněji odlišovat v devět hodin. K největším rozdílům dochází v hodině dvanácté a po ní se k sobě shluky začínají přibližovat. Shluky k sobě mají nejblíže mezi 1. a 7. hodinou. V tomto časovém intervalu můžeme pozorovat, že centoriody shluků s průměrnými a podprůměrnými dny jsou téměř totožné. Dokonce můžeme v hodině čtvrté a páté pozorovat vyšší průměrnou hodnotu u shluku podprůměrných dní než u shluku průměrných. To je vlastnost, která by mohla naznačovat, že shlukování metodou K-průměrů není zcela vhodné.

2.2.2. PAM - Shlukování Kolem Medoidů

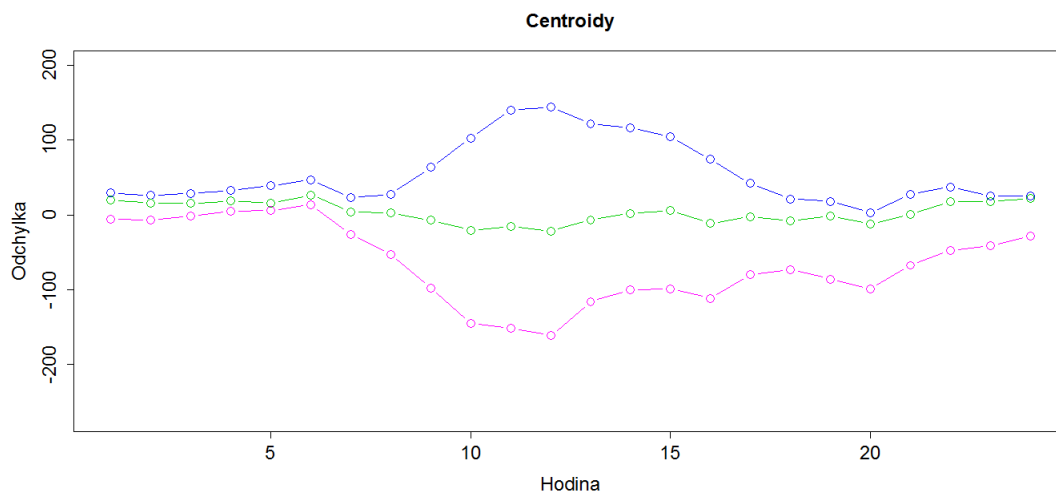
Shlukování dat do k shluků “*kolem medoidů*” (Partitioning Around Medoids), je více robustní verzí metody K-průměrů. Narozdíl od metody K-průměru, kde se pracuje s centroidy (průměry), shlukování touto metodou nebude ovlivněno odlehlými pozorováními. PAM-algoritmus je založený na nalezení k reprezentativních objektů (medoidů) tak, aby co nejlépe vystihovaly strukturu dat. Po nalezení k medoidů se shluky konstruují přiřazením pozorování do shluku, od jehož medoidu má nejmenší vzdálenost. Cílem je nalézt k medoidů, které minimalizují součet vzdáleností pozorování od jejich nejbližšího medoidu.

Obrázek 10: Denní průběhy systémové odchylky pro 3 shluky vytvořené metodou PAM



Metoda PAM roční data rozčlenila na 62 nadprůměrných dní, 180 dní s průměrnou hodnotou systémové odchylky a 124 podprůměrných dní.

Obrázek 11: Průběhy centroidů v jednotlivých shlucích vytvořených metodou PAM

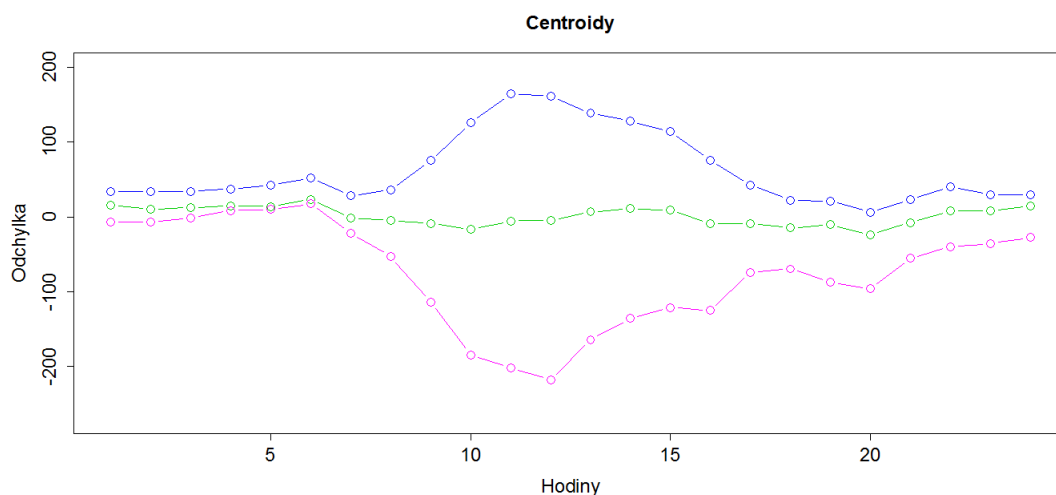


Metoda PAM vytvoří shluky tak, že nadprůměrný shluk má v průměru ve všech hodinách hodnotu systémové odchylky vyšší než shluk průměrný a podprůměrný. To stejné platí pro shluk průměrný a podprůměrný. To je vítaná vlastnost, a proto bych z nehierarchických metod zvolil metodu PAM.

2.3. Volba shlukové metody

Z hierarchických metod se jako nejvhodnější ukázala Wardova metoda, z nehierarchických to je metoda PAM. Tyto dvě metody porovnám v několika bodech a vyberu tu, která bude vhodnější pro další analýzu. Nejprve se podívám, jaké centroidy shluků získám Wardovou metodou.

Obrázek 12: Průběhy centroidů v jednotlivých shlucích vytvořených Wardovou metodou



Stejně jako u metody PAM, i zde nedochází k protnutí centroidů a ty jsou si pro obě metody hodně podobné. Tímto porovnáním se neukázalo, která metoda je vhodnější.

Tabulka 5: Porovnání metod na základě velikosti shluků

	WARD	PAM
Shluk 1	89	62
Shluk 2	196	182
Shluk 3	81	124

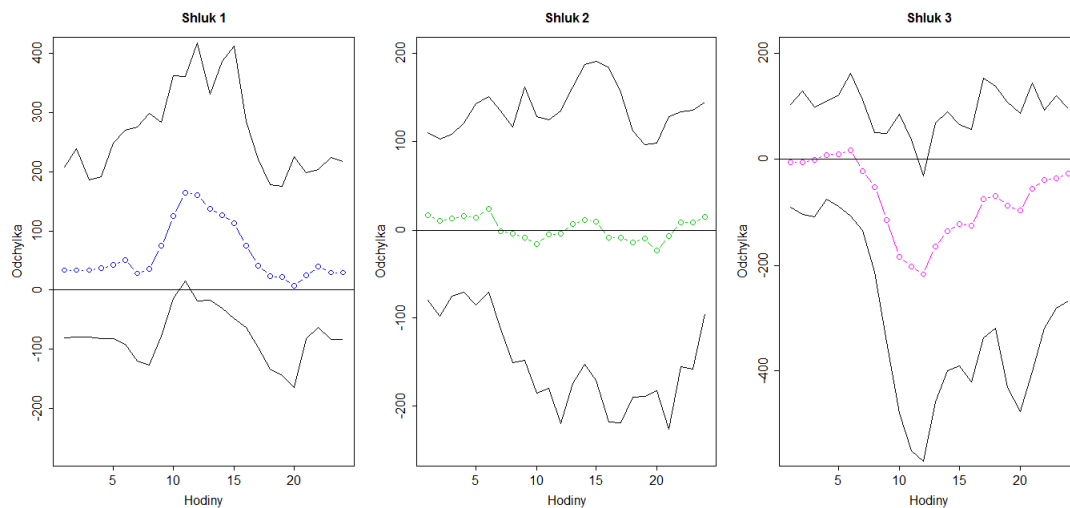
Shluk průměrných dní se výrazně neliší, ale ve velikostech shluků nadprůměrných a podprůměrných jsou velké rozdíly. Jako rozumnější volba se jeví Wardova metoda, která tyto shluky vytvoří s podobnou velikostí než volba PAM, u níž je shluk podprůměrných dní dvakrát větší než shluk nadprůměrných dní.

