

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A
INFORMATIKY



ANALÝZA KORELÁCIE SPOTREBY
ZEMNÉHO PLYNU A SHENOVÝCH ZRÁŽOK

Diplomová práca

2013

Róbert Staniek

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY
A INFORMATIKY



ANALÝZA KORELÁCIE SPOTREBY
ZEMNÉHO PLYNU A SHENOVÝCH ZRÁŽOK

Diplomová práca

Študijný program:	Ekonomická a finančná matematika
Študijný odbor:	Aplikovaná matematika 1114
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci diplomovej práce:	Mgr. Matej Krušpán
Kód práce:	cbdd0e82-c525-42e9-bb70-4baedbeffe43

Bratislava, 2013

Bc. Róbert Staniek



Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Róbert Staniek
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský

Názov: Analýza korelácie spotreby zemného plynu a snehových zrážok

Cieľ: Cieľom práce je preskúmať závislosť spotreby zemného plynu od snehových zrážok v prostredí Slovenskej Republiky a v prípade potvrdenia závislosti navrhnúť zmenu stratégie pre uspokojenie zvýšeného dopytu po plyne zabezpečením dostatočného objemu plynu v zásobníkoch.

Vedúci: Mgr. Matej Krušpán

Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky

Dátum zadania: 25.01.2012

Dátum schválenia: 26.01.2012

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

.....
študent

.....
vedúci práce

ČESTNÉ PREHLÁSENIE

Čestne prehlasujem, že diplomovú prácu s názvom „Analýza korelácie spotreby zemného plynu a snehových zrážok“ som vypracoval samostatne, iba s použitím uvedenej literatúry a s pomocou vedúceho diplomovej práce. Svoje prehlásenie potvrdzujem podpisom.

V Bratislave, apríl 2013

.....
Róbert Staniek

POĎAKOVANIE

Chcel by som sa poďakovať vedúcemu diplomovej práce Mgr. Matejovi Krušpánovi za spoluprácu a vôbec príležitosť siahnuť si na dáta z praxe a analyzovať túto zaujímavú tému, ďalej Ing. Pavlovi Ondruškovi zo Slovenského hydrometeorologického ústavu SR za dôležité informácie z oblasti meteorológie a konzultácie k týmto dátam, Prof. Dr. Jarkovi Fidrmucovi za inšpiráciu pri konštrukcii modelov a v neposlednom rade mojej rodine za psychickú podporu a všetkým priateľom, ktorí mi boli oporou.

ABSTRAKT

STANIEK, R.: *Analýza korelácie spotreby zemného plynu a snehových zrážok.* [Diplomová práca] .– Univerzita Komenského v Bratislave. Fakulta matematiky, fyziky a informatiky: Katedra aplikovanej matematiky. – Vedúci práce: Mgr. Matej Krušpán, Bratislava: UK FMFI, 2013. 71 s.

Diplomová práca analyzuje vzťah medzi spotrebou zemného plynu a snehovými zrážkami. Na začiatku sa podrobne venuje téme meteorológie, nakoľko táto oblasť nie je triviálna a dáta, s ktorými sa pracuje majú rôznu povahu. Tieto informácie sa hneď aplikujú, už pri samotnej analýze dát, kde si treba dať pozor na niektoré rozdiely hlavne v časovej nekonzistencii dát. Následne sa na tieto dáta aplikujú rozmanité metódy. Tie sú popísané v teoretickej rovine a potom sú priamo aplikované na vybrané dáta. Konkrétne sa jedná o regresnú analýzu založenú na odhade metódou najmenších štvorcov, ďalej to sú autoregresné procesy a napokon panelová analýza s fixnými efektmi. Vždy sa dbá najmä o vysvetlenie vplyvu snehových zrážok v danom modeli, ktorý však nie je jediným vysvetľujúcim vplyvom.

Kľúčové slová: spotreba plynu, výška snehovej pokrývky, analýza dát, nesignifikantnosť

ABSTRACT

STANIEK, R.: *Analysis of correlation between natural gas consumption and snowfall.* [Diploma thesis] .- Comenius University in Bratislava. Faculty of Mathematics, Physics and Informatics: Department of Applied Mathematics and Statistics. - Supervisor: Mgr. Matej Krušpán, Bratislava: UK FMFI, 2013. 71 p.

This diploma thesis analyses a relationship between natural gas consumption and snowfall. At the beginning, it deals with meteorology in detail since this area is not trivial and the nature of the data is diverse. This information is applied in the analysis and one needs to beware of differences, mainly regarding the time inconsistency of the data. Afterwards, various methods are applied to the data. Detailed description of the methods is introduced first and then directly applied to the data. To be more precise, we use a regression analysis based on the least squares method, autoregressive models and finally the analysis of panel data with fixed effects. We always try to explain the influence of snowfall which is not the only one explanatory variable, however.

Key words: natural gas consumption, height of snow cover, data analysis, nonsignificance

Obsah

Úvod

1. Meteorológia a spotreba zemného plynu

1.1. Spotreba zemného plynu a dáta

1.2. Výber meteorologických staníc

1.3. Výber meteorologických charakteristík

1.3.1. Miestny vs. Stredoeurópsky čas v dátach (nekonzistentnosť)

1.3.2. Priemerná denná teplota

1.3.3. Celková výška snehovej pokrývky

1.3.4. Denná priemerná rýchlosť vetra

1.3.5. Denná dĺžka slnečného svitu

1.3.6. Denný úhrn zrážok

2. Dáta a ich štatistické vlastnosti

2.1. Úvodné úpravu dát a ich špecifiká

2.1.1. „Záhľadné“ skoky v spotrebe plynu

2.2. Deskriptívne štatistiky a vlastnosti dát

2.2.1. Denné spotreby plynu

2.2.2. Denné priemerné teploty

2.2.3. Celková snehová pokrývka

3. Teória a doterajšie poznatky

3.1. Lineárna regresia

3.1.1. Problém heteroskedasticity rezíduí

3.1.2. Problém autokorelácie rezíduí

3.1.3. Normalita rezíduí

3.1.4. Problém multikolinearity

3.1.5. Testovanie hypotéz o parametri

3.2. Časové rady a dynamické regresné modely

3.3. Panelová analýza

3.3.1. Model s fixnými efektmi

4. Praktická časť

4.1. Regresná analýza

4.2. Časové rady a dynamické regresné modely

4.3. Panelová analýza s fixnými efektmi

Záver

Zoznam použitej literatúry a internetových zdrojov

Úvod

Dodávky zemného plynu (ďalej len plyn) sa stali neodmysliteľnou súčasťou každodenného života fyzických ako i právnických osôb. Okrem výrobného sektora sa používa najmä na získanie tepelnej energie, použitej hlavne na vykurovanie bytových a domových jednotiek. Lenže na to, aby sme mohli túto tepelnú energiu využívať kontinuálne, musí sa zabezpečiť dostatočný objem plynu v systéme pre spotrebiteľskú sieť. O túto neľahkú úlohu sa stará množstvo odborníkov ako aj regulácia stanovujúca určité pravidlá a postupy.

Spotreba plynu (ďalej len spotreba) určená pre výrobné procesy je v priebehu roka málo volatilná. Takýto odberatelia dopredu vedia, koľko budú vyrábať, preto si u svojho dodávateľa objednávajú konkrétne množstvo plynu za výhodnejších podmienok a dodávateľ si nemusí robiť starosti s predikovaním budúcej spotreby. Na druhej strane existujú aj takí odberatelia, ktorých spotreba závisí predovšetkým od počasia. Pod týmito odberateľmi budeme myslieť teplárne, ktoré dodávajú teplú vodu do jednotlivých obytných panelákov. Koncoví užívatelia ju potrebujú pre vykurovanie a ohrev teplej pitnej vody. Kým sa však táto teplá voda dostane do paneláku, musí prejsť cestu potrubím od teplárne, čím dochádza k energetickým stratám.

