UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ANALÝZA VZŤAHOV MEDZI ČASOVÝMI RADMI METÓDAMI SIEŤOVEJ ANALÝZY A ZHLUKOVANIA

DIPLOMOVÁ PRÁCA

2018 Radka LITVAJOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ANALÝZA VZŤAHOV MEDZI ČASOVÝMI RADMI METÓDAMI SIEŤOVEJ ANALÝZY A ZHLUKOVANIA

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	9.1.9. Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	doc. RNDr. Beáta Stehlíková, PhD.

Bratislava 2018 **Bc. Radka LITVAJOVÁ**

52556936

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Radka Litvajová
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Analýza vzťahov medzi časovými radmi metódami sieťovej analýzy a zhlukovania.
Analysis of relations between time series using methods of network analysis and clustering.

Cieľ: Vzdialenosť medzi stacionárnymi časovými radmi sa dá merať využitím korelácie medzi nimi, z ktorej sa dá vytvoriť funkcia vzdialenosti. Takisto bol navrhnutý spôsob, ktorým sa dá merať vzdialenosť korelačných matíc. Tieto koncepty sa využijú na analýzu vybraných časových radov. Matica vzdialeností sa využije na zhlukovanie, ako aj na konštrukciu najlacnejšej kostry. V grafoch sa bude analyzovať existencia komunít, centralita jednotlivých vrcholov a stabilita týchto výsledkov v čase.

Vedúci: doc. RNDr. Beáta Stehlíková, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 26.01.2017

Dátum schválenia: 27.01.2017

prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou by som sa chcela poďakovať svojej vedúcej diplomovej práce doc. RNDr. Beáte Stehlíkovej, PhD. za usmernenie, trpezlivosť a odbornú pomoc pri písaní tejto práce. Ďalej by som sa chcela poďakovať mojej rodine, ktorej vďačím za morálnu podporu počas písania práce. Osobitné poďakovanie však patrí všemohúcemu Bohu, ktorý mi dával múdrosť počas celého štúdia.

Abstrakt v štátnom jazyku

LITVAJOVÁ, Radka: Analýza vzťahov medzi časovými radmi metódami sieťovej analýzy a zhlukovania [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: doc. RNDr. Beáta Stehlíková, PhD., Bratislava, 2018, 96 s.

Vzdialenosť medzi časovými radmi sa dá merať rôznymi metódami. Populárnym nástrojom na vyjadrenie týchto vzťahov sa stali korelácie. V poslednom období vzniklo viacero prác, ktoré sa zaoberajú týmto problémom. Cieľom práce je skúmanie vzdialeností časových radov inými prístupmi, a tak obohatiť a priniesť nový pohľad na túto problematiku. Preto sme sa v našej práci okrem klasických metód zamerali aj na techniky, ktoré nie sú v týchto prácach zahrnuté. Pomocou výpočtu parciálnych korelácií alebo použitím LASSO metódy sa dajú vzťahy medzi časovými radmi veľmi dobre analyzovať. Na vyjadrenie vlastností používame rôzne typy grafov, ako napríklad najlacnejšie kostry alebo grafy korelačných vzťahov. V práci sa tiež zaoberáme centralitou vrcholov, koeficientmi zhlukovania a v závere práce priblížime testy na zisťovanie zhody korelačných matíc, čo s touto témou úzko súvisí.

Kľúčové slová: časový rad, korelácie, LASSO metóda, zhluková analýza, najlacnejšia kostra grafu, graf korelačných vzťahov

Abstract

LITVAJOVÁ, Radka: Analysis of relations between time series using methods of network analysis and clustering [Diploma Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics, and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. RNDr. Beáta Stehlíková, PhD., Bratislava, 2018, 96 p.

The distance between time series is analyzed by different methods. A correlation has become the popular tool for expressing these relationships. Recently, a number of works has been done to address this issue. The aim of this thesis is to investigate the distance of time series by other approaches and by means of this work enrich and bring a new perspective on this issue. Therefore, in our work, apart from classical methods, we have also focused on techniques that are not included in these works. Using partial correlation or using the LASSO method, relationships between time series can be very well analyzed. To express the properties, we use different types of graphs, such as the minimum spanning tree or correlation graphs. We also deal with centrality indices, clustering coefficients, and at the end of the work, we approach tests of differences between correlation matrices, which are closely related to this topic.

Keywords: time series, correlation, LASSO method, cluster analysis, minimum spanning tree, correlation graph

Obsah

Úvod

Časový rad vyjadruje postupnosť po sebe nasledujúcich údajov, ktoré charakterizujú následnosť udalostí v rovnakom časovom rozostupe. Stacionárny časový rad je definovaný ako taký, ktorý pri posune na časovej osi nemení rozdelenie pravdepodobností, teda vlastnosti ako priemer, odchýlka a autokorelácie ostávajú rovnaké.

Vzdialenosť časových radov sa dá merať prostredníctvom korelácií. Toto metrické vyjadrenie vzdialenosti má široké využitie. Dá sa aplikovať v rôznych odvetviach matematiky. My budeme korelácie používať predovšetkým v zhlukovej analýze.

Obyčajné korelácie však nie sú jediným prostriedkom na vyjadrenie takýchto vzťahov, dokonca v niektorých prípadoch dochádza k nepresnostiam, ktoré môžu výrazne ovplyvniť interpretáciu výsledkov. Preto sme sa v práci zamerali aj na iné, menej bežné metódy.

Touto problematikou sa zaoberalo viacero autorov, ktorí svoje výsledky publikovali v článkoch. Niektoré z nich sa stali motiváciou pre našu prácu. V práci [18] sa autori zaoberali vizualizáciou zložitých vzťahov na devízových trhoch. Pomocou korelácií a najlacnejšej kostry grafu analyzovali rozmanité vzťahy európskych ako aj svetových mien. Publikácia [19] je venovaná úrokovým mieram na kapitálových a finančných trhoch. Úrokové miery a dlhopisy majú veľmi podobné správanie, čo vypovedá o vysokej korelácii. Autori skúmali vzťahy pomocou zhlukovej analýzy. Na vyhodnocovanie korelácií medzi časovými radmi využili hierarchické zhlukovanie.

V poslednom období sa populárnym nástrojom na vizualizáciu finančných vzťahov stali siete. V práci [20] autori využívajú korelačnú sieť na zachytenie dynamiky vzťahov medzi výnosmi cien akcií. Najlacnejšia kostra grafu poslúžila na zoskupenie akcií podľa vlastností a zatriedenie do skupín podľa hospodárskych sektorov.

V mnohých prácach sú vzťahy medzi časovými radmi vyjadrené koreláciami, čo sú základné postupy pri tvorení sietí. Na základe ich výpočtu sú časové rady analyzované najlacnejšou kostrou grafu. V práci [1] autor rozoberá talianske akciové trhy s využitím výnosov firiem a štyrmi rôznymi metódami zostrojuje najlacnejšie kostry, ktoré následne porovnáva. Vlastnosti devízových trhov, vo vzťahu k menovým krízam prístupom najlacnejšej kostry, sú analyzované v článku [2]. Ďalšia práca [3] sa zaoberá priemyslom obnoviteľných zdrojov. Pomocou najlacnejšej kostry rozoberá spoločnosti

v tomto odvetví, ich postavenie v rozvoji obnoviteľných zdrojov z hľadiska kapitálových trhov.

Obsahom diplomovej práce je skúmanie vzťahov medzi vybranými časovými radmi. Aplikáciou metód zhlukovej analýzy budeme vizualizovať naše výsledky pomocou grafov, ako je najlacnejšia kostra grafu alebo graf korelačných vzťahov. Našu pozornosť upriamime na tvorbu komunít v grafe rôznymi algoritmami, centralitu pozorovaných objektov a meranie zhody korelačných matíc.

Doterajšie práce, zaoberajúce sa problematikou vzdialeností časových radov a tvorbou komunít, využívajú ako základný stavebný prvok korelácie. Použitím parciálnych korelácií a metódy LASSO túto problematiku rozšírime a poskytneme nový pohľad na skúmanie vzťahov medzi časovými radmi.

Práca je rozdelená do piatich častí. V prvej kapitole priblížime výpočet obyčajných a parciálnych korelácií. Stručne vysvetlíme fungovanie niektorých metód, ktoré sa budú v práci používať. Následne tieto poznatky aplikujeme na jednoduchom názornom príklade o výmenných kurzoch. Podrobnejšie sa parciálnym koreláciám budeme venovať v druhej kapitole. V tejto časti tiež riešime problematiku centrality vrcholov a tendenciu zhlukovania sa uzla, vyjadrenú veličinou nazývajúcou sa koeficient zhlukovania. Tretia kapitola sa zameriava na porovnávanie dvoch období, pred svetovou finančnou krízou a počas krízy. Rozdiely medzi časovými radmi modelujeme na multifaktorových produktivitách. Pre lepšie pozorovanie zmien zostrojíme intervaly spoľahlivosti. V štvrtej časti, parciálne korelácie vyjadríme iným spôsobom, pomocou regularizačnej metódy LASSO. Jej aplikácia je modelovaná na názornom príklade, ktorý rozoberá vzťahy rastu HDP krajín Eurozóny. Posledná kapitola bude venovaná testom zhody korelačných matíc vypočítaných z časových radov, ktoré vyjadrujú rast zamestnanosti v európskych krajinách.

V práci budeme pracovať v prostredí softvéru R [4]. Všetky dáta, prostredníctvom ktorých budeme teóriu aplikovať na názorných príkladoch, čerpáme z internetovej stránky Organizácie pre hospodársku spoluprácu a rozvoj, *www.oecd.org* [5].

1 Výmenné kurzy

Na názornom príklade vysvetlíme výpočet korelácií a tiež aplikáciu rôznych metód, ktorých hlavným prvkom sú korelácie. Budeme pracovať so sadou ročných dát, ktoré sme čerpali z internetovej stránky Organizácie pre hospodársku spoluprácu a rozvoj [5]. Dáta pozostávajú z výmenných kurzov 18 mien sveta počas rokov 1992 až 2016. Výmenný kurz vyjadruje cenu jednej meny vo vzťahu k druhej mene. Tento ukazovateľ je meraný vzhľadom k národnej mene dolár.

1.1 Korelácie

Jeden z najvýznamnejších a najviac používaných prostriedkov na opísanie závislosti medzi pozorovanými premennými je korelácia. Korelačný koeficient označujeme premennou $\rho_{X,Y}$, kde X a Y sú náhodné premenné. Vo všeobecnosti vyjadruje numerickú mieru korelácie medzi premennými. Existuje viacero druhov korelácií, my budeme pracovať s jedným z najbežnejších typov, Pearsonovým korelačným koeficientom. Meria lineárnu závislosť. Tak ako ostatné, tiež nadobúda hodnoty z intervalu $[-1, 1]$. Ak sa jeho veľkosť blíži ku krajným hodnotám intervalu, hovoríme o silnej lineárnej korelácii. Naopak, korelácie približujúce sa k nule vyjadrujú čoraz väčšiu lineárnu nezávislosť, ale ani v tom prípade to nemusí nutne hovoriť o nezávislosti.

Pearsonov korelačný koeficient vypočítame nasledovne

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{D(X)D(Y)}}. \tag{1.1}$$

1.2 Stacionarita časových radov

Stacionárne časové rady, voľne povedané, sú tie, pri ktorých predpokladáme, že ich štatistické vlastnosti, ako stredná hodnota, disperzia, autokorelácie a iné, sa časom nemenia a ostávajú rovnaké v minulosti ako aj v budúcnosti. Väčšina metód slúžiacich na prognózu časových radov sú postavené na predpoklade ich stacionarity. Napríklad, pri zostrojovaní ARIMA modelov je táto požiadavka dôležitá [11]. Takisto pri opisovaní budúceho správania sa časových radov a pri zmysluplnej analýze štatistík a vzťahov

medzi nimi.

Stacionarita časového radu môže byť silná alebo slabá. Pri silnej stacionarite predpokladáme, že združená distribučná funckia $(X_{t_1}, X_{t_2}, \ldots, X_{t_k})$ pre časový rad $\{X_t, t \in Z\}$ je rovnaká ako posunutá distribučná funckia $(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_k+h})$. To znamená, že sa časom nemení a závisí len od posunu h a nie od času $(t_1, t_2, \ldots, t_k)$.

My budeme pracovať so slabou stacionaritou, ktorá má konštantnú strednú hodnotu a požaduje, aby autokorelácia bola daná ako

$$cov(x_t, x_s) = \gamma(|t - s|),$$

pričom γ je konštanta, a teda autokorelácia je závislá len od vzdialenosti t od s. Z toho vyplýva, že disperzia je tiež konštantná.

Nech premenná $\{u_t\}$ označuje biely šum, ktorý definujeme ako stacionárny proces s nulovou strednou hodnotou, bez autokorelácií a s konštantnou varianciou. Slúži na definíciu ostatných procesov a pri tvorbe modelov.

Každý proces, ktorý je stacionárny, sa dá zapísať v tvare, ktorý nazývame Woldova reprezentácia [17]

$$x_t = \mu + \sum_{j=0}^{\infty} \psi_j u_{t-j}, \quad \sum_{j=0}^{\infty} \psi_j^2 < \infty, \tag{1.2}$$

kde μ je konštanta a $\psi_0 = 1$. Pri zápise v tomto tvare platí:

$$E(x_t) = \mu, \qquad Var(x_t) = \sum \psi_j^2, \qquad Cov(x_t, x_{t+k}) = \sigma^2 \sum_j \psi_j \psi_{j+k}$$

Majme stacionárny časový rad daný predpisom

$$x_t = \beta + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} \cdots + \alpha_p x_{t-p} + u_t, \tag{1.3}$$

kde u_t je biely šum a β konštanta. Predpis sa dá upraviť do tvaru použitím operátora posunu L [17], ktorý vráti hodnotu procesu posunutú o jedno obdobie dozadu nasledovne

$$\alpha(L)x_t = \beta + u_t, \qquad \alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_p L^p.$$

Ďalej výraz upravíme do tvaru

$$x_t = \alpha^{-1}(L)(\beta + u_t),$$

kde $\alpha^{-1}(L)$ je inverzný operátor hľadaný v tvare $\alpha^{-1}(L) = 1 + \psi_1 L + \psi_2 L^2 + \ldots$, čo vyjadruje nekonečný rad s neznámymi koeficientami, ktoré rekurzívne dorátame metódou neurčitých koeficientov. Z rovnice

$$1 = \alpha(L)\alpha^{-1}(L)$$

porovnaním koeficientov pri L^j získame pre ψ_j nasledovnú diferenčnú rovnicu

$$\psi_k - \alpha_1\psi_{k-1} - \ldots - \alpha_p\psi_{k-p} = 0.$$

Ako sme spomínali vyššie, každý stacionárny rad sa dá zapísať v tvare (1.2), preto kvôli splneniu podmienky konvergencie $\sum \psi_j^2 < \infty$ vyplývajúcej z Woldovej reprezentácie, musia byť korene charakteristickej rovnice $\lambda^k - \alpha_1\lambda^{k-1} - \ldots - \alpha_p = 0$ vo vnútri jednotkového kruhu, a teda korene $\alpha(L) = 0$ sa musia nachádzať mimo jednotkového kruhu. Táto vlastnosť je podmienkou stacionarity časového radu (1.3). Ak nastane prípad, že proces má jednotkový koreň, potom je nestacionárny. Podrobnejšie o tejto problematike sa môžeme dočítať v článkoch [14, 15].

Uvedieme jednoduchý príklad. Uvažujme proces

$$x_t = x_{t-1} + u_t, \qquad x_0 = 0, \tag{1.4}$$

ktorý použitím operátora posunu L upravíme do tvaru

$$x_t - x_{t-1} = (1 - L)x_t = u_t.$$

Koreň charakteristického polynómu $(1-L) = 0$ je rovný 1, preto proces má jednotkový koreň.

Ak proces podrobne rozpíšeme, dostaneme nasledovný tvar

$$x_t = x_0 + \textstyle\sum_{j=1}^{t} u_j.$$

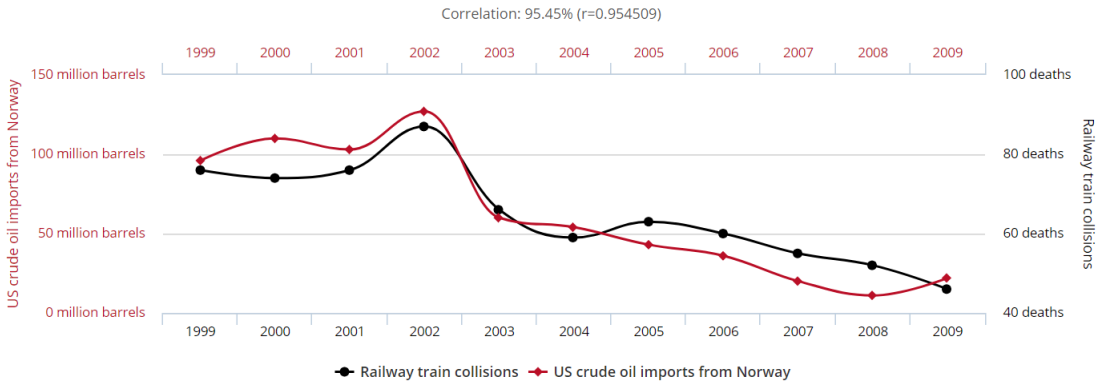
Ak predpokladáme, že u_t je biely šum, ktorý je charakterizovaný ako nekorelovaný stochastický proces s nulovou strednou hodnotou a konštantnou varianciou σ^2, potom pre varianciu procesu x_t platí

$$Var(x_t) = t\sigma^2.$$

Z toho vyplýva, že variancia procesu je závislá od t a s narastajúcim časom sa zväčšuje až donekonečna, preto je proces nestacionárny.

Majme dva nestacionárne časové rady X_t a Y_t pozorované v tom istom období. Ak podliehajú trendu, ten ovplyvňuje ich spomínané vlastnosti, ktoré už nemusia byť konštantné, ale závislé od času t. Preto, ak by sme sledovali ich koreláciu, potrebujeme stacionaritu časových radov, aby hodnoty neboli ovplyvnené spoločným trendom.

Uvedieme dva názorné príklady, v ktorých nestacionarita časových radov spôsobuje problémy. Prvý rozoberá koreláciu medzi dovozom ropy do USA z Nórska (*US crude oil imports from Norway*) a počtom vodičov zabitých pri kolízii s vlakom (*Railway train collisions*) [16]. Z Obr. 1.1 pozorujeme silnú závislosť aj napriek tomu, že tieto dve veličiny sa neovplyvňujú. Ich korelácia má veľkosť 0.9545.



Obr. 1.1: Porovanie dvoch nezávislých časových radov [16]

V druhom príklade použijeme naše dáta, s ktorými budeme v ďalšej časti pracovať. Na Obr. 1.2 je vykreslený priebeh dvoch časových radov - výmenných kurzov dvoch svetových mien voči mene dolár, CHF vľavo a JPY vpravo. Vývoj sa zdá byť odlišný, ale aj napriek tomu vypočítaná korelácia má hodnotu pomerne vysokú, rovnú 0.64. Lenže po stabilizácií časových radov diferencovním, aby sme z nich získali stacionárne časové rady, získame koreláciu, ktorá je takmer o polovicu menšia, s hodnotou 0.34.

Ak časové rady nie sú stacionárne, je nutné ich transformovať do tvaru, ktorý spĺňa túto požiadavku. Existuje viacero techník, ktoré riešia problém stacionarity. Diferencovanie je jedna z metód, ktorá vytvorí z nestacionárneho časového radu stacionárny eliminóciou trendu alebo odstránením jednotkového koreňa, a zároveň stabilizuje strednú

Obr. 1.2: Porovanie dvoch časových radov

hodnotu tohto radu.

Na zistenie stacionarity časového radu sa v štatistike používajú testy jednotkového koreňa. Nulová hypotéza vyjadruje prítomnosť jednotkového koreňa, čo znamená nestacionaritu. Existuje niekoľko takýchto testov, my použijeme *ADF test* (angl. *The Augmented Dickey-Fuller*) [6], ktorý je implementovaný v softvéri R [4] v balíčku *urca* [8] pod názvom *ur.df*. Naznačíme myšlienku tohto testu podľa [7].

Úpravou časového radu daného predpisom (1.3) získame nasledovný tvar (bez ujmy na všeobecnosti zvolíme $\beta = 0$)

$$\begin{aligned} x_t &= (\alpha_1 + \alpha_2 + \cdots + \alpha_p)x_{t-1} - (\alpha_2 + \cdots + \alpha_p)x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + u_t \\ &= (\alpha_1 + \alpha_2 + \cdots + \alpha_p)x_{t-1} - (\alpha_2 + \cdots + \alpha_p)(x_{t-1} - x_{t-2}) + \cdots + \alpha_p(x_{t-p-1} - x_{t-p}) + u_t. \end{aligned}$$

Kvôli zjednodušeniu preznačíme výrazy $(\alpha_2 + \cdots + \alpha_p) = -\theta_1, \ldots, \alpha_p = -\theta_{p-1}$, a rozdiel $x_{t-i} - x_{t-i-1}$ zapíšeme ako diferenciu Δx_{t-i} pre $i = 1, 2, \ldots, p$. Od obidvoch strán rovnice odčítame výraz x_{t-1} a daný predpis nadobudne nasledovný tvar

$$\Delta x_t = (\alpha_1 + \alpha_2 + \cdots + \alpha_p - 1)x_{t-1} + \theta_1 \Delta x_{t-1} + \cdots + \theta_p \Delta x_{t-p-1} + u_t. \tag{1.5}$$

Platí, že ak má uvedený proces jednotkový koreň, potom výraz $1 - \alpha_1 - \alpha_2 - \ldots - \alpha_p$ je rovný nule. Rovnosť ekvivalentne vyjadríme ako $\sum \alpha_i = 1$. Teda nulová hypotéza zodpovedajúca testu jednotkového koreňa má tvar: $H_0 : \sum \alpha_i = 1$.

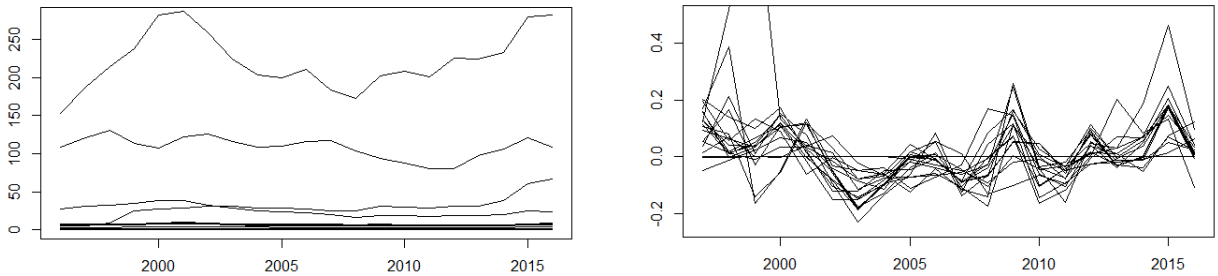
Hypotéza je odvodená z nasledujúcich úvah. Ak výraz (1.5) rozložíme do pôvodného tvaru, získame rovnicu

$$(1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_p L^p)x_t = u_t.$$

Jednotkový koreň v tomto prípade znamená, že ak operátok posunu L bude rovný 1, výraz $1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_p L^p = 0$, čiže $1 - \alpha_1 - \alpha_2 - \ldots - \alpha_p = 0$.

Ak sa vrátime k príkladu (1.4), nulová hypotéza v tomto prípade má tvar $H_0 : \alpha = 1$. Výraz $(1 - \alpha L) = 0$ spĺňa rovnicu len pre $L = 1$, pretože $\alpha = 1$. Z nulovej hypotézy teda vyplýva, že proces má jednotkový koreň.

Vráťme sa späť k príkladu o výmenných kurzoch. Pôvodné nestacionárne dáta musíme upraviť. Funkcia *ur.df* nezamieta prítomnosť jednotkového koreňa, preto dáta diferencujeme a zlogaritmujeme. Logaritmovanie slúži na stabilizáciu variancie. Obr. 1.3 zachytáva dva grafy. Prvý vykresľuje vývoj výmenných kurzov, druhý už upravené časové rady.

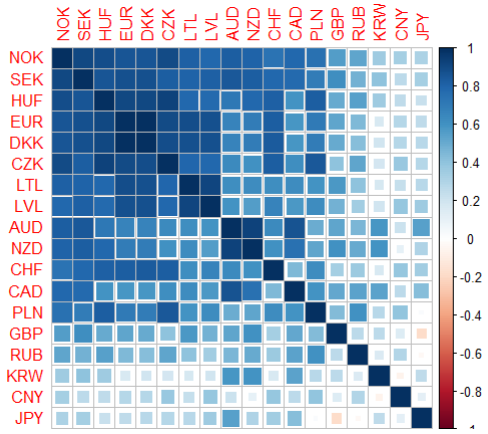


a) Vývoj výmenných kurzov v pôvodnom tvare

b) Vývoj výmenných kurzov v upravenom tvare

Obr. 1.3: Vykreslenie časových radov

Po odhade Pearsonových korelácií $\hat{\rho_{ij}}$ pomocou (1.1) môžeme korelačné vzťahy zobraziť graficky. Tieto vzťahy sú spoločne zachytené v korelačnej matici na Obr. 1.4.



Obr. 1.4: Korelačná matica

Z matice pozorujeme silnú závislosť niektorých výmenných kurzov. Lenže nevieme z

nej vyvodiť závery o tom, ktoré vzťahy sú signifikantné. Preto sa budeme ďalej zaoberať testovaním signifikancie korelačných koeficientov.

Prvým krokom takéhoto testovania je úprava korelačných koeficientov. Aby sme s koreláciami mohli ďalej pracovať, je dôležité ich upraviť do tzv. variačno-stabilizačného tvaru. Korelačný koeficient dosahuje ohraničené hodnoty z intervalu $[-1, 1]$. V práci budeme používať nasledovnú Fisherovu transformáciu

$$z_{ij} = tanh^{-1}(\hat{\rho}_{ij}) = \frac{1}{2} \log \frac{(1 + \hat{\rho_{ij}})}{(1 - \hat{\rho_{ij}})}. \tag{1.6}$$

Slúži na konvertovanie rozdelenia Pearsonových korelačných koeficientov na hodnoty s približne normálnym rozdelením. Podrobnejšie informácie o transformácií môžeme nájsť napr. v práci [39].

Ďalším dôležitým krokom je správny výpočet p-hodnoty, ktorá hovorí o sile hypotézy a jej signifikancii. Ak by sme testovali každú hypotézu nezávisle, ale pri rovnakej úrovni signifikancie, výsledná pravdepodobnosť nesprávneho zamietnutia by bola oveľa vyššia ako nominálna, ktorú sme použili pri samostatných testoch. Keďže chceme testovať viacero hypotéz súčasne, je nutné použiť jej upravenú hodnotu (angl. *adjusted p-value*) [22], ktorá berie do úvahy viacnásobné testovanie.

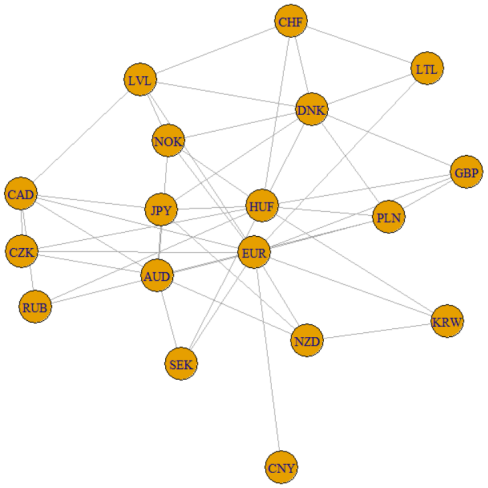
Korelácie vyjadrujú vzťahy medzi objektmi. Vzťahy môžu byť pozitívne alebo negatívne, záležia od znamienka korelácie. Následne tieto vzťahy môžeme zakresliť do spoločného grafu, kovariačného (korelačného) grafu $G = (V, E)$, kde objekty predstavujú vrcholy V a korelácie hrany medzi nimi

$$E = \{\{i, j\} \in V^{(2)} : \rho_{ij} \neq 0\}.$$

Okrem tohto spomínaného grafu existuje viacero tried grafov. Napríklad, ak sú hrany vážené, pričom váhy môžu byť dané koreláciami, hovoríme o vážených grafoch. Ďalším typom sú orientované a neorientované grafy, ktoré sa určujú podľa toho či hrany majú určenú orientáciu alebo nie.

My budeme používať vyššie opísaný korelačný graf. Pri zostrojovaní jeho hrán je dôležité rozoznať, ktoré vzťahy sú signifikantné. Je to ekvivalentné s testovaním hypotézy

$$H_0 : \rho_{ij} = 0 \qquad vs. \qquad H_1 : \rho_{ij} \neq 0. \tag{1.7}$$



Obr. 1.5: Korelačný graf

Ak sledujeme závislosť na N objektoch, testujeme $\frac{N(N-1)}{2}$ hypotéz simultánne. Toto číslo predstavuje množstvo potenciálnych hrán v grafe G.

V našom prípade pozorujeme 18 výmenných kurzov, ktoré môžu mať medzi sebou 153 významných vzťahov. Vypočítaním upravenej p-hodnoty a následným testovaním hypotéz (1.7), získame 46 signifikantných korelácií.

Vďaka vypočítaným signifikantným koreláciám vieme vykresliť graf korelačných vzťahov (Obr. 1.5), ktorých počet je výrazne menší než množstvo potenciálnych hrán v grafe.

1.3 Vzdialenosti vypočítané z korelácií

Dôležitou veličinou používanou v matematike je matica vzdialeností. V teórii grafov má široké využitie, napríklad v analýze hierarchického zhlukovania (angl. *Hierarchical Cluster Analysis*), čo je jedna z metód zhlukovej analýzy. Vlastnosti, ktorými sa vyznačuje sú: štvorcovosť, symetrickosť a nulová diagonála. Hodnoty v matici vyjadrujú vzdialenosti medzi pozorovanými objektami a môžu byť vypočítané pomocou rôznych metrík.