Zajímavou věc lze také pozorovat v grafech denních průběhů jednotlivých shluků. Absolutní minimum systémové odchylky je -860.104, tato hodnota byla naměřena 18.2.2012 ve 20:00. Celkový průměr podprůměrného shluku je -74.86

a pro tento den -102.32 . Metoda PAM tento den zařadila do shluku průměrných dní a Wardova metoda do shluku podprůměrných dní.

V dalších výpočtech budu pracovat s výsledky Wardovy metody. Jako poslední charakteristika výsledku, s kterým budu dále pracovat je variabilita ve shlucích. Určil jsem 95% interval spolehlivosti pro systémovou odchylku v jednotlivých hodinách pro každý shluk zvlášť. Postupoval jsem neparametricky, data v každé hodině seřadil a meze intervalu stanovil na 2.5% nejmenší a 97.5% největší hodnotě. Při neparametrickém postupu se ne vždy podaří dosáhnout přesně stanovené hladiny. Postupoval jsem tedy tak, aby hladina vždy dosahovala minimálně 5%.

Obrázek 13: Variabilita ve shlucích



Ve shluku s průměrnými dny jsem žádnou užitečnou informaci nezískal. Ve shluku s nadprůměrnými dny víme, že s minimálně 95% pravděpodobností bude hodnota SO v jedenáctou hodinu kladná a ve shluku s podprůměrnými dny víme, že s minimálně 95% pravděpodobností bude hodnota SO ve dvanáctou hodinu záporná.

3. Predikční model

V této kapitole se na základě výsledku shlukové analýzy a informací o systematické odchylce pokusím vytvořit návrh, jak predikovat tuto veličinu. Moji predikci zaměřím pouze na směr systematické odchylky, její absolutní hodnota nebude mít tedy na výsledek žádný efekt. Moje analýza bude souviset s Tabulkou 1, cílem bude predikce znaménka SO. Jestliže budeme v pozici SZ a budeme chtít na trhu s elektřinou obchodovat, pak na základě informací z minulosti, určím sázku na hodnotu SO tak, aby pravděpodobnost zisku byla co největší a naopak ztráty, co nejmenší.

Zde jsem postupoval pomocí knih Základy matematické statistiky (Anděl, J.) a Matematika náhody (Anděl, J.).

Budu pracovat především s podmíněnou pravděpodobností.

Definice: Nechť $A, B \in \mathcal{A}$ jsou dva náhodné jevy a nechť $P(B) > 0$. Podmíněná pravděpodobnost jevu A za podmínky, že nastal jev B , se definuje vzorcem

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Důležitým výstupem shlukových metod je vektor, který obsahuje označení příslušnosti objektů ke shlukům. V tomto případě je tento vektor permutací 89 jedniček 196 dvojek a 81 trojek.

Pro další postup zavedu náhodnou veličinu K , která představuje příslušnost dne ke shluku. Tato veličina nabývá hodnot z množiny $S = \{1, 2, 3\}$.

Jak bylo řečeno v úvodu této kapitoly, dále mě bude zajímat pouze znaménko SO. Proto upravím matici $\mathbf{M}$, se kterou jsem pracoval ve shlukové analýze tak, že za kladnou hodnotu SO dosadím 1 a za zápornou 0 a tuto nově vzniklou matici

označím $\mathbf{N}$. Pozorování v matici $\mathbf{N}$ jsou realizacemi náhodné veličiny s alternativním rozdělením pravděpodobnosti. Tuto veličinu budu dále označovat Y . Dále budu uvažovat náhodné veličiny Z_j pro $j=1, \dots, 24$. Tyto veličiny popisují chování znaménka v jednotlivých hodinách a mají také alternativní rozdělení pravděpodobnosti, $Z_j \sim \text{Alt}(p_j)$.

Realizace veličin K a Y budu chápat jako posloupnost pozorování v diskrétních časových okamžicích (dnech a hodinách). Bude mě zajímat predikce pro aktuální časový okamžik, který označím K_t pro den a Y_t pro hodinu, na základě minulých pozorování těchto veličin.

Cílem tedy bude určit odhad parametru p náhodné veličiny Y_t a Z_j na základě předchozích realizací veličin Y , K , Z_j . Na základě těchto odhadnutých pravděpodobností bych měl být schopný odpovědět na otázku, kdy je vhodné vsadit na kladné a kdy na záporné znaménko. Budu tedy popisovat dva různé přístupy. První popíše situaci, ve které se zajímám pouze o předpověď pro aktuální hodinu a nebude pro mě podstatné, o kterou denní hodinu jde. Při druhém přístupu budu odhadovat parametry p_j pro jednotlivé hodiny. V první situaci budu tedy pracovat s náhodnou veličinou Y_t a jejím parametrem p , zatímco u druhé to budou náhodné veličiny Z_j s parametry p_j .

Pro odhad parametru náhodné veličiny s alternativním rozdělením pravděpodobnosti použiji odhad metodou maximální věrohodnosti. Ten je obecně ve tvaru:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n Y_i$$

Jde o určení proporce kladného znaménka na daném pravděpodobnostním prostoru. Pravděpodobnost záporného znaménka už určím lehce, protože společně s kladným znaménkem tvoří množinu elementárních jevů a tyto jevy jsou disjunktní.

Dále také budu chtít znát přesnost svého odhadu, tu popíšu pomocí intervalů spolehlivosti. Pro určení odhadu parametru p je nutné určit

$$X_n = \sum_{i=1}^n Y_i,$$

kde $Y_1, \dots, Y_n$ jsou nezávislé náhodné veličiny se stejným rozdělením pravděpodobnosti jako veličina Y . Náhodná veličina X_n má binomické rozdělení pravděpodobnosti s parametry p a n .

Věta (Moivreova - Laplaceova): Nechť $X_n \sim \text{Bi}(n, p)$, kde $p \in (0, 1)$. Pak pro $n \rightarrow \infty$ platí:

$$\frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0, 1)$$

Budu pracovat s velkým počtem odhadů parametrů $p_1, p_2, \dots, p_{24}$ a pro každý parametr určím interval spolehlivosti. Nepoužiji *standartní Waldův interval spolehlivosti* a to z důvodu vlastnosti špatného dodržení předepsané hladiny spolehlivosti. Místo toho budu pracovat s intervalem, který je uveden v práci Brown a kol. (2001) na základě článku Agresti, Coull (1998).

Označme

$$\tilde{X} = X_n + z^2/2, \quad \tilde{n} = n + z^2/2, \quad \tilde{p} = \tilde{X}/\tilde{n}, \quad \tilde{q} = 1 - \tilde{p},$$

kde z je 2.5 % kvantil $N(0, 1)$

Stanovil jsem hladinu spolehlivosti $\alpha = 0.05$, budu tedy konstruovat 95% interval spolehlivosti pro parametr p ve tvaru:

$$I_{AC} = \left[\tilde{p} - z_{0.975} \sqrt{\tilde{p}\tilde{q}/\tilde{n}}, \tilde{p} + z_{0.975} \sqrt{\tilde{p}\tilde{q}/\tilde{n}} \right]$$

3.1. Matice pravděpodobností přechodu

Jako první mě zajímá, jaká je pravděpodobnost přechodu mezi stavy z množiny S , neboli mezi dny z jednotlivých shluků.