V tejto diplomovej práci (ďalej len práci) sme sa pokúsili preskúmať vplyv snehovej pokrývky na spotrebu plynu. Sneh by totižto mohol poslúžiť ako izolant, ktorý by znížil tepelné straty pri distribúcii teplej vody potrubím pri silných mrazoch. Navyše by rovnako mohol izolovať aj strechy budov. Preto sa zameriame na spotrebu hlavne v zimnom období. Keďže na spotrebu vplyvajú aj iné meteorologické veličiny, budeme ich v našich modeloch takisto využívať. Nesporne sa jedná o teplotu, ale pozrieme sa aj na vplyv vetra a dĺžku slnečného svitu.

V prvej kapitole sme sa snažili vybrať vhodné dáta pre naše ďalšie analýzy. Ich výber spočíval v priradení čo najbližších meteorologických staníc k našim odberateľom, ktoré nám dalo k dispozícii SPP. Ďalej sme sa museli rozhodnúť pre konkrétne meteorologické charakteristiky (veličiny), ktorými budeme vysvetľovať spotrebu plynu. V tejto kapitole je dosť dopodrobna rozobratá aj téma meteorológie, nakoľko sme chceli ilustrovať jednotlivé detaily týkajúce sa zberu týchto dát, pretože sa domnievame, že často detaily o poznaní dát môžu napomôcť k správnejšej interpretácii výsledkov ekonometrických analýz.

V druhej kapitole sme sa zamerali na jednotlivé špeciálne prípady v dátach, ktoré bolo nevyhnutné si uvedomiť ešte pred začiatkom analyzovania modelov. Načrtli sme tu hneď aj možné riešenie problému. Dôležitým zistením bol fakt, že niektorí odberatelia typu tepláreň sa nesprávajú štandardne, t.j. podľa počasia. Preto sme si definovali odhad Pearsonovho korelačného koeficientu a na jeho základe sme vylúčili niektorých odberateľov z nášho skúmania. Túto kapitolu uzatvárame konkrétnymi príkladmi deskriptívnych štatistík vybraných lokalít, aby sme dali čitateľovi základnú intuíciu o povahe týchto dát.

Tretia kapitola je v rovine teoretického základu, pričom sme sa snažili zhrnúť niektoré základné matematické formuly a princípy. Je rozčlenená do troch častí, a to podľa prístupu k modelovaniu dát.

Štvrtá kapitola zachytáva praktickú časť práce, kedy sme odhadovali jednotlivé modely pomocou troch rôznych pohľadov na riešenie problému (členenie ako v kapitole 3). Táto kapitola má dosť dokumentačnú povahu, nakoľko v nej prikladáme množstvo výstupov z jednotlivých odhadnutých modelov zo štatistického programu EViews 5.0. Snažili sme sa popísať vývoj v našich modeloch, pričom sme skúšali obrovské množstvo rôznych modelov, z ktorých sme vybrali niekoľko zaujímavých výstupov.

Poslednou časťou je záver práce, kde sme sa snažili zhrnúť výsledky nášho úsilia.

1. Meteorológia a spotreba zemného plynu

Motivácia k napísaniu tejto kapitoly bola veľmi silná, nakoľko sa domnievam, že pri práci s dátami o ktorých človek nemá veľa informácií môže dôjsť k nesprávnym postupom. Navyše vyslovenie záverov na základe modelov, ktoré by nemuseli byť správne formulované kvôli nedostatočnej informácii týkajúcej sa použitých dát, by mohlo byť nazvané „šarlatánstvom“.

Tak napríklad obyčajná hodnota priemernej dennej teploty je na prvé zamyslenie niečo triviálne a zřejmé. Lenže keby som sa Vás opýtal na postup výpočtu tejto veličiny, tak odpoveď už nie je celkom jasná. Niekomu by sa zdalo logické odmerať hodnotu teploty každú celú hodinu behom jedného dňa (24 hodín) a urobiť aritmetický priemer. Alebo by to mohlo byť vypočítané na základe spojitého údajov cez integrál ako obsah pod krivkou za daný deň predelený dĺžkou intervalu. Ak si aj Vy myslíte, že sa to v praxi takto počíta, tak Vás asi sklámam, pretože sa to vypočíta celkom iným spôsobom. Ako? Správnu odpoveď Vám prezradím až za chvíľu, len čo uvediem niekoľko ďalších dôležitých informácií. A mimochodom, priemerná denná teplota je vypočítaná na základe miestneho času a nie stredoeurópskeho, čo spôsobí ďalšiu zmenu v jej ponímaní.

1.1. Spotreba zemného plynu a dáta

Táto práca vznikla na základe potreby z praxe (Slovenský plynárenský priemysel, a.s., ďalej len SPP), preskúmať vzťah medzi spotrebou plynu a počasím. Preto všetky dáta týkajúce sa spotreby plynu budú získané od tejto spoločnosti.

Ako sme už spomínali v úvode, svoju pozornosť upriamime na dáta týkajúce sa odberateľov plynu typu tepláreň. Tieto dáta majú charakter dennej bázy. To je to čo potrebujeme pre našu analýzu. Asi ťažko by sme analyzovali dáta mesačných alebo ročných spotrieb, ktoré by bolo možné získať hoci aj od jednotlivých odberateľov (domácnosti), keď počasie sa za takú dlhú dobu veľmi zmení a my nevieme jednoznačne povedať čo zapríčinilo nárast resp. pokles spotreby.

Ďalšou výhodou je, že poznáme presnú polohu daného odberateľa. Na rozdiel od toho, ako sa na dáta spotrieb vo väčšine analýz doteraz pozerali, t.j. spotreba a počasie pre celé Slovensko, my budeme pracovať s údajmi zodpovedajúcimi konkrétnej oblasti. Očakávame, že takto analyzované dáta presnejšie popíšu realitu.

Spotreba sa udáva v m³ za 24 hodín. Je meraná od 8:00 ráno do 8:00 rána nasledujúceho dňa. Takto získaná hodnota je zaznamenaná automaticky systémom, ktorý sa riadi stredoeurópskym časom (SEČ) a zodpovedá prvému dňu (realizovala sa v ňom väčšina spotreby).

Odberateľov tohto typu máme k dispozícii konečné množstvo a pre túto prácu sme získali dáta spotrieb pre odberateľov uvedených v nasledujúcej tabuľke (tab.01).

Tab.01 – Zoznam odberateľov plynu, zdroj: vlastný

Poradové číslo	Ozn. odberateľa SPP (posledné 3 číslice)	Obec	Okres	Kraj
1	019	Bratislava - Devínska Nová Ves	Bratislava IV	Bratislavský
2	925	Bratislava - Dúbravka	Bratislava IV	Bratislavský
3	160	Bratislava - Nové Mesto	Bratislava III	Bratislavský
4	320	Bratislava - Nové Mesto	Bratislava III	Bratislavský
5	194	Bratislava - Petržalka	Bratislava V	Bratislavský
6	219	Bratislava - Petržalka	Bratislava V	Bratislavský
7	220	Bratislava - Petržalka	Bratislava V	Bratislavský
8	276	Bratislava - Petržalka	Bratislava V	Bratislavský
9	640	Skalica	Skalica	Trnavský
10	936	Dubnica nad Váhom	Ilava	Trenčiansky
11	150	Považská Bystrica	Považská Bystrica	Trenčiansky
12	915	Považská Bystrica	Považská Bystrica	Trenčiansky
13	078	Púchov	Púchov	Trenčiansky
14	127	Trenčín	Trenčín	Trenčiansky
15	483	Komárno	Komárno	Nitriansky
16	776	Nitra	Nitra	Nitriansky
17	268	Nové Zámky	Nové Zámky	Nitriansky
18	316	Štúrovo	Nové Zámky	Nitriansky
19	203	Topoľčany	Topoľčany	Nitriansky
20	836	Martin	Martin	Žilinský
21	386	Nižná	Tvrdošín	Žilinský
22	206	Ružomberok	Ružomberok	Žilinský
23	025	Banská Bystrica	Banská Bystrica	Banskobystrický
24	803	Rimavská Sobota	Rimavská Sobota	Banskobystrický
25	314	Humenné	Humenné	Prešovský
26	294	Prešov	Prešov	Prešovský
27	323	Prešov	Prešov	Prešovský
28	825	Svit	Poprad	Prešovský
29	186	Košice - Juh	Košice IV	Košický