Našu maticu vzdialeností získame z korelačnej matice transformáciou korelačných koeficientov, vypočítaním euklidovských vzdialeností nasledovným spôsobom

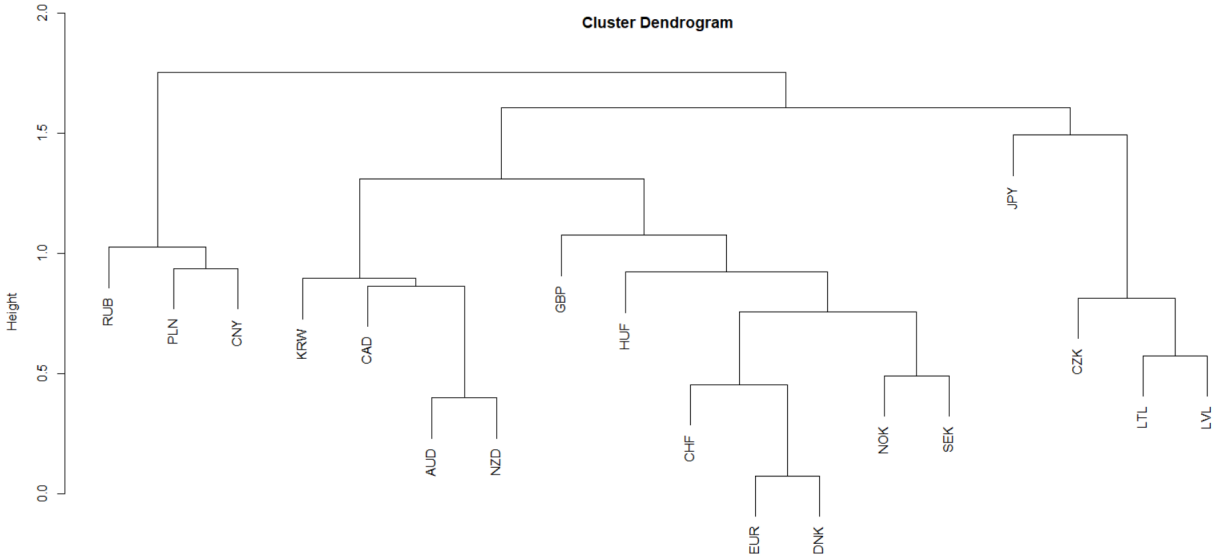
$$d_{ij} = \sqrt{2(1 - \rho_{ij})}.$$

Dôvod používania práve tejto vzdialenosti je odvodený v práci [20, 21]. Výhodou zvolenia tejto miery je, že všetky vzdialenosti d_{ij} ležia medzi hodnotami 0 a 2, keďže korelačný koeficient nadobúda hodnoty z intervalu $[-1, 1]$.

Usporiadané zhluky, získané pomocou hierarchického zhlukovania, sa zvyčajne graficky vyobrazujú pomocou špeciálnych stromových grafov, nazývaných dendrogramy. Tie obsahujú ramená spájajúce údaje - uzly, a tak vytvárajú strom určitej výšky. Táto výška vyjadruje vzdialenosť objektov, čím sú dva uzly od seba viac vzdialené, tým majú medzi sebou menšiu koreláciu.

Dendrogram získaný z vypočítaných hodnôt korelácií výmenných kurzov je zachytený na Obr. 1.6. Spomedzi všetkých vzdialeností má dvojica výmenných kurzov - dánska koruna (DKK) a euro (EUR), nachádzajúca sa v spodnej časti, najmenšiu vzdialenosť.

Na druhej strane, japonský jen (JPY) je v grafe umiestnený najvyššie, čo značí o jeho veľkej vzdialenosti od ostatných mien. Z definície vzdialenosti d_{ij} jeho postavenie hovorí o malej hodnote korelácie voči ostatným menám.



Obr. 1.6: Dendrogram

1.4 Stromy a najlacnejšie kostry

Teória grafov [24, 27] slúži na definovanie objektov a vzťahov medzi nimi. Každý graf $G = (V, E)$ sa skladá z množiny vrcholov V, ktoré predstavujú objekty a množiny hrán E vyjadrujúcich ich vzájomné vzťahy.

Strom je súvislý neprázdny graf, ktorý neobsahuje kružnicu. Je to grafické vyjadrenie členenia množiny a jej podmnožín. Označuje sa písmenom $T = (V, E)$.

Určiť či daný graf je strom môže byť v niektorých prípadoch zložité. Preto sa používajú rôzne definície a charakteristiky. Dôležitým pozorovaním je, že každý strom s aspoň dvomi vrcholmi obsahuje aspoň dva vrcholy stupňa jeden, t.j. koncové vrcholy alebo tiež nazývané *listy* (pozorovanie platí len pre konečné stromy). V tejto časti práce budeme pracovať s literatúrou [24].

Nasledujúca veta uvádza hlavné vlastnosti stromu v tvare piatich ekvivalencií.

Veta 1. *(Charakterizácia stromu podľa [24]) Nech $G = (V, E)$ je graf, pre ktorý sú nasledujúce podmienky ekvivalentné:*

i) G je strom.

ii) (jednoznačnosť cesty) Pre každé dva vrcholy $x, y \in V$ existuje práve jediná cesta z x do y.

iii) (minimálna súvislosť) Graf G je súvislý. Vynechaním ľubovoľnej hrany vznikne nesúvislý graf.

iv) (maximálny graf bez kružníc) Graf G neobsahuje kružnicu a pridaním novej hrany vznikne nový graf, ktorý už obsahuje kružnicu.

v) (Eulerov vzorec) Graf G je súvislý a platí $|V| = |E| + 1$.

Kostra grafu

Každý súvislý graf má kostru. Teda ak $G = (V, E)$ je graf, potom kostra grafu G je strom $T = (V, E')$, kde $E' \subseteq E$. Zároveň je podgrafom grafu G a obsahuje všetky vrcholy pôvodného grafu G. Platí vlastnosť, že medzi každými dvomi vrcholmi existuje práve jedna cesta.

Na nájdenie kostry grafu existujú rôzne algoritmy. My uvedieme dva z nich.

Algoritmus 1.

Nech $G = (V, E)$ je graf obsahujúci n vrcholov a m hrán, ktoré sú ľubovoľne zoradené do postupnosti $\{e_1, e_2, \ldots, e_m\}$.

Algoritmus postupným pridávaním hrán vytvára množiny $E_0 \subseteq E_1, \cdots \subseteq E$, kde $E_0 = \emptyset$. Končí, ak po pridaní ďalšej hrany vznikne kružnica, teda obsahuje $n-1$ hrán alebo sa už prebrali všetky hrany grafu G, teda $i = m$.

Algoritmus je daný nasledovnou formulkou

$$E_i = \begin{cases} E_{i-1} \cup \{e_i\} & \text{ak graf } \quad (V, E_{i-1} \cup \{e_i\}) \quad \text{neobsahuje kružnicu} \\ E_{i-1} & \text{inak.} \end{cases}$$

Tvrdenie spolu s dôkazom o správnosti algoritmu je popísaný v literatúre [24].

Algoritmus 2.

Rovnako ako v predchádzajúcom algoritme uvažujme, že graf $G = (V, E)$ je zložený z n vrcholov a m hrán. Postupne budeme vytvárať množiny vrcholov $V_0 \subseteq V_1 \cdots \subseteq V$ a množiny hrán $E_0 \subseteq E_1, \cdots \subseteq E$, kde $V_0 = \{v\}$ je ľubovoľný vrchol a $E_0 = \emptyset$.

Algoritmus funguje nasledovne: po vytvorení množín V_{i-1} a E_{i-1} nájdeme hranu $e_i = \{x_i, y_i\} \in E(G)$, kde $x_i \in V_{i-1}$ a $y_i \in V \backslash V_{i-1}$. Vzniknú nové množiny $V_i = V_{i-1} \cup \{y_i\}$ a $E_i = E_{i-1} \cup e_i$. V prípade, že taká hrana neexistuje, algoritmus končí. Algoritmus ako aj tvrdenie o jeho správnosti nájdeme v knihe [24].

Najlacnejšia kostra grafu

V tejto časti požadujeme graf $G = (V, E)$ s ohodnotenými hranami. To znamená, že každej hrane $e \in E$ priradíme určitú váhu $w(e)$ (nezápornú hodnotu), čo v niektorých prípadoch môže reprezentovať vzdialenosť medzi vrcholmi . Takýto graf s ohodnotením hrán $w : E \to R$ nazývame *sieť*.

V teórii grafov najlacnejšou kostrou grafu (angl. *Minimum Spanning Tree*) sa označuje špeciálny druh grafu, ktorý spája všetky vrcholy bez vytvorenia kružnice a zároveň veľkosť stromu, teda dĺžka hrán, je minimálna.

Pod problémom najlacnejšej kostry grafu rozumieme úlohu nájsť podgraf (V, E'), kde súčet ohodnotených hrán bude minimálny, čo môžeme zapísať, že uvedený súčet

$$w(E') = \sum_{e \in E'} w(e)$$

bude čo najmenší.

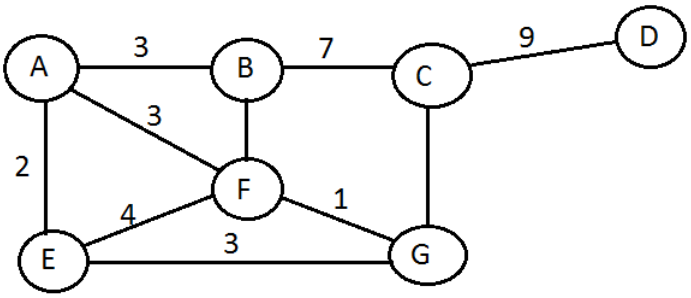
Graf môže obsahovať viacero kostier. Existuje niekoľko metód na riešenie tohto problému. V našej práci budeme používať Kruskalov algoritmus. Ostatné algoritmy (ako Borůvkov, Primov, ale aj zložitejšie) nájdeme v prácach [24, 25, 26].

Kruskalov algoritmus

Vstupom algoritmu je súvislý graf $G = (V, E)$ s ohodonotenými hranami w a ich usporiadaním podľa váhy $w(e_1) \leq w(e_2) \leq \cdots \leq w(e_m)$. Po aplikovaní algoritmu 1. získame Kruskalov algoritmus, tiež nazývaný pažravý algoritmus.

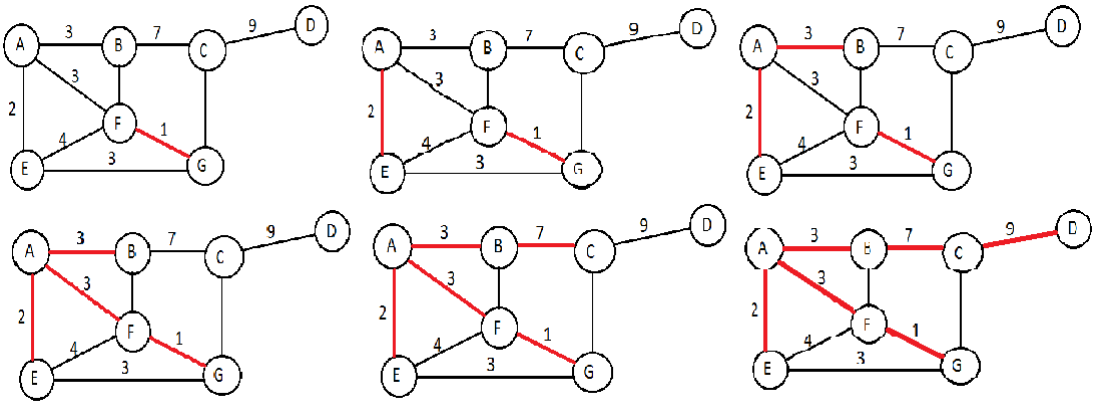
Uvedený algoritmus rieši problém najlacnejšej kostry, čo je podrobne rozobraté v literatúre [24].

Pre lepšie pochopenie uvedieme príklad Kruskalovho algoritmu. Pokúsime sa nájsť najlacnejšiu kostru nasledujúceho grafu.



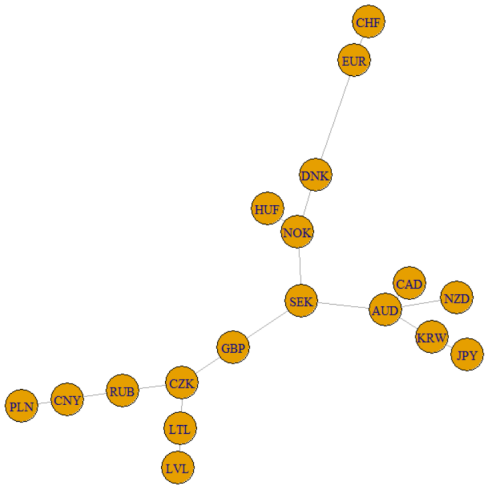
Kruskalov algoritmus postupne vyberá hrany s najmenšou váhou. Týmto spôsobom spojí všetky vrcholy. Podmienkou ale je, že spojením dvoch vrcholov nemôže vzniknúť kružnica.

V našom príklade ako prvú hranu vyberie s váhou 1, ktorá sa nachádza medzi vrcholmi F a G, druhú hranu medzi vrcholmi A a B s ohodonotením 2, atď. Postup, ako aj najlacnejšia kostra grafu, je vyznačený na nasledujúcom obrázku.



Najlacnejšia kostra a výmenné kurzy

Vráťme sa k príkladu o výmenných kurzoch. Na tieto dáta aplikujeme výpočet najlacnejšej kostry. Z vypočítanej matice vzdialeností odvodíme maticu dosažiteľnosti (angl. *Adjacency matrix*). Matica slúži na reprezentáciu konečného grafu a jej členy vyjadrujú či vrcholy v grafe spolu susedia alebo nie. Následne ju využijeme na výpočet najlacnejšej kostry. Pomocou funkcie *minimum.spanning.tree* z balíčka *igraph* [9], ktorej vstupom je práve táto matica, získame nasledujúcu kostru.



Obr. 1.7: Najlacnejšia kostra grafu

1.5 Zhlukovanie a tvorba komunít

Úlohou sieťovej analýzy je zoskupenie objektov do spoločných zhlukov, pričom tieto objekty zdieľajú spoločné znaky a sú si navzájom viac podobné ako s objektami z iných

zhlukov. Týmto spôsobom sa snaží nájsť štruktúru dát a na tvorbu skupín využíva len informácie získané z dát.

Existuje viacero spôsobov ako zoskupiť dáta do zhlukov. My sme v našej práci využívali balíček *igraph* [9], ktorý sa používa pri práci so sieťovou analýzou a vizualizáciou grafov. Sú v ňom zahrnuté algoritmy, ktoré sú schopné počítať zhluky v grafoch a mnoho iných funkcií.

V práci sme knižnicu a v nej implementované algoritmy využívali predovšetkým na zisťovanie komunít a ich tvorbu v najlacnejšej kostre. Ich procesy sa navzájom líšia, pretože sú založené na odlišných princípoch. Nižšie poskytujeme stručné vysvetlenie niektorých z nich [13].

- **cluster_edge_betweenness** - hierarchická metóda, ktorá na vytvorenie komunít využíva blízkosť (angl. *betweenness*) vrcholov, čo je jeden z centrálnych indexov. Algoritmus je zostavený z niekoľkých krokov. Najskôr z grafu postupne odstraňuje hrany s najvyššou hodnotou blízkosti. V druhom kroku znovu vypočíta blízkosť vrcholov a opakuje prvý krok. Proces sa opakuje, až kým nie sú odstránené všetky hrany z grafu.

- **cluster_fast_greedy** - nedeterministická aglomeratívna metóda, ktorá priamo optimalizuje výsledky modularity. Nájde optimálne lokálne riešenia, čo ale v niektorých prípadoch môže spôsobiť komplikácie. Algoritmus je schopný spracovať aj veľmi veľké siete a nájsť ich štruktúru.

- **cluster_label_prop** - algoritmus využíva metódu detekcie komunity. To znamená, že na začiatku je každý vrchol označený a v každom kroku vrcholy získavajú také označenie ako má väčšina susedných vrcholov. Algoritmus je schopný pracovať tiež s váženými hranami v grafe.

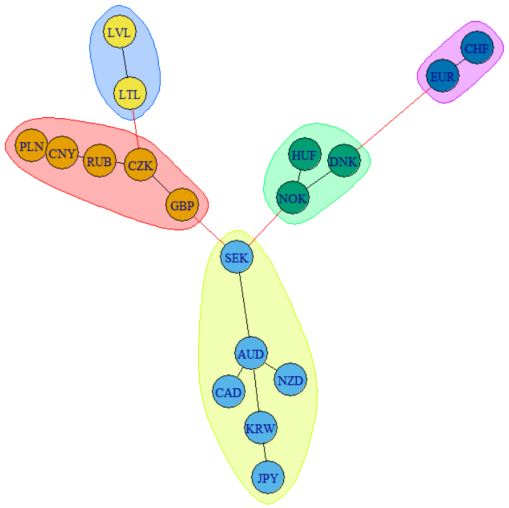
- **cluster_louvain** - algoritmus je založený na modularitnej optimalizácii. V každom kroku sa priradia do komunity objekty s najväčšou vzájomnou modularitou. Proces priraďovania prebieha dovtedy, kým nevznikne jeden veľký zhluk zložený zo všetkých objektov alebo pokiaľ sa modularita už nezvyšuje a ostáva rovnaká.

- **cluster_optimal** - hlavnou úlohou tejto metódy je transformácia modulárnej optimalizácie na problém celočíselného programovania. V tomto štádiu algorit-

mus pracuje s knižnicou GLPK, čo je softvér na riešenie rozsiahleho lineárneho programovania.

- **cluster_walktrap** - proces je založený na metóde náhodnej prechádzky, pomocou ktorej vrcholom priraďuje podobnosť. Tento algoritmus patrí medzi najrýchlejšie, je veľmi efektívny a schopný vhodne zachytiť štruktúru zhlukov.

Aplikáciou funkcie *cluster_walktrap* na vypočítanú najlacnejšiu kostru na Obr. 1.7, získame zhluky výmenných kurzov na Obr. 1.8.

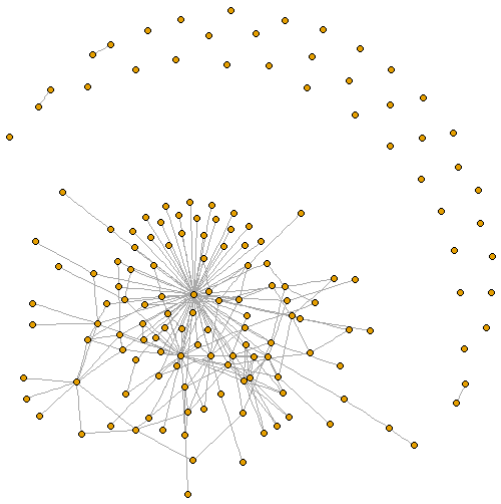


Obr. 1.8: Najlacnejšia kostra grafu a zhluky

1.6 Parciálne korelácie

Uvedieme príklad podľa [23]. K dispozícii máme výber 153 génových expresií bektérie *Escherichia.coli*, známej ako *E.coli*, so 40 meraniami v rôznych podmienkach. Tieto dáta sú obsiahnuté v balíčku *sand* [10]. Budeme skúmať ich správanie počas experimentov a hľadať asociácie medzi nimi. Na Obr. 1.9 sú vykreslené známe vzťahy medzi všetkými génmi, aj tými, ktoré nie sú v našom výbere. Celkovo týchto vzťahov - hrán je 209.

Budeme postupovať podľa [23], kde bola táto metóda spracovaná. Keďže neuvádzame všetky génove expresie, budeme skúmať $\frac{153\times152}{2} = 11628$ potenciálnych vzťahov, ktoré môžu mať tieto vybrané gény medzi sebou a vykonáme rovnaký počet simultánnych testov na získanie signifikantných korelácií. Ku každému vypočítame p-hodnoty



Obr. 1.9: Sieť vzťahov medzi génmi baktérie E.coli

a porovnáme s príslušnou päť percentnou kritickou hodnotou normálneho rozdelenia, pretože ak vektor $(X_i, X_j)^T$ má viacrozmerné normálne rozdelenie, potom aj hustotu korelácie ρ_{ij} za platnosti nulovej hypotézy, ktorá hovorí o nezávislosti vzťahov, vieme aproximovať hustotou normálneho rozdelenia.

Testovaním získame 5 227 signifikantných vzťahov medzi génmi, ktoré tvoria hrany v grafe. Táto hodnota predstavuje nadmerne veľké číslo, ktoré je pochybné, pretože také množstvo vzťahov medzi génmi nie je v skutočnosti možné. O tomto závere sa môžeme presvedčiť tiež z Obr. 1.9, kde je počet hrán medzi génmi podstatne menší, pritom sú na ňom vykreslené všetky vzťahy a nie len tie, ktoré my skúmame.

Kvôli nedôveryhodným predchádzajúcim výsledkom sme motivovaní skúmať vzťahy génových expresií iným spôsobom. Využijeme parciálne korelácie, a po ich aplikácií na dáta baktérie E.coli, získame 25 podmienených p-hodnôt, ktoré sa vyznačujú ako štatisticky významné. Je to podstatne menšie číslo ako v predchádzajúcom prípade. Ak by sme si vypísali dvojice so signifikantnou p-hodnotou a porovnali s databázou alebo literatúrou, kde sú tieto vzťahy uvedené, ľahko by sme sa presvedčili o korektnosti vzťahov.

Ak majú dva vrcholy $i, j \in V$ silnú koreláciu medzi charakteristikami X_i a X_j, nutne to neznamená, že musia byť silno korelované. Tento vzťah môže byť ovplyvnený iným, tretím vrcholom $k \in V$, s ktorým sú tieto dva atribúty silno korelované. Pomocou

parciálnych korelácií môžeme tieto nepriaznivé vzťahy odstrániť.

Nech množina $S_m = \{k_1, \ldots, k_m\}$. Potom parciálne korelácie medzi X_i a X_j vzhľadom k $X_{S_m} = (X_{k_1}, \ldots, X_{k_m})^T$ definujeme nasledovne

$$\rho_{ij|S_m} = \frac{\sigma_{ij|S_m}}{\sqrt{\sigma_{ii|S_m}\sigma_{jj|S_m}}}.$$

Po výpočte parciálnych korelácií môžu byť vzťahy vizualizované pomocou sietí. Každý vrchol zodpovedá jednej premennej a hrana závislému vzťahu, kde parciálna korelácia vyjadruje váhu hrany. Dva objekty nezdieľajú medzi sebou žiadnu hranu, ak ich parciálna korelácia je rovná nule.

Pri zostrojovaní korelačného grafu vytvárame hrany iným spôsobom ako pri obyčajných koreláciách, nasledovne

$$E = \{\{i,j\} \in V^{(2)} : \rho_{ij|S_m} \neq 0, \text{ pre všetky } S_m \in V^{(m)}_{\setminus\{i,j\}}\}.$$

Pri testovaní, či daná hrana existuje alebo nie, testujeme nové hypotézy

$$H_0 : \rho_{ij|S_m} = 0 \text{ pre nejaké } S_m \in V^{(m)}_{\setminus\{i,j\}} \quad \text{vs.} \quad H_1 : \rho_{ij|S_m} \neq 0 \text{ pre všetky } S_m \in V^{(m)}_{\setminus\{i,j\}},$$

ktoré môžeme rozložiť na množinu menších podtestov s hypotézami

$$H_0 : \rho_{ij|S_m} = 0 \qquad vs. \qquad H_1 : \rho_{ij|S_m} \neq 0. \tag{1.8}$$

Ako v predchádzajúcom prípade, korelácie $\rho_{ij|S_m}$ nahradíme výberovými koreláciami $\widehat{\rho}_{ij|S_m}$, ktoré transformujeme Fisherovou transformáciou [23]

$$z_{ij|S_m} = \tfrac{1}{2} log[\tfrac{1+\widehat{\rho}_{ij|S_m}}{1-\widehat{\rho}_{ij|S_m}}].$$

Príslušné p-hodnoty vypočítame ako

$$p_{ij,max} = max\{p_{ij|S_m} : S_m \in V^{(m)}_{\setminus\{i,j\}}\},$$

kde $p_{ij|S_m}$ sú p-hodnoty z menších podtestov.

Existujú dva najbežnejšie postupy výpočtu parciálnych korelácií. Prvý z nich [28], keď parciálne korelácie získame ako inverznú maticu kovariančnej matice Σ spočítanej zo súboru dát $\mathbf{X}$, s ktorými pracujeme. Predpokladáme, že výber má rozdelenie s variančnou maticou Σ. Korelácie zapíšeme nasledovne

$$K = \Sigma^{-1},$$

kde κ_{ij} po štandardizácii vyjadruje parciálnu koreláciu medzi x_i a x_j. Matica K sa nazýva *koncentračná matica* (angl. *Concentration matrix*) a opisuje úplnú parciálnu koreláciu medzi všetkými objektami

$$\rho(x_i, x_j | x_{-(i,j)}) = \frac{-\kappa_{ij}}{\sqrt{\kappa_{ii}\kappa_{jj}}}.$$

Druhý prístup [29] je pomocou regresie, kedy je jedna premenná modelovaná pomocou ostatných premených

$$y_i = \beta_{i0} + \beta_{i1} y_1 + \beta_{i2} y_2 \ldots + \varepsilon_i, \quad i = 1, 2, \ldots, N.$$

Platí, že ak $\beta_{ij} \neq 0$ pre nejaké $i \neq j$, tak potom i a j sú parciálne korelované.

Ak štandardnú odchýlku pre nejakú premennú x označíme $D(x)$, potom korelácie vypočítame nasledovne

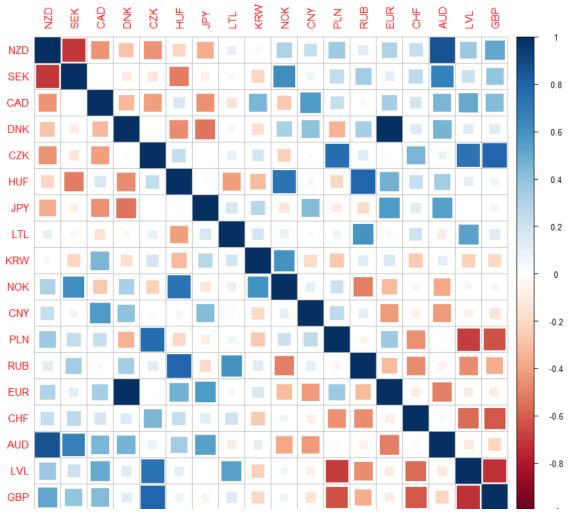
$$\rho(x_i, x_j | x_{-(i,j)}) = \frac{\beta_{ij} D(\varepsilon_j)}{D(\varepsilon_i)},$$

čo vyjadruje ekvivalenciu medzi tým, že x_i a x_j sú parciálne korelované a lineárnou koreláciou medzi rezíduami x_i a x_j.

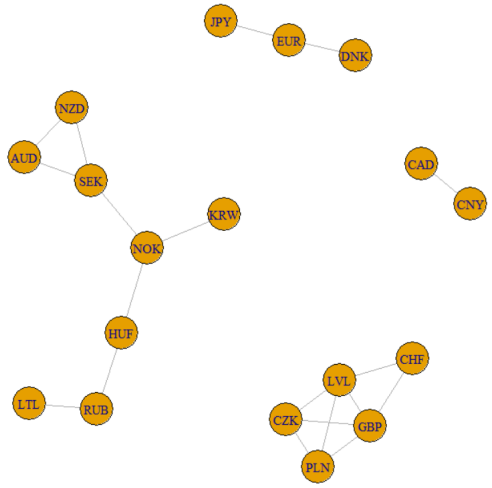
Teóriu parciálnych korelácií aplikujeme na časových radoch obsahujúcich informácie o výmenných kurzoch. Vypočítame ich pomocou funkcie *pcor* z balíčka *ppcor* [12] a spoločne vykreslíme v korelačnej matici na Obr. 1.10.

Pri porovnaní matice s obyčajnými koreláciami na Obr.1.4 si všimneme výrazné rozdiely. Okrem pozitívnych vzťahov, vyznačených modrou farbou, sa tu objavuje väčší počet negatívnych korelácií.

Po otestovaní hypotéz (1.8) získame 19 signifikantných parciálnych korelácií, ktoré sú v grafe na Obr. 1.11 reprezentované hranami. Štruktúra siete sa odlišuje od grafu obyčajných korelácií. Delí sa na štyri menšie podgrafy. Je to spôsobené vplyvom silných negatívnych korelácií, ktoré spôsobujú separáciu výmenných kurzov a menším počtom štatisticky významných vzťahov. Výmenné kurzy v každom podgrafe sú navzájom pozitívne korelované a v porovnaní s ostatnými vrcholmi, ktoré sa v danom podrafe nevyskytujú, sú tieto vzťahy silnejšie. Toto platí pre každú skupinu až na jednu



Obr. 1.10: Matica parciálnych korelácií

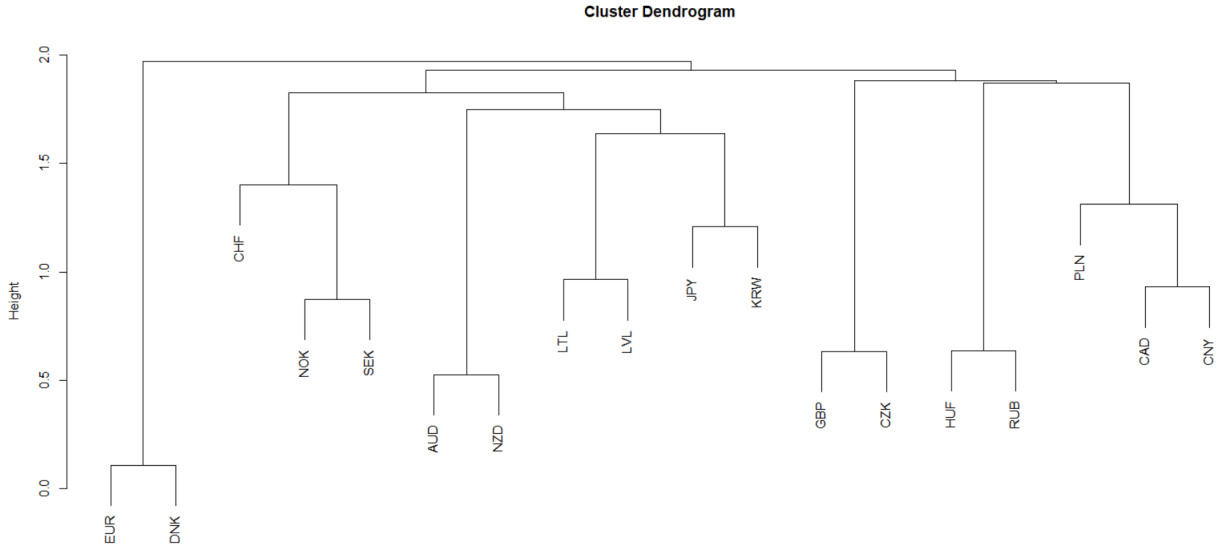


Obr. 1.11: Graf parciálnych korelácií

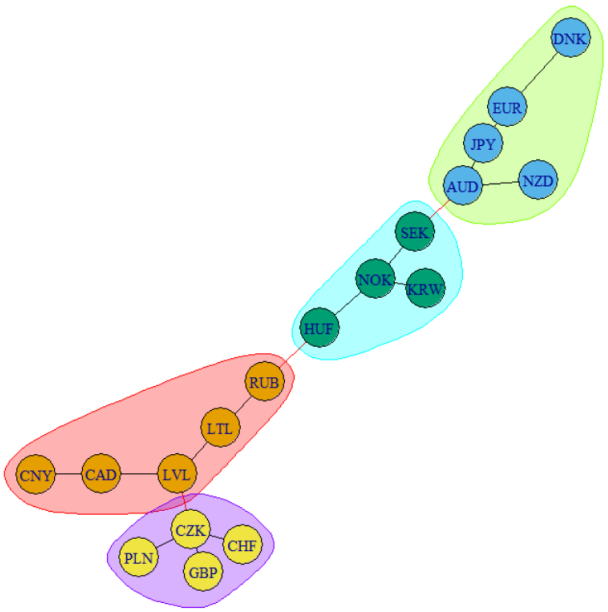
výnimku, ktorá obsahuje CZK. Ostatné meny sú v podgrafe navzájom silno negatívne korelované, len so CZK vykazuje pozitívnu koreláciu.