Spočítám pravděpodobnosti $P(K_{t+1} = m | K_t = n)$, kde $m, n \in S$ a určím matici pravděpodobností přechodu.

Tabulka 7: Matice pravděpodobností přechodu mezi shluky

	Shluk 1	Shluk 2	Shluk 3
Shluk 1	0.284	0.625	0.091
Shluk 2	0.209	0.546	0.245
Shluk 3	0.272	0.420	0.309

Jak se dalo očekávat, ať je dnes den z libovolného shluku, s největší pravděpodobností můžu zítra očekávat průměrný den. Zajímavé je, že po podprůměrném dni můžu očekávat den z nadprůměrného shluku s třikrát větší pravděpodobností než by tomu bylo naopak.

3.2. Predikce – obecný přístup

V této podkapitole budu určovat $\hat{p}$ a $(1 - \hat{p})$ náhodné veličiny Y za různých podmínek. Zvolil jsem tři různé strategie, jak predikovat znaménko v aktuální hodině.

1. Pomocí informace o včerejších dni

Budu predikovat znaménko na libovolnou hodinu dnes na základě toho, jaký den byl včera.

Určím tedy pravděpodobnosti $P(Y_t = l | K_{t-1} = m)$, kde $l = 0, 1$ a $m \in S$.

Tabulka 8: Včerejší den byl nadprůměrný

	Kladné	Záporné
Dolní mez intervalu spolehlivosti	0.541	0.417
$P(Y_t K_{t-1} = 1)$	0.562	0.438
Horní mez intervalu spolehlivosti	0.583	0.459

Tabulka 9: Včerejší den byl průměrný

	Kladné	Záporné
Dolní mez intervalu spolehlivosti	0.518	0.454
$P(Y_t K_{t-1} = 2)$	0.532	0.468
Horní mez intervalu spolehlivosti	0.546	0.482

Tabulka 10: Včerejší den byl podprůměrný

	Kladné	Záporné
Dolní mez intervalu spolehlivosti	0.467	0.489
$P(Y_t K_{t-1} = 3)$	0.489	0.511
Horní mez intervalu spolehlivosti	0.511	0.533

2. Pomocí informace o předchozí hodině

Zde půjde o predikci znaménka v aktuální hodině pomocí informace o znaménku v předchozí hodině.

Určím pravděpodobnosti $P(Y_t = h|Y_{t-1} = l)$, kde $l, h = 0, 1$.

Tabulka 11: V předchozí hodině SO měla záporné znaménko

	Kladné	Záporné
Dolní mez intervalu spolehlivosti	0.218	0.756
$P(Y_t Y_{t-1} = 0)$	0.231	0.769
Horní mez intervalu spolehlivosti	0.244	0.782

Tabulka 12: V předchozí hodině SO měla kladné znaménko

	Kladné	Záporné
Dolní mez intervalu spolehlivosti	0.785	0.192
$P(Y_t Y_{t-1} = 1)$	0.795	0.203
Horní mez intervalu spolehlivosti	0.808	0.215

3. Pomocí informace o předchozí hodině a včerejším dni

U této strategie zkombinuji dvě předchozí a bude mě zajímat, jak to zvýší šanci na správný tip.

Určím pravděpodobnosti $P(Y_t = h|Y_{t-1} = l, K_{t-1} = m)$, kde $l, h = 0, 1$ a $m \in S$.

Tabulka 13: V předchozí hodině SO měla záporné znaménko a včerejší den byl nadprůměrný

	Kladné	Záporné
Dolní mez intervalu spolehlivosti	0.218	0.726
$P(Y_t Y_{t-1} = 0, K_{t-1} = 1)$	0.245	0.755
Horní mez intervalu spolehlivosti	0.274	0.782

Tabulka 14: V předchozí hodině SO měla kladné znaménko a včerejší den byl nadprůměrný

	Kladné	Záporné
Dolní mez intervalu spolehlivosti	0.788	0.168
$P(Y_t Y_{t-1} = 1, K_{t-1} = 1)$	0.811	0.189
Horní mez intervalu spolehlivosti	0.832	0.212

Tabulka 15: V předchozí hodině SO měla záporné znaménko a včerejší den byl průměrný

	Kladné	Záporné
Dolní mez intervalu spolehlivosti	0.210	0.755
$P(Y_t Y_{t-1} = 0, K_{t-1} = 2)$	0.227	0.773
Horní mez intervalu spolehlivosti	0.245	0.790

Tabulka 16: V předchozí hodině SO měla kladné znaménko a včerejší den byl průměrný

	Kladné	Záporné
Dolní mez intervalu spolehlivosti	0.784	0.185
$P(Y_t Y_{t-1} = 1, K_{t-1} = 2)$	0.799	0.201
Horní mez intervalu spolehlivosti	0.815	0.217

Tabulka 17: V předchozí hodině SO měla záporné znaménko a včerejší den byl podprůměrný

	Kladné	Záporné
Dolní mez intervalu spolehlivosti	0.203	0.745
$P(Y_t Y_{t-1} = 0, K_{t-1} = 3)$	0.228	0.772
Horní mez intervalu spolehlivosti	0.255	0.797

Tabulka 18: V předchozí hodině SO měla kladné znaménko a včerejší den byl podprůměrný

	Kladné	Záporné
Dolní mez intervalu spolehlivosti	0.741	0.205
$P(Y_t Y_{t-1} = 1, K_{t-1} = 3)$	0.769	0.231
Horní mez intervalu spolehlivosti	0.795	0.259

U první strategie je vidět, že není o moc lepší než se rozhodnout hodem mincí. Při pohledu na druhou možnost je jasné, že pro předpověď je určující znaménko v minulé hodině. To nám říká že, s velkou pravděpodobností po kladném můžu

očekávat kladné a po záporném znaménku záporné. Přidám-li informaci o shluku včerejšího dne, je vidět malé zlepšení mé předpovědi. Pokud byl včerejší den průměrný, tak si nijak nepomůžu. Ale v případě nadprůměrného dne jsou mé předpovědi lepší zhruba o 1.5%. V případě, že včerejší den byl podprůměrný, očekávám po kladném znaménku záporné s větší pravděpodobností než bez této podmínky.

3.3. Predikce - hodinový přístup

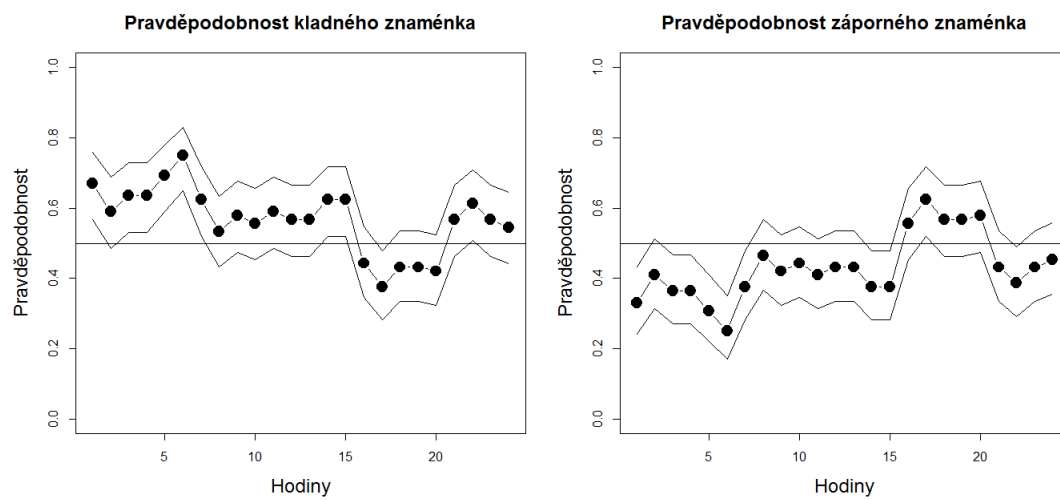
U tohoto přístupu uvažuji stejné tři strategie, ale na data nyní nahlížím hodinově. Pracovat budu s náhodnými veličinami Z_j a odhadnu jejich parametry p_j pro $j = 1, \dots, 24$. Pro danou podmínku nebudu odhadovat jeden, ale dvacet čtyři parametrů. Proto tyto odhady a jejich konfidenční intervaly nezobrazím jako tabulku, ale graficky. V druhém a třetím bodě této podkapitoly ještě přidám tabulku odhadnutých parametrů $\hat{p}_j$, abych viděl, jestli jsem svoji predikci zlepšil. Přesnost odhadů v tomto případě bude menší než u předchozího přístupu. To je způsobené menším počtem pozorování, pomocí nichž odhady parametrů p_j určuji.