1.2. Výber meteorologických staníc

Odberateľov plynu a ich lokalizáciu už poznáme. Teraz by sme radi získali potrebné meteorologické charakteristiky, ktoré sa viažu k týmto miestam. Slovenský hydrometeorologický ústav (SHMÚ) má na svojej internetovej stránke¹ prehľad o staničnej sieti, kde sa nachádza aj zoznam meteorologických staníc (ďalej len staníc) s ich presnou polohou. [2]

Tuto narazíme na prvý menší problém, a to je existencia viacerých typov staníc. Delia sa na klimatologické, zrážkomerné a fenologické. Fenologické stanice nás nebudú zaujímať, keďže fenológia je odbor, ktorý sa zaoberá vplyvom klímy a zmien počasia v jednotlivých ročných obdobiach na životné prejavy rastlín a živočíchov. My sa budeme sústrediť na sieť klimatologických a zrážkomerných staníc. Kým sieť zrážkomerných staníc je naozaj veľmi hustá (viď. obr. 01), sieť klimatologických staníc je výrazne redšia (viď. obr. 02). Je to logické, nakoľko klimatologická stanica má oproti zrážkomernej oveľa viac prístrojov a náklady na jej zariadenie a prevádzku sú mnohonásobne vyššie. V podstate sa na zrážkomerné stanice môžeme pozrieť ako na podmnožinu klimatologických staníc z hľadiska dostupných údajov. Z hľadiska množstva staníc je to presne naopak.

Tento fakt má za následok to, že sme nútení zamerať svoju pozornosť na zoznam klimatologických staníc, ktoré poskytujú okrem informácií o zrážkach aj mnoho iných veličín (teplota, vietor, ...). Ak by sme chceli regresne preskúmať vplyv nejakej meteorologickej premennej na spotrebu plynu, budeme potrebovať aj hodnoty priemernej dennej teploty, nakoľko táto je logicky rozhodujúca. To bolo potvrdené aj v diplomovej práci² na rovnakej katedre pred 2 rokmi. Priradenie klimatologickej stanice, ktorej poloha je priamo v obci alebo jej tesnej blízkosti (10 km) vybraného odberateľa plynu je veľmi ľahká. Otázkou však ostáva, akú stanicu treba priradiť k odberateľovi v ktorého tesnej blízkosti sa nenachádza žiadna stanica. Rozhodli sme sa pozrieť na najbližšiu možnú. Tu by sa dalo vymyslieť taktiež niečo sofistikovanejšie, napr. že by sme zobrali 2 až 3 stanice dosť blízke k odberateľovi, pričom by sme mohli priradiť týmto staniciam konkrétne váhy podľa ich blízkosti, čím by sme získali hodnoty virtuálnej stanice pre daného odberateľa. Informácie o nami priradených staniaciach nájdete v nasledovnej tabuľke (tab.02). V tabuľke 03 je samostatný zoznam meteorologických staníc, pre ktoré máme k dispozícii

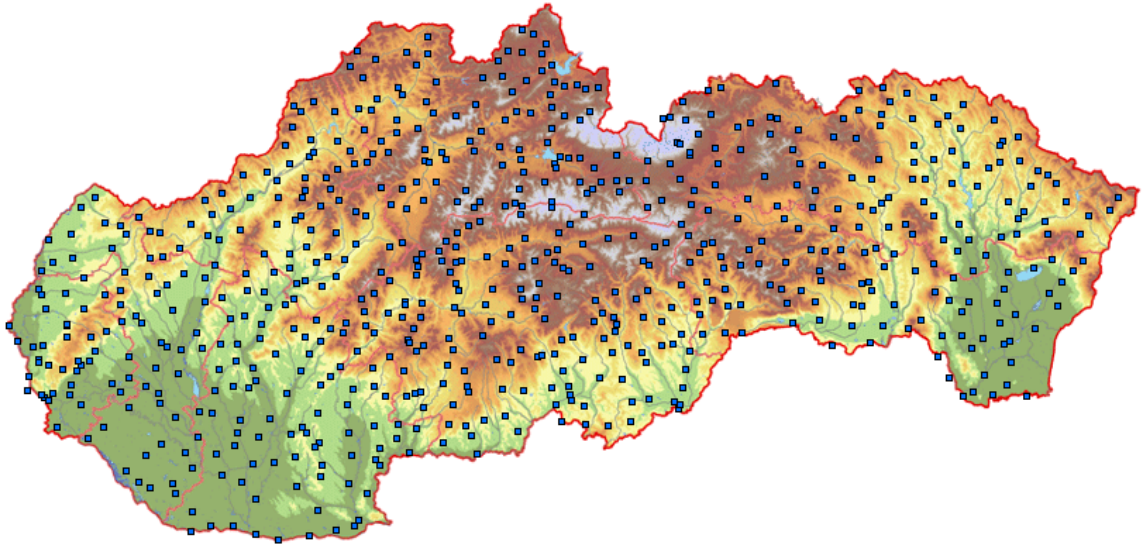
¹ <http://www.shmu.sk/sk/?page=1603>

² Spracovanie predikčných modelov pre prognózy spotreby zemného plynu v predajnom portfóliu SPP na slovenskom území - Zdenka Zubáčová, 2011

vybrané meteorologické charakteristiky. Ich zobrazenie je na obr. 03.

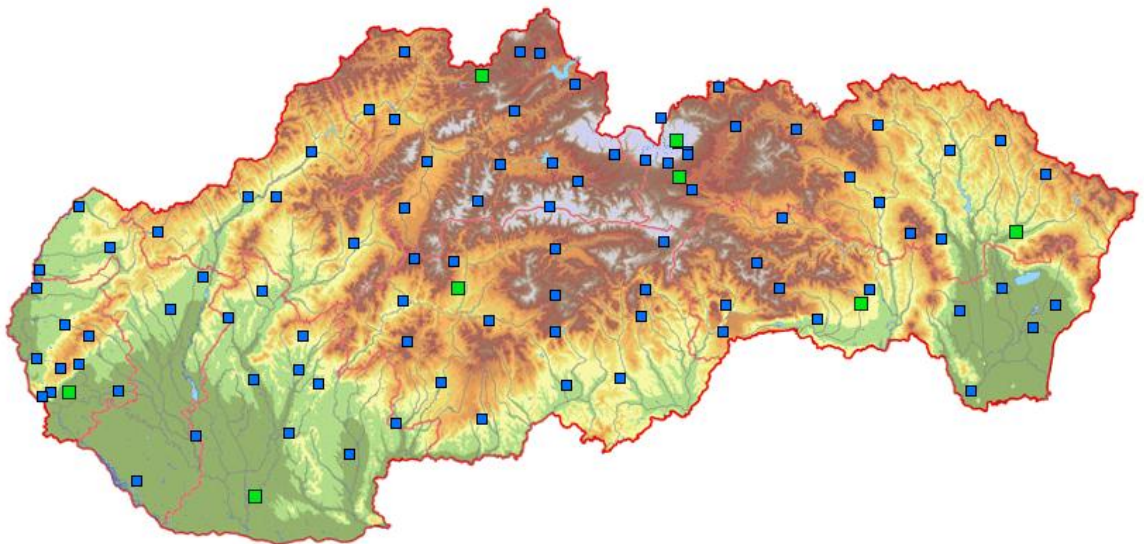
Obr.01 – Zrážkomerná sieť, zdroj: <http://www.shmu.sk/sk/?page=1605>