Z parciálnej korelačnej matice získame maticu vzdialeností rovnakým výpočtom ako pri obyčajných koreláciách, pomocou euklidovských vzdialeností $d_{ij} = \sqrt{2(1-\rho_{ij})}$. Vykreslíme prislúchajúci stromový diagram - dendrogram. Rovnako platí, čím je korelácia medzi dvomi vrcholmi menšia, tým je ich vzdialenosť väčšia a naopak.

Pri hľadaní najlacnejšej kostri pre parciálne korelácie budeme pokračovať ako v prípade obyčajných korelácií a na vypočítané parciálne korelácie uplatníme rovnaké funkcie. Získaná najlacnejšia kostra spolu so zhlukmi je vyobrazená na Obr. 1.13.



Obr. 1.12: Dendrogram parciálnych korelácií



Obr. 1.13: Najlacnejšia kostra pre parciálne korelácie

Všimnime si umiestnenie vrcholov CZK, PLN, CHF, GBP, ktoré sa v korelačnom grafe na Obr. 1.11 vyskytujú v spoločnom podgrafe aj napriek silným vzájomným negatívnym parciálnym koreláciám. Algoritmus na určenie komunít v najlacnejšej kostre ich zatriedil do spoločného zhluku. CZK, ktorá má silné pozitívne korelácie s ostatnými menami, sa nachádza uprostred a s každou menou zdieľa hranu. Ostatné spomenuté

meny nezdieľajú medzi sebou žiadne hrany. Dôvodom, prečo táto štvorica štátov tvorí samostatnú skupinu, môže byť silná pozitívna korelácia týchto výmenných kurzov s CZK. Funkcia *minimum.spanning.tree* vytvorila hrany CZK s ostatnými spomenutými menami, ale vzťahy medzi PLN, CHF a GBP sa už v najlacnejšej kostre neobjavujú.

2 Výnosové rozpätie

Jedným z meradiel, ktoré používajú investori na meranie výdavkov konkrétnych dlhopisov, sa nazýva výnosové rozpätie (angl. *Yield Spread*) [38]. Vo všeobecnosti je definovaný ako rozdiel medzi výnosmi dvoch finančných nástrojov s rôznou dobou splatnosti, teda ho spočítame odrátaním výnosu do splatnosti jedného produktu od výnosu do splatnosti druhého produktu. Vyjadruje rizikovú prémiu jedného produktu nad druhým [37]. Napríklad, ak dlhopis s maturitou 30 rokov má výnos 10 % a iný dlhopis s maturitou 5 rokov vynáša 5 %, potom výnosové rozpätie medzi nimi sa vypočíta ako: 10 % - 5 % = 5 %.

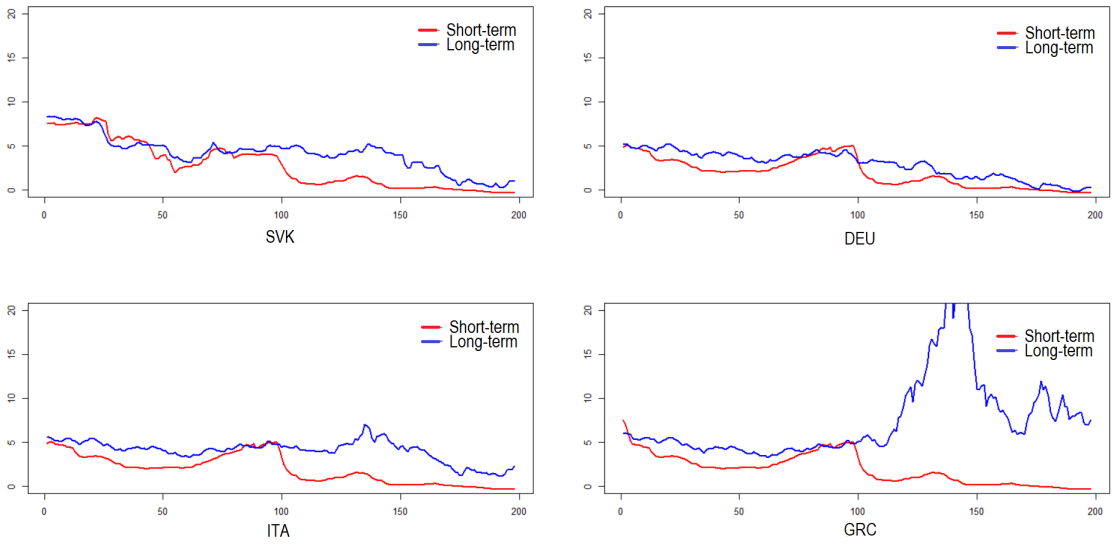
Výnosové rozpätie môže mať v súčasnosti mnohostrannejšie využitie. Ako je uvedené v článku [36], autor sa zaoberal predikovaním budúcich výskytov simultánnych recesií v niektorých krajinách sveta. Výskyt recesií skúmal pomocou výnosového rozpätia, ktoré použil v probit modeloch.

Kvôli zaujímavému využitiu a mnohotvárnosti výnosového rozpätia sme sa aj my zamerali na túto veličinu a prostredníctvom nej skúmame vzťahy európskych krajín a ich vzájomných korelácií.

V tejto časti budeme pracovať s dlhodobými a krátkodobými úrokovými mierami súčasne. Dlhodobé úrokové miery sa vzťahujú na štátne dlhopisy s dlhšou dobou splatnosti. Sú vyvodené z cien, s ktorými sú štátne dlhopisy obchodované na finančných trhoch a ich splatenie zaručuje vláda. Krátkodobé úrokové miery sú sazby, za ktoré obchodujú finančné inštitúcie.

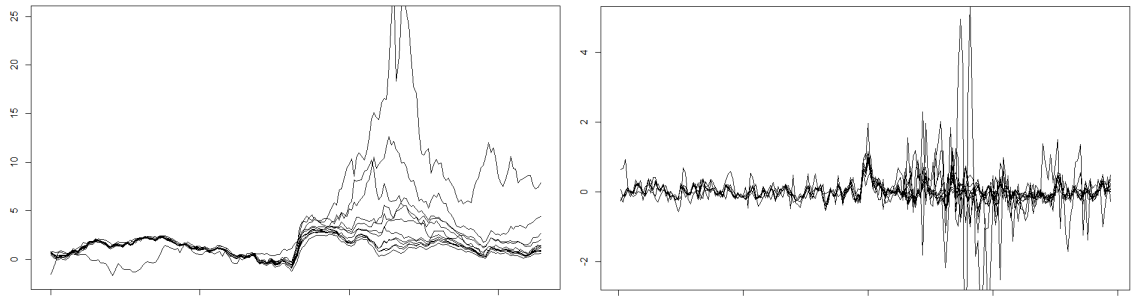
Dáta budeme čerpať z internetovej stránky OECD [5]. V našom prípade budeme pracovať s úrokovými mierami dvanástich krajín, ktoré sú súčasťou Eurozóny. Dlhodobé úrokové miery majú splatnosť desať rokov a krátkodobé úrokové miery expiračnú dobu tri mesiace. Dáta sú mesačné, v časovom rozpätí od januára 2001 do februára 2017. Na Obr. 2.1 je znázornený vývoj krátkodobých a dlhodobých úrokových mier niektorých štátov.

Odrátaním krátkodobých úrokových mier od dlhodobých získame výnosové rozpätie. ADF testy, zahrnuté v balíčku *urca* [8], potvrdia nestacionaritu časových radov. Diferencovaním dát získame stacionárne časové rady, čo je podstatný krok pre ďalšiu korektnú prácu s dátami, ako sme vysvetlili v predchádzajúcej kapitole. V nasledujúcej



Obr. 2.1: Krátkodobé a dlhodobé úrokové miery krajín: SVK, DEU, ITA, GRC

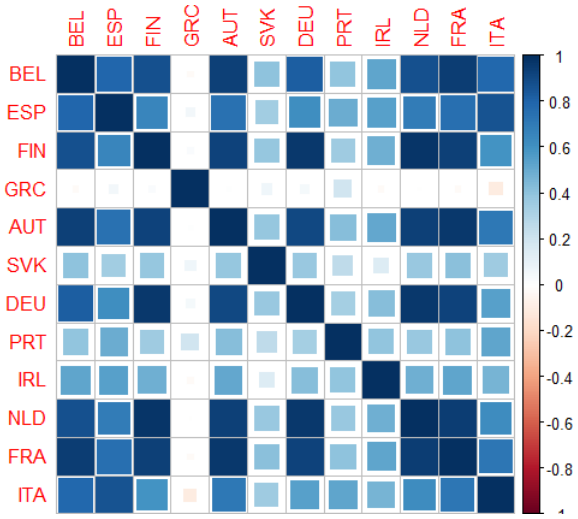
časti sa budeme zaoberať vzťahmi medzi výnosovými rozpätiami štátov, ich koreláciami a parciálnymi koreláciami. Obr. 2.2 znázorňuje vývoj výnosového rozpätia v jednotlivých krajinách. Prvý obrázok pred diferencovaním radov, na druhom sú vykreslené stacionárne časové rady.



Obr. 2.2: Vývoj časových radov (vľavo pred diferencovaním)

Výber štátov tvoria členovia Eurozóny. Nezvolili sme všetky krajiny, ale len tie, ku ktorým sme mali dostatok dát k dispozícii. Krajiny Eurozóny majú spoločné niektoré finančné inštitúcie (napr. Európsku centrálnu banku), ktoré rozhodujú o mnohých dôležitých otázkach. Ich ekonomiky sú poprepájané, čo sa odzkradľuje na vypočítaných koreláciách. Obr. 2.3 zachytáva korelačnú maticu. Vzájomné vzťahy výnosových rozpätí štátov sa zdajú byť silno korelované až na jednu výnimku - Grécko, ktorého korelácie sú takmer všetky nulové.

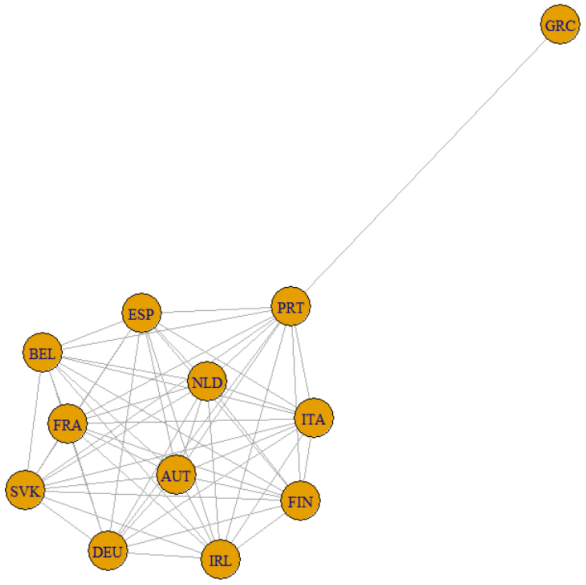
Dáta sú tvorené 12 európskymi štátmi, ktoré môžu medzi sebou nadobudnúť maxi-



Obr. 2.3: Korelačná matica

málne 66 korelácií. Všetky korelácie však nemusia byť signifikantné.

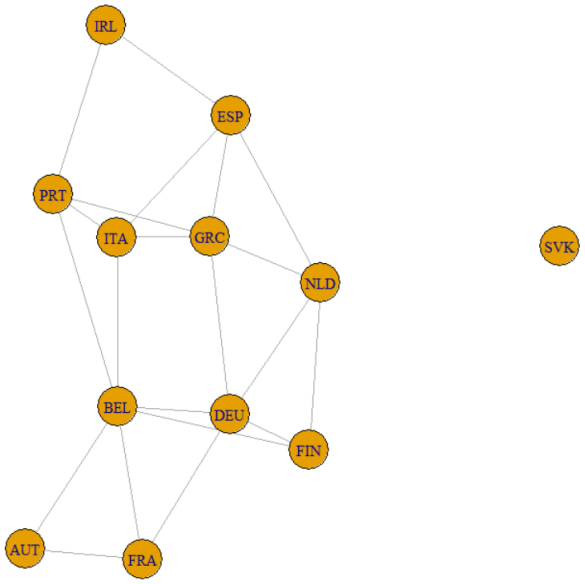
Budeme postupovať ako v predchádzajúcej kapitole 1. Korelácie stabilizujeme pomocou Fisherovej transformácie (1.6), otestujeme hypotézy o nulovosti korelácií (1.7) a vypočítame p-hodnoty, ktoré porovnáme s päť percentými kritickými hodnotami normálneho rozdelenia. Týmto spôsobom získame 56 štatisticky významných korelácií. Signifikantné vzťahy vykreslíme do grafu (Obr. 2.4). Sieť vzťahov je hustá, pretože obsahuje veľa hrán - signifikantných korelácií.



Obr. 2.4: Graf obyčajných korelačných vzťahov

Spočítame parciálne korelácie. To nám umožňuje vylúčiť vzťahy dvoch štátov, ktoré

sú ovplyvnené výnosovým rozpätím tretieho štátu, s ktorým majú tieto dve krajiny vzájomnú koreláciu. Otestujeme hypotézy o nulovosti podmienených korelácií (1.8) a následne vypočítané p-hodnoty porovnávame s päť percentnými kritickými hodnotami normálneho rozdelenia. Postupne sa dopracujeme k výsledku s 21 signifikantnými parciálnymi koreláciami, ktoré sú vyobrazené na Obr. 2.5. Rozdiel v počte korelácií s predošlým výsledkom je nezanedbateľne výrazný.



Obr. 2.5: Graf parciálnych korelačných vzťahov

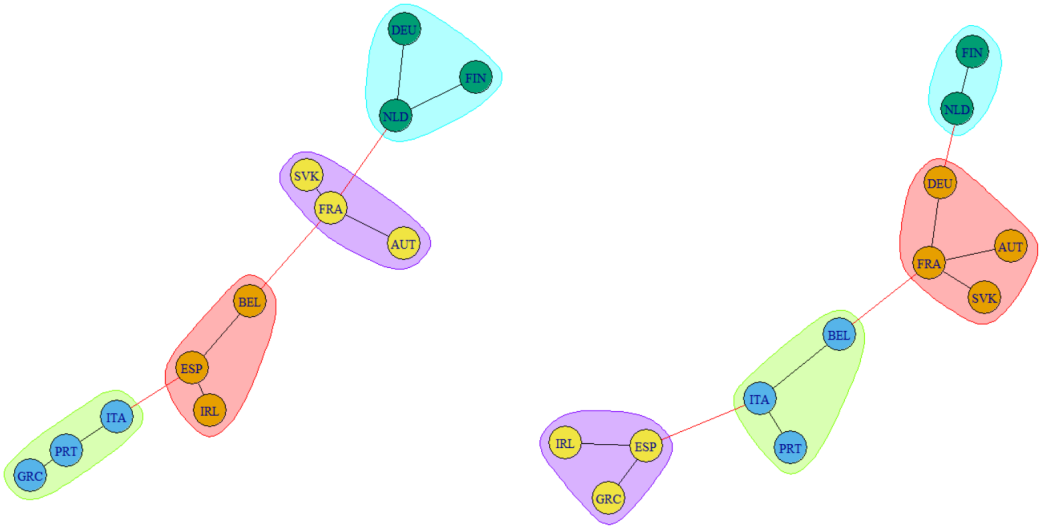
Na predchádzajúcich výsledkoch pozorujeme dôležitosť počítania parciálnych korelácií najmä pri práci s väčším počtom objektov, ktoré sa môžu navzájom ovplyvňovať. Tak odstránime neexistujúce vzťahy, ktoré by mohli výrazne ovplyvniť interpretáciu výsledkov.

Porovnanie obyčajných a parciálnych korelácií

Pri počítaní obyčajných korelácií vyšlo 56 signifikantných z celkového počtu 66. Ak vypočítame parciálne korelácie, toto množstvo sa zredukuje na 21 významných korelácií, čo je podstatný rozdiel. Porovnaním grafov korelačných vzťahov (Obr. 2.4 a Obr. 2.5) si všimneme, že v prípade parciálnych korelácií sa veľa hrán v grafe už neobjavuje.

Ďalší prístup, ktorý poslúži na analýzu vzťahov výnosového rozpätia, je najlacnej-

šia kostra grafu. Jej grafickým vyjadrením pozorujeme rozdiely medzi obyčajnými a pariálnymi koreláciami.



Obr. 2.6: Najlacnejšie kostry grafu pre obyčajné (vľavo) a parciálne (vpravo) korelácie

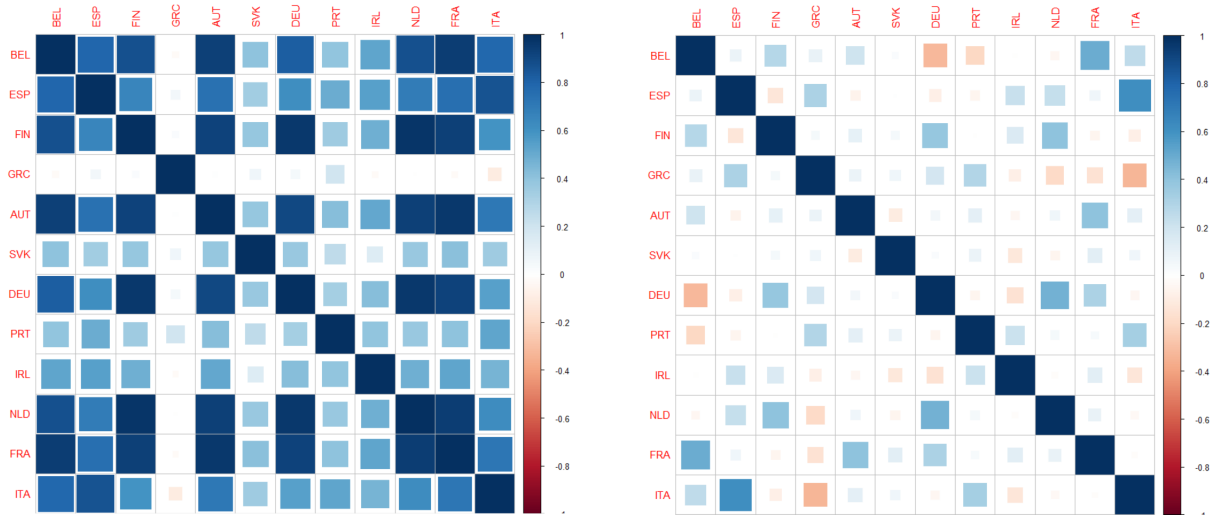
Grafy na Obr. 2.6 predstavujú najlacnejšie kostry grafu pre výnosové rozpätia krajín. Graf vľavo vyjadruje obyčajné korelácie, vpravo parciálne korelácie. Sú vykreslené pomocou funkcie *minimum.spanning.tree* a zhluky sú zhotovené funkciou *walktrap.community*, obidve funkcie sú z balíčka *igraph* [9].

Ako sme vyššie spomínali, pri obyčajných koreláciách môžu vznikať vzťahy, ktoré v skutočnosti neexistujú. Tieto dôsledky pozorujeme ako rozdiely medzi najlacnejšími kostrami obyčajných a parciálnych korelácií a vytvorení zhlukov.

Pri porovnaní korelačných matíc parciálnych a obyčajných korelácií na Obr. 2.7 si všimneme výrazné zmeny. Ako prvé, veľký úbytok signifikantných parciálnych korelácií. Súvislosti vzťahov sú tiež pozorovateľné na Obr 2.6. Napríklad, krajina FRA má podľa matíc výraznejšie vzťahy s BEL, AUT a DEU, a zároveň tieto vzťahy pozorujeme v najlacnejšej kostre parciálnych korelácií, kde FRA priamo zdieľa hrany s týmito krajinami.

Ak upriamime pozornosť na štát GRC, všimneme si druhú zmenu. V obidvoch grafoch sa síce nachádza na okrajoch, čo vypovedá o jeho slabých koreláciách v porovnaní s ostatnými krajinami, ale v prípade parciálnych korelácií pozorujeme v matici ich nárast. V kladnom smere s krajinami ESP a PRT a v zápornom vzťahu s ITA a NLD.

Pomocou matíc na Obr. 2.7 môžeme lepšie porozumieť ako vznikla najlacnejšia



Obr. 2.7: Porovnanie korelačných matíc pre výnosové rozpätie: vľavo obyčajné korelácie, vpravo parciálne korelácie

kostra parciálnych korelácií. Krajiny s najvyšším počtom signifikantných korelácií, ako ITA, FRA a ESP, majú viac vzťahov s ostatnými krajinami, čo sa prejavuje väčším počtom hrán v grafe.

Zaujímavá je zmena korelácií pri dvojiciach štátov DEU-BEL a ITA-GRC. V prípade druhej dvojice sa negatívna korelácia, ktorú sme pozorovali len miernu pri obyčajných koreláciách, zvýšila dosť výrazne. Na druhej strane, štáty DEU-BEL mali medzi sebou silnú pozitívnu koreláciu. V prípade parciálnych korelácií sa zmenila na výrazne negatívnu.

V najlacnejšej kostre niektoré hrany nevystupujú ako v grafe korelačných vzťahov na Obr. 2.5, čo je dôsledkom stavby tohto typu stromu. Preto môžeme podrobnejšie skúmať súvislosť výnosového rozpätia štátov, ich vzájomnú prepojenosť a pozorovať, ktoré štáty sú významnejšie alebo menej významné v porovnaní s ostatnými.

2.1 Centralita vrcholov

Dôležitou súšasťou sieťovej analýzy a teórie grafov je skúmanie centrality vrcholov. Znamená to identifikáciu, ktoré uzly sú v sieti vplyvnejšie než ostatné. V práci sme sa zaoberali tromi základnými mierami centrality, ktoré vypočítame pomocou vhodných funkcií, ktoré sú zahrnuté v balíčku *igraph* [9].

- *Centralita stupňa* udáva absolútny počet hrán, ktoré sú priamo spojené s daným uzlom. Väčší počet väzieb evokuje zvýhodnenú pozíciu, teda väčšiu centralitu. Počítame ju pomocou funkcie *degree*.

- *Centralita blízkosti* matematicky vyjadruje prevrátenú hodnotu súčtu vzdialeností [1]od jedného vrchola ku všetkým ostatným. Inak povedané, ako daný vrchol môže jednoducho dosiahnúť ostatné vrcholy bez toho, aby musel byť s nimi priamo spojený. Čím je blízkosť vrchola väčšia, tým väčšia je možnosť sa z neho dostať ku ostatným vrcholom. Tieto hodnoty vypočítame funkciou *closeness*.

- *Centralita stredovej medzipolohy* udáva počet prípadov, kedy je vrchol najkratšou cestou medzi ostatnými vrcholmi. Väčšia hodnota vyjadruje viac ciest, ktoré budú daným vrcholom prechádzať. Funkcia *betweenness* slúži na výpočet centrality stredovej medzipolohy.

Budeme pracovať s parciálnymi koreláciami a našu pozornosť upriamime na dva typy grafov, graf korelačných vzťahov a najlacnejšiu kostru. Ich vykreslením na Obr.2.5 a Obr.2.6 (vpravo najlacnejšia kostra) a vzájomným porovnaním výsledkov potvrdíme ich odlišnú štruktúru. Preto predpokladáme, že aj hodnoty ich mier centrality sa budú líšiť. V tabuľke 2.1 uvádzame porovnanie mier centralít vrcholov pre tieto dva typy grafov.

Podľa výpočtov, najcentrálnejším vrcholom v najlacnejšej kostre je štát FRA. Je umiestnený v strede stromu a spojený s ďalšími štyrmi štátmi. Ostatné miery centrality tiež nadobúdajú vysoké hodnoty. Medzi ďalšie centrálnejšie štáty patrí napr. BEL a ITA. Na druhej strane, presná polovica štátov zdieľa práve jednu hranu s iným štátom. Z týchto vrcholov má najnižšie hodnoty ostatných mier FIN, IRL a GRC.

V korelačnom grafe je za najvplyvnejší štát jednoznačne určené BEL so šiestimi zdieľanými hranami. Medzi ďalšie "vplyvné"vrcholy patria DEU a GRC. Na druhej strane, SVK ako jediný vrchol nie je prepojené s ostatnými krajinami. Jeho centralita stupňa a centralita stredovej medzipolohy sú nulové a centralita blízkosti veľmi nízka.

Z definície najlacnejšej kostry vieme, že tento druh stromu sa snaží vytvoriť graf s minimálnou dĺžkou hrán. Graf korelačných vzťahov zobrazuje všetky signifikantné

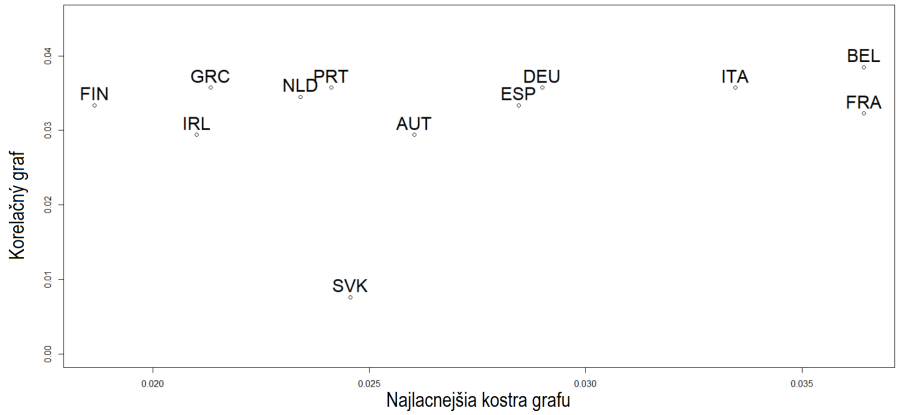
[1]Vzdialenosť dvoch vrcholov v konečnom grafe je definovaná ako počet hrán v najkratšej ceste, ktorá spája dané vrcholy [47].

	Najlacnejšia kostra grafu			Graf korelačných vzťahov		
	Stupeň	Blízkosť	Medzipoloha	Stupeň	Blízkosť	Medzipoloha
BEL	2	0.0364	30	6	0.0384	14.50
ESP	3	0.0284	19	4	0.0333	3.50
FIN	1	0.0186	0	3	0.0333	0.83
GRC	1	0.0213	0	5	0.0357	4.16
AUT	1	0.0260	0	2	0.0294	0.00
SVK	1	0.0245	0	0	0.0075	0.00
DEU	2	0.0289	18	5	0.0357	5.66
PRT	1	0.0241	0	4	0.0357	5.58
IRL	1	0.0210	0	2	0.0294	0.33
NLD	2	0.0234	10	4	0.0344	3.08
FRA	4	0.0364	37	3	0.0322	1.08
ITA	3	0.0334	31	4	0.0357	3.25

Tabuľka 2.1: Miery centrality vrcholov

parciálne korelácie medzi štátmi. O tom, že grafy sú od seba výrazne rozdielne sa môžeme presvedčiť či už z obrázkov, ale aj z výsledkov. Najväčší rozdiel je v určení najmenej centrálneho bodu. Zatiaľ čo v korelačnom grafe je SVK zreteľne odčlenené od celého grafu, v najlacnejšej kostre je najmenej centrálnym štátom GRC. Zaujímavé je, že v korelačnom grafe tento štát patrí medzi tie najvplyvnejšie a nachádza sa v strede grafu.

Kvôli rozdielnym štruktúram grafov môžeme pozorovať odlišné hodnoty indexov. Na Obr. 2.8 sú vykreslené hodnoty blízkosti uzlov. Na x-ovej osi sú hodnoty pre najlacnejšiu kostru a na y-ovej osi hodnoty blízkosti pre uzly z korelačného grafu. Ako si môžeme všimnúť z grafu, nevyskytuje sa tu žiadna závislosť. Preto si môžeme klásť otázku, na ktoré výsledky by sme sa mali zamerať a ktoré ignorovať? Odpoveď nie je jasná, pretože nevieme určiť, ktorý graf je lepší alebo horší, alebo ktoré vypočítané centrality vrcholov sú tie správne.



Obr. 2.8: Porovnanie blízkosti uzlov pre korelačný graf a najlacnejšiu kostru

2.2 Rôzne algoritmy tvorby zhlukov

Na dátach výnosových rozpätí krajín zopakujeme postup tvorby zhlukov v najlacnejšej kostre pomocou algoritmov z kapitoly 1. Výsledky sú zachytené v tabuľke zhlukov na Obr. 2.9. Zistíme, že až na pár výnimiek algoritmy zatriedili krajiny do rovnakých skupín. Prvých šesť algoritmov vytvorilo takmer identické zhluky. Posledné dve funkcie patria medzi algoritmy zhlukovej analýzy. Princíp tvorby zhlukov je odlišný od predchádzajúcich, preto vytvorenie zhlukov sa tiež líši. Ich podrobnejším popisom sa budeme zaoberať v ďalších častiach práce.

	ClustOpt	ClustFastGr	ClustLouv	ClustLabPr	ClustEdgeBt	ClustWalk	K-mean	HClust
GRC								
IRL								
ESP								
FIN								
NLD								
DEU								
AUT								
SVK								
FRA								
ITA								
BEL								
PRT								

Obr. 2.9: Tabuľka zhlukov, použité funkcie: cluster_optimal, cluster_fast_greedy, cluster_louvain, cluster_label_prop, cluster_edge_betweenness, cluster_walktrap, kmeans, hclust

Vo všeobecnosti hlavnou vstupnou zložkou algoritmov je graf, v našom prípade naj-

lacnejšia kostra grafu. Hrany a celý strom sú zostavené na základe parciálnych korelácií vyjadrujúcich vzťahy výnosových rozpätí. Pomocou tohto vstupu sa funkcie snažia vytvoriť subgrafy, nazývané komunity. Prvých šesť algoritmov počíta prostredníctvom princípu, ktorý sme stručne priblížili v predchádzajúcej kapitole 1. Postupne priblížime posledné dva algoritmy.