1. Pomocí informace o včerejších dni

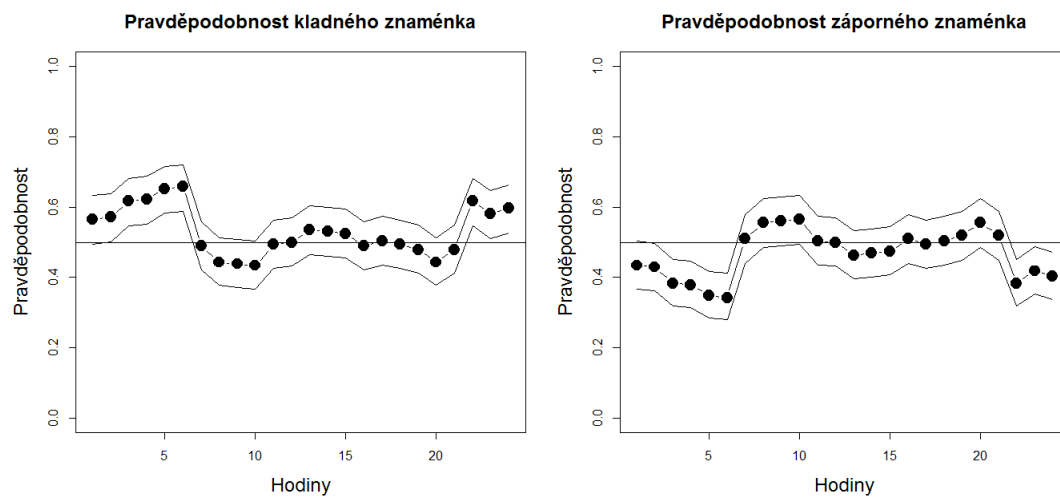
Budu predikovat znaménko na konkrétní hodinu dnes na základě toho, jaký den byl včera.

Určím pravděpodobnosti $p_j = P(Z_j = 1 | K_{t-1} = m)$ pro $j=1, \dots, 24$, kde $m \in S$.

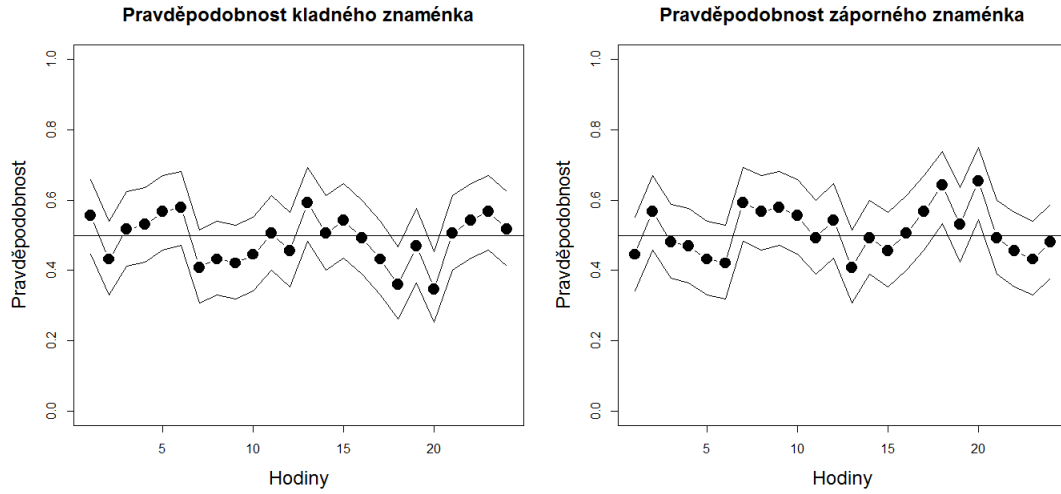
Obrázek 14: Včerejší den byl nadprůměrný



Obrázek 15: Včerejší den byl průměrný



Obrázek 16: Včerejší den byl podprůměrný



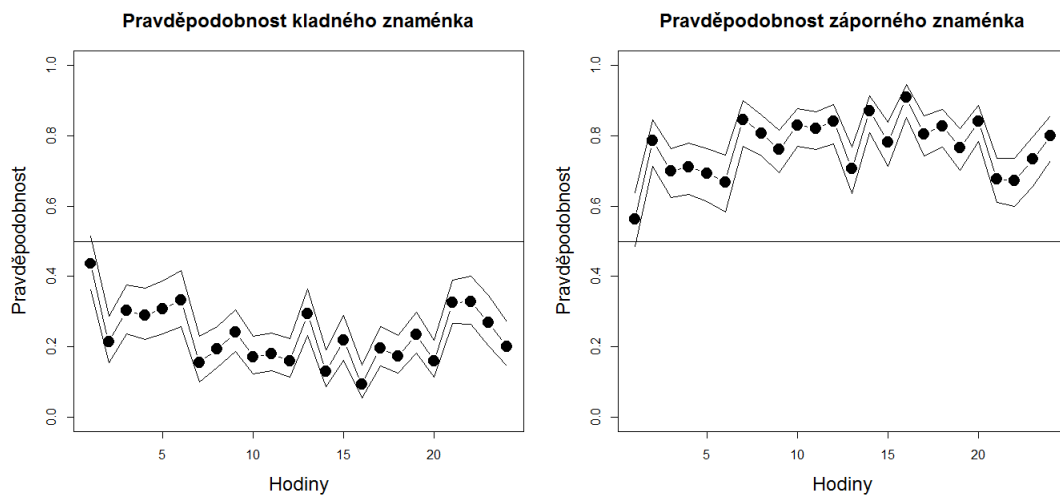
2. Pomocí informace o minulé hodině

Zde půjde o predikci znaménka v aktuální konkrétní denní hodině pomocí informace o znaménku v předchozí $j - 1$ hodině.

Určím pravděpodobnosti $p_j = P(Z_j = h | Z_{j-1} = l)$ pro $j = 1, \dots, 24$, kde $l, h = 0, 1$.

Přechod mezi jednotlivými dny řeším tak, že se podívám jaké je znaménko v poslední hodině ve včerejším dni a znaménko v první hodině dnes. Tuto pravděpodobnost představuje parametr $p_1 = (Z_1 = h | Z_0 = l)$. Tuto úvahu jsem použil i v předchozích výpočtech, kdy jsem určoval pravděpodobnost na základě informace z předchozí hodiny.