Zrážkomerná sieť



Obr.02 – Klimatologická sieť, zdroj: <http://www.shmu.sk/sk/?page=1604>

Klimatologická sieť



Tab.02 – Priradenie klimatologickej stanice k odberateľovi, zdroj: vlastný

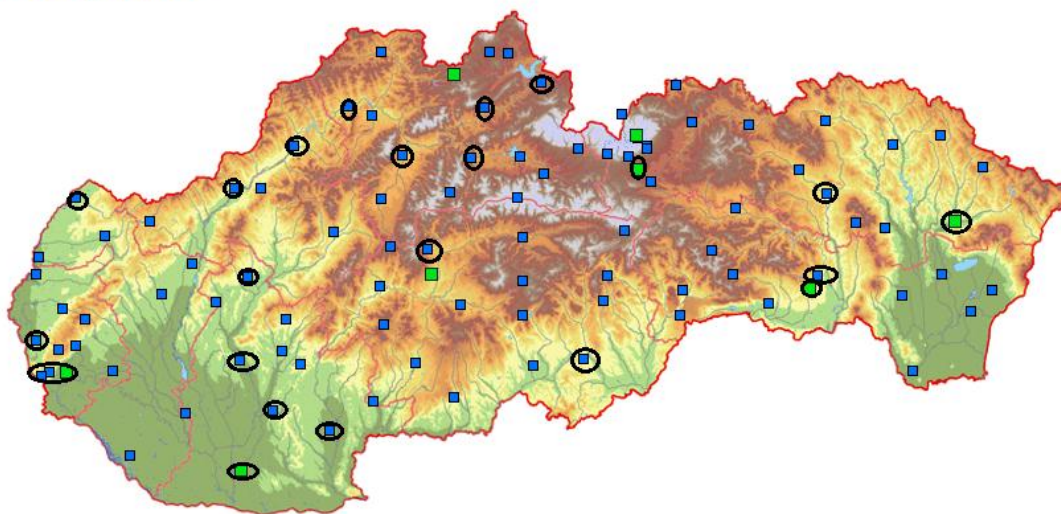
Odberatelia SPP			PRIRADENIE KLIM. STANICE (SHMÚ)		
obec	okres	kraj	číslo kraja	shmú klimatologická stanica 1	shmú klimatologická stanica 2
Bratislava - Dev. Nová Ves	Bratislava IV	Bratislavský	1	Stupava	NIE
Bratislava - Dúbravka	Bratislava IV	Bratislavský	1	BA - Mlynská dolina	NIE
Bratislava - Nové Mesto	Bratislava III	Bratislavský	1	BA - Koliba	BA - Letisko
Bratislava - Nové Mesto	Bratislava III	Bratislavský	1	BA - Koliba	BA - Letisko
Bratislava - Petržalka	Bratislava V	Bratislavský	1	BA - Mlynská dolina	BA - Letisko
Bratislava - Petržalka	Bratislava V	Bratislavský	1	BA - Mlynská dolina	BA - Letisko
Bratislava - Petržalka	Bratislava V	Bratislavský	1	BA - Mlynská dolina	BA - Letisko
Bratislava - Petržalka	Bratislava V	Bratislavský	1	BA - Mlynská dolina	BA - Letisko
Skalica	Skalica	Trnavský	2	Holíč	NIE
Dubnica nad Váhom	Ilava	Trenčianský	3	Trenčín	NIE
Považská Bystrica	Považská Bystrica	Trenčianský	3	Beluša	Dolný Hričov
Považská Bystrica	Považská Bystrica	Trenčianský	3	Beluša	Dolný Hričov
Púchov	Púchov	Trenčianský	3	Beluša	NIE
Trenčín	Trenčín	Trenčianský	3	Trenčín	NIE
Komárno	Komárno	Nitrianský	4	Hurbanovo	NIE
Nitra	Nitra	Nitrianský	4	Nitra - Velké Janíkovce	NIE
Nové Zámky	Nové Zámky	Nitrianský	4	Hurbanovo	Podhájska
Štúrovo	Nové Zámky	Nitrianský	4	Želiezovce	Hurbanovo
Topoľčany	Topoľčany	Nitrianský	4	Topoľčany	NIE
Martin	Martin	Žilinský	5	Martin MS	NIE
Nižná	Tvrdošín	Žilinský	5	Oravský Podzámok	Liesek
Ružomberok	Ružomberok	Žilinský	5	Ružomberok	NIE
Banská Bystrica	Banská Bystrica	Banskobystrický	6	Banská Bystrica - Zelená	NIE
Rimavská Sobota	Rimavská Sobota	Banskobystrický	6	Rimavská Sobota	NIE
Humenné	Humenné	Prešovský	7	Kamenica nad Cir.	NIE
Prešov	Prešov	Prešovský	7	Prešov - vojsko	NIE
Prešov	Prešov	Prešovský	7	Prešov - vojsko	NIE
Svit	Poprad	Prešovský	7	Poprad	NIE
Košice - Juh	Košice IV	Košický	8	Košice - Mesto	Košice - Letisko

Tab.03 – Zoznam meteorologických staníc (pre ktoré máme dáta vybraných meteorologických charakteristík) , zdroj: vlastný

Poradové číslo	Obec	Okres	Kraj
1	Bratislava - Kolyba	Bratislava	Bratislavský
2	Bratislava - letisko	Bratislava	Bratislavský
3	Bratislava - Mlynská dolina	Bratislava	Bratislavský
4	Stupava	Malacky	Bratislavský
5	Holíč	Skalica	Trnavský
6	Trenčín	Trenčín	Trenčiansky
7	Beluša	Púchov	Trenčiansky
8	Nitra - Veľké Janíkovce	Nitra	Nitriansky
9	Topoľčany	Topoľčany	Nitriansky
10	Hurbanovo	Komárno	Nitriansky
11	Podhájska	Nové Zámky	Nitriansky
12	Želiezovce	Levice	Nitriansky
13	Dolný Hričov	Žilina	Žilinský
14	Ružomberok	Ružomberok	Žilinský
15	Oravský Podzámok	Dolný Kubín	Žilinský
16	Liesek	Tvrdošín	Žilinský
17	Martin MS	Martin	Žilinský
18	Rimavská Sobota	Rimavská Sobota	Banskobystrický
19	Banská Bystrica - Zelená	Banská Bystrica	Banskobystrický
20	Prešov - vojsko	Prešov	Prešovský
21	Poprad	Poprad	Prešovský
22	Kamenica nad Cirochou	Humenné	Prešovský
23	Košice - Mesto	Košice	Košický
24	Košice - Letisko	Košice	Košický

Obr.03 – Krúžkami označené nami vybrané klimatologické stanice, zdroj: SHMÚ a naša úprava (krúžky)

Klimatologická sieť



1.3. Výber meteorologických charakteristík

Meteorologické dáta sú väčšinou zaznamenávané priebežne počas celého dňa, takže požiadavka na dáta denného charakteru je bezproblémová.

Čo sa týka priestoru, resp. lokalizácie, túto otázku sme vyriešili taktiež na predchádzajúcich stranách. Ešte raz však podotýkam že, ak už máme analyzovať spotrebu od počasia, tak by bolo dobré mať veľmi presné informácie o počasi, kde sa daný odberateľ nachádza. Lebo vychádzať z meteorologickej situácie charakterizujúcej celé územie Slovenskej Republiky by bolo veľmi všeobecné. Ved' z praxe vieme, že jednotlivé regióny, hoci aj takej malej krajiny ako je Slovensko, sú veľmi odlišné z pohľadu meteorologickej charakterizácie. Preto sme priradili k danému odberateľovi konkrétnu stanicu.

Teraz nám ostáva za úlohu vybrať tie správne meteorologické charakteristiky, ktoré sú nevyhnutné pre naše analýzy (teplota, sneh) ako aj tie, ktoré by mohli byť istým spôsobom modely zlepšujúce. Tie sme vyberali z ponuky SHMÚ. [2]

Samotná práca na meteorologickej stanici spočíva v neustálom pozorovaní počasia. Meteorologické merania sprevádzané okamžitými pozorovaniami predpísaných prvkov (charakteristík / veličín) a javov sa však konajú v určitých termínoch, keďže ich vykonanie napr. uprostred noci by bolo dosť nákladné (málokto by bol ochotný pracovať na stanici aj v noci). Tieto termíny sú o 7:00, 14:00 a 21:00 miestneho času. Aby meteorologické charakteristiky, ktorých okamžité hodnoty sú závislé na polohe slnka (majú tzv. denný chod) boli v celej staničnej sieti dobre porovnateľné, je nevyhnutné používať miestny čas. Takže pozorovania sú prevádzané za určitej, pritom však na všetkých staniaciach rovnakej polohe slnka nad obzorom vzhľadom k pozorovaciemu miestu. [1]

1.3.1. Miestny vs. Stredoeurópsky čas v dátach (nekonzistentnosť)

Miestny čas je daný miestnym poludníkom, pričom sa odvíja od času platného pre isté časové pásmo. Slovensko sa nachádza v stredoeurópskom časovom pásme, ktorého teoretickým stredom prechádza 15. poludník východnej zemepisnej dĺžky (v.z.d.), kde je o 1 hodinu viac než v pásme základného (nultého / Greenwichského) poludníka. Keďže obvod Zeme má 360° a deň má 24 hodín, na jedno pásmo tak teoreticky pripadá 15° . Na miestach 15.-ho poludníka v.z.d sa miestny čas rovná presne stredoeurópskemu času (SEČ), kým na miestach východne od tohto poludníka je skutočná poloha slnka na obzore západnejšie a preto je miestny čas väčší. Konkrétne je to o 4 minúty na 1 dĺžkový stupeň.