2.3 Hierarchické zhlukovanie

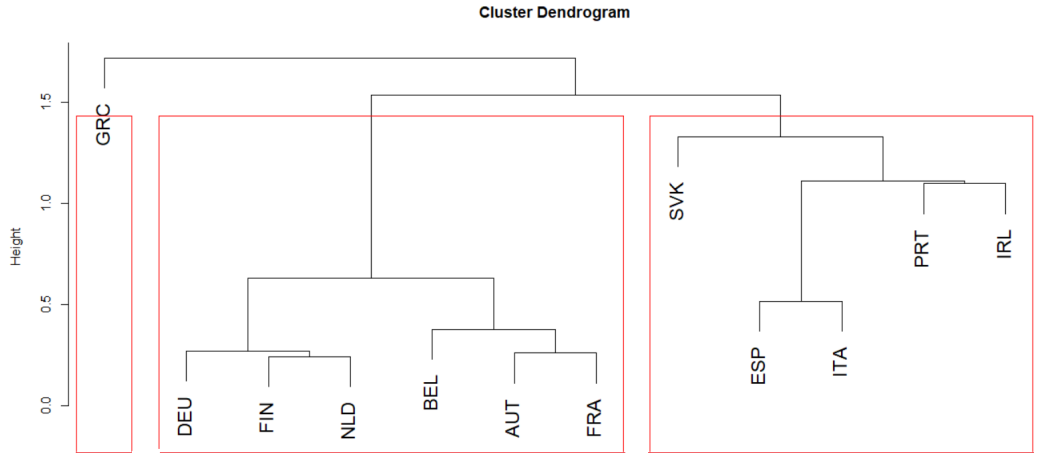
Zhlukovú analýzu môžeme vo všeobecnosti definovať ako procedúru, ktorá spája objekty do spoločných zhlukov podľa určitých logických podobností a zároveň oddeľuje objekty od seba odlišné [51]. Metódy, ktoré zahrňuje, rozdeľujeme do dvoch základných skupín, hierarchické a nehierarchické. V predchádzajúcich častiach práce sme sa zaoberali hlavne nehierarchickými metódami.

Analýza hierarchického zhlukovania zahŕňa metódy, ktoré vytvárajú systém zhlukov usporiadaných hierarchicky. Po vykreslení dendrogramu na Obr. 2.10, kde sme využívali obyčajné korelácie, môžeme pozorovať vzťahy výnosových rozpätí jednotlivých krajín. Skupiny štátov sú zoskupené v spoločných rámikoch. Vo funkcii *hclust* z balíčka *stats* [50], ktorá slúži na vytvorenie dendrogramu, sme počet zhlukov nadstavili manuálne pre $k = 3$.

Pri porovnaní týchto dvoch grafov, dendrogramu na Obr. 2.10 a najlacnejšej kostri na Obr.2.6 vľavo, ktoré sú vytvorené z obyčajných korelácií, si všimneme rozdielnu štruktúru zhlukov. Krajiny v komunitách sú zoskupené úplne odlišne, čo je dôsledok rozdielnych metód, ktoré tieto dve funkcie používajú pri tvorbe zhlukov. Obidva prístupy vychádzajú z korelačnej matice upravenej do tvaru matice vzdialeností. V tabuľke 2.2 sú zachytené porovnania zhlukov vytvorených v najlacnejšej kostre a dendrograme.

Pre lepšie pochopenie vzniku dendrogramu (Obr. 2.10) uvedieme vzťah dvoch štátov - FIN a NLD. Majú silné vzájomné korelácie, preto sa nachádzajú blízko seba a v rovnakej výške. O ich vzťahu sa môžeme presvedčiť z korelačnej matice na Obr. 2.11.

Z korelačných vzťahov sa dájú vyčítať ďalšie usporiadania krajín. AUT má najsilnejší vzťah s FRA, čo sa takisto prejavuje spoločným prepojením v dendrograme. Štvorica krajín sa nachádza na samom spodku grafu, pretože ako vidíme z korelačnej matice, majú väčšie množstvo silných korelácií s ostatnými krajinami.

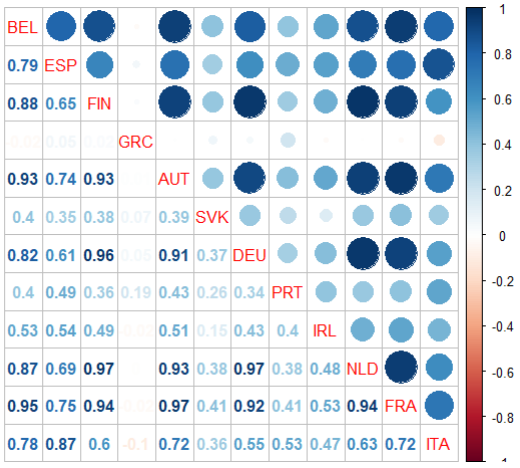


Obr. 2.10: Dendrogram obyčajných korelácií

	MST	Dendrogram
BEL	3	3
FRA	3	3
FIN	3	1
NLD	3	4
AUT	3	2
SVK	2	2
DEU	2	1
PRT	2	4
IRL	2	3
ITA	2	4
ESP	2	2
GRC	1	1

Tabuľka 2.2: Porovnanie zhlukov v dendrograme a v najlacnejšej kostre pre obyčajné korelácie

Prvá dvojica je ďalej spojená s DEU vďaka silným koreláciám obidvoch štátov s touto krajinou (s FIN 0,96 a s NLD 0,97) a druhá dvojica so štátom BEL. Nakoniec je šestica štátov prepojená do spoločného subgrafu, ktorý tvorí podstatnú časť dendrogramu. Viac o tejto problematike a metódach pri zostavovaní dendrogramov sa môžeme dočítať v článku [52].



Obr. 2.11: Matica obyčajných korelácií

Ako základný materiál pre tvorbu zhlukov v hierarchickom zhlukovaní slúži matica vzdialeností. Zhotovenie takejto matice môže byť založený na rôznych metódach, my sme používali vzdialenosť vypočítanú nasledovne $d_{ij} = \sqrt{2(1-\rho_{ij})}$ [20].

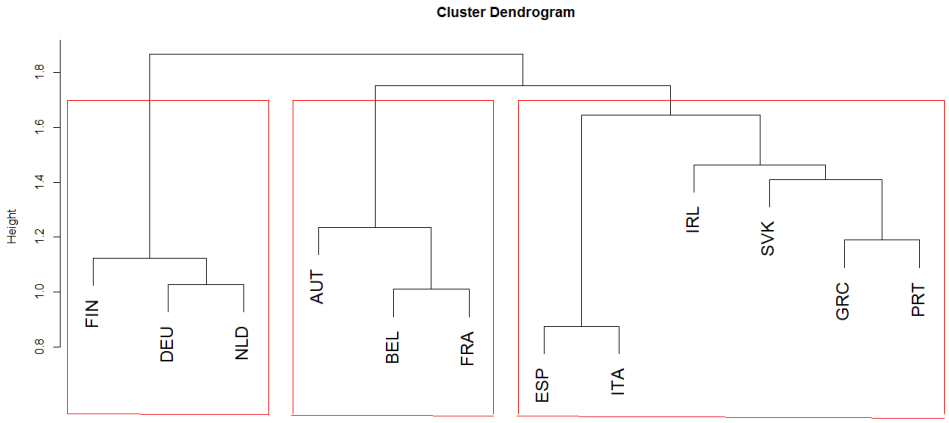
Obr. 2.12 zobrazuje maticu vzdialeností pre naše dáta vypočítanú funkciou *as.dist* z balíčka *stats* [50]. Dvojice štátov s najmenšími vzdialenosťami (FIN-NLD, AUT-FRA) sú v dendrograme priamo prepojené. Tieto vzťahy sme mohli pozorovať z korelačnej matice. Krajina DEU má najmenšiu vzdialenosť s hodnotou veľkosti 0,258 s NLD, lenže NLD má najmenšiu vzdialenosť s FIN a to 0,2437. Na druhej strane, ak upriamime pozornosť na GRC, z korelačnej matice pozorujeme veľmi nízke korelácie. V matici vzdialeností sa prejavujú ako veľké hodnoty vzdialenosti GRC voči ostatným štátom.

	BEL	ESP	FIN	GRC	AUT	SVK	DEU	PRT	IRL	NLD	FRA
ESP	0.6423883										
FIN	0.4926871	0.8314778									
GRC	1.4301698	1.3772825	1.3969895								
AUT	0.3703565	0.7178344	0.3747730	1.4080109							
SVK	1.0929771	1.1423282	1.1112143	1.3642974	1.1085324						
DEU	0.5938041	0.8801784	0.2703222	1.3797725	0.4334069	1.1183868					
PRT	1.0973336	1.0093191	1.1333216	1.2704491	1.0689240	1.2194457	1.1513687				
IRL	0.9701426	0.9568563	1.0103874	1.4283664	0.9896977	1.3053448	1.0691165	1.0987483			
NLD	0.5035513	0.7843364	0.2437742	1.4161354	0.3709212	1.1165401	0.2581814	1.1178629	1.0160986		
FRA	0.3304428	0.7024873	0.3481711	1.4287094	0.2604144	1.0823955	0.3971424	1.0865078	0.9708940	0.3333321	
ITA	0.6570295	0.5152383	0.8963595	1.4849683	0.7487587	1.1356387	0.9517309	0.9717442	1.0315442	0.8614018	0.7479007

Obr. 2.12: Matica vzdialeností

Po vykreslení dendrogramu pre parciálne korelácie (Obr. 2.13) a porovnaní s predchádzajúcim dendrogramom si všimneme odlišnosť štruktúry. Medzi uzlami sú vzdialenosti menšie a roztriedenie krajín do jednotlivých orámovaných skupín je rovnomer-

nejšie. Je to spôsobené menšími rozdielmi v koreláciách, teda tým, že v parciálnych koreláciách nevystupujú také výrazné vzťahy ako v obyčajných.



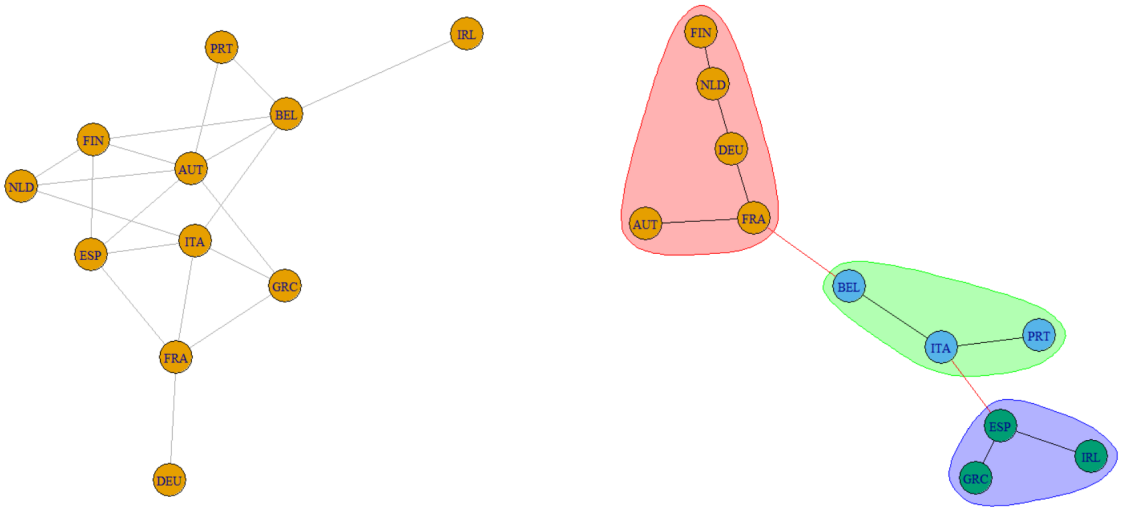
Obr. 2.13: Dendrogram parciálnych korelácií

Ak existuje krajina, ktorá nemá vzťahy s ostatnými krajinami významné alebo sú ich korelácie blízke nule, nezdieľa v grafe žiadnu hranu s inou krajinou. V našom prípade pri práci s parciálnymi koreláciami je takou krajinou SVK. Jej odstránením z dát vytvoríme nové grafy (Obr. 2.14). Zoradenie krajín v najlacnejšej kostre ani graf korelačných vzťahov to neovplyvnilo, pretože SVK nemalo silný vplyv na tvorbu grafov a siete. V prípade korelačného grafu, ako sme sa už presvedčili, je výnosové rozpätie tohto štátu najmenej centrálne. V najlacnejšej kostre patrí tiež medzi slabšie vrcholy.

2.4 K-means algoritmus

Ďalšia metóda, ktorú stručne priblížime, je *K-means* algoritmus. Tak ako ostatné metódy zhlukovej analýzy, tiež vyjadruje podobnosť pozorovaných objektov, ktoré zaraďuje do zhlukov, v ktorých sú si tieto objekty viac podobné v porovnaní s objektami v ostatných zhlukoch.

K-means slúži na skúmanie štruktúry množiny, ale nie na hĺbkové analyzovanie. Záleží to od pozorovaných dát. Hlavnou nevýhodou tohto algoritmu je, že na začiatku procedúry požaduje určenie počtu k zhlukov. Výhodami sú jednoduchosť algoritmu a jeho konvergencia pri konečnom počte krokov. O ďalších výhodách a nevýhodách sa môžeme dočítať v práci [49].



Obr. 2.14: Graf parciálnych korelačných vzťahov a najlacnejšej kostry

Hlavnou myšlienkou algoritmu k-means je rozdelenie množiny n objektov na k zhlukov, kde $k \leq n$. Cieľom je minimalizácia vnútornej variability zhlukov, ktoré sú definované ťažiskom (tiež nazývané ako centroida zhluku). Klasický prístup na určenie variability môžeme vyjadriť nasledovne [48]

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2,$$

kde x_i je objekt z dát, μ_k stredná hodnota objektov priradených k zhluku C_k. Celkový súčet súčtu štvorcov, ktorý minimalizujeme, formálne vyjadríme ako

$$\sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2.$$

Viac o tomto algoritme a načrtnutých krokoch postupu metódy sa môžeme dozvedieť v spomínaných prácach [48, 49].

Túto metódu počítame pomocou funkcie *kmeans*, ktorá je zahrnutá v balíčku *stats* [50].

Vráťme sa späť k príkladu z časti 2.2, v ktorom sme vytvárali zhluky rozdielnymi algoritmami v najlacnejšej kostre. Všimneme si, že k-means algoritmus zatriedil krajiny rovnako ako algoritmus pri hierarchickom zhlukovaní.

2.5 Koeficient zhlukovania

V teórii grafov vieme pre každý uzol vyjadriť istý stupeň v grafe. Určuje, do akej miery má tento uzol tendenciu sa zhlukovať s ostatnými uzlami. Veličina vyjadrujúca túto mieru stupňa sa nazýva *koeficient zhlukovania*.

Existuje viacero spôsobov na výpočet tejto hodnoty, lenže všetky metódy výpočtov predpokladajú neúplné grafy. Bližší popis týchto metód môžeme nájsť napríklad v článkoch [53, 54, 55]. My sa v tejto časti práce zameriame na úplné vážené neorientované grafy. To znamená, že budeme predpokladať existenciu hrán medzi každým párom vrcholov. Dôvodom upriamenia pohľadu na úplné vážené neorientované grafy je ten, že práve tento typ grafu dobre vystihuje korelácie. Každý pozorovaný objekt má vzťah s inými, pričom váhy medzi nimi sú vyjadrené koreláciami. Hlavnou inšpiráciou sa pre nás stal článok [56].

Predpokladajme, že premenná $w_{ij} \in [0, 1]$ vyjadruje hranu s určitou váhou medzi vrcholmi i a j. Podľa [56], ak hrana prislúchajúca dvom vrcholom má veľkosť blízku jednej, potom sa tieto vrcholy považujú za *silných* susedov. V opačnom prípade sú vrcholy *slabými* susedmi. Uzly s hranou, ktorej veľkosť sa nachádza niekde medzi nulou a jednotkou, sa označujú ako *mierni* susedia.

Koeficient zhlukovania pre ľubovoľný vrchol vo váženom neorientovanom grafe by mal spĺňať nasledovné charakteristiky:

- Koeficient zhlukovania pre vrchol i nadobúda veľkú hodnotu, ak jeho silné susedné vrcholy sú si navzájom silnými susednými bodmi. Naopak, hodnota koeficientu vrchola i klesá, ak podiel jeho silných susedov, ktoré sú si navzájom slabé, stúpa.

- Ak váhy väzieb zahŕňajúce ostatných susedov vrchola i narastajú, potom aj koeficient zhlukovania tohto vrchola by mal narastať proporcionálne.

- Koeficient zhlukovania uzla i by mal byť nízky, ak uzol má len slabých susedov alebo najviac jedného nie slabého suseda.

Aby koeficient zhlukovania spĺňal tieto znaky aj pre úplný vážený neorientovaný graf

je potrebné, aby sa zaviedli označenia a definície dôležité na výpočet a presnú definíciu koeficientu.

1. Maticu susednosti pre graf N_t definujeme ako $A_t = [1\{w_{ij} \geq t\}]$, kde $t \in [0, 1]$. Graf N_t vzniká pridelením hrán medzi všetkými pármi uzlov, ktorých váha medzi sebou je aspoň taká veľká ako zvolená hodnota t. Prvok matice A_t nachádzajúci sa v i-tom riadku a v j-tom stĺpci označujeme a_{ij}^t.

2. Nech $\gamma_i(t)$ označuje počet trojuholníkov, kedy jeden z uzlov je uzol i a ostatné dva uzly sú ľubovolné susedné body tohto uzla. Túto veličinu vypočítame nasledovne

$$\gamma_i(t) = \sum_{j \neq i} \sum_{k \neq i,j} a_{ij}^t a_{jk}^t a_{ik}^t = [A_t^3]_{ii}.$$

   Nech N je počet uzlov a $\Gamma_i(t)$ označuje množstvo trojuholníkov, kde jeden z vrcholov je uzol i a ostatné dva sú susedné vrcholy tohto uzla a zároveň sú si susednými uzlami navzájom. Vypočítame

$$\Gamma_i(t) = \sum_{j \neq i} \sum_{k \neq i,j} a_{ij}^t a_{ik}^t = [A_t O A_t]_{ii},$$

   kde $O = 1.1^T - I$, pričom 1 vyjadruje jednotkový vektor dĺžky N a I je $N \times N$ identická matica.

   Z týchto dvoch hodnôt definujeme koeficient zhlukovania pre vrchol i a graf N_t nasledovne

$$C_i(t) = \frac{\gamma_i(t)}{\Gamma_i(t)} = \frac{[A_t^3]_{ii}}{[A_t O A_t]_{ii}},$$

   ak $[A_t O A_t]_{ii} \neq 0$, inak $C_i(t) = 0$.

3. Koeficient zhlukovania zodpovedajúci uzlu i vo váženom neorientovanom grafe vypočítame ako

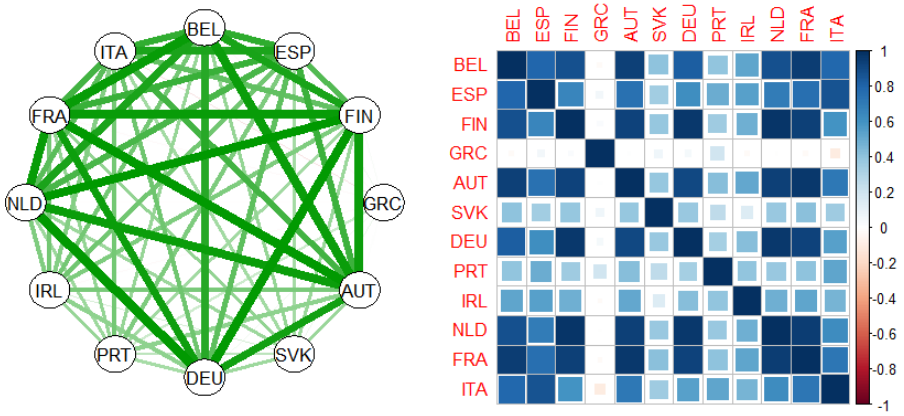
$$C_i = \int_0^1 C_i(t) dt.$$

4. Koeficient zhlukovania pre graf N je definovný ako priemerná hodnota koeficientov zhlukovania nad všetkými uzlami

$$C = N^{-1} \sum_{i=1}^{N} C_i.$$

V našom prípade budú váhy predstavovať korelácie. Aj napriek tomu, že niektoré korelácie sú veľmi malé, bude splnený predpoklad o nenulovosti hrán a teda o úplnom grafe.

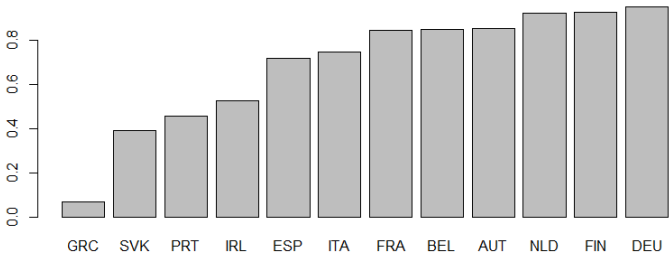
Výpočet koeficientu zhlukovania vrcholov, ktorý sme definovali vyššie, ilustrujeme na príklade výnosového rozpätia krajín Eurozóny. Korelácie medzi krajinami zakreslíme do úplného váženého neorientovaného grafu na Obr. 2.15. Farba a hrúbka hrán reprezentujú veľkosť korelácií medzi výnosovými rozpätiami. Červená farba vyjadruje negatívne korelácie, zelená farba pozitívne korelácie. Niektoré hrany medzi krajinami nie sú viditeľné, čo vyjadruje koreláciu blížiacu sa k nule medzi pozorovanými objektmi.



Obr. 2.15: Úplný vážený neorientovaný graf obyčajných korelácií pre výnosové rozpätie a korelačná matica

Po aplikácií vyššie uvedeného postupu na dáta získame hodnotu koeficientu pre každý uzol. Hodnoty sú vyznačené na diagrame na Obr. 2.16.

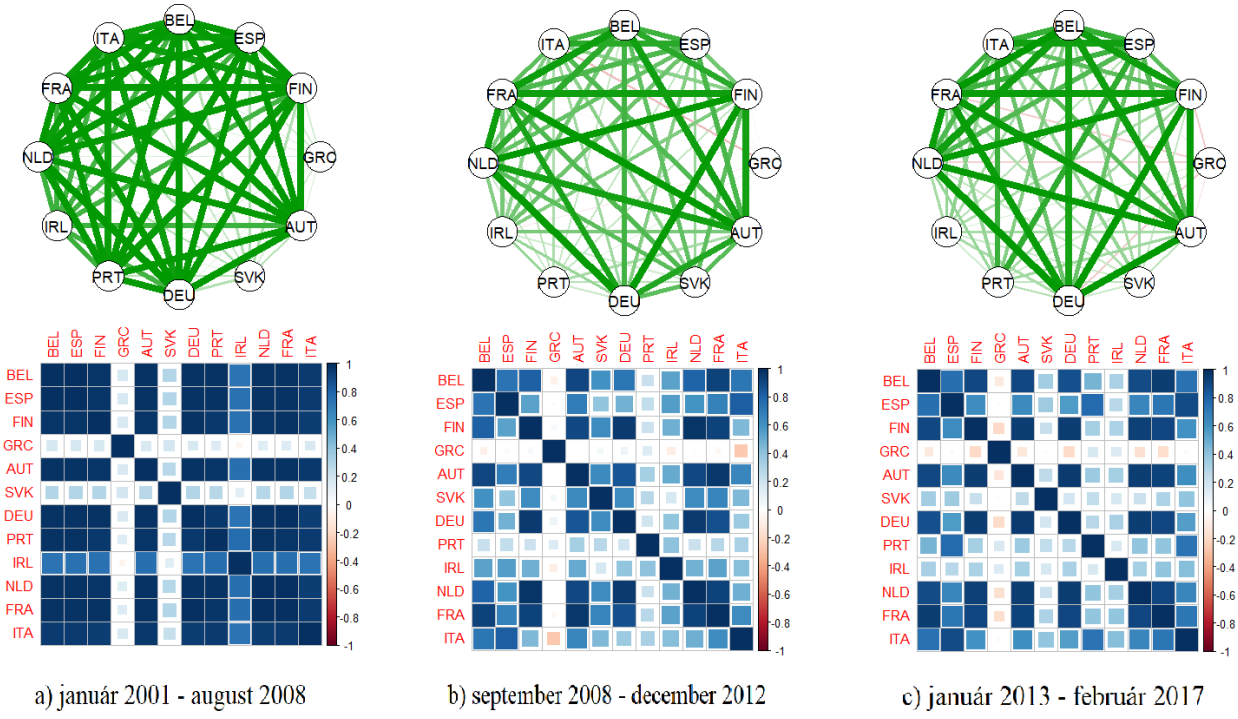
Pre zaujímavosť, pomocou koeficientu zhlukovania porovnáme tri časové obdobia vývoja výnosového rozpätia v Eurozóne. Ako medzník zvolíme globálnu finančnú krízu, ktorá vypukla v septembri v roku 2008. Preto výnosové rozpätie, ktorý analyzujeme od januára 2001 po február 2017, rozdelíme nasledovne: prvé obdobie - január 2001 až august 2008, druhé obdobie - september 2008 až december 2012, tretie obdobie - január 2013 až február 2017.



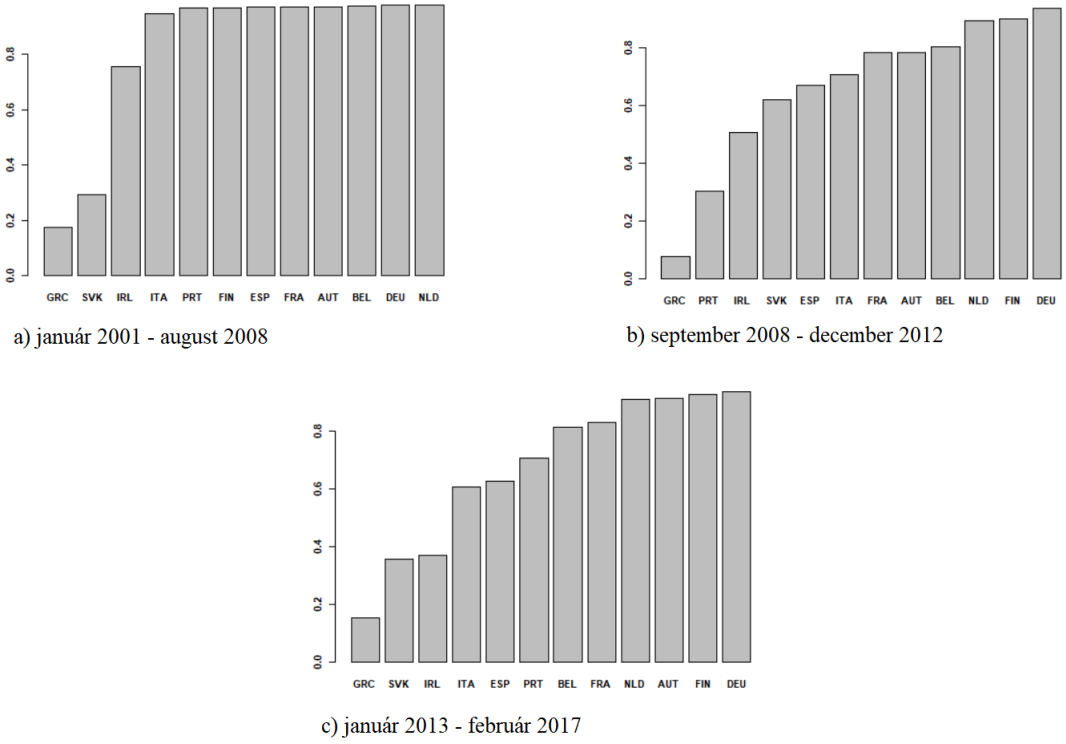
Obr. 2.16: Stĺpcový graf koeficientov zhlukovania

Na Obr. 2.17 sú zachytené grafy s korelačnými maticami pre jednotlivé obdobia a na Obr. 2.18 stĺpcové diagramy koeficientov zhlukovania. Už z obrázkov sa dá vyčítať, ktoré uzly majú nízke hodnoty koeficientov. V predkrízovom období korelácie dosahujú vysoké hodnoty až na tri výnimky - GRC, SVK a IRL. Tieto tri uzly majú v porovnaní s ostatnými krajinami najnižšie hodnoty koeficientov. Na stĺpcovom diagrame sa to prejavuje ako veľký schodkovitý rozdiel medzi jednotlivými koeficientami. V posledných dvoch obdobiach nie je kontrast v koreláciách tak veľmi výrazný ako v prvom období, čo sa tiež prejaví na tvare stĺpcových diagramov.

Postavenie uzlov v grafoch sa výrazne nemení. Najnižšiu tendenciu zhlukovania sa má krajina GRC vo všetkých troch obdobiach. Ostatné krajiny s nízkym koeficientom zhlukovania sú SVK, IRL a počas krízy PRT. Na druhej strane, DEU patrí medzi krajiny s veľkou hodnotou tejto veličiny, z čoho môžeme tvrdiť, že zohráva dôležitú úlohu v zostrojených grafoch.



Obr. 2.17: Úplné vážené neorientované grafy s príslušnými korelačnými maticami



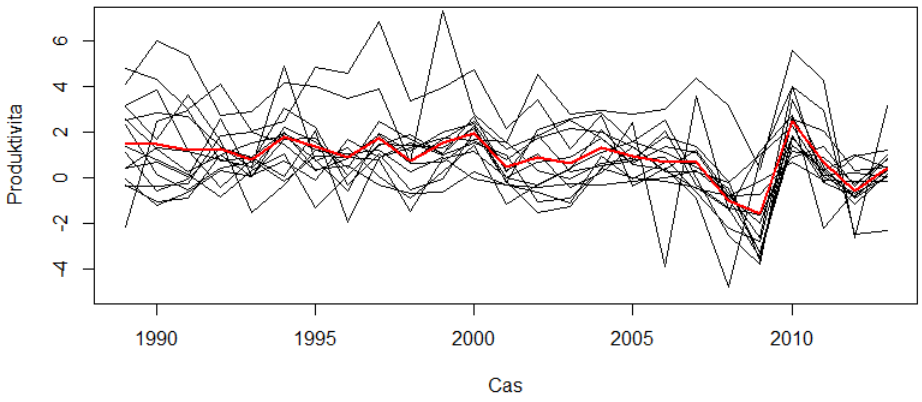
Obr. 2.18: StĺpcovÉ grafy koeficientov zhlukovania pre tri časové obdobia

3 Multifaktorová produktivita

Multifaktorová produktivita je ukazovateľ, ktorý hovorí o celkovej efektívnosti, s akou sa vstupy, ako práca, kapitál, energia, materiál a iné spolu používajú vo výrobe. Akékoľvek zmeny v multifaktorovej produktivite sa prejavujú ako zmeny napríklad v nákladoch, organizačných zmenách alebo v úsporách. Ak sa vstupy medzi dvomi obdobiami nezmenia, potom multifaktorová produktivita odráža všetky zmeny v produkcii. Tento faktor je meraný v ročnej miere rastu [5].