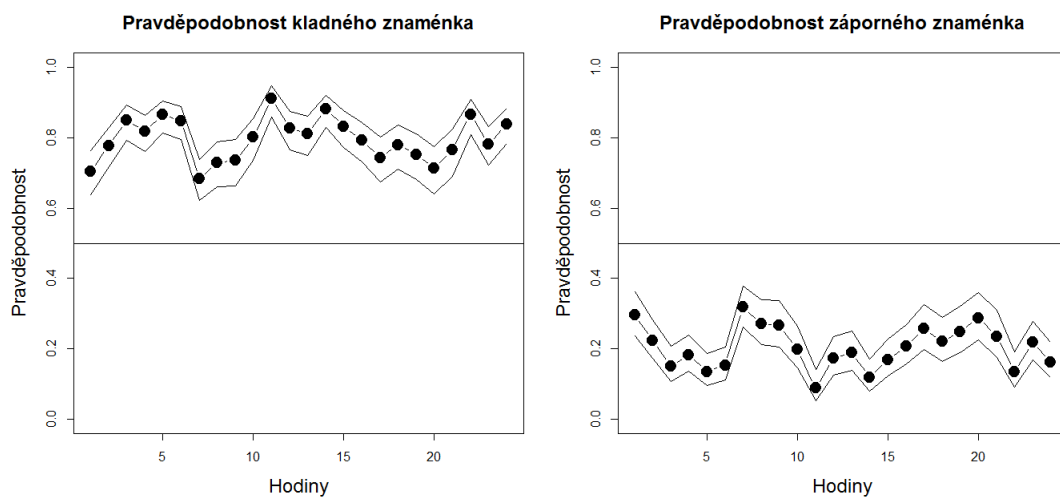
Obrázek 17: V předchozí $j - 1$ hodině SO měla záporné znaménko



Tabulka 19: Hodnoty odhadnutých parametrů

j	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{p}_j$	0.44	0.21	0.30	0.29	0.31	0.33	0.15	0.19	0.24	0.17	0.18	0.16
$1 - \hat{p}_j$	0.56	0.79	0.70	0.71	0.69	0.67	0.85	0.81	0.76	0.83	0.82	0.84
j	13	14	15	16	17	18	19	20	21	22	23	24
$\hat{p}_j$	0.29	0.13	0.22	0.09	0.20	0.17	0.23	0.16	0.32	0.33	0.27	0.20
$1 - \hat{p}_j$	0.71	0.87	0.78	0.91	0.80	0.83	0.77	0.84	0.68	0.67	0.73	0.80

Obrázek 18: V předchozí $j - 1$ hodině SO měla kladné znaménko



Tabulka 20: Hodnoty odhadnutých parametrů

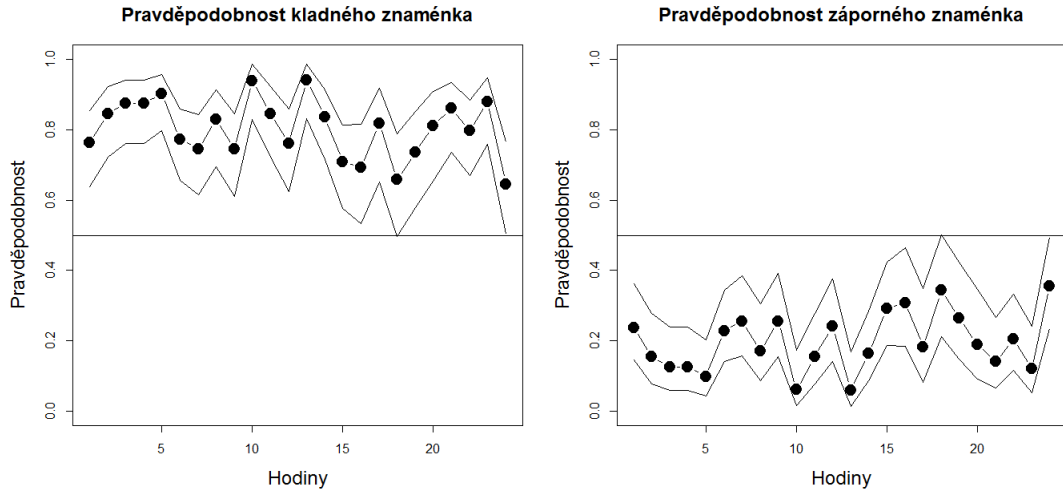
j	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{p}_j$	0.70	0.78	0.85	0.82	0.86	0.85	0.68	0.73	0.74	0.80	0.91	0.83
$1 - \hat{p}_j$	0.30	0.22	0.15	0.18	0.14	0.15	0.32	0.27	0.26	0.20	0.09	0.17
j	13	14	15	16	17	18	19	20	21	22	23	24
$\hat{p}_j$	0.81	0.88	0.83	0.79	0.74	0.78	0.75	0.71	0.76	0.87	0.78	0.84
$1 - \hat{p}_j$	0.19	0.12	0.17	0.21	0.26	0.22	0.25	0.29	0.24	0.13	0.22	0.16

3. Pomocí informace o předchozí hodině a včerejším dni

Předchozí odhady vylepším o informaci získanou ze shlukové analýzy.

Určím pravděpodobnosti $p_j = P(Z_j = h | Z_{j-1} = l, K_{t-1} = m)$ pro $j = 1, \dots, 24$, kde $l, h = 0, 1$ a $m \in S$.

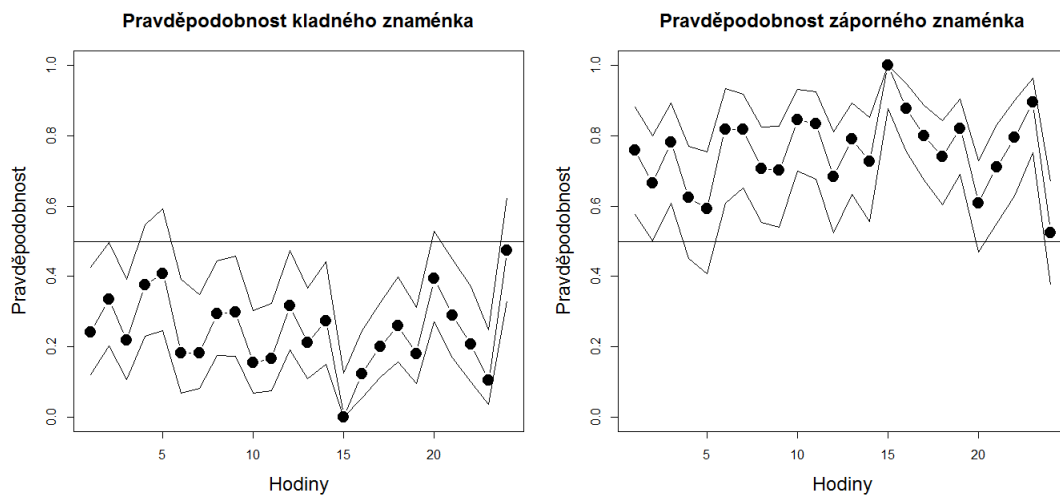
Obrázek 19: V předchozí $j - 1$ hodině SO měla kladné znaménko a včerejší den byl nadprůměrný



Tabulka 21: Hodnoty odhadnutých parametrů

j	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{p}_j$	0.76	0.85	0.88	0.88	0.90	0.77	0.74	0.83	0.74	0.94	0.85	0.76
$1 - \hat{p}_j$	0.24	0.15	0.12	0.12	0.10	0.23	0.26	0.27	0.26	0.06	0.15	0.24
j	13	14	15	16	17	18	19	20	21	22	23	24
$\hat{p}_j$	0.94	0.84	0.71	0.69	0.82	0.66	0.74	0.81	0.86	0.80	0.88	0.65
$1 - \hat{p}_j$	0.06	0.16	0.19	0.31	0.18	0.24	0.26	0.19	0.14	0.20	0.12	0.35