V praxi to znamená, že stanice ležiace východne od tohto 15.-ho poludníka konajú termínové pozorovania skôr, než udáva SEČ. Napr. stanica ležiaca na 17. ° v.z.d. (okres Bratislava) koná termínové merania o 6:52, 13:52 a 20:52 SEČ a stanica ležiaca na 22. ° v.z.d. (okres Michalovce) koná termínové merania o 6:32, 13:32 a 20:32 SEČ. Rozdiel v miestnom čase medzi východnou a západnou časťou Slovenska je teda približne 20 minút. [1]

Miestny čas sme nedefinovali len preto, aby sme chápali pojmu termínové meranie, ale chceme upozorniť na to, že by mohlo byť zaujímavé v neskoršej analýze rozdeliť dáta z jednotlivých staníc do viacerých oblastí, napr. západné, stredné a východné Slovensko. To by však bolo opodstatnené najmä vtedy, kedy sme mali dáta za spotrebu plynu a napr. teploty v hodinových intervaloch. Vtedy by 20 minútový rozdiel v čase mohol hrať významnú úlohu, nakoľko dáta za spotrebu plynu podliehajú SEČ. Teoreticky keby mám hodnoty spotreby plynu dvoch rovnako veľkých odberateľov za hodinový interval od 7:00 do 8:00 SEČ, pričom poznám aj priemerné teploty avšak v miestnom čase, a prvý odberateľ je z miesta na 17. ° v.z.d. a druhý na 22. ° v.z.d., tak tieto priemerné teploty by zodpovedali SEČ 6:52-7:52 a 6:32-7:32. To by znamenalo, že by reálne iba 52/60 t.j. 86,7 % z intervalovej hodnoty spotreby prvého odberateľa prislúchalo k údaju zodpovedajúcej intervalovej priemernej teplote. V prípade druhého odberateľa by to bolo iba 32/60 t.j. 53,3 % z intervalovej hodnoty spotreby. Teda údaje týkajúce sa počasia by neboli konzistentné k údajom spotreby plynu.

Ak budeme uvažovať 24 hodinové časové intervaly (ktoré máme my), pričom dáta charakterizujúce počasia sú uvedené v miestnom čase a dáta o spotrebe plynu v SEČ, tak by táto časová nezrovnalosť predstavovala v prvom prípade iba 1432/1440 t.j. 99,4 % a v druhom prípade 1412/1440 t.j. 98,1 %. Toto sú už naozaj nepatrné rozdiely a preto o časovej nekonzistencii v zmysle SEČ a miestneho času nebudeme ďalej uvažovať. V realite sa aj tak tieto rozdiely vyrušia tým spôsobom, že na západe Slovenska bude spotreba o 7:00 SEČ relatívne vyššia než na východe Slovenska (tam svieti dlhšie slnko), ale pri spotrebe o 21:00 SEČ to bude presne naopak. Ďalej si to môžeme dovoliť urobiť aj preto, že spotreba má istý druh zotrvačnosti v tom zmysle, že ak aj nastane práve teraz zmena počasia, potrvá to niekoľko minút resp. hodín, kým na túto zmenu zareaguje spotrebiteľ a zmení svoju spotrebu.

Ak budeme ďalej v práci pojednávať o časovej nekonzistencii, budeme mať na mysli nekonzistentnosť v zmysle výpočtu priemernej dennej teploty a uzatvárania spotreby plynu. Keďže priemerná teplota je vychádzajúca z meteorologických termínov 7:00 ,14:00

a 21:00, teploty ktoré budú nasledovať po 21:00 už nie sú zahrnuté v údajoch o priemernej teplote ale v skutočnosti začínajú ovplyvňovať dennú spotrebu plynu uzatváranú o 8:00 ráno nasledujúceho dňa. Lenže ako sme už vyššie spomenuli, že spotreba reaguje na zmeny počasia s istým oneskorením, je otázne či sa táto časová nekonzistentnosť stihne prejavíť. Ale keďže sa jedná o rozdiel až 11 hodín, predpokladáme že bude potrebné túto nekonzistentnosť prešetriť. V praxi to bude znamenať to, že k danému dňu spotreby priradíme priemernú dennú teplotu získanú vážením teploty daného a nasledujúceho dňa. Napr.: $\frac{2}{3} * \text{teplota dnes} + \frac{1}{3} * \text{teplota zajtra} = \text{nová dnešná teplota}$.

1.3.2. Priemerná denná teplota

Priemerná denná teplota (ďalej len teplota) je nesporne najvýznamnejší regresor zo všetkých meteorologických charakteristík vplyvujúcich na spotrebu plynu. Ostatné regresory ako napr. vietor či snehová pokrývka len tlmia alebo umocňujú jej efekt. Keďže je to priemerná hodnota, vznikla na základe nejakej konkrétnej formuly. Jej hodnota je vypočítaná ako vážený priemer teplôt termínových meraní (všetky udávané v stupňoch Celzia, zn.: °C), pričom váha v termíne 7:00 a 14:00 je 1 a váha pre termíne 21:00 je 2. Preto takto vážená suma 3 hodnôt je delená 4. [1], [3]

$$\text{tepl} = \frac{t_7 + t_{14} + 2 \cdot t_{21}}{4}$$

tepl ... priemerná denná hodnota teploty

t_7 ... termínová hodnota teploty meranej o 7:00 miestneho času

t_{14} ... termínová hodnota teploty meranej o 14:00 miestneho času

t_{21} ... termínová hodnota teploty meranej o 21:00 miestneho času

1.3.3. Celková výška snehovej pokrývky

Táto premenná je z hľadiska tejto práce kľúčová, keďže práca má za cieľ preskúmať vplyv tejto meteorologickej charakteristiky na spotrebu. Okrem tejto veličiny boli k dispozícii údaje aj o novej snehovej pokrývke, t.j. iba snehu napadnutom za každých 24 hodín od 7:00 predošlého dňa do 7:00 dnešného dňa miestneho času, alebo v jednotlivých intervaloch medzi termínovými pozorovaniami. My sme sa logicky rozhodli pre celkovú výšku snehovej pokrývky, nakoľko keby nenapadol nový sneh v danom dni a napriek tomu by existovala snehová pokrývka z predchádzajúcich dní, neodzrkadľovalo by to skutočnosť, resp. reálny vplyv na spotrebu. A navyše z tejto veličiny si vieme

späťne vypočítať aká je denná zmena v jej výške. Tieto hodnoty sa udávajú v centimetroch (zn.: cm). Vždy to sú celé čísla, teda sú štandardne zaokrúhlené na celé cm. [1], [3]

Výška celkovej snehovej pokrývky sa meria snehomernou tyčou. Poznáme 2 základné typy: prenosné a neprenosné. Štandardne majú dĺžku 1m. Prenosnými tyčami meriame na miestach s nestálou a nízkou snehovou pokrývkou (cca. do 400m nadmorskej výšky). Keďže sneh neleží v takýchto staniaciach väčšinou rovnomerne, meranie sa uskutočňuje viac krát (min. 3x) a to na miestach kde nie je naviaty alebo odviaty. Za výsledný údaj sa berie priemer z týchto meraní. Meranie sa uskutočňuje ak snehová pokrývka pokrýva aspoň polovicu územia v okolí stanice a to vrstvou snehu vysokej aspoň 1cm. Tu sa však uplatňuje zaokrúhľovanie, t.j. výška 0,5 až 0,9 sa zaokrúhli nahor na 1cm. Inak sa charakterizuje špeciálnymi kódmi typu nesúvislá snehová pokrývka alebo tzv. poprašok (vysvetlené v 2.kap.). [1]