Budeme pracovať s časovými radmi, ktoré predstavujú ročné dáta multifaktorovej produktivity 15 štátov sveta v časovom rozpätí od roku 1989 do roku 2013. Budeme ich analyzovať počas celého časového obdobia, ale tiež dáta rozdelíme na menšie časové úseky. Prvú skupiny budú tvoriť dáta do roku 2007 vrátane, pretože v roku 2008 nastala vo svete finančná kríza. Druhá časť dát predstavuje krízové obdobie a tretiu časť tvorí len jeden rok, konkrétne 2013, čo je pokrízové obdobie. S jedným pozorovaním sa však nedajú robiť žiadne výpočty, preto tento rok v tomto delení nebudeme používať. V závere výsledky navzájom porovnáme. Rovnako ako v predchádzajúcich častiach práce, dáta čerpáme zo stránky OECD [5].

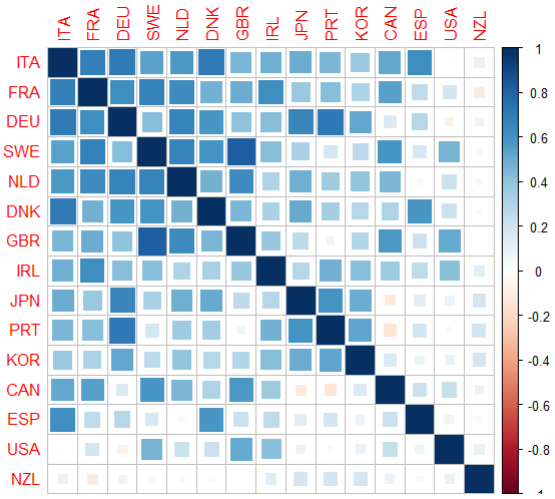
Na Obr. 3.1 je znázornený vývoj produktivít krajín. Hodnoty dát predstavujú percentuálnu zmenu produktivity voči predchádzajúcemu roku. Červenou farbou je vyznačený ich priemer. Nedá sa prehliadnúť veľký pokles okolo roku 2008, kedy vo svete prepukla finančná kríza, ktorá mala veľký dopad na všetky ekonomiky sveta.



Obr. 3.1: Vývoj multifaktorovej produktivity

Vzťahy medzi produktivitami štátov spočítame pomocou korelačných koeficientov ako v kapitole 1.1. Celkovo je týchto vzťahov 105. Podľa predpisu (1.6) upravíme korelácie pomocou Fisherovej transformácie, aby sme ich signifikantnosť mohli otestovať rovnakým postupom ako v kapitole 1.1. Týmto spôsobom získame 37 štatisticky významných vzťahov.

Na Obr. 3.2 je graficky znázornená korelačná matica. Modrá farba vyjadruje pozitívnu koreláciu, negatívne korelácie sú vykreslené červenou farbou.

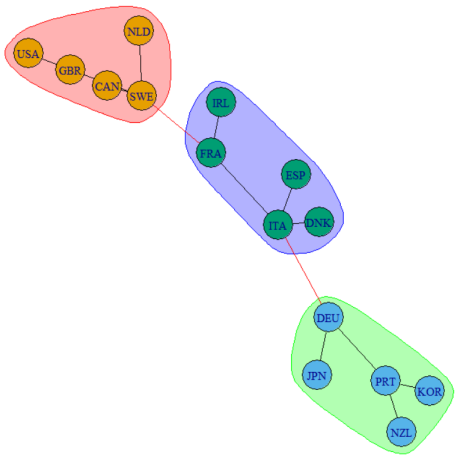


Obr. 3.2: Korelačná matica

Ako si môžeme všimnúť z korelačnej matice, vzťahy medzi časovými radmi sú len pozitívne a tiež sa dá pozorovať prevažne silná korelácia. Keďže väčšina štátov sa nachádza v Európe alebo patrí do Európskej únie, ekonomiky štátov sú navzájom závislé a prepojené, preto vysoká korelácia nie je veľmi prekvapujúca. Jedine Nový Zéland je slabo korelovaný s ostatnými štátmi. Táto takmer nulová korelácia je spôsobená tým, že jeho ekonomika je prepojená s inými krajinami ako tu uvádzame, predovšetkým kvôli geografickému umiestneniu krajiny.

Korelačnú maticu transformujeme do matice vzdialeností a výsledky graficky vykreslíme. Funkciou *walktrap.community* z balíčka *igraph*[9] vytvoríme komunity priamo v diagrame najlacnejšej kostri na Obr. 3.3. Sú tvorené štátmi s podobným vývojom produktivity počas celého pozorovaného obdobia.

Skupina grafov na Obr. 3.4 vyobrazuje priebeh multifaktorových produktivít štátov, ktoré boli spolu zaradené do rovnakých skupín. Na prvom grafe NZL, KOR, JPN, DEU, PRT, na druhom grafe IRL, DNK, ESP, ITA, FRA a na treťom grafe SWE, CAN, NLD,



Obr. 3.3: Najlacnejšia kostra grafu

GBR, USA. Z obrázka môžeme pozorovať, že vývoj štátov v spoločných zhlukoch je podobný.

3.1 Multifaktorová produktivita do roku 2008

Pri analyzovaní dát z časového rozpätia 1989 - 2008 sa výsledky vzájomných vzťahov líšia od výsledkov z celých dát. Tento rozdiel je dobre pozorovateľný na korelačnej matici, ktorá je vyobrazená na Obr. 3.5.

Môžeme si všimnúť, že okrem pozitívnych vzťahov sa tu objavujú aj negatívne závislosti, vyjadrené červenou farbou. Produktivita štátov je menej korelovaná v porovnaní s predchádzajúcou korelačnou maticou na Obr. 3.2. Tiež ubudlo štatisticky dôležitých korelácií, ktorých je len 10. Ich počet získame rovnakým postupom ako v predchádzajúcej časti tejto kapitoly alebo v 1.1.

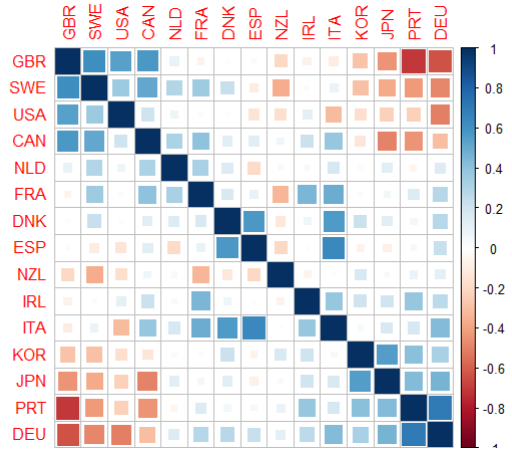
Kvôli týmto rozdielom vznikli iné zhluky krajín. Sú vykreslené pomocou najlacnejšej kostry na diagrame 3.6. Vystupuje tu oveľa viac zhlukov, čo môže byť dôsledkom negatívnych korelácií. Niektoré centrálne umiestnené štáty, ako ITA, FRA a DEU, nezmenili svoje pozície. Na druhej strane, keď berieme do úvahy celé dáta, SWE patrí medzi centrálne krajiny a je prepojené so štyrmi inými štátmi. V redukovaných dátach sa nachádza na okraji a je prepojené len s GBR.

Grafy na Obr. 3.7 poskytujú náhľad vývoja produktivít. Môžeme pozorovať ekonomické situácie štátov z minulosti po zlukoch, do ktorých boli zatriedené.

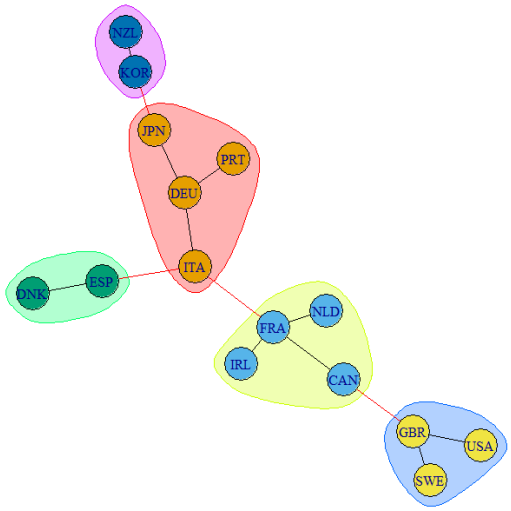
Obr. 3.4: Grafy vývoja multifaktorových produktivít krajín zadelených do spoločných zhlukov [5]

Na grafe c) je vykreslená skupina FRA, GBR, CAN a IRL. Aj napriek tomu, že produktivita IRL je vyššia, tieto krajiny majú veľmi podobný vývoj produktivít až na druhú polovicu 90. rokov. V tomto období má IRL voči ostatným krajinám výrazný rast produktivity. Medzi príčiny spôsobujúce tento výrazný rozdiel patrí napríklad zlepšenie konkurencie schopnosti Írska, atraktívne zahraničné investície, členstvo v EÚ/EMÚ, zlepšenie flexibility, investície do vzdelania a iné [66, 67].

Vývoj produktivity v ESP, znázornený na prvom grafe b) vpravo, je takmer konštantný v porovnaní s DNK, s ktorým je v spoločnej skupine. Podľa [43] bolo v Dánsku



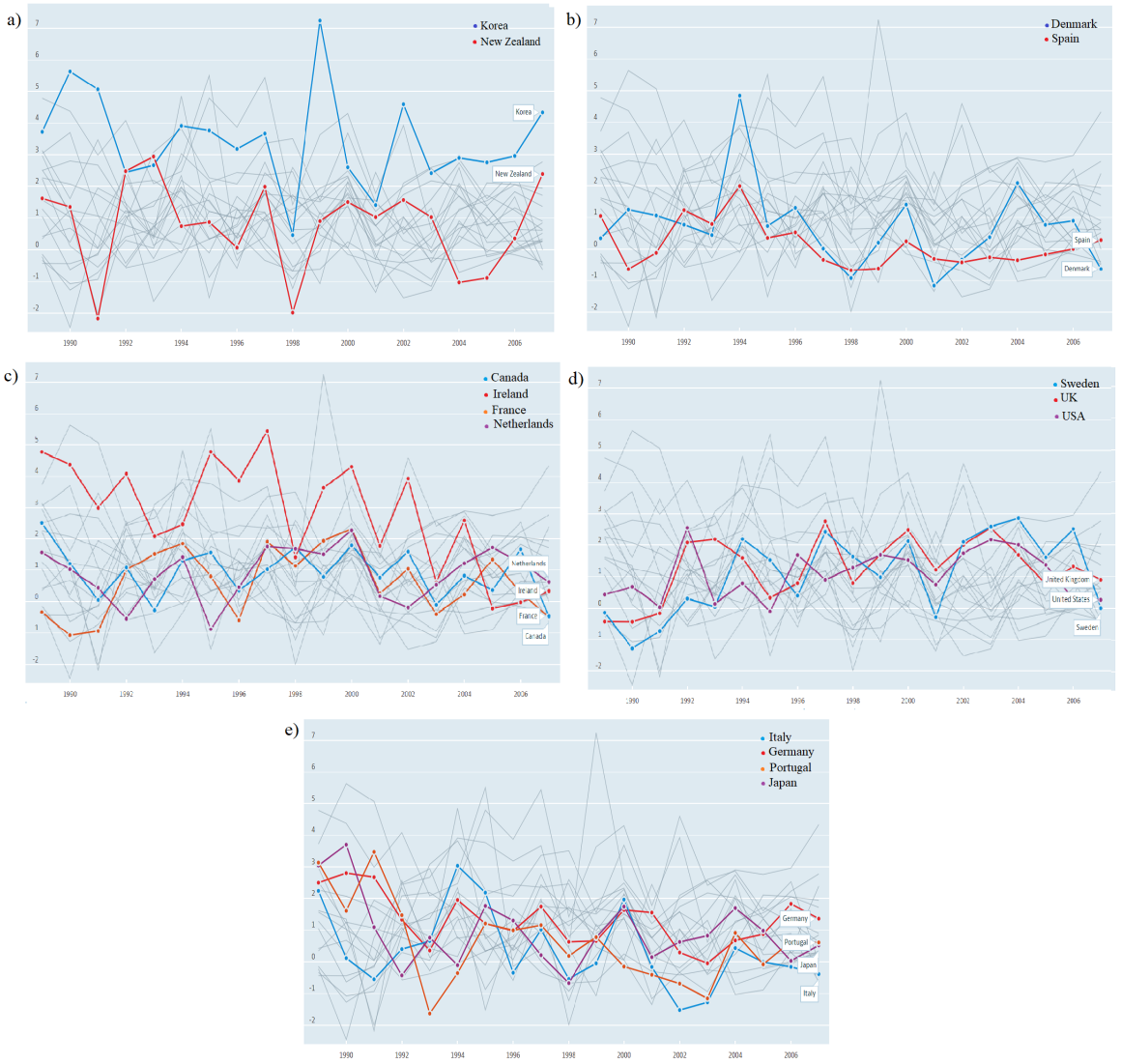
Obr. 3.5: Korelačná matica z dát do roku 2008



Obr. 3.6: Najlacnejšia kostra grafu z dát do roku 2008

od 90. rokov minulého storočia vykonaných niekoľko sektorových analýz a rozvojových programov s cieľom zvýšiť produktivitu, predovšetkým v stavebníctve a priemysle. Výrazný skok vo vývoji môžeme pozorovať na grafe hlavne okolo roku 1994.

Kórea a Nový Zéland sú najvzdialenejšie krajiny od ostatných štátov, ktorými sa zaoberáme. Preto nie je veľmi prekvapujúce, že sa nachádzajú v spoločnom zhluku. Z najlacnejšej kostry na Obr. 3.6 je tiež jasne vidno, že ich produktivita je rozdielna od ostatných, lebo sa nachádzajú na samotnom konci a sú oddelené od ostatných ekonomík. Na grafe a) na Obr. 3.7 pozorujeme výrazný pokles v rokoch 1997-1998. V prípade Kórei bola príčinou Ázijská finančná kríza spôsobená krízou komerčného bankovníctva. Jej následkom bolo tiež zvýšenie nezamestanosti na Novom Zélande [65],

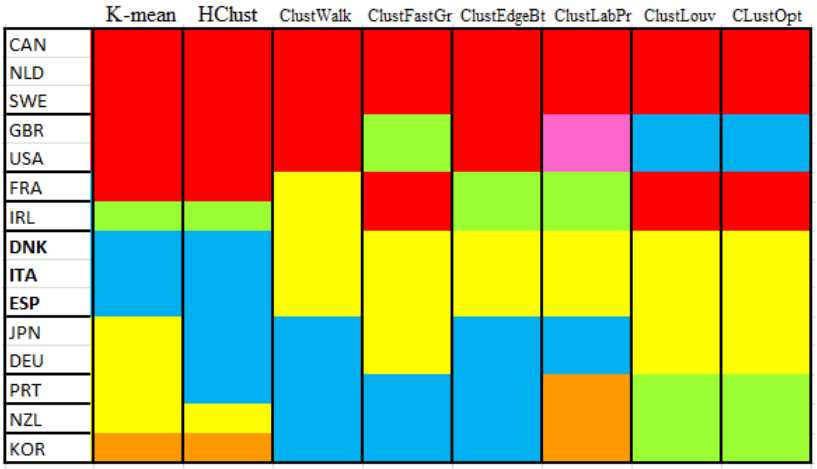


Obr. 3.7: Vývoj produktivít [5]: a)NZL,KOR; b)ESP,DNK; c)IRL,GBR,FRA,CAN; d)USA,SWE,GBR; e)DEU,JPN,PRT,ITA

a teda aj celkové zníženie produktivity Nového Zélandu. Kríza zasiahla predovšetkým ázijské krajiny, preto jej následky vo forme zníženej produktivity pozorujeme výrazne len na týchto dvoch ekonomikách.

3.2 Rôzne algoritmy na porovnanie zhlukov

Na celých dátach vyskúšame algoritmy, s ktorými sme sa zaoberali v podkapitolách 1.5, 2.3 a 2.4, a pomocou nich vytvoríme zhluky. Výsledky sú zaznamenané v nasledujúcej tabuľke na Obr. 3.8. Každý stĺpec zodpovedá jednému algoritmu.



Obr. 3.8: Tabuľka pre rôzne algoritmy, všetky dáta

Môžeme si všimnúť, že niektoré algoritmy tvorili zhluky veľmi podobne, niektoré sa od seba viditeľne líšia. Dôvodom je princíp, na akom tieto algoritmy fungujú.

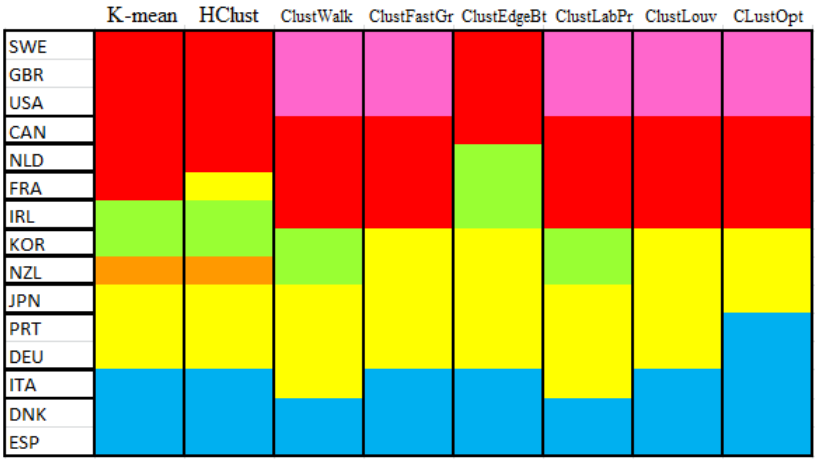
Skupiny krajín, ktoré každý jeden algoritmus zaradil spolu do rovnakej skupiny, sú vyznačené v spoločnom čiernom rámiku alebo zachytené v tabuľke 3.1. Ostatné štáty, ktoré nie sú zaznamenané v tabuľke, tvoria každý osobitne samostatnú skupinu.

CAN, NLD, SWE
DNK, ITA, ESP
JPN, DEU

Tabuľka 3.1: Skupiny krajín

Ak porovnáme výsledky s tabuľkou na obrázku 3.9, ktorá vyjadruje zhluky pre dáta do roku 2008, všimneme si rozdiely. Ak zoberieme do úvahy všetky dáta, tretina krajín sa nachádza v samostatnom čiernom rámiku. Na druhej strane, pre dáta do roku 2008 platí, že osem krajín, čo je viac ako polovica, sú v samostatnom rámiku. Z toho môžeme pozorovať aj nižšiu koreláciu medzi krajinami pred finančnou krízou a rovnako nárast negatívnych korelácií.

Pri porovnávaní tabuliek dôjdeme k zaujímavým pozorovaniam, a tiež k potvrdeniu predchádzajúcich výsledkov. Z prvej tabuľky zistíme, že máme dve veľké skupiny krajín, každá sa skladá z troch štátov. Prvú trojprvkovú skupinu tvorí CAN, NLD a SWE. Aj z korelačnej matice na Obr. 3.2 je zrejmé, že tieto štáty sú silno korelované. Pri porovnaní s druhou tabuľkou, SWE sa nachádza v inej skupine. CAN a NLD sa



Obr. 3.9: Tabuľka pre rôzne algoritmy, redukované dáta

nenachádzajú v rovnakom rámčeku, lebo algoritmus *cluster_edge_betweenness* ich zatriedil do rozdielnych zhlukov. SWE sa nachádza v skupine spolu s GBR a USA. Tieto dva štáty tvorili samostatnú skupinu, keď sme brali do úvahy všetky dáta. Rovnako sú spolu zoradené aj v druhej tabuľke, teda tieto štáty sú silno korelované pred finančnou krízou, ale aj počas celého pozorovaného obdobia.

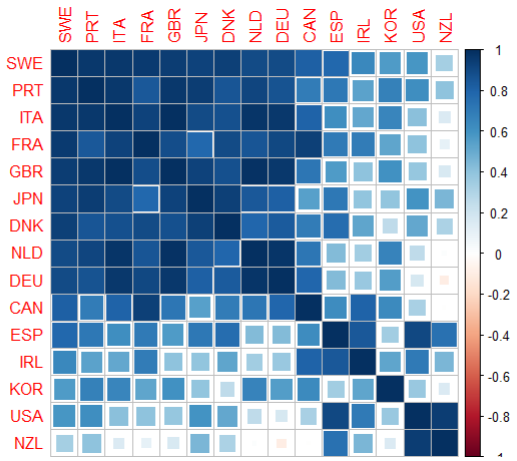
V predchádzajúcej časti sme spomínali, že v dátach do roku 2008 sú medzi štátmi aj záporné korelácie. Tieto vzťahy sa dajú pozorovať v druhej tabuľke 3.9 prostredníctvom farieb. Zatiaľ čo v prvej polovici tabuľky prevláda červená farba s ružovou, v spodnej časti modrá so žltou. Keďže v prvej korelačnej matici nemáme záporné vzťahy, tak prvá tabuľka nie je výrazne rozdelená farbami ako druhá. To znamená, že v pásme červenej farby v hornej časti sa vyskytujú aj farby ako modrá zo spodnej časti tabuľky.

Z druhej tabuľky 3.9 si môžeme všimnúť ešte jednu zaujímavosť. Štáty, ktoré sú zaradené podľa algoritmov do viac ako troch zhlukov, teda obsahujú viac farieb v riadku, majú malé takmer nulové korelácie. V redukovaných dátach sem patria štáty FRA a IRL, ktoré sa nachádzajú v strednom pásme tabuľky.

3.3 Porovnanie produktivít

Časové rady, s ktorými pracujeme, vyjadrujú ročné zmeny multifaktorovej produktivity krajín sveta v rozmedzí rokov 1989 až 2013. Z dát vyčleníme roky 2008 až 2012, ktoré vyjadrujú obdobie finančnej krízy vo svete. Vypočítaním korelačných koeficien-

tov získame silné pozitívne závislosti medzi pozorovanými krajinami. Pri testovaní ich signifikancie získame vysoké číslo, 82 štatisticky významných vzťahov. Graficky sú vykreslené pomocou korelačnej matice na Obr. 3.10.



Obr. 3.10: Korelačná matica pre obdobie počas finančnej krízy

Korelácie sa výrazne líšia, ak ich porovnáme s korelačnou maticou z predkrízového obdobia. Negatívne korelácie sa zmenili na pozitívne. Najvýraznejší rozdiel pozorujeme pri krajinách GBR, PRT, DEU SWE a NLD. Z toho dôvodu sa budeme zaoberať otázkou či je štatisticky významný rozdiel v koreláciách pred krízou a počas nej. Výpočtom intervalu spoľahlivosti pre korelačné koeficienty z krízového obdobia zistíme vzťah ku korelačným koeficientom pre predkrízové roky. Budeme postupovať ako v práci [44]. Výpočet 95%-ného intervalu spoľahlivosti plynie z transformácie (1.6) a jeho tvar je nasledovný

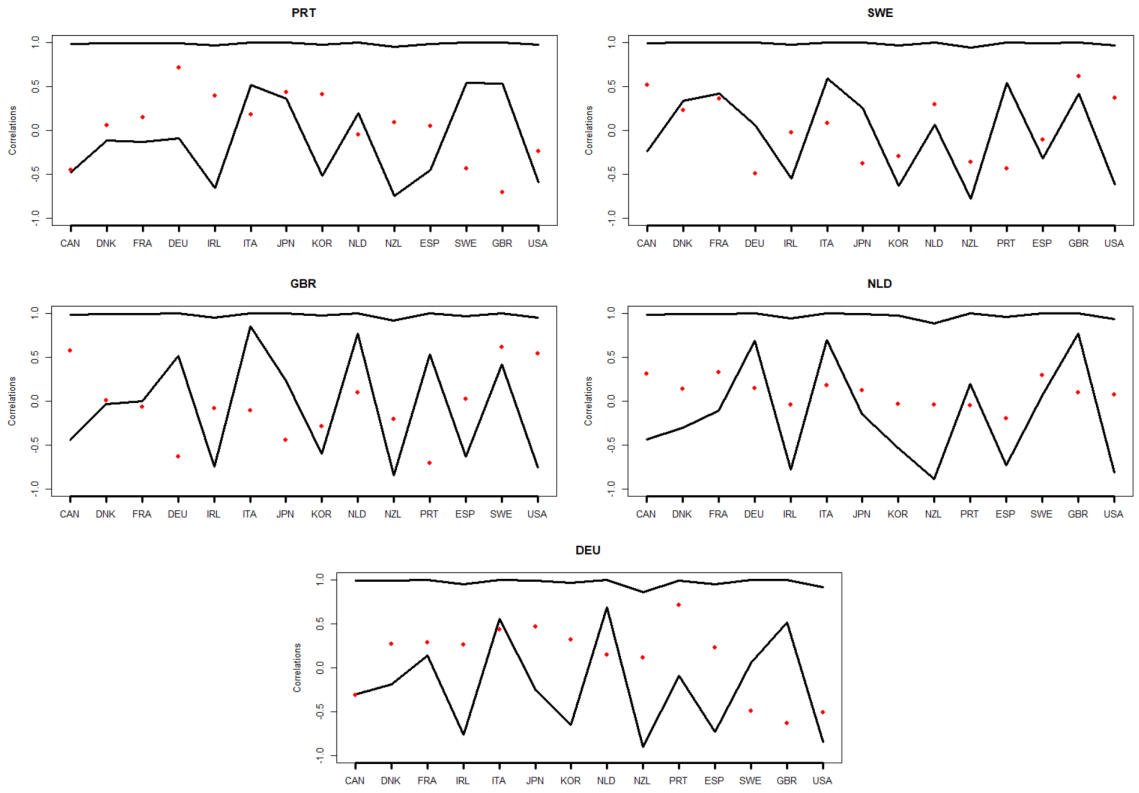
$$\rho \in [tanh(atanh(r) + z_{2.5\%}SE), tanh(atanh(r) - z_{2.5\%}SE)].$$

Premennou ρ označujeme korelačný koeficient a premennou r jeho odhad. SE vyjadruje štandardnú odchýlku definovanú ako $SE = \frac{1}{\sqrt{n-3}}$, kde n je počet pozorovaných objektov a *z-score* je pre 95%-ný interval rovné hodnote 1,96.

Vykreslením intervalov spoľahlivosti vytvorených z dát počas krízy a zaznačením korelácií vypočítaných z predkrízového obdobia, získame 15 grafov, pre každú jednu krajinu zvlášť.

Uvedieme grafy vyššie spomenutých krajín, ktorým sa najvýraznejšie zmenila silná negatívna korelácia na silno pozitívnu alebo neutrálnu. Body na x-ovej osi prislúchajú

jednotlivým krajinám. Z grafov na Obr. 3.11 potvrdíme naše predošlé pozorovania. Vzťahy medzi krajinami sú vyjadrené červenými bodmi. Mimo intervalu spoľahlivosti sa nachádzajú tie, ktorých korelácie sa zmenili veľmi výrazne. Vo všetkých prípadoch sú odhadnuté korelácie nižšie, čo je v súlade s korelačnými maticami. Zatiaľ čo v predkrízovom období sú tieto korelácie veľmi nízke, v období krízy nadobúdajú vysoké hodnoty blízke jednej.

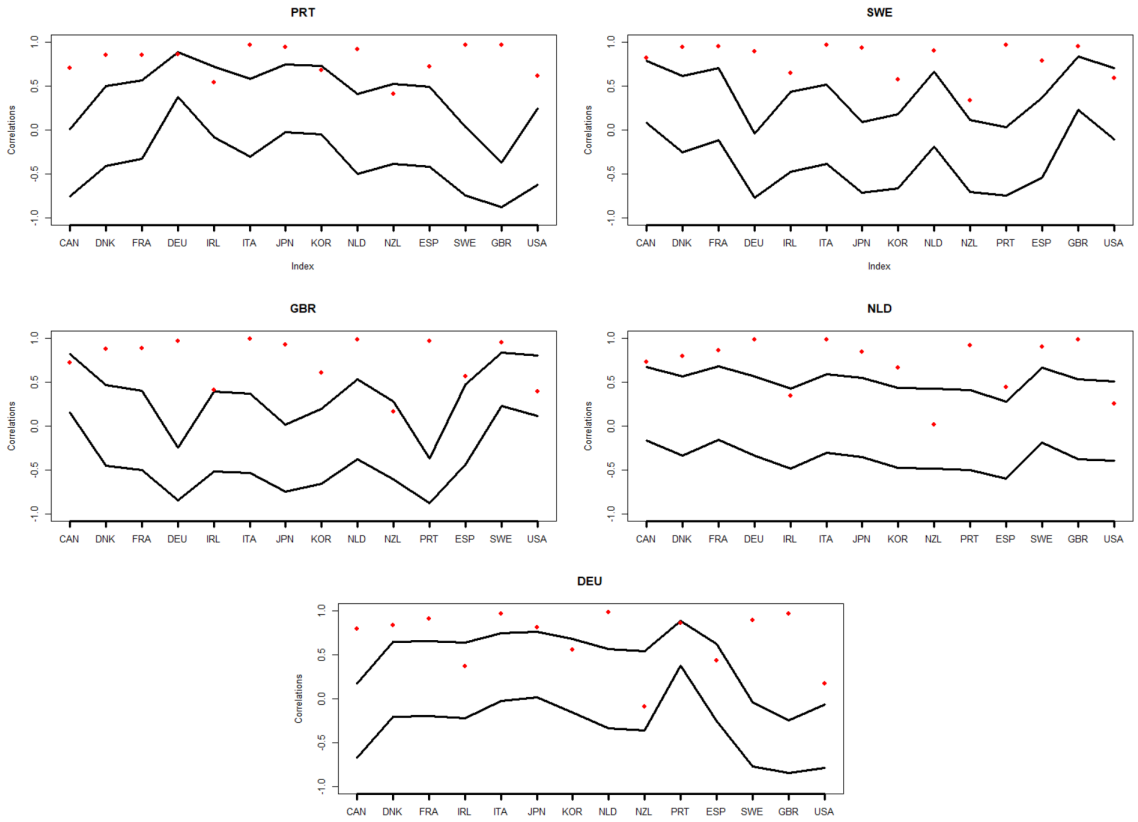


Obr. 3.11: 95%-né intervaly spoľahlivosti

Intervaly spoľahlivosti zvyšných krajín takisto obsahovali korelácie mimo intervalov, ale nie až tak výrazne alebo sa korelácie nachádzali na ich hranici. Vo výsledkoch sa tiež objavili výnimky, ktorých všetky korelácie spadali medzi hornú a dolnú hranicu intervalov. Sú to krajiny CAN, IRL, KOR a FRA. Z korelačnej matice sa môžeme presvedčiť, že produktivita prvých troch krajín počas krízy patrí medzi menej korelované v porovnaní s ostatnými. Vzťahy pred finančnou krízou nie sú až tak výrazné, preto zapadajú do intervalov.