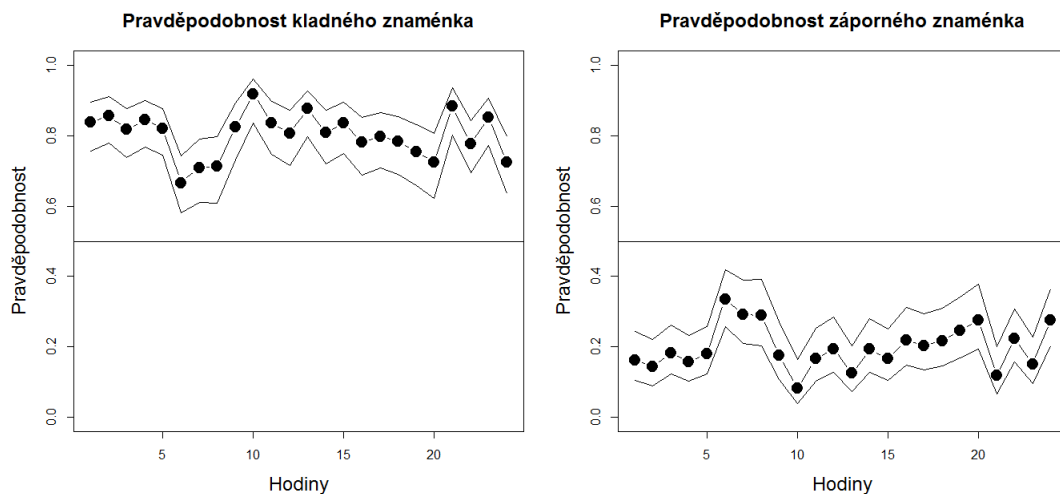
Obrázek 20: V předchozí $j - 1$ hodině SO měla záporné znaménko a
včerejší den byl nadprůměrný



Tabulka 22: Hodnoty odhadnutých parametrů

j	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{p}_j$	0.24	0.33	0.22	0.38	0.41	0.18	0.18	0.29	0.30	0.15	0.17	0.32
$1 - \hat{p}_j$	0.76	0.67	0.78	0.62	0.59	0.82	0.82	0.71	0.70	0.85	0.83	0.68
j	13	14	15	16	17	18	19	20	21	22	23	24
$\hat{p}_j$	0.21	0.27	0.00	0.12	0.20	0.26	0.18	0.39	0.29	0.21	0.10	0.48
$1 - \hat{p}_j$	0.79	0.73	1.00	0.88	0.80	0.74	0.82	0.61	0.71	0.79	0.90	0.52

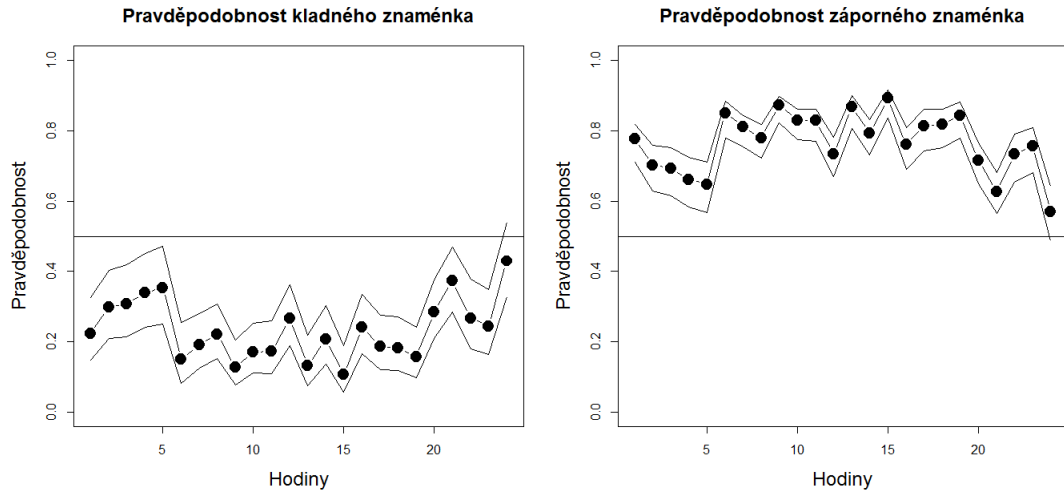
Obrázek 21: V předchozí $j - 1$ hodině SO měla kladné znaménko a
včerejší den byl průměrný



Tabulka 23: Hodnoty odhadnutých parametrů

j	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{p}_j$	0.84	0.86	0.82	0.84	0.82	0.67	0.71	0.71	0.83	0.92	0.84	0.81
$1 - \hat{p}_j$	0.16	0.14	0.18	0.16	0.18	0.33	0.29	0.29	0.17	0.08	0.16	0.19
j	13	14	15	16	17	18	19	20	21	22	23	24
$\hat{p}_j$	0.88	0.81	0.84	0.78	0.80	0.78	0.76	0.72	0.88	0.78	0.85	0.72
$1 - \hat{p}_j$	0.12	0.19	0.16	0.22	0.20	0.22	0.24	0.18	0.12	0.22	0.15	0.28

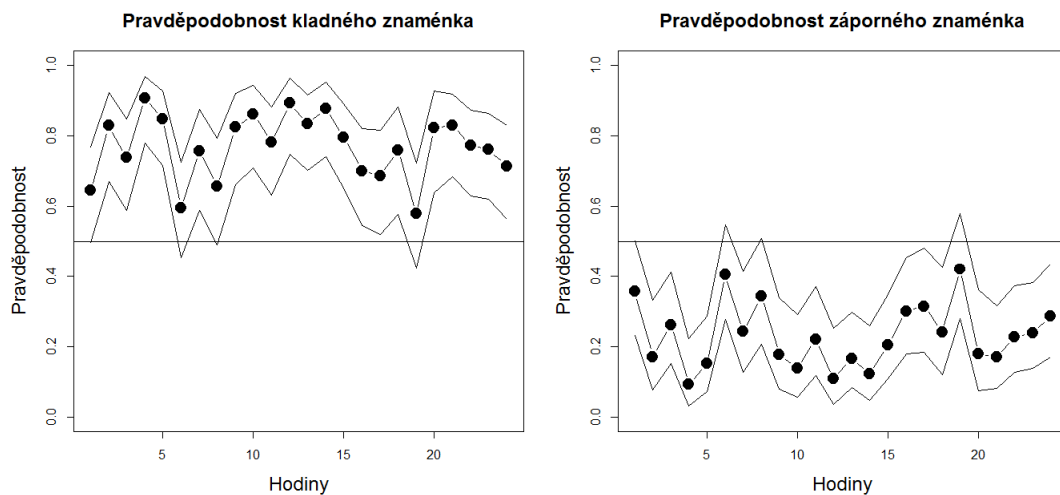
Obrázek 22: V předchozí $j - 1$ hodině SO měla záporné znaménko a včerejší den byl průměrný



Tabulka 24: Hodnoty odhadnutých parametrů

j	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{p}_j$	0.22	0.30	0.31	0.34	0.35	0.15	0.19	0.22	0.13	0.17	0.17	0.27
$1 - \hat{p}_j$	0.78	0.70	0.69	0.66	0.65	0.85	0.81	0.78	0.87	0.83	0.83	0.73
j	13	14	15	16	17	18	19	20	21	22	23	24
$\hat{p}_j$	0.13	0.21	0.11	0.24	0.19	0.18	0.16	0.28	0.37	0.27	0.24	0.43
$1 - \hat{p}_j$	0.87	0.79	0.89	0.76	0.81	0.82	0.84	0.72	0.63	0.73	0.76	0.57