1.3.4. Denná priemerná rýchlosť vetra

Veľmi zaujímavou a podľa nás dôležitou premennou by mohol byť aj vietor, nakoľko jeho pôsobením by mohlo dôjsť k výraznejšiemu vplyvu vonkajšej teploty na teplotu v budovách. V podstate vietor tu zohráva úlohu akéhosi urýchľovača tepelnej výmeny, keďže pri napr. mínusových teplotách spôsobuje intenzívnejšie ochladzovanie vonkajších múrov budov. K dispozícii sú údaje o smere vetra ako aj jeho rýchlosti v známych termínoch, avšak pre nás zaujímavou bude hodnota dennej priemernej rýchlosti vetra. Tá je však na rozdiel od dennej priemernej teploty vypočítaná ako rovnovážny priemer hodnôt v termínoch 7:00, 14:00 a 21:00.

$$viet = \frac{v_7 + v_{14} + v_{21}}{3}$$

viet ... denná priemerná rýchlosť vetra

v_7 ... termínová hodnota rýchlosti vetra meranej o 7:00 miestneho času

v_{14} ... termínová hodnota rýchlosti vetra meranej o 14:00 miestneho času

v_{21} ... termínová hodnota rýchlosti vetra meranej o 21:00 miestneho času

Pre meranie priemernej rýchlosti vetra sa používa anemometer. Meranie sa zakladá na účinku vetra na ľahké misky, pričom vietor pôsobí na dutú stranu misiek väčším tlakom, než na vypuklú. Misky sú pochopiteľne upevnené a to na tzv. Robinzonový kríž, ktorý následne začne rotovať. Rýchlosť pohybu misiek je v určitom vzťahu k rýchlosti vetra. Toto meranie je nezávislé na smere vetra. Samotné meranie je vykonávané na

voľnom priestranstve vo výške 2-3 m nad zemou po dobu 1-2 min. , kedy intenzita vetra kolíše a preto sa berie do úvahy priemerný údaj. [1], [3]

1.3.5. Denná dĺžka slnečného svitu

Ďalšou premennou, ktorú by bolo zaujímavé zahrnúť do modelu je dĺžka slnečného svitu. Motivácia k preskúmaniu aj tejto premennej je z dôvodu vylepšenia modelov a následne predikcií spotreby, i keď táto premenná nie je priamym cieľom analyzovania. Myslíme si, že by mohla priamo pôsobiť na teplotu v budovách, či už zohrievaním ich povrchu alebo prenikaním priamo cez okná a následne zohrievať interiér budovy.

Domnievame sa, že je z časti prepojená alebo zachytená v údajoch o priemernej dennej teplote, keďže slnečné lúče zohrievajú zemský povrch a následne teplota narastá. V modeli regresie by to mohlo spôsobiť tzv. multikolinearitu, ktorej sa budeme venovať v príslušnej kapitole.

Prístroj na zaznamenávanie dĺžky slnečného svitu sa nazýva slnkomer. Princíp na ktorom funguje je založený na prechode slnečných lúčov cez špecifickú sklenenú guľu za ktorou je pripevnená špeciálna papierová páska, do ktorej sa následne pôsobením týchto lúčov vypaľuje stopa (princíp ako pri lupe). Ak slnko svieti nepretržite, táto stopa je spojitá, v opačnom prípade je prerušovaná. Ak je intenzita slnečného žiarenia veľká, táto páska sa dokonca prepáli. Ak je naopak slabá, na páske je viditeľná stopa bledej farby. Intenzita slnečného žiarenia sa však nezaznamenáva, preto o nej nevieme nič povedať z informácie o dĺžke slnečného svitu.

Pri vyčísľovaní záznamov nás zaujíma dĺžka stopy, zanechanej na papierovej páske slnečnými lúčmi. Páska je delená čiarami nielen po celých hodinách, ale aj po pol hodinách. Pre odhad trvania svitu sa započítajú aj najslabšie stopy, nakoľko sa dajú na páske zistiť. Súčet dĺžok všetkých stôp, spôsobených na papieri slnečným žiarením medzi susednými hodinovými čiarami, pozorovateľ zapíše celým číslom. Napr. 7 znamená, že v dotyčnej hodine záznam všetkých stôp slnečného svitu spolu činil 7/10 vzdialenosti medzi hodinovými čiarami. Teda slnko svietilo približne 42/60 hodiny, čo je 42 minút. Napr. 10 znamená nepretržitý svit po celú hodinu. [1], [3]

1.3.6. Denný úhrn zrážok

Poslednou premennou o ktorú sme prejavili záujem je denný úhrn zrážok, ktorý je meraný od 7:00 rána daného dňa do 7:00 rána nasledujúceho dňa. Je udávaný

v milimetroch (zn.: mm). Táto premenná má skôr kontrolný charakter, aby sme vedeli lepšie charakterizovať priebeh počasia v daný deň a nemyslíme si, že by mohla pomôcť zlepšiť naše modely, i keď vplyv teploty je rôzny pri inej vlhkosti vzduchu. Keďže však veličinu charakterizujúcu vlhkosť vzduchu nemáme (tento nápad vznikol až po zrealizovaní objednávky na meteorologické dáta), nie sme celkom istý do akej miery určuje množstvo zrážok vlhkosť vzduchu. Toto by bolo návrhom pre budúcu prácu. [3]

2. Dáta a ich štatistické vlastnosti

V úvode tejto kapitoly by som rád poznamenal faktickú poznámku, že získanie meteorologických dát nebolo veľmi jednoduché a rozhodne ani rýchle (niekoľko mesiacov). Keďže sa jedná o diplomovú prácu (vedecké účely), SHMÚ by malo tieto údaje poskytnúť bezodplatne vzhľadom na právnu legislatívu SR. To sa aj stalo, keď sme podali žiadosť na daný objem dát, len sme však museli ešte splniť podmienku, že dáta musia byť staršie ako 3 roky. V čase vybavovania týchto dát, t.j. začiatkom roku 2012 sme mohli požiadať o dáta do konca roku 2008. Lenže aby sme v nich mali obsiahnuté obdobie aspoň 2 zím, požiadali sme o dáta za obdobie december 2006 – december 2008. Teda v dátach máme 2 celé zimy, 2006/2007 a 2007/2008.

2.1. Úvodné úpravy dát a ich špecifiká

Po doručení dát sme ich začali analyzovať. Prázdne bunky bolo potrebné nahradiť nulami, bunky s pomlčkami znamenali nedostupnosť daných údajov (napr. porucha zariadenia) a v prípade snehovej pokrývky sme sa museli rozhodnúť ako správne nahradiť špeciálne kódy. Tieto kódy sú 995 a 999. Kód 995 znamená nesúvislú snehovú pokrývku. Podľa zdroja [1] to nastáva vtedy, ak vrstva snehu vysoká najmenej 1cm (zaokrúhlená) pokrýva menej ako polovicu plochy staničného okolia (oblasť, ktorú vidí z merného miesta). Inak povedané ak na niektorých miestach máme výšku snehu niekoľko cm (aj 1cm stačí), ale na väčšine plochy nie je žiadny resp. max. 0,4 cm. Ďalší kód 999 znamená tzv. poprašok. Znamená to výšku snehovej pokrývky menej ako 0,5 cm (na celej ploche). Preto by sme mohli výšku snehovej pokrývky vzostupne usporiadať takto:

$0\text{cm} < 999 < 995 < 1\text{cm} < 2\text{cm} < \dots \text{cm}.$

Pri kóde 999 sme sa rozhodli takéto údaje nahradiť nulou. Pri kóde 995 sme sa z počiatku rozhodli takisto nahradiť nulou tieto bunky, ale neskôr sme ho skúsili charakterizovať ako vrstvu 1cm. Ale keďže sa jedná o veľmi nízku vrstvu snehu, je otázne či sa už aj tu prejavujú nejaké účinky snehu na spotrebu.

Pri veličine o snehovej pokrývke sme v dátach narazili na jednu podstatnú záležitosť. Jedná sa o časový posun, resp. zmenu dátumu. Pre ilustráciu sa pozrime na nasledovnú tabuľku, ktorá posluží ako názorná ukážka na reálnych dátach.