Zopakujeme predchádzajúci postup. Vytvoríme intervaly spoľahlivosti pre dáta z

obdobia pred finančnou krízou a do grafov zakreslíme korelácie z krízového obdobia. Ako vidíme z grafov niektorých krajín na Obr. 3.12, nemajú také široké rozpätie ako v predchádzajúcich výsledkoch. Výsledok je porovnateľný s príslušnou korelačnou maticou. Vzťahy medzi krajinami nie sú až tak výrazné ako v prípade vzťahov počas krízy. Veľa krajín má medzi sebou neutrálne korelácie, o čom svedčí aj otestovanie ich vzájomnej signifikantnosti.



Obr. 3.12: 95%-né intervaly spoľahlivosti

V tomto prípade sa takmer väčšina korelácií z krízového obdobia nachádza mimo intervalov. Korelácie nadobúdajú vyššie hodnoty ako sú intervalové hranice, čo je tiež podmienené tvarom korelačnej matice.

Z výsledkov môžeme vyvodiť záver, že vzťahy medzifaktorových produktivít krajín v predkrízovom období sa výrazne líšia od vzťahov produktívít počas finančnej krízy.

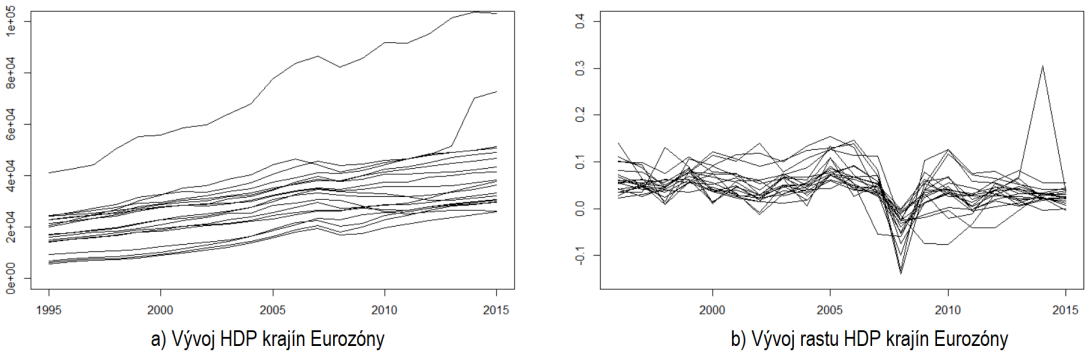
4 HDP krajín Eurozóny

V práci sa zaoberáme vzdialenosťami stacionárnych časových radov. Tieto vzťahy vieme vyjadriť pomocou korelácií. Dáta vyjadrujú ročné HDP na jedného obyvateľa krajín patriacich do Európskej menovej únie počas rokov 1996 - 2006.

Aby sme s dátami mohli pracovať, potrebujeme, aby tieto časové rady boli v stacionárnom tvare. Pomocou funkcie *ur.df* z balíčka *urca* [8] otestujeme ich stacionaritu. Pôvodné časové rady sú nestabilné, preto HDP upravíme do nasledujúceho tvaru

$$x_t := log(HDP_t), \quad r_t := x_t - x_{t-1},$$

a ďalej budeme používať rýchlosť rastu HDP vyjadrenú ako r_t, kde t je čas. Obr. 4.1 obsahuje dva grafy. Prvý vyjadruje časové rady pred diferencovaním a druhý vykresľuje časové rady v upravenom tvare, ktoré sú už vhodné pre ďalšie spracovanie.

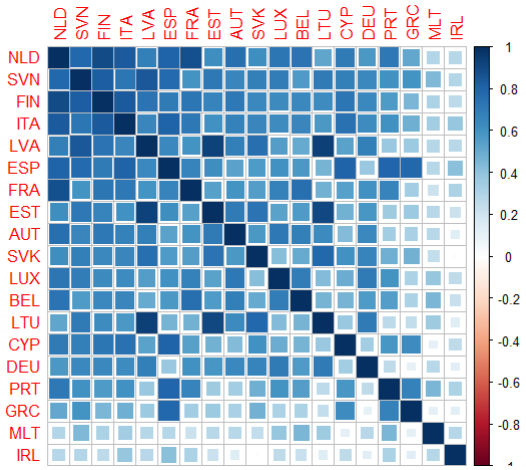


Obr. 4.1: Grafy vývoja HDP a rastu HDP krajín Eurozóny

4.1 Korelačné vzťahy HDP krajín

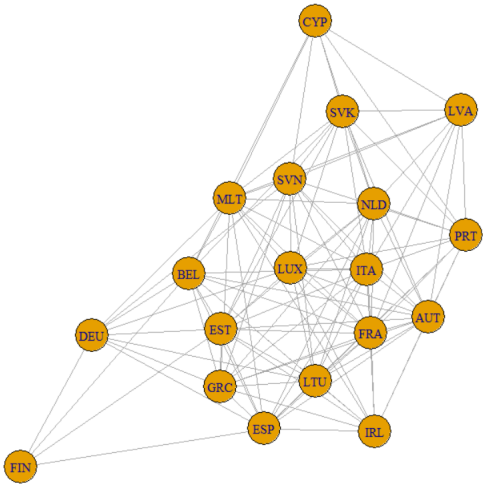
Pomocou korelácií vyjadríme vzťahy medzi rýchlosťami rastu HDP jednotlivých štátov. Na výpočet korelačných koeficientov použijeme Pearsonovu metódu ako v predchádzajúcich kapitolách. Vzájomné závislosti vyobrazíme graficky prostredníctvom korelačnej matice na Obr. 4.2. Každé jedno políčko zodpovená jednému vzťahu. Čím je tmavšej farby, tým je korelácia silnejšia.

Rast HDP krajín Eurozóny sa zdá byť silno korelovaný až na dve výnimky - MLT a IRL. Všetky tieto vzťahy, ktorých sila je meraná pomocou korelácií, nemusia byť významné. Preto z dôvodu odstránenia nepodstatných vzťahov je nutné ich otestovať. Na to poslúži p-hodnota, dôležité meradlo na posúdenie štatistickej signifikancie.



Obr. 4.2: Matica korelačných koeficientov

Prvým krokom testovania je úprava korelácií do stabilizačného tvaru (1.6) a následný vypočet p-hodnoty. Celkový možný počet signifikantných korelácií je $\frac{N(N-1)}{2}$, teda 171, kde N je počet objektov, v našom prípade 19 krajín Eurozóny. Ak označíme X_i a X_j ako dva objekty, ktoré pozorujeme, potom testujeme hypotézy s nulovou hypotézou, že korelácia $\rho_{i,j}$ je nulová. Po aplikácii testov na dáta získame 111 signifikantných korelácií. Graf na Obr. 4.3 vyjadruje tieto vzťahy.



Obr. 4.3: Graf korelačných vzťahov

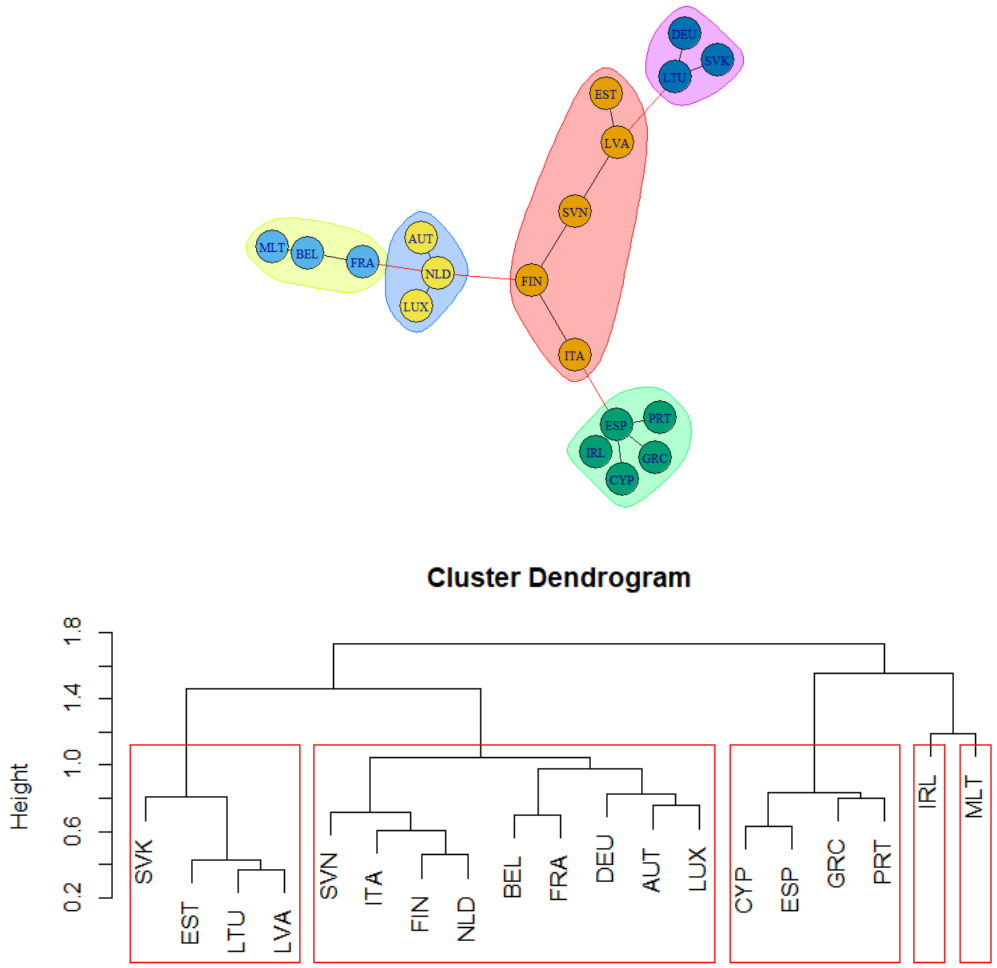
Najlacnejšia kostra grafu a dendrogram

Európsku menovú úniu tvoria štáty, ktoré sú súčasťou EÚ a používajú spoločnú menu - euro. Ako sme predpokladali, krajiny sú silno prepojené, o čom svedčí korelačná

matica. V nasledujúcej časti práce sa budeme zaoberať najlacnejšou kostrou grafu a dendrogramami, s ktorými sme pracovali už v predchádzajúcich kapitolách.

Najlacnejšia kostra grafu je nástroj, pomocou ktorého sa dá množina štátov rozdeliť do zhlukov podľa ich podobností a zároveň poskytuje vizualizáciu vytvorených komunít. Zhluky v najlacnejšej kostre sa môžu zostrojiť pomocou rôznych metód. My sme na tomto konkrétnom príklade použili funkciu *walktrap.community* z balíčka *igraph* [9].

Na Obr. 4.4 môžeme pozorovať dva grafické útvary. Prvý zobrazuje najlacnejšiu kostru vyrátanú pomocou Kruskalovho algoritmu spolu so zatriedením krajín do zhlukov. Druhý obrázok ilustruje dendrogram spolu so zhlukmi pre $k = 5$. Počet zhlukov sme zvolili na základe počtu zhlukov vytvorených v najlacnejšej kostre.



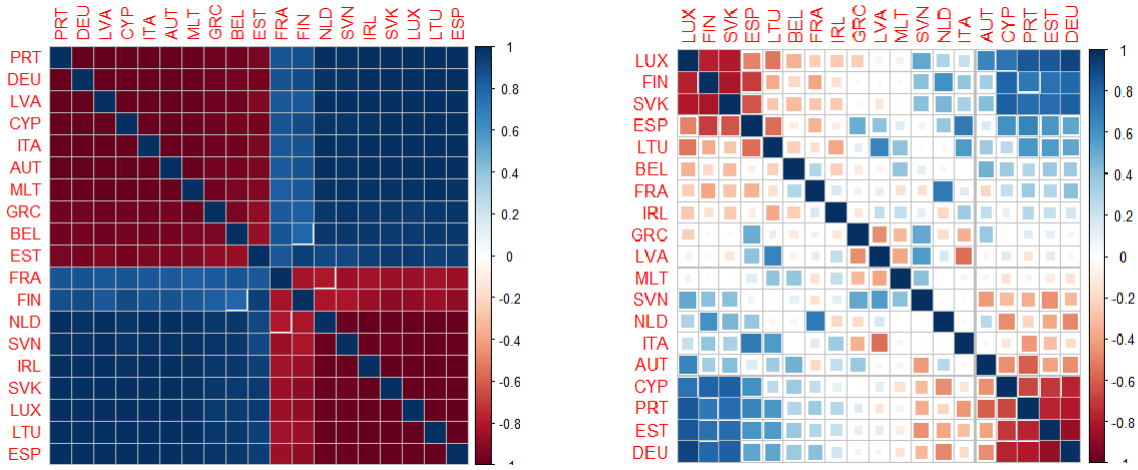
Obr. 4.4: Najlacnejšia kostra grafu a dendrogram hierarchického zhlukovania

4.2 Parciálne korelácie HDP krajín

Parciálna korelácia meria silu korelácie medzi dvomi objektmi, pričom kontroluje vplyv jedného alebo skupiny iných objektov.

Po výpočte parciálnych korelácií môžu byť tieto vzťahy vizualizované pomocou sietí. Každý vrchol zodpovedá jednej premennej a hrana závislému vzťahu, kde parciálna korelácia vyjadruje váhu hrany. Dva objekty nezdieľajú medzi sebou žiadnu hranu, ak ich parciálna korelácia je rovná nule.

Ak vypočítame parciálne korelácie pre rast HDP európskych krajín a vykreslíme ich pomocou korelačnej matice (Obr. 4.5, vľavo), získame nečakané výsledky. Každý jeden vzťah sa zdá byť silno korelovaný. Okrem pozitívnych korelácií, vykreslených modrou farbou, sú tu vo veľkom množstve zastúpené aj negatívne korelácie, vykreslené červenou farbou. Tiež sa nedá zanedbať, že štáty sa úplne jasne oddeľujú a tvoria akoby tri bloky.



Obr. 4.5: Matica parciálnych korelácií

Jeden z možných dôvodov tohto tvaru korelačnej matice je nedostatok dát. Počet pozorovaných premenných je 19, zatiaľ čo počet pozorovaní - rokov vývoja rastu HDP, je len 21. Po pridaní roku 1995 k pôvodným dátam (pozorované obdobie bude 1995 - 2016, spolu 22 rokov) sa matica zmení a nadobudne nový tvar zachytený na Obr. 4.5 vpravo. Veľa predchádzajúcich vzťahov, ktoré sa spočiatku zdali veľmi silné, sa zredukovalo, ale aj napriek tomu si matica ponecháva svoju štruktúru a delí sa na bloky, aj keď menej výrazne.

Kvôli nedostupnosti ďalších dát pre všetky štáty nie je možné analyzovať parciálne

korelácie a vzťahy medzi rastom HDP spôsobom ako v predchádzajúcich príkladoch. Keďže interpretácia výsledkov nie je úplne jasná, snažili sme sa nájsť metódu, ktorou by bolo možné korektne sledovať parciálne korelácie rýchlosti rastu HDP krajín Eurozóny.

4.3 Regularizovaná sieť parciálnych korelácií

Hlavným predpokladom odhadu siete parciálnych korelácií pomocou regularizácie je tzv. riedkosť. To znamená, že sieť nie je úplná a očakávame existenciu vrcholov, ktoré nie sú navzájom spojené pomocou hrany.

Sieť vieme úplne definovať pomocou koncentračnej matice K a matice susednosti A. Riedkosť sa prejavuje ako veľký počet nulových prvkov v týchto maticiach.

LASSO regularizácia

V nasledujúcej časti textu sa pod pojmom korelácie budú vždy brať do úvahy parciálne korelácie. Priblížime výpočet parciálnych korelácií iným spôsobom ako v predchádzajúcich kapitolách. Čerpať budeme predovšetkým z článku [30].

Po odhade korelácií takmer nikdy nenastane prípad, dokonca ani pri podmienenej nezávislosti [2], že by premenné vykazovali presne nulovú koreláciu. Dôsledkom toho sa v grafe vyskytujú hrany, ktoré sú veľmi slabé a značia tzv. falošné korelácie (angl. *spurious* alebo *false positives*). Aby sme sa vyhli zlej interpretácii výsledkov, potrebujeme eliminovať čo najviac týchto korelácií. Na to slúžia tzv. regularizačné metódy, jednou z nich je metóda LASSO (angl. *Least Absolute Shrinkage and Selection Operator*), ktorou sa budeme podrobnejšie zaoberať. Tento model je odvodený z metódy najmenších štvorcov.

Hlavným cieľom metódy LASSO je ohraničenie súčtu korelácií, pričom sa tieto korelácie zmenšujú a niektoré nadobudnú presne nulovú hodnotu. Táto vlastnosť je dôležitá najmä kvôli získaniu riedkej siete. Zo zložitého grafu s veľkým počtom hrán sa stáva čoraz jednoduchší, ktorý obsahuje len dôležité vzťahy vyjadrené v parciálnych koreláciách.

[2] Dva objekty A a B sú nezávisle podmienené vzhľadom k objektu C len a len vtedy, ak pravdepodobnostné rozdelenie A pre B je rovnaké a pravdepodobnostné rozdelenie B pre A je tiež rovnaké, matematicky to vyjadríme nasledovne $P(A \cap B|C) = P(A|C)P(B|C)$ [57].

Cieľom metódy je odhadnutie matice K maximalizáciou penalizovanej vierohodnostnej funkcie

$$L(\lambda) = \log det(K) - tr(SK) - \lambda \sum_{<i,j>} |\kappa_{ij}|,$$

kde premennou S označujeme výberovú korelačnú maticu, K vyjadruje koncentračnú maticu a κ_{ij} prvky matice K. Funkcia sa penalizuje priamo prostredníctvom ladiaceho parametra λ, ktorý určuje mieru kontroly. Pri nízkej hodnote parametra zo siete odbudne málo hrán. Na druhej strane, ak je jeho hodnota príliš vysoká, je odstránených veľa hrán a okrem falošných môže dôjsť k zmiznutiu aj tých dôležitých. Preto výber optimálnej hodnoty tohto parametra je podstatným krokom v LASSO regularizácii.

Výber parametra lambda

Existuje viacero spôsobov ako vybrať najvhodnejšiu hodnotu parametra λ. Napríklad ako je uvedené v článku [35], λ určíme z logaritmicky rozloženého intervalu, kde jedna krajná hodnota zodpovedá maximálnej absolútnej korelácii, čo je maximálne λ, a druhá hodnota je jej skalárnym násobkom.

Princíp metódy LASSO je založený na tom, že vytvorí viacero grafov pre rôzne hodnoty λ, usporiada ich podľa hustoty sietí, a potom vyberie tú najvhodnejšiu. Na určenie vhodnej siete, a ku nej prislúchajúcej optimálnej hodnote λ, sa využíva informačné kritérium. Hlavnou myšlienkou hľadania siete je minimalizácia informačného kritéria a pomocou toho optimalizácia prispôsobenia siete dátam. Najbežnejšie používané kritérium je rozšírené Bayesovo informačné kritérium (EBIC, angl. *Extended Bayesian Information Criterion*), ktoré ako sa ukázalo, patrí medzi najvhodnejšie hlavne pri práci s riedkymi sieťami. Viac o tomto kritériu môžeme nájsť napríklad v prácach [31, 32].

Nástroj, ktorý využíva EBIC kritérim, je hyperparameter γ. Tento parameter slúži na kontrolu ostatných parametrov a jeho hodnota sa nadstavuje manuálne. Vyjadruje preferencie kritéria, nakoľko EBIC uprednostňuje jednoduchší model pred zložitejším. Rozsah parametra sa nadstavuje medzi 0 a 0,5. Ak nadstavíme γ rovné nule, je odhadnuté väčšie množstvo hrán, medzi nimi aj tie falošné. Naopak, pre γ veľkosti 0,5 preferujeme jednoduchší model s menším počtom hrán, ale odstránené môžu byť aj

dôležité vzťahy. EBIC vypočítame nasledovne

$$EBIC(E)_\gamma = -2\hat{L} + E\log(N) + 4\gamma E\log(P),$$

kde N je veľkosť výberovej vzorky, E množstvo nenulových hrán, P množstvo vrcholov a $\hat{L}$ vyjadruje maximum vierohodnostnej funkcie v logaritickom tvare.

Numerická implementácia LASSO úlohy

Vo všeobecnosti môžeme LASSO odhad definovať ako

$$\hat{\theta_\lambda} = \arg\min_\theta \sum_{i=1}^{n}(y_i - X_i\theta)^2 + \lambda\sum_{j=1}^{p}|\theta_j|, \qquad \lambda \geq 0,$$

kde $y \in R^n$ vyjadruje vektor vysvetľovanej premennej a $X \in R^{n\times p}$ maticu vysvetľujúcich premenných.

Ekvivalentne môžeme LASSO odhad vyjadriť pomocou noriem nasledovne

$$\hat{\theta_\lambda} = \arg\min_\theta \|y - X\theta\|_2^2 + \lambda\|\theta\|_1,$$

kde $\|.\|_1$ vyjadruje l_1-normu. Zo zadania úlohy vyplýva, že výpočet odhadu tejto metódy regresnej analýzy patrí medzi úlohy konvexného programovania.

Interpretácia LASSO odhadu je podobná ako pri odhade metódou najmenších štvorcov s tým rozdielom, že pri minimalizácií súčtu štvorcov chýb berie do úvahy ladiaci parameter λ. Výpočet odhadu patrí medzi úlohy kvadratického programovania a je možné ho riešiť štandardnými algoritmami numerickej analýzy.

Softvér R obsahuje vstavané algoritmy na riešenie úloh LASSO. Medzi ne patria dva základné balíčky, *Glmnet* [58] a *Lars* [59]. Glmnet (angl. *Lasso and elastic-net regularized generalized linear models*) odhaduje všeobecný model penalizáciou maximálnej vierohodnostnej funkcie. Regularizačná cesta je počítaná vo forme mriežky, v ktorej sú uložené hodnoty regularizačného parametra. Lars (angl. *Least Angle Regression, Lasso and Forward Stagewise*) je metóda, ktorá funguje na princípe, že z množiny možných vysvetľujúcich premenných vyberie tú, ktorá nadobúda najväčšiu absolútnu koreláciu s vysvetľovanou premennou y. Viac o týchto balíčkoch a ich porovnaní sa môžeme dozvedieť v práci [45].

Balíček *CVRX* [46] je odvodený od toolboxu CVX v softvéri Matlab. Takisto je zameraný na konvexnú optimalizáciu. Prostredníctvom modelovacieho jazyka, určeného pre disciplinované konvexné programovanie, umožňuje formulovať optimalizačný problém prirodzenými matematickými pravidlami.

4.4 Aplikácia LASSO regularizácie

V predchádzajúcej časti sme stručne načrtli fungovanie LASSO metódy. V nasledujúcom texte aplikujeme túto regularizačnú metódu na reálnych dátach.

Na vizualizáciu LASSO metódy slúži grafický LASSO algoritmus, nazývaný *glasso*. Parciálne korelácie odhaduje invertovaním výberovej korelačnej matice. V nasledujúcej časti budeme pracovať s balíčkom *glasso* [60]. Funkcia využívajúca EBIC informačné kritérium je zahrnutá v balíčku *qgraph* [61] a *bootnet* [62].

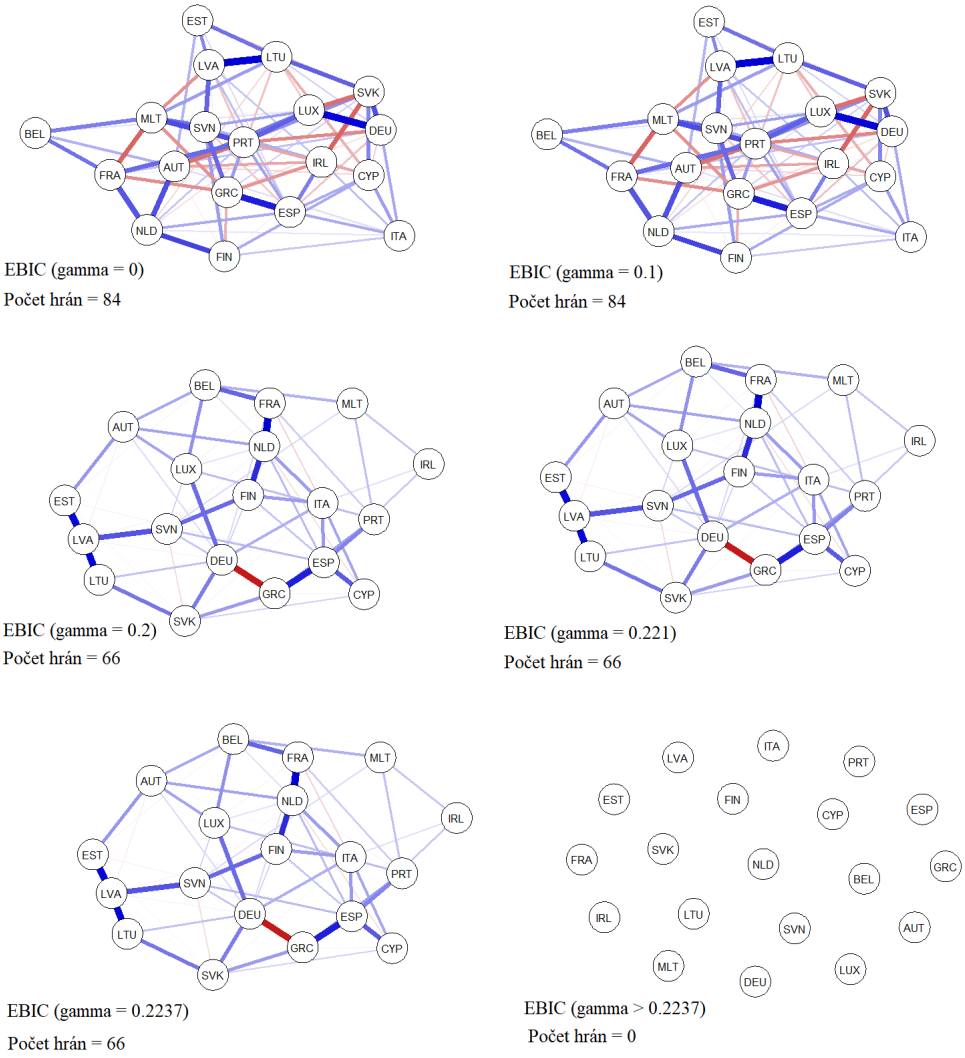
V tejto časti budeme pokračovať s príkladom o rýchlosti rastu HDP. Aplikáciou metódy LASSO na dáta odhadneme sieť parciálnych korelácií a zhluky krajín ako pri práci s obyčajnými koreláciami.

Priamo vo funkcii *estimateNetwork* z balíčka *bootnet* [62] zvolíme hodnotu hyperparametra γ. Funkcia odhadne parciálne korelácie, ktoré môžu byť následne vizualizované vo váženej sieti. Každý vrchol v grafe vyjadruje premennú a hrana závislosť medzi dvomi premennými. Ak sú dva objekty nekorelované, t.j. korelácia je presne nulová, nezdieľajú žiadnu spoločnú hranu. Farba a sýtosť hrany vyjadrujú silu korelácie. Modrou farbou sú vyznačené pozitívne korelácie, červenou farbou negatívne. Na Obr. 4.6 sú vykreslené grafy pre rôzne hodnoty hyperparametra.

Podľa očakávania, sieť, ktorej hyperparameter mal najmenšiu t.j. nulovú hodnotu, obsahuje najväčší počet hrán. Postupným zväčšovaním parametra odbúdajú hrany, až pre hodnoty väčšie než 0,2237 sa v sieti neobjavujú žiadne hrany.

V prvých dvoch grafoch pozorujeme väčší počet záporných korelácií. Postupnou elimináciou hrán ostane len jedna silnejšia a to medzi DEU a GRC.

Miera dôležitosti vrcholov v grafe sa dá merať pomocou indexov centrality. Máme tri základné indexy, ktorými sme sa zaoberali v kapitole 2.1: centralita stupňa, blízkosti a stredovej medzipolohy. Všetky uvedené miery sa dajú graficky vyobraziť pomocou

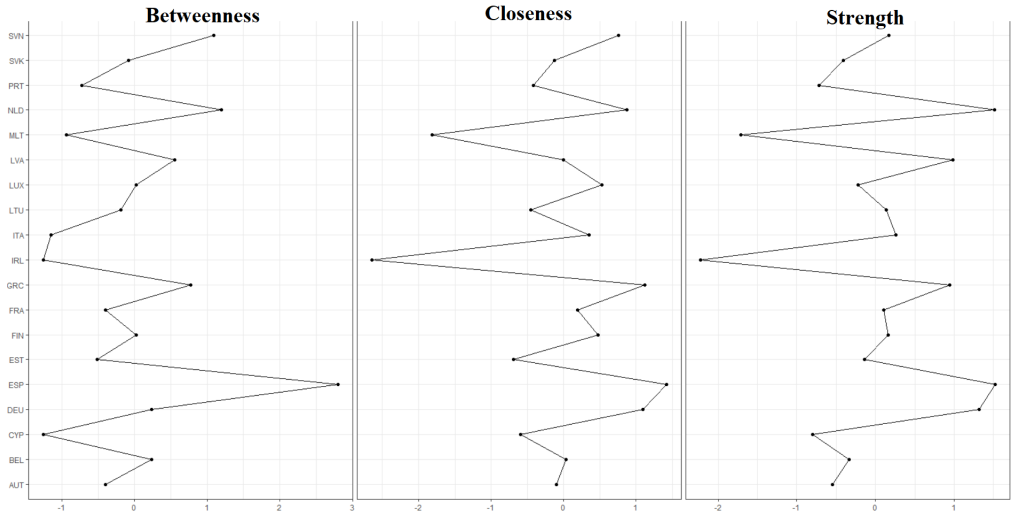


Obr. 4.6: Siete parciálnych korelácií pre rôzne γ

funkcií z balíčka *qgraph* [61]. Čím sú hodnoty indexov pre daný objekt väčšie, tým sa zväčšuje dôležitosť vrcholu v grafe.