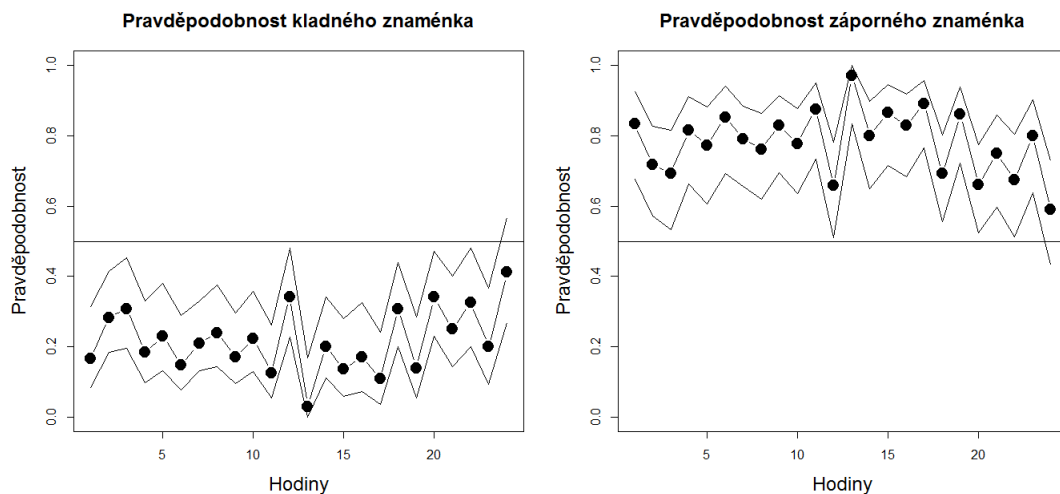
Obrázek 23: V předchozí $j - 1$ hodině SO měla kladné znaménko a
včerejší den byl podprůměrný



Tabulka 25: Hodnoty odhadnutých parametrů

j	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{p}_j$	0.64	0.83	0.74	0.91	0.85	0.60	0.76	0.66	0.82	0.86	0.78	0.89
$1 - \hat{p}_j$	0.36	0.17	0.26	0.09	0.15	0.40	0.24	0.34	0.18	0.14	0.22	0.11
j	13	14	15	16	17	18	19	20	21	22	23	24
$\hat{p}_j$	0.83	0.88	0.80	0.70	0.69	0.76	0.58	0.82	0.83	0.77	0.76	0.71
$1 - \hat{p}_j$	0.17	0.12	0.20	0.30	0.31	0.24	0.42	0.18	0.17	0.23	0.24	0.29

Obrázek 24: V předchozí $j - 1$ hodině SO měla záporné znaménko a
včerejší den byl podprůměrný



Tabulka 26: Hodnoty odhadnutých parametrů

j	1	2	3	4	5	6	7	8	9	10	11	12
$\hat{p}_j$	0.17	0.28	0.31	0.18	0.23	0.15	0.21	0.24	0.17	0.22	0.12	0.34
$1 - \hat{p}_j$	0.83	0.72	0.69	0.82	0.77	0.85	0.79	0.76	0.83	0.78	0.88	0.66
j	13	14	15	16	17	18	19	20	21	22	23	24
$\hat{p}_j$	0.03	0.20	0.14	0.17	0.11	0.31	0.14	0.34	0.25	0.32	0.20	0.41
$1 - \hat{p}_j$	0.97	0.80	0.86	0.83	0.89	0.69	0.86	0.66	0.75	0.68	0.80	0.59

Pokud se rozhoduji pouze na základě informace o včerejším dni, z obrázků 14, 15 a 16 je vidět, že pravděpodobnost správného tipu není moc velká a zřídka překročí 60 %. Rozhodnu-li se podle znaménka v minulé hodině, moje šance se výrazně zvýší a ve většině případů je větší než 70 %. Vždy můžu po kladném znaménku očekávat kladné a po záporném záporné. Je tomu tak i po přidání informace o včerejším dni. A jaký má tedy vliv přidaná podmínka na sledované pravděpodobnosti? Ideální by bylo, abych kladnou nebo zápornou systémovou odchylku v dané hodině očekával s pravděpodobností blízko jedné. Nebo jinak, aby absolutní hodnota rozdílu těchto pravděpodobností byla co nejblíže jedničce. A přidání podmínky o včerejším dni tento efekt přináší. Nelze konkrétně říct, v kterou denní dobu se toto děje. Ale podmíním-li daný jev nadprůměrným nebo podprůměrným dnem, začne se tato absolutní hodnota rozdílu zvyšovat ve velké většině případů.

4. Ověření modelu

Provedu posouzení kvality mnou navrženého predikčního modelu. K tomu využiji data o systémové odchylce z roku 2013. Tyto dny roztřídím do tří shluků, podle vzdálenosti od centroidů. Centoridy mám vypočtené ze shlukování Wardovou metodou a den zařadím do toho shluku, od jehož centroidu má nejmenší vzdálenost. Takto jsem získal shluk 100 nadprůměrných dní, shluk 200 průměrných dní a shluk podprůměrných dní s 65 dny. Strategii, kterou zakládám pouze na informaci o včerejším dni nebudu zkoumat podrobně. Pro obecný přístup byla její úspěšnost těsně pod 50 %, pro ten hodinový to bylo 57 %. Pro strategie založené na znaménku v minulé hodině a včerejším dni víme, že po kladném vždy sázím na kladné znaménko a po záporném na záporné. Takto se budu řídit při obecném i při hodinovém přístupu. Úspěšnost těchto strategií je 78 %. Pro bližší srovnání mezi sebou porovnáám očekávané pravděpodobnosti predikčního modelu se skutečností, kterou představuje rok 2013. Od pravděpodobnosti predikčního modelu, odečtu pravděpodobnost určenou z dat roku 2013. Pokud bude tento rozdíl kladný, byla má předpověď příliš optimistická. Bude-li záporný, měl jsem při svém tipu větší jistotu než jakou mi dal predikční model. Také mě bude zajímat, o kolik procent se liším od skutečnosti, a jestli tyto rozdíly pokrývají intervaly spolehlivosti. Nebude-li interval spolehlivosti daný rozdíl pokrývat, přidám k této hodnotě *. Jinak tato hodnota bude intervalem pokryta.

1. Pomocí informace o předchozí hodině – obecný přístup

Tabulka 27: V předchozí hodině SO měla záporné znaménko

Sázka	Rozdíl
$P(Y_t = 0 Y_{t-1} = 0)$	0.025*

Tabulka 28: V předchozí hodině SO měla kladné znaménko

Sázka	Rozdíl
$P(Y_t = 1 Y_{t-1} = 1)$	-0.009

2. Pomocí informace o předchozí hodině a včerejším dni – obecný přístup

Tabulka 29: V předchozí hodině SO měla záporné znaménko a včerejší den byl nadprůměrný

Sázka	Rozdíl
$P(Y_t = 0 Y_{t-1} = 0, K_{t-1} = 1)$	0.027*

Tabulka 30: V předchozí hodině SO měla kladné znaménko a včerejší den byl nadprůměrný

Sázka	Rozdíl
Sázka $P(Y_t = 1 Y_{t-1} = 1, K_{t-1} = 1)$	-0.006

Tabulka 31: V předchozí hodině SO měla záporné znaménko a včerejší den byl průměrný

Sázka	Rozdíl
$P(Y_t = 0 Y_{t-1} = 0, K_{t-1} = 2)$	0.039*

Tabulka 32: V předchozí hodině SO měla kladné znaménko a včerejší den byl průměrný

Sázka	Rozdíl
$P(Y_t = 1 Y_{t-1} = 1, K_{t-1} = 2)$	-0.008

Tabulka 33: V předchozí hodině SO měla záporné znaménko a včerejší den byl nadprůměrný

Sázka	Rozdíl
$P(Y_t = 0 Y_{t-1} = 0, K_{t-1} = 3)$	-0.014

Tabulka 34: V předchozí hodině SO měla kladné znaménko a včerejší den byl nadprůměrný