Tab. 04 – Zrážky a snehová pokrývka, zdroj: tab. - vlastný, dáta – SHMÚ

SHMÚ - Klimatologická stanica	dátum	teplota	zrážky	sneh
BB-Zelená	1.3.2007	3,8	6,2	8
BB-Zelená	2.3.2007	5,7	0,4	4
BB-Zelená	3.3.2007	1	12,6	0
BB-Zelená	4.3.2007	3,2	0	3
BB-Zelená	5.3.2007	2	0	0
BB-Zelená	6.3.2007	6,8	0	0

BB-Zelená ... Banská Bystrica-Zelená

Teplota ... denná priemerná teplota

Zrážky ... denný úhrn zrážok, od 7:00 do 7:00

Sneh ... výška celkovej snehovej pokrývky o 7:00

Na prvý pohľad sa zdá, že v tabuľke musí existovať nejaká chyba, lebo z riadku 3.3.2007 na 4.3.2007 pribudol sneh, ale pritom 4.3.2007 nie sú vykázané žiadne zrážky. Taktiež dosť zvláštne je aj to, že pribudol sneh, hoci teploty sú stále kladné.

Vysvetlenie:

Všimnime si, že 1.3.2007 sme na tejto meteorologickej stanici zaznamenali 8cm snehu. Tento stav bol spozorovaný v daný deň v termíne 7:00. Keďže ale bola priemerná denná teplota +3,8 °C, dá sa očakávať, že na druhý deň bude nižšia snehová pokrývka. To sa aj naozaj stalo, lebo jej výška poklesla z 8cm na 4cm. Takisto aj 2.3.2007 bola priemerná teplota kladná a to až +5,7 °C, takže sa dá očakávať, že 3.3.2007 bude opäť menej snehu. A naozaj už nebol žiadny sneh. Ďalej 3.3.2007 poklesla priemerná teplota na +1°C a denný úhrn zrážok bol 12,6 mm. Musíme si uvedomiť, že nevieme jednoznačne vylúčiť ten fakt, že v tento deň mohlo mrznúť. Pretože síce **priemerná** denná teplota vyšla kladná, ale hociktorá jej **termínová** zložka mohla byť záporná, za predpokladu že aspoň jedna bola kladná. Podľa toho priebehu teplôt by sme to najskôr tipovali na teplotu v termíne 21:00. Ale aj keby všetky 3 termínové teploty boli kladné, nevieme nič povedať o teplote, ktorá bola po 21:00 do 7:00 nasledujúceho dňa. Takže v noci z 3.3.2007 na 4.3.2007 mohlo mrznúť, ale z priebehu priemerných denných teplôt to nemáme šancu zistiť. Hodnota zrážok pripísaná ku dňu 3.3.2007 mohla byť podobne zrealizovaná celá behom noci a 4.3.2007 o 7:00 sa hodnota uzavrela na výsledných 12,6 mm zrážok. Ale presne v tom istom termíne, t.j. 4.3.2007 o 7:00 sme odmerali celkovú výšku snehovej pokrývky (3 cm) a pripísali ju ale k dátumu 4.3.2007, ako stav pre tento deň. Otázkou teda ostáva, že by nebolo lepšie celý stĺpec hodnôt výšky snehovej pokrývky posunúť o 1 riadok nahor, t.j.

stavom z daného dňa popísať stav resp. priebeh dňa minulého. Rozhodli sme sa, že tento posun v dátumoch nebudeme realizovať, nakoľko vplyv snehovej pokrývky by sa rozhodne neprejavil hneď a navyše tento efekt očakávame pri vyššej vrstve snehu a dlhodobom trvaní, t.j. krátkodobý sneh a (alebo) s nízkou výškou nebude mať izolačné vlastnosti.

2.1.1. „Záhľadné“ skoky v spotrebe plynu

Ďalším dôležitým zistením, tento krát pri dátach o spotrebe plynu, je fakt, že v spotrebe sa nachádzajú aj dosť výrazne skoky. Je zrejmé, že sa nejedná o nárast spotreby spôsobený zmenou počasia, keďže tieto regresory sa v daných skokoch výrazne nemenili. Jedná sa preto o z nášho pohľadu „záhadné“ skoky v spotrebe. Ich príčiny tu nebudeme veľmi rozoberať, ale musíme na tento fakt neustále myslieť pri kontrole správnosti našich modelov. Aby sme boli korektní, uvidíme príklad na vysvetlenie týchto skokov v spotrebe. Významný poznatok je ten, že niektorí odberatelia typu tepláreň sú v tesnej blízkosti a pracujú systémom vzájomného korigovania výroby teplej vody. Jeden z nich zohráva úlohu viac menej konštantnej výroby a teda aj spotreby v danom období, kým druhý uspokojuje výkyvy v dopyte po teplej vode. Sčítaním ich spotrieb dostávame reálnejší obraz o vplyve počasia na spotrebu.

V praxi sa takéto skoky v spotrebe stratia, lebo predikcie sú robené pre celé Slovensko a nie jednotlivých odberateľov, ktorých spotreba voči celoslovenskej spotrebe tvorí väčšinou menej ako 1%. Lenže v prípade práce s dátami za jediného odberateľa budeme musieť tieto skoky nejako odstrániť. Samozrejme budeme pracovať s takými odberateľmi, ktorých spotreba závisí v čo najväčšej miere na počasi a v prípade, že aj tu objavíme dosť výrazne skoky v spotrebe (rádovo nad 10%), tieto skoky nahradíme priemerom spotrieb deň pred a deň po tomto skoku. Keby to neurobíme, tak máme v dátach pozorovanie označované ako tzv. outlier, a tento by veľmi vplýval na odhadnuté parametre modelu a zväčšoval tak jeho nepresnosť, resp. chybovosť v predikovaní.

Pre analýzu odberateľov, na ktorých si treba dať pozor v zmysle skokových spotrieb sme odhadli pearsonové korelačné koeficienty pre jednotlivých odberateľov, pričom túto koreláciu sledujeme výlučne medzi spotrebou a teplotou, nakoľko teplota je najvýznamnejší regresor. Robíme tak na základe úvahy, že ak by spotreba mala skákať dosť výrazne, tak to nie je spôsobené vplyvom počasia, a preto tu očakávame nízku mieru závislosti medzi spotrebou a teplotou. Táto korelácie je najprv odhadnutá pre obdobie JAN 2007 - DEC 2008, kde sú však zahrnuté aj obdobia ako leto (menšia závislosť),

a ďalej pre samostatné zimné obdobia, ktoré nás zaujímajú najviac. Výsledky odhadov korelačných koeficientov dokumentujeme v nasledujúcej tabuľke (tab. 05). Tieto odhady sú urobené na základe nasledujúceho vzťahu (všeobecne platný), ktorý upravíme pre použitie dát t.j. odhad pre Pearsonov korelačný koeficient.

$$\rho = \frac{\text{kov}(X, Y)}{\sqrt{D(X)D(Y)}}$$

ρ ... Pearsonov korelačný koeficient

$\text{kov}(X, Y)$... kovariancia náhodných premenných X a Y

$D(X)$... disperzia (rozptyl) náh. prem. X

$D(Y)$... disperzia (rozptyl) náh. prem. Y

$$\hat{\rho} = \frac{\sum_{i=1}^n (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

$\hat{\rho}$... odhad pre Pearsonov korelačný koeficient

\bar{X} ...odhad strednej hodnoty náh. pre. X

\bar{Y} ...odhad strednej hodnoty náh. pre. Y

Tab. 05 – Odhady pre Pearsonove korelačné koeficienty,
zdroj: tab. - vlastný, dáta – SHMÚ, SPP