Nasledujúca tabuľka 4.1 poskytuje výstup funkcie *centrality* z balíčka *qgraph* [61], ktorá spočíta hodnoty dvoch indexov centrality pre zvolený hyperparameter $\gamma = 0,2$. Výsledky sú graficky zachytené na Obr. 4.7.

Z tejto tabuľky a z obrázka sa dá vyčítať dôležitosť uzlov - štátov v grafe. Štáty, ako ESP, NLD a GRC dosahujú vo všetkých troch indexoch väčšie hodnoty v porovnaní s ostatnými, preto ich môžeme pokladať za významnejšie v zmysle centrality. Na druhej strane, MLT a IRL nadobúdajú najnižšie hodnoty. Na Obr. 4.8 sa tieto krajiny nachádzajú na okraji siete so slabou intenzitou hrán, čo tiež svedčí o ich menej významnej



Obr. 4.7: Indexy centrality pre $\gamma = 0,2$

	Betwenneess	Closeness		Betwenneess	Closeness
AUT	16	0.0054	ITA	2	0.0058
BEL	28	0.0055	LTU	20	0.0052
CYP	0	0.0050	LUX	24	0.0059
DEU	28	0.0064	LVA	34	0.0055
EST	14	0.0050	MLT	6	0.0041
ESP	76	0.0066	NLD	46	0.0062
FIN	24	0.0059	PRT	10	0.0052
FRA	16	0.0057	SVK	22	0.0054
GRC	38	0.0064	SVN	44	0.0061
IRL	0	0.0034			

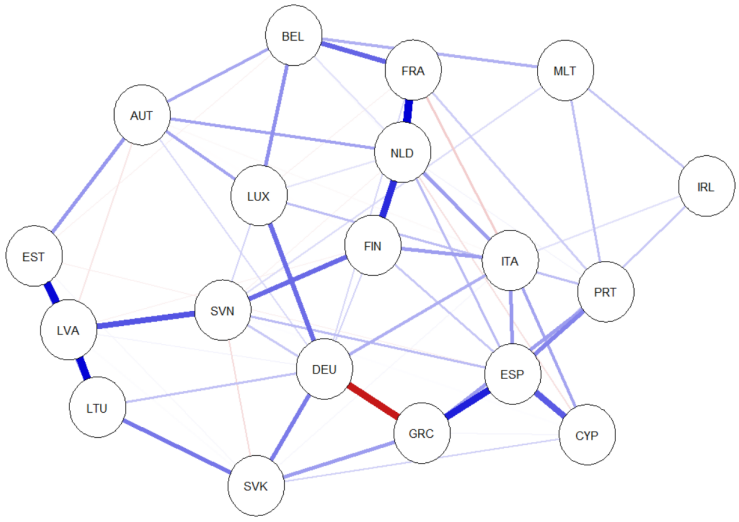
Tabuľka 4.1: Hodnoty indexov cetrality

centralite.

Zluky a LASSO algoritmus

Parciálne korelácie získané výpočtami pomocou LASSO algoritmu vieme aplikovať na tvorbu zhlukov krajín. Ďalej budeme pracovať so sieťou, ktorú sme získali pri nadstavení hyperparametra na veľkosť $\gamma = 0,2$. Voľba tejto konkrétnej siete je podmienená tým, že neobsahuje príliš veľa hrán ako v prípadoch pre menšiu hodnotu γ, a tak odstránime falošné korelácie, ktoré sa v grafe objavujú. Na druhej strane, skladá sa z dostatočného

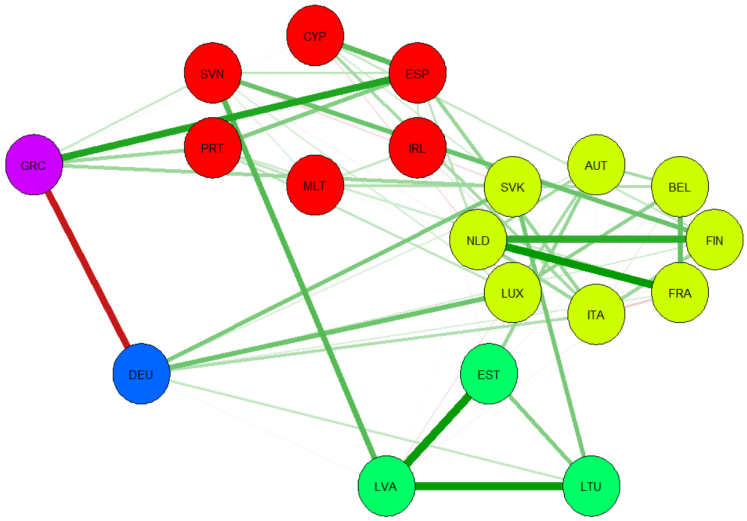
množstva korelácií vhodných na pozorovanie vzťahov, tvorbu komunít a interpretáciu výsledkov.



Obr. 4.8: Sieť parciálnych korelácií pre $\gamma = 0,2$

Získaná sieť je vykreslená na Obr. 4.8. Graf obsahuje 66 hrán. Z celkového počtu vzťahov pozorujeme len málo silno vyznačených korelácií a až na jednu výraznú výnimku, len kladné vzťahy.

Použitím vhodných funkcií z balíčka *igraph* [9] a *qgraph* [61] extrahujeme priradenia krajín do jednotlivých komunít. Výsledky následne vykreslíme do grafu na Obr. 4.9.

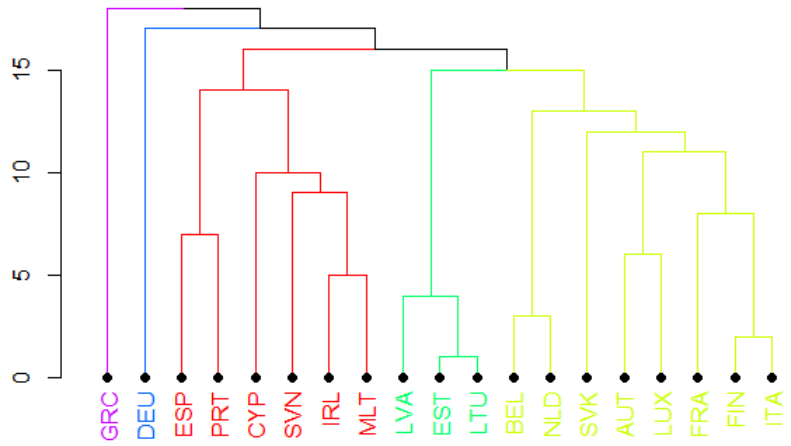


Obr. 4.9: Zatriedenie krajín do zhlukov

Vznikne päť zoskupení, z toho dve sú tvorené len jedným štátom - DEU a GRC.

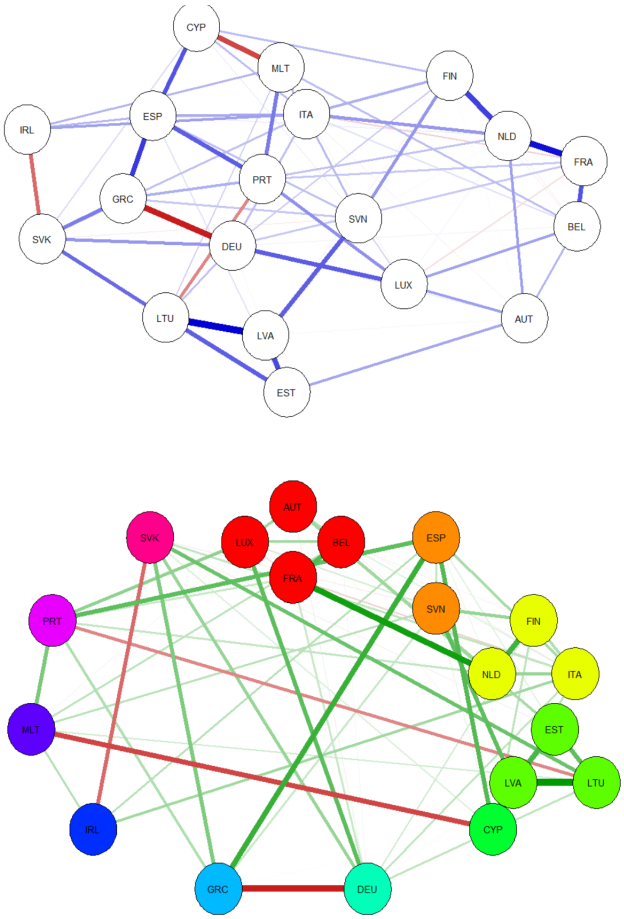
Táto dvojica krajín sa vyznačuje silnou zápornou koreláciou s hodnotou -0,44905 (určená podľa LASSO metódy). Mohli sme ju pozorovať tiež na predchádzajúcich grafoch. Tretiu skupinu tvoria štáty LVA, LTU a EST. Tieto tri krajiny spolu tvoria štáty Pobaltia. Zdieľajú spoločné vlastnosti a históriu, keďže do deväťdesiatych rokov minulého storočia boli súčasťou Sovietskeho zväzu. Spolu tvoria integrovanú ekonomickú oblasť. Prepojenosť týchto ekonomík je podrobnejšie rozpracovaná v [34].

Posledné dve skupiny sú početnejšie. Prečo sú krajiny usporiadané práve do týchto komunít už nie je až tak zrejmé, pretože sú zlúčené krajiny s ekonomikami na rôznych stupňoch úrovní. Hierarchické usporiadanie zachytáva dendrogram na Obr. 4.10.



Obr. 4.10: Dendrogram hierarchických zhlukov

V nasledujúcej časti sa zameriame na stabilitu metódy v zmysle či aj v tomto prípade sú výsledky ovplynené malým počtom dát ako sme skúmali na začiatku tejto kapitoly. To znamená, ak pridáme rok 1995 k pôvodným dátam či sa výrazne zmenia zhluky krajín. Budeme teda pracovať s rastom HDP v časovom období 1995 - 2016. Ako sa môžeme presvedčiť z Obr. 4.11, výsledky sa zmenili, čo sa dá pozorovať hlavne v počte zhlukov. Vytvorilo sa sedem komunít obsahujúcich len jeden vrchol, konkrétne CYP, DEU, GRC, IRL, MLT, PRT a SVK. V sieti pribudlo viacero negatívnych korelácií, čo môže byť dôvodom vytvorenia väčšieho počtu zhlukov.



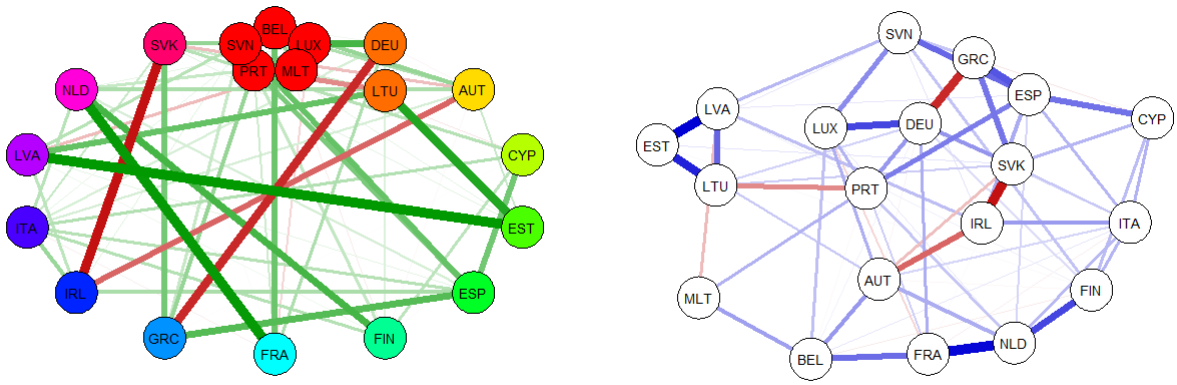
EST	3	3
LTU	3	3
LVA	3	3
AUT	1	1
BEL	1	1
LUX	1	1
FRA	1	1
FIN	1	7
ITA	1	7
NLD	1	7
SVK	1	11
MLT	2	9
PRT	2	10
CYP	2	6
IRL	2	8
ESP	2	2
SVN	2	2
DEU	4	4
GRC	5	5

Obr. 4.11: Sieť paciálnych korelácií pre $\gamma = 0,2$ a tabuľka porovnania zhlukov pre obdobia 1996-2016 a 1995-2016

Podľa [42] mala svetová finančná kríza, ktorá vypukla v roku 2008, v rámci Európskej únie najväčší vplyv na členov Eurozóny. Začala sa prejavovať v prvom štvrťroku 2008, kedy sa rast veľmi spomalil a inflácia prudko narastala. Tento rok môže mať výrazný vplyv na výpočet korelácií a silno ovplyvniť vzťahy medzi krajinami. Dôsledky krízy sa dajú pozorovať tiež z Obr. 4.1.

Ak v dátach ponecháme pridaný rok 1995 a odstránime rok 2008, spočítame nové hodnoty korelácií, získame výsledky, ktoré sa líšia od pôvodných. Otestovaním signifikantnosti korelácií, ako v kapitole 1.1, získame len 48 štatisticky významných korelácií. V porovnaní s predchádzajúcim výsledkom je to výrazný rozdiel. Dôsledkom menšieho počtu týchto vzťahov pri výpočte parciálnych korelácií metódou LASSO a ponechaní hyperparametra γ rovné hodnote 0,2, získame graf, ktorý neobsahuje žiadne hrany. Preto musíme parameter zmenšiť. Nižšou hodnotou γ pripúšťame zložitejšiu sieť s väč-

ším počtom hrán. Pri nadstavení γ rovné 0,1 získame grafy na Obr. 4.12. Ako si môžeme všimnúť, tak aj v tomto prípade sa medzi štátmi objavuje väčšie množstvo zhlukov, ktoré je podmienené väčším výskytom záporných korelácií.



Obr. 4.12: Sieť paciálnych korelácií pre $\gamma = 0,1$ a obdobie 1995-2016 s vynechaným rokom 2008 z dát

Po pridaní jedného roku k pôvodným dátam sa výsledky čiastočne zmenili. Aj napriek tomu sa metóda LASSO ukázala efektívnejšia ako výpočet parciálnych korelácií podľa postupu z kapitoly 1.1, pretože sme boli schopní analyzovať rast HDP krajín a zatriediť ich do zhlukov, čo predtým nebolo možné.

5 Zamestnanosť

Doposiaľ sme sa v práci venovali koreláciám, ktoré sme používali na tvorbu sietí. Vypočítané korelačné koeficienty sme zapisovali do spoločnej tabuľky - korelačnej matice. Tú sme následne používali pri ďalších výpočtoch a analýzach. V tejto kapitole sa budeme zaoberať len samotnými korelačnými maticami. Tak ako v predchádzajúcich častiach, aj tu budeme pozorovať zmeny vo výsledkoch pri použití dát z rôznych období.

Mnohé situácie si vyžadujú štatistické porovnania korelačných koeficientov alebo korelačných matíc, ktoré sú merané na tých istých pozorovaných objektoch. Vo všeobecnosti je snahou testovať nulovú hypotézu v tvare

$$H_0 : P^{(1)} = P^{(2)} = \cdots = P^{(k)}, \quad k \geq 2,$$

kde premennou $P^{(i)}$ označujeme korelačnú maticu.

Existuje viacero testov vyvinutých konkrétne pre túto problematiku, my stručne načrtneme tri z nich, ktoré sú implementované v balíčku *psych* [63].

Jednou z možností je hľadať a porovnávať súčet druhých mocnín korelácií alebo ich Fisherových transformácií (1.6). Nulová hypotéza skúma či sú tieto hodnoty rovné nule, teda či sú prvkami identickej matice. Za platnosti nulovej hypotézy majú hodnoty chí-kvadrát rozdelenie a korelácie sú nezávislé. Tento postup je užitočný hlavne pri skúmaní otázky či sú prvky v korelačnej matici blízke nule, alebo pri porovnávaní matíc navzájom. Metódou sa zaoberal J. Steiger (1980) v článku [40]. V balíčku *psych* môžeme túto funkciu nájsť pod názvom *cortest* alebo *cortest.normal*.

Autorom druhého postupu je R. Jennrich (1970), ktorý sa tiež zameral na testovanie rovnosti dvoch matíc prostredníctvom chí-kvadrát testu. Ako uvádza vo svojej práci [41], počítanie LRT (angl. *Likelihood Ratio Test*) štatistiky môže byť v niektorých prípadoch zložité, preto uviedol novú testovaciu štatistiku s vylepšenými výpočtovými a distribučnými vlastnosťami. Nevýhodou ale je, že je určená pre veľké dáta. Testovacia štatistika má za platnosti nulovej hypotézy asymptoticky chí-kvadrát rozdelenie. Funkcia je v balíku implementovaná pod názvom *cortest.jennrich*. Vo svojej práci tiež uvádza vylepšenú testovaciu štatistiku určenú pre malé výbery dát, ktorá sa riadi normálnym rozdelenením za platnosti nulovej hypotézy.

Poslednou funkciou, ktorá sa zaoberá porovnávaním matíc, je *cortest.mat*. Využíva

princíp podobný faktorovej analýze, v ktorej odhad metódou maximálnej vierohodnosti (angl. *Maximum Likelihood Estimation*) má tvar

$$f = \log(tr((FF^T + U_2)^{-1}R)) - \log(|(FF^T + U_2)^{-1}R|) - n,$$

pričom $(FF^T + U_2)$ je odhad korelačnej matice modelovaný pomocou faktorovej analýzy (F je $p \times k$ rozmerná matica nákladov), R pôvodná korelačná matica a n počet pozorovaných objektov.

Funkcia vypočítané f následne upraví do tvaru, ktorý sa riadi chí-kvadrát rozdelením

$$\chi^2 = \frac{m-1-(2p+5)}{6} - \frac{2*factors}{3} f.$$

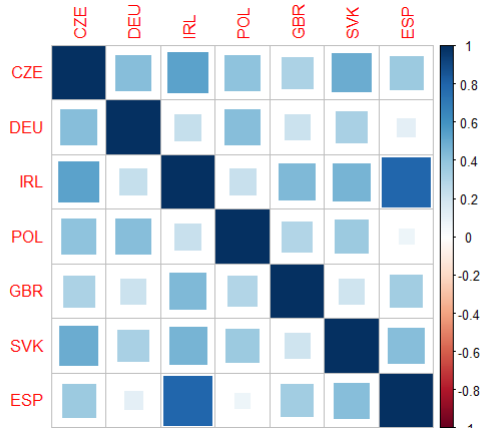
Vo všeobecnosti premenná m určuje počet pozorovaní, p vyjadruje hodnoty pravdepodobností určených pri zostrojovaní intervalov spoľahlivosti a *factors* počet faktorov, ktoré majú byť odhadnuté [64].

5.1 Zamestnanosť

V tejto časti budeme pracovať s dátami, ktoré predstavujú zamestnanosť v európskych krajinách. Dáta sú kvartálne od roku 2000 po rok 2016. Rovnako ako v predchádzajúcich príkladoch ich budeme čerpať zo stránky organizácie OECD [5]. Tú sú dostupné dáta pre mieru zamestnanosti. Tento indikátok vyjadruje, do akej miery sú využívané dostupné pracovné sily, ľudia schopní pracovať. Je závislá od hospodárskych cyklov, v dlhodobom horizonte je významne ovplyvnená napr. výškou vzdelania alebo politikou štátu. Počítame ju ako pomer zamestaných ľudí k celkovému počtu ľudí v produktívnom veku, teda v rozmedzí od 15 do 64 rokov. Tento indikátor je očistený od sezónnych vplyvov a je vyjadrený buď v percentách pracujúceho obyvateľstva, alebo v tisíckach ľudí nad 15 rokov. V práci použijeme dáta vyjadrujúce práve počet ľudí v produktívnom veku. Tieto hodnoty následne upravíme, aby sme získali percentuálny rast zamestnanosti.

Najskôr budeme testovať sedem európskych štátov. Výber nie je náhodný. Snahou bola voľba takých, o ktorých vieme, že z pohľadu zamestnanosti sa navzájom ovplyvňujú, a tak by sa na nich dala pozorovať vysvetliteľná korelácia. Grafy na Obr. 5.2

zachytávajú vývoj zamestnanosti v týchto európskych krajinách. Na Obr. 5.1 je vykreslená korelačná matica zamestnanosti pre celé časové obdobie.



Obr. 5.1: Korelačná matica zamestnanosti

Zameriame sa na porovnanie viacerých korelačných matíc tých istých štátov, vypočítaných počas rôznych časových období. Pracujeme v časovom rozmedzí od roku 2000 po rok 2016. V roku 2008 vypukla svetová finančná kríza, preto chceme pozorovať či toto obdobie ovplyvní zamestnanosť v krajinách a ak áno, do akej miery to ovplyvní tvar korelačných matíc.

Časové obdobie rozdelíme na šesť rovnakých častí, každá bude obsahovať 11 údajov - kvartálnych dát zamestnanosti, čo je presne 2.75 roka. Pre časové úseky vypočítame korelačné matice, ktoré sú zachytené na Obr. 5.3.

Z jednotlivých korelačných matíc pozorujeme, že negatívne korelácie vystupujú prevažne v prvej a šiestej matici, čo sú matice z dvoch období najviac vzdialených od finančnej krízy. Zaujímavosťou ale je, že mnoho pozitívnych korelácií sa zmenilo na negatívne a naopak. Štvrtá matica zachytáva obdobie vypuknutia finančnej krízy, ktoré sa zdá byť silno pozitívne korelované.

V Tabuľke 5.1 uvádzame p-hodnoty vypočítané z dvoch spomínaných testovacích funkcií, kde navzájom porovnávame korelačné matice z dvoch období. Prvok r_{ij} pre $i < j$ vyjadruje p-hodnotu z *cortest.jennrich* testu a prvok r_{ij} pre $i > j$ vyjadruje p-hodnotu z *cortest.mat* testu pri porovnaní korelačných matíc z obdobia i a j. Jennrichov test prijíma nulovú hypotézu vo viacerých prípadoch. Lenže funkcia *cortest.mat* zamieta podobnosť všetkých porovnaných matíc okrem dvojice $R3 - R5$, kde premennou Rk

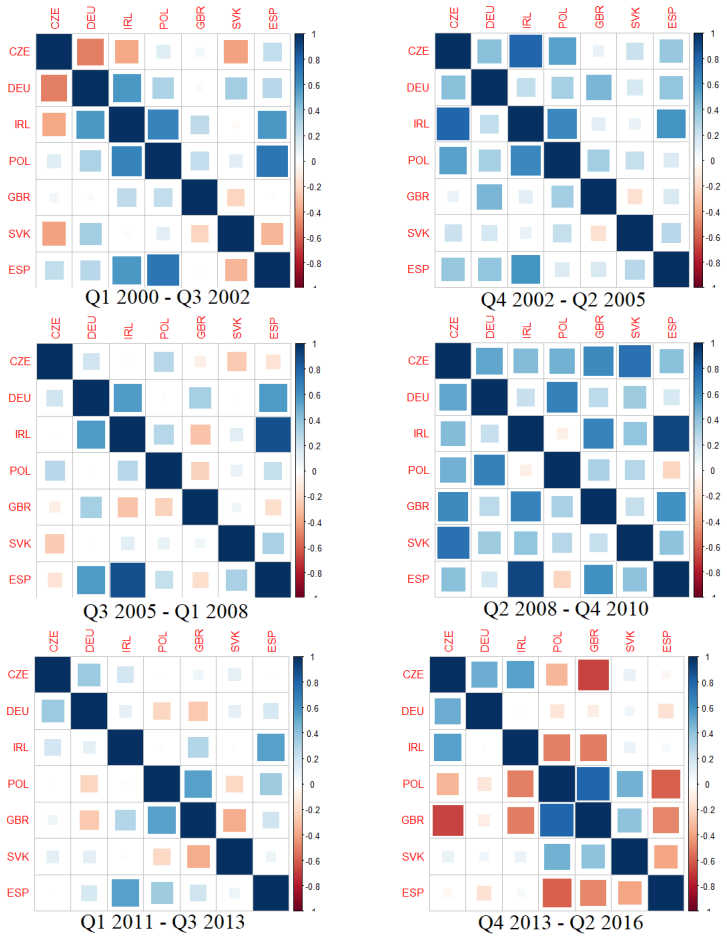
Obr. 5.2: Vývoj zamestnanosti v pozorovaných krajinách

vyjadrujeme korelačnú maticu z k-teho obdobia. Preto zo získaných výsledkov môžeme tvrdiť, že matice, ktoré vyjadrujú zamestnanosť v siedmych európskych krajinách počas pozorovaného obdobia, nie sú podobné kvôli takmer nulovým p-hodnotám v druhej funkcii až na jednu výnimku, kedy je p-hodnota vyššia než päť percent v obidvoch funkciách.

5.2 Vyšehradská štvorka

Vytvoríme novú skupinu tvorenú štátmi Vyšehradskej štvorky (ozn. V4). Krajiny sú stáročia úzko spojené, pretože vždy boli súčasťou jednej civilizácie a zdieľajú spoločné hodnoty a korene. Ich ekonomiky sa navzájom ovplyvňujú [68].

Prvé štyri grafy na Obr. 5.4 vykresľujú vývoj zamestnanosti, ktorý má rastúci trend.



Obr. 5.3: Korelačné matice zamestnanosti pre rôzne časové úseky

		cortest.jennrich					
		R1	R2	R3	R4	R5	R6
cortest.mat	R1	1	$7.91e^{-05}$	$1.1e^{-04}$	$6e^{-04}$	0.098	0.0039
	R2	$1.4e^{-18}$	1	0.0032	0.8216	0.3983	0.04998
	R3	$4.8e^{-09}$	0.0026	1	0.0337	0.4897	0.0046
	R4	$3.3e^{-43}$	$5.5e^{-11}$	$3.6e^{-20}$	1	0.799	$1.2e^{-4}$
	R5	$1.6e^{-07}$	0.0042	0.065	$1.4e^{-10}$	1	0.0662
	R6	$5.1e^{-52}$	$5.9e^{-15}$	$1.7e^{-20}$	$6.6e^{-41}$	$1.4e^{-12}$	1

Tabuľka 5.1: Porovnanie korelačných matíc: p-hodnoty z testov *cortest.jennrich* a *cortest.mat*

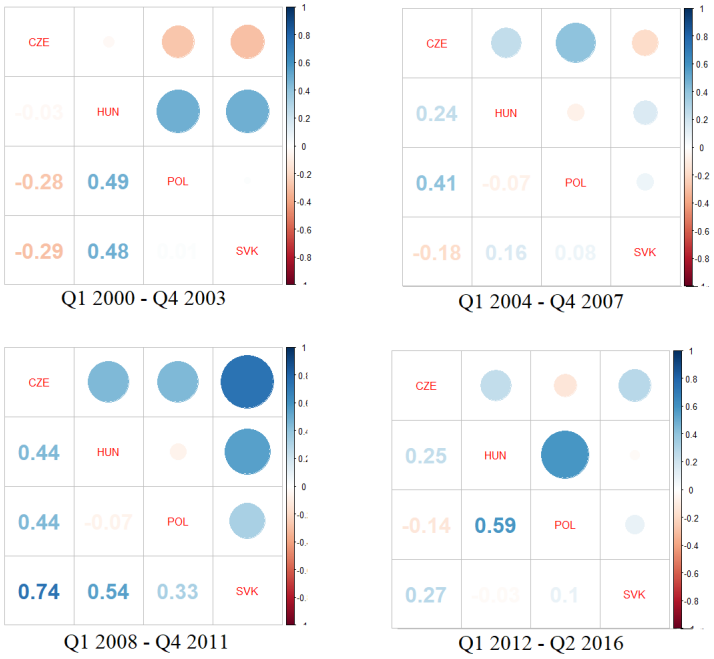
Posledný graf zobrazuje rast zamestnanosti týchto štátov. Môžeme vidieť, že obdobie finančnej krízy (od roku 2008) malo dopad na zmenu zamestnanosti, ale nie veľmi výrazne. Príčinou poklesu množstva pracujúcich ľudí je aj to, že finančná kríza rovnako

zasiahla krajiny západnej Európy. To malo za následok návrat pracujúcich ľudí zo zahraničia do vlastných krajín. V krajinách SVK a HUN kríza nespôsobila dlhotrvajúci upadajúci trend zamestnanosti. V POL zamestnanosť počas celého obdobia neklesala, skôr od roku 2008 len stagnovala. Čo sa týka CZE, zamestnanosť sa udržiavala približne na rovnakej hodnote počas celého obdobia.

Obr. 5.4: Vývoj zamestnanosti v krajinách V4

V tejto časti budeme pozorovať závislosť korelačných matíc a navzájom ich porovnávať. Najskôr časové obdobie rozdelíme na štyri časti, každá zahŕňa štyri roky.

V tabuľke 5.2 sú uvedené p-hodnoty pre jednotlivé testy, ak porovnávame dve po sebe nasledujúce obdobia. Hodnoty testov sa výrazne líšia, p-hodnota metódy *cor-test.mat* nadobúda veľmi nízke hodnoty. Dokonca v každom jednom prípade zamieta podobnosť korelačných matíc ako v predchádzajúcom príklade. Jennrichova metóda v prvých dvoch prípadoch prijíma nulovú hypotézu o podobnosti matíc. Tieto výsledky nie sú prekvapivé. Ak si všimneme tvary matíc a hodnoty korelačných koeficientov na



Obr. 5.5: Korelačné matice pre Vyšehradskú štvorku

Obr. 5.5, tak už na prvý pohľad bez testovania sa zdajú byť matice odlišné.

	cortest.jennrich	cortest.mat
$R_1 vs R_2$	0.0651	0.0023
$R_2 vs R_3$	0.1435	0.0084
$R_3 vs R_4$	0.0117	6.3e-06

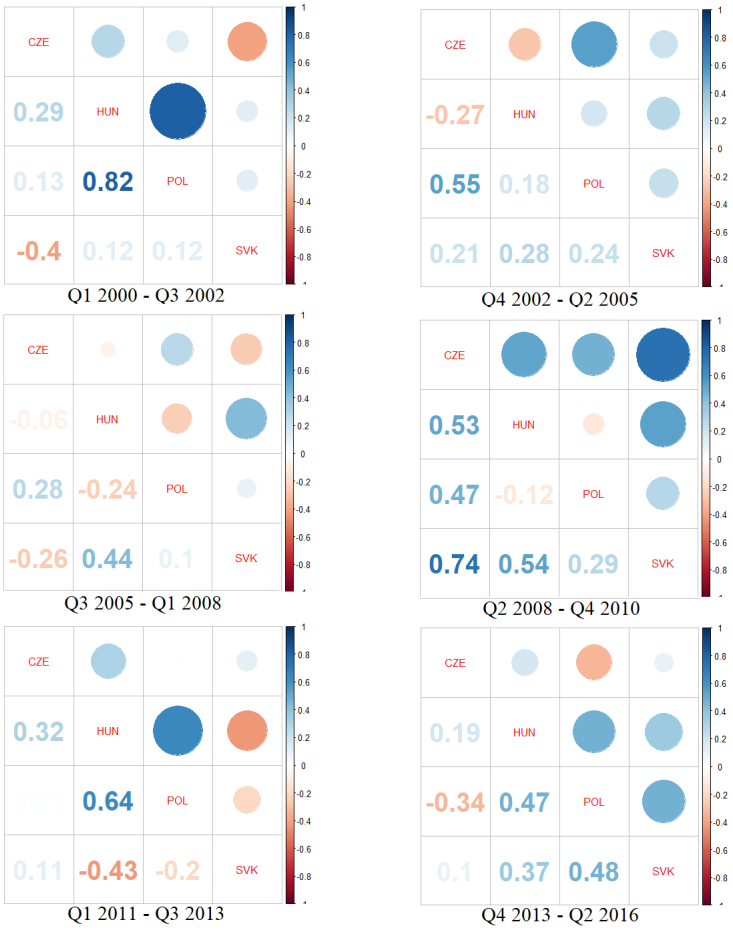
Tabuľka 5.2: Porovnanie korelačných matíc

V tabuľke 5.3 poskytujeme p-hodnoty ako výstupy z testov, ktoré získame pri porovnaní všetkých korelačných matíc navzájom. Prvok r_{ij} v tabuľke vyjadruje p-hodnotu z testov pri porovnaní matice R_i a R_j.