Sázka	Rozdíl
$P(Y_t = 1 Y_{t-1} = 1, K_{t-1} = 3)$	0.006

3. Pomocí informace o předchozí hodině – hodinový přístup

Tabulka 35: V předchozí $j - 1$ hodině SO měla záporné znaménko

j	1	2	3	4	5	6	7	8	9	10	11	12
Rozdíl	-0.13*	0.09*	-0.09*	0.03	0.06	0.19*	0.14*	0.02	-0.07*	0.04	-0.02	0.04
j	13	14	15	16	17	18	19	20	21	22	23	24
Rozdíl	-0.13*	0.09*	0.00	0.12*	-0.01	0.11*	0.01	0.12*	-0.08	-0.06	-0.08*	0.22*

Tabulka 36: V předchozí $j - 1$ hodině SO měla kladné znaménko

j	1	2	3	4	5	6	7	8	9	10	11	12
Rozdíl	-0.12*	-0.09*	0.01	0.01	0.01	-0.04	-0.16*	-0.02	-0.05	-0.01	0.07*	0.02
j	13	14	15	16	17	18	19	20	21	22	23	24
Rozdíl	-0.02	0.03	-0.04	-0.01	-0.03	0.04	-0.04	0.02	0.05	0.18*	-0.03	0.02

4. Pomocí informace o předchozí hodině a včerejším dni – hodinový přístup

Tabulka 37: V předchozí $j - 1$ hodině SO měla záporné znaménko a včerejší den byl nadprůměrný

j	1	2	3	4	5	6	7	8	9	10	11	12
Rozdíl	0.28*	-0.09	0.14	0.20*	-0.01	0.14	-0.03	-0.02	-0.02	0.10	-0.01	-0.04
j	13	14	15	16	17	18	19	20	21	22	23	24
Rozdíl	0.00	-0.08	0.10	0.10	0.00	-0.01	-0.02	-0.09	0.02	0.14	0.20*	-0.07

Tabulka 38: V předchozí $j - 1$ hodině SO měla kladné znaménko a včerejší den byl nadprůměrný

j	1	2	3	4	5	6	7	8	9	10	11	12
Rozdíl	-0.01	0.00	0.05	-0.04	0.03	0.09	0.01	-0.01	-0.13*	0.19*	-0.10*	-0.09
j	13	14	15	16	17	18	19	20	21	22	23	24
Rozdíl	0.05	0.00	-0.06	-0.11	-0.01	-0.07	-0.04	0.11	0.02	-0.01	0.01	-0.20*

Tabulka 39: V předchozí $j - 1$ hodině SO měla záporné znaménko a včerejší den byl průměrný

j	1	2	3	4	5	6	7	8	9	10	11	12
Rozdíl	0.01	0.01	-0.05	0.01	0.04	0.04	-0.04*	0.02	0.12*	0.08*	0.06*	-0.06*
j	13	14	15	16	17	18	19	20	21	22	23	24
Rozdíl	0.12*	0.03	0.09*	-0.06*	0.07	0.15*	0.08*	0.00	-0.07*	0.17*	0.10*	-0.07

Tabulka 40: V předchozí $j - 1$ hodině SO měla kladné znaménko a včerejší den byl průměrný

j	1	2	3	4	5	6	7	8	9	10	11	12
Rozdíl	-0.02	0.03	-0.10*	-0.03	-0.06*	-0.03	-0.04	-0.04	0.01	0.07	0.00	-0.02
j	13	14	15	16	17	18	19	20	21	22	23	24
Rozdíl	0.09*	-0.02	0.05	0.00	-0.01	0.04	0.00	0.02	0.15*	-0.04	-0.07*	-0.06

Tabulka 41: V předchozí $j - 1$ hodině SO měla záporné znaménko a včerejší den byl podprůměrný

j	1	2	3	4	5	6	7	8	9	10	11	12
Rozdíl	-0.02	0.03	-0.10	-0.03	-0.06	-0.03	-0.04	-0.04	0.01	0.07	0.00	-0.02
j	13	14	15	16	17	18	19	20	21	22	23	24
Rozdíl	0.09	-0.02	0.05	0.00	-0.01	0.04	0.00	0.02	0.15	-0.04	-0.07	-0.06

Tabulka 42: V předchozí $j - 1$ hodině SO měla kladné znaménko a včerejší den byl podprůměrný

j	1	2	3	4	5	6	7	8	9	10	11	12
Rozdíl	-0.17*	0.06	-0.13*	0.14*	0.10	0.03	0.13	-0.05	0.02	0.10	0.07	0.09
j	13	14	15	16	17	18	19	20	21	22	23	24
Rozdíl	-0.13*	0.14*	-0.13*	-0.09	-0.14*	0.05	-0.09	0.14	0.08	0.04	-0.12*	-0.03

U obecného přístupu je moje předpověď příliš optimistická sázím-li na záporné znaménko. Jestliže přidám informaci o včerejším dni, tak toto přecenění pro nadprůměrný a průměrný den ještě narůstá. Pro podprůměrný včerejší den tato vlastnost odpadá. Na kladné znaménko za všech situací sázím velmi přesně s rozdílem do 1 %. Budu-li k predikci přistupovat hodinově, nelze tvrdit, že bych se v konkrétní hodiny dopouštěl chyby systematicky. Toto se mění v závislosti na podmínce a přidání podmínky o včerejším dni se tato informace podstatně mění. Po nadprůměrném dni při sázce na kladné znaménko se chyby větší než 10 % dopustím pouze ve dvou hodinách. Po podprůměrném dni při sázce na záporné znaménko se chyby větší než 10 % dopustím pouze v jedné hodině. Informace o včerejším dni tedy přináší zlepšení mé znalosti toho, kdy se nejčastěji při svém tipu dopouštím chyby.

Závěr

Jako první jsem představil analyzovanou veličinu a popsal, jakým způsobem je získávána. Dále jsem uvedl určitá pravidla na trzích s elektřinou, která jsou závislá na hodnotě systémové odchylky a říkají, jak s touto veličinou obchodovat.

V další části jsem se věnoval shlukové analýze, od níž jsem si sliboval informaci, která pomůže vylepšit mé predikce. Vyzkoušel jsem několik shlukových metod a ukázal jejich výhody a nedostatky. Výstupy těchto výpočtů jsem prezentoval především graficky a na základě určitých kritérií vybral jednu pro další postup.

Nejvýznamnější část mé práce je uvedena ve třetí kapitole, kde navrhuji predikční model pro znaménko systémové odchylky. Popisuji několik různých strategií založených na informaci ze shlukové analýzy a znaménku systémové odchylky. Model otestuji na datech z roku 2013 a sleduji úspěšnost jednotlivých strategií. Ukázalo se, že strategie založená na informaci ze shlukové analýzy není efektivní. Jako určující se ukázala informace o znaménku systémové odchylky. Přidání informace o shlucích prakticky moji sázku nijak nemění a vždy tipuju znaménko z předchozí hodiny. Jako užitečná se ukazuje především při hodinovém přístupu, kde mi pomůže určit situace, kdy se dopouštím největší chyby ve své predikci. Shluková analýza tedy nezvýší můj zisk, ale pokud bych se rozhodl neobchodovat v každou hodinu, mohla by minimalizovat moji ztrátu.

Cílem práce bylo předvést metody shlukové analýzy na reálných datech a řešení praktického problému pomocí statistických metod.

Literatura

- [1] Anděl, J.: *Základy matematické statistiky*, Matfyzpress, 2011
- [2] Anděl, J.: *Matematika náhody*, Matfyzpress, 2007
- [3] Clustering in R: The Cluster Package, <http://cran.r-project.org/web/packages/cluster/index.pdf>
- [4] Hebák, P. a kol.: *Vícerozměrné statistické metody [3]*, Informatorium, 2007
- [5] Johnson, R.A. a Wichern, D.V.: *Applied multivariate statistical analysis*, Springer, 2007
- [6] OTE, a.s.. In: O společnosti [online]. Praha. [cit. 2014-04-07]. Dostupné z: http://www.ote-cr.cz/o-spolecnosti/files-statutarni-organy/Product_Sheet_CZ.pdf