V druhom prípade rozdelíme dáta na šesť období ako v predchádzajúcom príklade a otestujeme podobnosť korelačných matíc. Obrázky sú zachytené na Obr. 5.6 a tabuľka 5.4 poskytuje výsledky p-hodnôt z testov, kde porovnávame dve obdobia nasledujúce po sebe. Z obrázkov na prvý pohľad nevieme usúdiť, v ktorých obdobiach by zamestnanosť mohla byť podobná. V prípade, kedy ani jeden z testov nezamietol nulovú hypotézu, sú úseky vyjadrené korelačnými maticami $R2 - R3$ a $R5 - R6$.

		cortest.jennrich			
		R1	R2	R3	R4
cortest.mat	R1	1	0.0651	0.0012	0.0352
	R2	0.0023	1	0.1435	0.06816
	R3	$8.7e^{-09}$	0.0084	1	0.0117
	R4	0.024	0.0006	$6.3e^{-06}$	1

Tabuľka 5.3: Porovnanie korelačných matíc: p-hodnoty z testov *cortest.jennrich* a *cortest.mat*



Obr. 5.6: Korelačné matice pre Vyšehradskú štvorku

Ak by sme porovnali všetky dvojice matíc navzájom, ďalšie podobnosti matíc vyjdú len v prípade R1-R5 a R1-R6. Dvojica R1-R6 vyjadruje korelačné matice z období najviac vzdialených od vypuknutia finančnej krízy. Zdá sa, akoby sa zamestnanosť po krízovom období dostala do podobného stavu ako bola pred finančnou krízou. Výsledky porovnaní všetkých korelačných matíc obsahuje tabuľka 5.5. Takisto aj v tomto prípade

	cortest.jennrich	cortest.mat
$R_1 vs R_2$	0.0414	0.0008
$R_2 vs R_3$	0.4783	0.45
$R_3 vs R_4$	0.2552	0.028
$R_4 vs R_5$	0.0265	1.9e-05
$R_5 vs R_6$	0.2996	0.31

Tabuľka 5.4: P-hodnoty z porovnaní korelačných matíc pre V4

platí, že prvok tabuľky r_{ij} vyjadruje p-hodnoty z testov pri porovnaní korelačných matíc z období i a j.

		cortest.jennrich					
		R1	R2	R3	R4	R5	R6
cortest.mat	R1	1	0.0414	0.0897	0.0012	0.1444	0.1865
	R2	0.0008	1	0.4783	0.0633	0.0740	0.1461
	R3	$3.3e^{-04}$	0.45	1	0.2552	0.0788	0.2202
	R4	$1.6e^{-08}$	0.018	0.028	1	0.0265	0.0558
	R5	0.34	0.0025	$8.7e^{-05}$	$1.9e^{-05}$	1	0.2996
	R6	0.22	0.012	0.031	$2.3e^{-04}$	0.31	1

Tabuľka 5.5: Porovnanie korelačných matíc: p-hodnoty z testov *cortest.jennrich* a *cortest.mat*

V predchádzajúcich prípadoch sme porovnávali zhodu korelačných matíc pre obdobia rozdelené na rovnako dlhé úseky. Ak rozdelíme pozorované obdobie na rôzne dlhé časy, aj v tomto prípade testy zhody zamietnu nulové hypotézy o podobnosti korelačných matíc, čo svedčí o rozdielnom vývoji zamestnanosti v krajinách V4 vo zvolených obdobiach.

Ako sme už spomínali, v roku 2008 vypukla svetová finančná kríza odštartovaná pádom banky Lehman Brothers v USA. Konkrétne bol bankrot vyhlásený 15. septembra 2008. Tento dátum zapadá do tretieho kvartálu roku 2008. Pri testoch, kedy sme analyzovali rôzne dlhé obdobia, sme roky 2000 až 2016 rozdelili nasledovne: prvé obdobie 2000 Q1 - 2008 Q2, druhé obdobie 2008 Q3 - 2012 Q4, kedy rok 2012 je považovaný

za koniec finančnej krízy a tretie obdobie 2013 Q1 - 2016 Q2. Ani v tomto prípade nenastala zhoda korelačných matíc, žiadny test nepotvrdil nulovú hypotézu.

V druhom prípade sme sa zamerali na dve obdobia. Prvé - pred vypuknutím finančnej krízy, teda 2000 Q1 až 2008 Q2, a druhé obdobie predstavuje zvyšné roky, 2008 Q3 - 2016 Q2. Lenže ani v tomto prípade nenastala zhoda korelácií.

Z jednotlivých analýz vyplýva, že ak sme delili obdobia buď na rovnaké časti, alebo podľa udalostí vo svete, len v pár prípadoch vyšla zhoda korelačných matíc.

5.3 Intervaly spoľahlivosti

V predošlej časti sme testovali zhodu korelačných matíc krajín tvoriacich V4 v šiestich obdobiach. Až na pár výnimiek testy zhody vyšli negatívne. Na potvrdenie zhody a odlišnosti korelačných matíc, ktoré vyjadrujú vzťahy medzi zamestnanosťou krajín, použijeme iný prístup. Aplikujeme intervaly spoľahlivosti a budeme postupovať rovnako ako v kapitole 3.3.

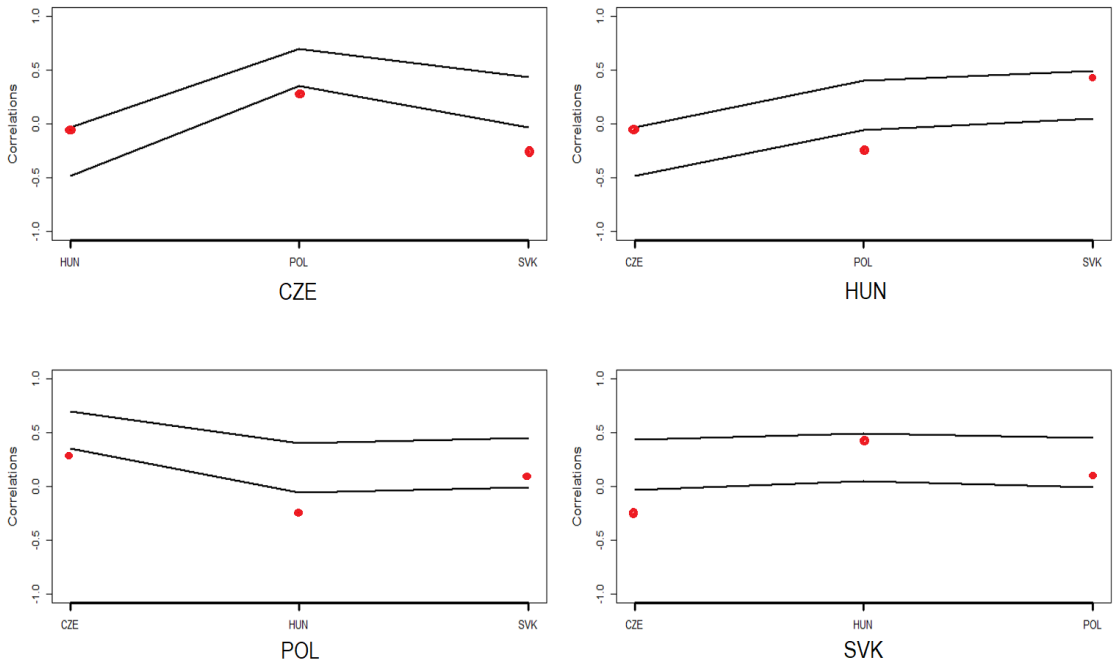
V prvom prípade sa zameriame na obdobia, kedy testy potvrdili zhodu korelačných matíc. Dokopy boli štyri, my vyberieme tú etapu, v ktorej vyšla zhoda najvyššia.

Najskôr vypočítame 95%-né intervaly spoľahlivosti pre korelácie z obdobia 2002 Q4 - 2005 Q2 a zakreslíme do nich korelácie z rokov 2005 Q3 - 2008 Q1. Na Obr. 5.7 sú vyobrazené príslušné grafy pre všetky štyri krajiny V4. Môžeme si všimnúť, že všetky korelácie nezapadajú do intervalov. Každý graf obsahuje koreláciu nachádzajúcu sa pod spodným rozhraním intervalu a ostatné tesne pri hraniciach.

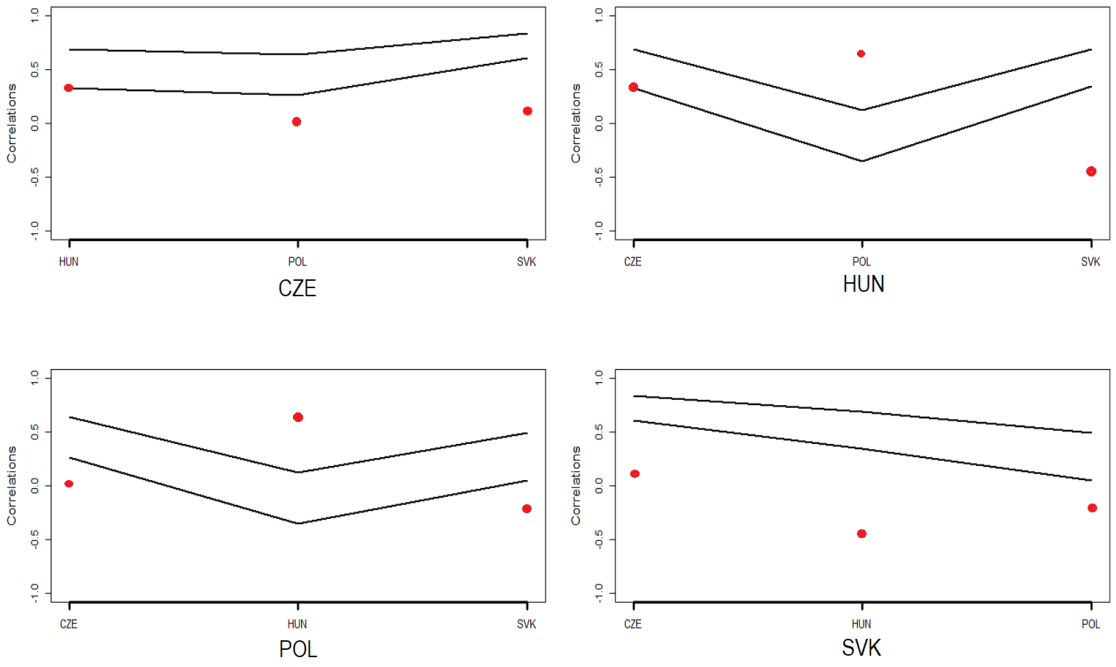
Ak výsledky porovnáme s korelačnými maticami prislúchajúcich období na Obr. 5.5, tak pozorovania sa potvrdia. Ako príklad uvedieme CZE. Najvýraznejšie sa prejaví korelácia s krajinou SVK, pôvodne pozitívny vzťah sa zmení na negatívny. To v grafe pozorujeme ako červenú bodku nachádzajúcu sa pod spodnou hranicou intervalu.

S POL sa pozitívna korelácia zmení na slabšiu, preto sa korelácia z druhého obdobia, zakreslená tiež červenou bodkou, nachádza tesne pod spodnou hranicou intervalu.

V druhom prípade upriamime pozornosť na dva časové úseky, kedy korelačné matice nevykazujú zhodu. Zvolíme si obdobia 2008 Q2 - 2010 Q4 a 2011 Q1 - 2013 Q3. Stav korelácií v tomto prípade opisujú grafy na Obr. 5.8.



Obr. 5.7: Intervaly spoľahlivosti pri zhode korelačných matíc z časových období: 2002 Q4 - 2005 Q2 a 2005 Q3 - 2008 Q1



Obr. 5.8: Intervaly spoľahlivosti pri odlišných korelačných maticiach z obdobia: 2008 Q2 - 2010 Q4 a 2011 Q1 - 2013 Q3

Hodnoty odhadnutých korelácií sa nenachádzajú v intervaloch. Najvýraznejšie sa to prejavuje medzi dvojicami CZE a SVK, HUN a SVK, POL a HUN. Výsledky sú

v súlade s maticami na Obr. 5.5. Korelácie medzi zamestnanoťou v krajinách V4 v prvom období sú výrazne odlišné od korelácií z druhého obdobia ako usudzujeme podľa výsledkov z grafov a zakreslených intervalov. Preto sa nemôžu zhodovať ani matice, ktoré pozostávajú z týchto korelačných koeficientov.

Záver

V práci sme sa zaoberali vzdialenosťami medzi stacionárnymi časovými radmi, ktorá sa dá merať koreláciami. Prostredníctvom rôznych metód sme analyzovali vybrané časové rady. Na vyjadrenie vzťahov používame prevažne grafy, ako je najlacnejšia kostra alebo graf korelačných vzťahov.

V úvode práce stručne približujeme základné pojmy, ako korelácia, stacionarita časových radov, najlacnejšia kostra, ale aj postup testovania signifikantnosti korelácií. Následne sme teóriu aplikovali na konkrétnych dátach výmenných kurzoch a získané výsledky sme zobrazili graficky. Vzťahy medzi časovými radmi sa dajú skúmať aj iným spôsobom, pomocou parciálnych korelácií. Tomu sme sa venovali v druhej časti prvej kapitoly, čím sme chceli predchádzajúce výsledky vylepšiť, a tak zredukovať pôvodné vzťahy o tie nesignifikantné. Najskôr opisujeme ich výpočet dvomi možnými spôsobmi a následne koncepty aplikujeme na príklade.

Druhá kapitola, sa okrem výpočtov korelácií, zameriava na postavenie vrcholov v grafe. Tieto vlastnosti vyjadrujú indexy centrality. Pomocou získaných výsledkov, ktoré sme testovali na dátach výnosového rozpätia, porovnávame dva typy grafov, najlacnejšiu kostru grafu s grafom korelačných vzťahov. Pomocou algoritmov na tvorbu zhlukov v grafe sme zisťovali stabilitu riešení. Koeficient zhlukovania tiež vyjadruje postavenie vrchola v grafe. Tejto veličine sme sa venovali v závere druhej časti práce.

V ďalšej kapitole sme skúmali vplyv svetovej finančnej krízy na korelácie. Pracovali sme s multifaktorovou produktivitou štátov. Metódami, ktorými sme sa zaoberali v predchádzajúcich kapitolách, sme porovnávali predkrízové obdobie s pokrízovým. Nový pohľad na danú problematiku priniesli intervaly spoľahlivosti pre korelácie, čím sme potvrdili výsledky z predchádzajúcich testovaní.

V štvrtej časti práce pracujeme s HDP krajín Európskej menovej únie. V tomto prípade však nastal problém pri vyjadrení parciálnych korelácií, ktorý sme sa pokúšali riešiť hľadaním novej metódy. Pomocou LASSO regularizácie sme vyjadrili vzájomné vzťahy, vytvorili grafy a zhluky. Hlavným motívom zamerania sa na novú metódu LASSO bola nedôveryhodnosť vypočítaných parciálnych korelácií kvôli nedostatku dát. Vďaka jej aplikácii sme mohli štáty rozdeliť do zhlukov, pričom analýza výsledkov sa zdala byť v tomto prípade spoľahlivejšia.

V poslednej kapitole testujeme zhodu korelačných matíc na dátach zamestnanosti. Pozorované obdobie sme rozdelili na viaceré časové úseky, pre ktoré sme vypočítali korelačné matice a následne sme ich medzi sebou porovnávali. Najskôr sme pracovali s viacerými európskymi štátmi, potom sme sa zamerali na krajiny Vyšehradskej štvorky. V závere kapitoly sa výsledky odtestovali pomocou intervalov spoľahlivosti ako v tretej kapitole.

Prínosom práce bolo obohatenie témy merania vzdialeností časových radov o iné metódy, ktoré sú schopné vylepšiť výsledky, a tak skvalitniť analýzu vzťahov. Obyčajné korelácie neboli častokrát postačujúce na vyjadrenie vzdialeností časových radov, pretože vznikali nedôveryhodné vzťahy. Zamerali sme sa na metódy, ktoré sú schopné spracovať informácie iným spôsobom, a tiež sa vysporiadať s menším počtom dát. Dôležitou súčasťou práce bola grafická interpretácia výsledkov.

Literatúra

[1] Colleti, P.: *Comparing Minimum Spanning Trees of the Italian Stock Market Using Returns and Volumes*, Faculty of Economics, Free University of Bozen Bolzano, Italy, Physica A 463 (2016) ,246–261, dostupné na internete (15.3.2018): `https://www.sciencedirect.com/science/article/pii/S0378437116304605`

[2] Jang, W., Lee, J., Chang, W.: *Currency Crises and the Evolution of Foreign Exchange Market: Evidence from Minimum Spanning Tree*, Physica A 390 (2011), 707–718, dostupné na internete (15.3.2018): `https://www.sciencedirect.com/science/article/pii/S0378437110008861?via%3Dihub`

[3] Kazemilari, M., Mardani, A., Streimikiene, D., Zavadskas E. K.: *An Overview of Renewable Energy Companies in Stock Exchange: Evidence from Minimal Spanning Tree Approach*, Renewable Energy 102 (2017), 107-117, dostupné na internete (15.3.2018): `https://www.sciencedirect.com/science/article/pii/S0960148116308953?via%3Dihub`

[4] R Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013, dostupné na internete (15.3.2018): `http://www.R-project.org/`

[5] *Organizácia pre hospodársky rozvoj a spoluprácu*, dostupné na internete (18.1.2018): `www.oecd.org`

[6] Fuller, W. A.: *Introduction to Statistical Time Series*, John Wiley and Sons Ltd, New York, United States, 1995

[7] Kirchgässner, G., Wolters, J., Hassler U.: *Introduction to Modern Time Series Analysis*, Springer, 2008

[8] Pfaff, B.: *Analysis of Integrated and Cointegrated Time Series with R*, Texts in Business and Economics, Second Edition, Springer, New York, 2008, dostupné na internete (8.3.2018): `http://www.pfaffikus.de`

[9] Csardi, G., Nepusz, T.: *The igraph Software Package for Complex Network Research*, InterJournal, Complex Systems, 2006, dostupné na internete (15.3.2018)http://igraph.org

[10] Kolaczyk, E. D., Csárdi, G.: (2017). *sand: Statistical Analysis of Network Data with R*, R package version 1.0.3., 2017, dostupné na internete (18.4.2018): https://CRAN.R-project.org/package=sand

[11] *OTexts*, ARIMA models, dostupné na internete (18.4.2018): https://www.otexts.org/fpp/8

[12] Seongho, K.: *ppcor: Partial and Semi-Partial (Part) Correlation*, R package version 1.1, 2015, dostupné na internete (15.3.2018): https://CRAN.R-project.org/package=ppcor

[13] *R igraph Manual Pages*, dostupné na internete (15.3.2018): http://igraph.org/r/doc/

[14] Eshel, G.: *The Yule Walker Equations for the AR Coefficients*, dostupné na internete (19.4.2018): http://www-stat.wharton.upenn.edu/~steele/Courses/956/ResourceDetails/YWSourceFiles/YW-Eshel.pdf

[15] *Real Statistics Using Excel*, Time Series Analysis, dostupné na internete (19.4.2018): http://www.real-statistics.com/time-series-analysis/

[16] *Spurious Correlations*, dostupné na internete (17.1.2018): http://www.tylervigen.com/spurious-correlations

[17] Holá, Z.: *Modelovanie ekonomických a finančných časových radov*, Diplomová práca, FMFI UK, Bratislava, 2013, 16-17

[18] Rešovský, M., Horváth, D., Gazda, V., Siničáková, M.: *Minimum Spanning Tree Application in the Currency Market*, Technical University, Faculty of Economics, Košice, 2013

[19] Di Matteo, T., Aste, T., Mantegna, R.N.: *An Interest Rates Cluster Analysis*, Physica A 339 (2004), 181 – 188

[20] Mandere, E.: *Financial Networks and Their Applications to the Stock Market*, Bowling Green State University, Physics, 2009

[21] Gower, J. C.: *Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis*, Biometrika 53(1966), 325-338

[22] Wright, P. S.: *Adjusted P-values for simultaneous inference*, Biometrics 48(1992), 1005-1013

[23] Kolaczyk, E., Csárdi, G.: *Statistical Analysis of Network Data with R*, Springer, New York, 2014, 116-125

[24] Matoušek, J., Nešetřil, J.: *Kapitoly z diskrétní matematiky*, Nakladateľství Karolinum, Univerzita Karlova, Praha, 2002

[25] Karger, D., Klein, P., Tarjan, R: *A randomized linear-time algorithm to find minimum spanning trees*, Journal of the Association for Computing Machinery, 42 (2), 1995, 321–328,

[26] Chazelle, B.: *A minimum spanning tree algorithm with inverse-Ackermann type complexity*, Journal of the Association for Computing Machinery, 47 (6), 2000, 1028–1047,

[27] Bondy, J. A., Murty, U. S. R.: *Graph Theory with Applications*, Macmillan Press Ltd., Department of Combinatorics and Optimization, University of Waterlo, Ontario, Canada, 1976, dostupné na internete (15.3.2018): `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.721.3161&rep=rep1&type=pdf`

[28] Chen, X. B., Xu, M., Biao, W. B.: *Covariance and Precision Matrix Estimation for High-dimensional Time Series*, Ann. Statist.41 (2013), 2994-3021

[29] Meinshausen, N., Bühlmann, P.: *Highdimensional Graphs and Variable Selection with the Lasso*, The annals of statistics, 2006, 1436–1462

[30] Epskamp, S., Fried, E.: *A Tutorial on Regularized Partial Correlation Networks*, University of Amsterdam, Department of Psychological Methods, 2017, dostupné na internete (7.11.2017): `https://www.researchgate.net/profile/`

Sacha_Epskamp/publication/304859642_A_Tutorial_on_Regularized_Partial_Correlation_Networks/links/59c47b890f7e9bd2c0fe2bac/A-Tutorial-on-Regularized-Partial-Correlation-Networks.pdf

[31] Foygel, R., Drton, M.: *Extended Bayesian information criteria for Gaussian graphical models*, Advances in Neural Information Processing Systems, 2010.

[32] Jiahua, Ch., Zehua, Ch.: *Extended Bayesian Information Criteria for Model Selection with Large Model Spaces*, Biometrika, Volume 95 (2008), Issue 3, 759–771, dostupné na internete (7.11.2017): http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.505.2456&rep=rep1&type=pdf

[33] Alarcón, O., Gordi, A.: *Partial Correlation Network Analysis*, Universitat Pompeu Fabra, 2013, dostupné na internete (7.11.2017): https://repositori.upf.edu/bitstream/handle/10230/22202/Correlation%20analysis%20FYP%20UPF%20version%20(1).pdf;sequence=1

[34] Poissonnier, A.: *The Baltics: Three Countries, One Economy?*, European Economy, 2017. ISBN 978-92-79-64841-0, dostupné na internete (7.11.2017): https://ec.europa.eu/info/sites/info/files/eb024_en.pdf

[35] Zhao, P., Yu, B.: *On Model Selection Consistency of Lasso*, The Journal of Machine Learning Research, 2006, 2541–2563

[36] Christiansen, CH.: *Predicting Severe Simultaneous Recessions Using Yield Spreads as Leading Indicators*, Journal of International Money and Finance, 2013, 1032-1043

[37] *The Balance*, What Every Investor Should Know About Yield Spread, dostupné na internete (22.3.2018): https://www.thebalance.com/what-is-a-yield-spread-417077

[38] *Investopedia*, dostupné na internete (22.3.2018): https://www.investopedia.com/terms/y/yieldspread.asp

[39] *Fisher's Transformation of the Correlation Coefficient*, dostupné na internete (14.11.2017): https://blogs.sas.com/content/iml/2017/09/20/fishers-transformation-correlation.html

[40] Steiger, J.: *Testing Pattern Hypotheses on Correlation Matrices: Alternative Statistics and Some Empirical Results.*, Multivariate Behavioral Research 15 (1980), 335-352

[41] Jennrich, R.: *An Asymptotic χ^2 Test for the Equality of Two Correlation Matrices.*, Journal of the American Statistical Association 65 (1970), 904-912

[42] Euractiv: *Finančná kríza*, dostupné na internete (30.12.2017): https://euractiv.sk/section/podnikanie-a-praca/linksdossier/financna-kriza-000227/

[43] Bertelsen, S., Nielsen, J.: *The Danish Experience from 10 Years of Productivity Development Proceedings of the 2nd International Conference on Construction Industry*, Singapore, 1999

[44] Marečáková, B.: *Aplikácie modelu Creditmetrics na odhad kreditného rizika krajín Eurozóny*, FMFI UK, Bratislava, 2015, 39-42

[45] Fonti, V.: *Feature Selection using LASSO*, VU Amsterdam, 2017, 10-11

[46] *Convex Optimization in R*, dostupné na internete (23.2.2018) https://cvxr.rbind.io/index.html

[47] *Wolfram MathWorld*, Graph Distance, dostupné na internete (22.3.2018): http://mathworld.wolfram.com/GraphDistance.html

[48] *DataCamp*, K-Means Clustering in R Tutorial, dostupné na internete (22.3.2018): https://www.datacamp.com/community/tutorials/k-means-clustering-r

[49] Harman, R.: *Cluster Analysis*, učebné texty, FMFI UK, Bratislava, 2011, dostupné na internete (22.3.2018): http://www.iam.fmph.uniba.sk/ospm/Harman/VSAclust.pdf

[50] R Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017, dostupné na internete (22.3.2018): `https://www.R-project.org/`

[51] *Wikipedia*, Cluster Analysis, dostupné na internete (22.3.2018): `https://en.wikipedia.org/wiki/Cluster_analysis`

[52] Murtagh, F., Legendre, P.: *Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?*, Journal of Classification 31 (2014), 274-295, dostupné na internete (23.3.2018): `https://link.springer.com/content/pdf/10.1007/s00357-014-9161-z.pdf`

[53] Barrat, A., Barthelemy, M., Pastor-Satorras, R., Vespignani, A.: *The architecture of complex weighted networks*, Proceedings of the National Academy of Sciences 101(11), 2004, 3747–3752

[54] Onnela, J.P., Saramäki, J., Kertész, J., Kaski, K.: *Intensity and coherence of motifs in weighted complex networks*, Physical Review E, 71(065103), 2005, 1–4

[55] Zhang, B., Horvath, S.: *A general framework for weighted gene co-expression network analysis*, Statistical Applications in Genetics and Molecular Biology 4, 2005, 17

[56] MCassey, M., Bijma, F.: *A clustering coefficient for complete weighted networks*, Network Science, 3(2), 2015, 183-195

[57] *Wikipedia*, Conditional independence, dostupné na internete (29.3.2018): `https://en.wikipedia.org/wiki/Conditional_independence`

[58] Friedman. J., Hastie, T., Tibshirani, R.: *Regularization Paths for Generalized Linear Models via Coordinate Descent*, Journal of Statistical Software, 33(1), 2010, 1-22, dostupné na internete (29.3.2018): `http://www.jstatsoft.org/v33/i01/`

[59] Hastie, T., Efron, B.,: *Least Angle Regression, Lasso and Forward Stagewise*, 2013, dostupné na internete (29.3.2018): `http://www-stat.stanford.edu/~hastie/Papers/#LARS`

[60] Friedman. J., Hastie, T., Tibshirani, R.: *glasso: Graphical lasso- estimation of Gaussian graphical models*, R package version 1.8., 2014, dostupné na internete (29.3.2018): `https://CRAN.R-project.org/package=glasso`

[61] Epskamp, S., Cramer, A., Waldorp, L., Schmittmann, V., Borsboom, V.: *qgraph: Network Visualizations of Relationships in Psychometric Data*, Journal of Statistical Software, 48(4), 2012, 1-18, dostupné na internete (29.3.2018) `http://www.jstatsoft.org/v48/i04/`

[62] Epskamp, S., Borsboom, D., Fried, E. I.: *Estimating Psychological Networks and Their Accuracy: A Tutorial Paper*, Behavior Research Methods, 2017, dostupné na internete (29.3.2018), `https://arxiv.org/abs/1604.08462`

[63] Revelle, W.: *psych: Procedures for Psychological, Psychometric, and Personality Research*, R package version 1.7.8, Northwestern University, Evanston, Illinois, USA, 2017, `https://CRAN.R-project.org/package=psych`

[64] *RDocumentation*, Exploratory Factor Analysis Using MinRes (Minimum Residual) As Well As EFA By Principal Axis, Weighted Least Squares Or Maximum Likelihood, dostupné na internete (3.4.2018): `https://www.rdocumentation.org/packages/psych/versions/1.7.8/topics/fa`

[65] *How bad is the Current Recession? Labour Market Downturns since the 1960s*, Ministry of Business, Innovation and Employment, 2014, dostupné na internete (9.4.2018): `https://web.archive.org/web/20141215000257/http://www.dol.govt.nz/publications/discussion-papers/current-recession/`

[66] Murphy, A.: *The 'Celtic Tiger' - An Analysis of Ireland's Economic Growth Performance*, European University Institute, Italy, 2000

[67] Hennessy, T., Kinsella, A.: *40 years of Irish farming since joining the European Union : a journey with the Teagasc National Farm Survey 1972 to 2012*, Rural Economy Research Centre, Teagasc, 2013

[68] *Visegrad Group*, About the Visegrad Group, dostupné na ineternete (28.4.2018): `http://www.visegradgroup.eu/about`