Ozn. odberateľa SPP (posledné 3 číslice)	Obec	Priemerná spotreba 2007-2008	Priemerná spotreba za január 2007	Test	Odhad pre pearsonov korelačný koeficient za 2007-2008	Odhad pre pearsonov korelačný koeficient za január 2007
019	Bratislava - Dev. Nová Ves	15 743	24 548	1	-0,92	-0,87
925	Bratislava - Dúbravka	122 997	227 664	1	-0,93	-0,95
160	Bratislava - Nové Mesto	655 076	1 013 717	1	-0,80	-0,80
320	Bratislava - Nové Mesto	57 191	70 964	1	-0,11	-0,89
194	Bratislava - Petržalka	12 838	22 665	1	-0,93	-0,97
219	Bratislava - Petržalka	12 366	21 769	1	-0,93	-0,96
220	Bratislava - Petržalka	15 656	29 032	1	-0,93	-0,96
276	Bratislava - Petržalka	12 727	23 250	1	-0,93	-0,96
640	Skalica	8 932	14 029	1	-0,93	-0,92
936	Dubnica nad Váhom	28 138	37 885	1	-0,65	-0,72
150	Považská Bystrica	10 106	17 246	1	-0,95	-0,94
915	Považská Bystrica	9 263	8 948	0	-0,10	-0,66
078	Púchov	15 428	28 679	1	-0,94	-0,94
127	Trenčín	6 638	15 498	1	-0,80	-0,95
483	Komárno	25 249	40 910	1	-0,93	-0,90
776	Nitra	9 640	17 620	1	-0,92	-0,93
268	Nové Zámky	46 950	68 779	1	-0,94	-0,84
316	Štúrovo	9 950	17 127	1	-0,94	-0,91
203	Topoľčany	20 957	39 450	1	-0,94	-0,90
836	Martin	9 767	8 455	0	-0,29	0,04
386	Nížna	14 689	20 923	1	-0,93	-0,88
206	Ružomberok	288 336	289 857	1	-0,32	-0,35
025	Banská Bystrica	35 855	64 847	1	-0,94	-0,86
803	Rimavská Sobota	16 911	31 213	1	-0,95	-0,90
314	Humenné	20 103	19 489	0	-0,07	0,34
294	Prešov	8 993	16 137	1	-0,94	-0,91
323	Prešov	20 154	34 334	1	-0,95	-0,93
825	Svit	28 860	51 378	1	-0,93	-0,85
186	Košice - Juh	163 493	290 690	1	-0,70	-0,28

V tejto tabuľke sa nachádza aj stĺpec „Test“, ktorý hovorí o tom, že ak je priemerná spotreba za január 2007 vyššia ako priemer za obdobie 2007-2008, tak test vypíše hodnotu „1“. Tento stav pokladáme za prirodzený, lebo priemerná spotreba za január by rozhodne mala byť vyššia ako napr. priemer za celé 2 roky, kde sú aj obdobia kedy nebolo potrebné takmer vôbec kúriť. Ak test vypíše „0“, toto je pre nás dosť paradoxná situácia pre bežného odberateľa typu tepláreň a signalizuje to odberateľa, ktorý sa nespráva veľmi podľa očakávaní, t.j. počasia (je zvýraznený oranžovou farbou). Preto budeme očakávať, že

modely a analýzy pre týchto spotrebiteľov nebudú veľmi vierohodné a teda nie vhodné pre skúmanie vplyvu snehovej pokrývky na spotrebu.

Posledné 2 stĺpce informujú o vyššie spomínaných odhadoch pre korelačné koeficienty pre jednotlivých odberateľov za obdobie 2 rokov ako aj január 2007. Ak je tento odhad v intervale $< -0,5 ; -0,8 >$, signalizuje to zvýšenú pozornosť pri analýze týchto odberateľov. V tabuľke sú označený žltou farbou. Ak je tento odhad väčší ako $-0,5$, signalizuje to pre nás veľmi neštandardného odberateľa a v týchto analýzach neočakávame uspokojujúce výsledky. Títo sú zvýraznení oranžovou farbou.

Pri odhadoch korelačných koeficientov pre jednotlivé teplárne nás dosť prekvapila tá skutočnosť, že tieto odhady vyšli väčšinou ešte nižšie, ako odhad pre korel. koef. pre spotreby za celé Slovensko a priemernú teplotu za celé Slovensko (počítaná na základe teplôt z 11-tich staníc rozmiestnených na celom území SR). Tu nám vyšiel odhad pre obdobie 2007-2008 $\hat{\rho} = -0,95$ a pre január 2007 $\hat{\rho} = -0,94$.

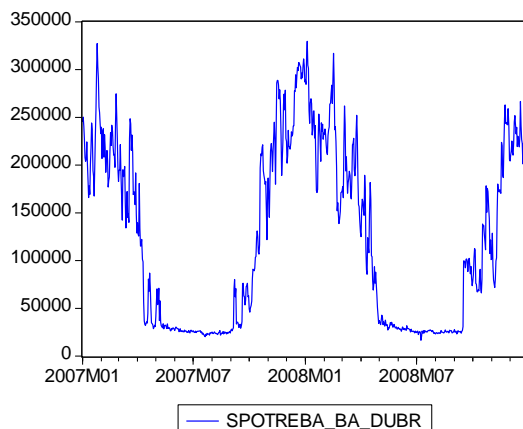
2.2. Deskriptívne štatistiky a vlastnosti dát

V tejto podkapitole si priblížime dáta, s ktorými budeme pracovať. Niektoré vlastnosti dát budú vyjadrené číselne a iné graficky, preto tu budeme mať viacero obrázkov. Týmto by si mal čitateľ utvoriť lepšiu predstavu o priebehu dát.

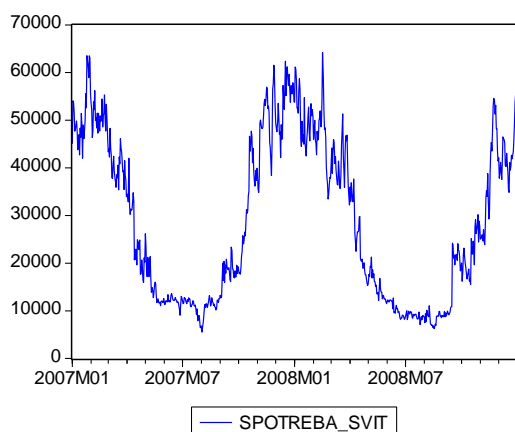
2.2.1. Denné spotreby plynu

Ako prvý si zobrazíme grafický priebeh spotreby za obdobie 2007-2008 pre 3 nami vybratých odberateľov (Bratislava - Dúbravka, Svit, Považská Bystrica_915). Prvé 2 priebehy sú pre odberateľov, ktorý dosť výrazne reagujú na zmeny počasia. Tretí z nich sa riadi predovšetkým inými vplyvmi. Štvrtý je pre celé SK. Na osi „Y“ je spotreba v m^3 .

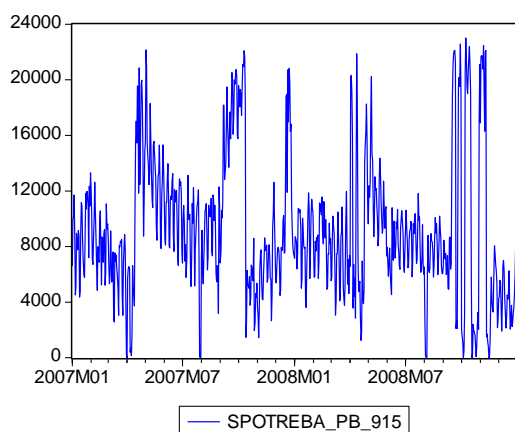
Obr. 04 – Denný priebeh spotreby plynu (BA - Dúbravka), zdroj: dáta SPP



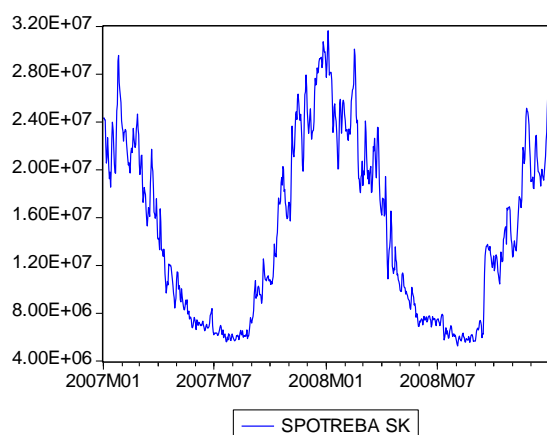
Obr. 05 – Denný priebeh spotreby plynu (Svit), zdroj: dáta SPP



Obr. 06 – Denný priebeh spotreby plynu (Považská Bystrica-915), zdroj: dáta SPP



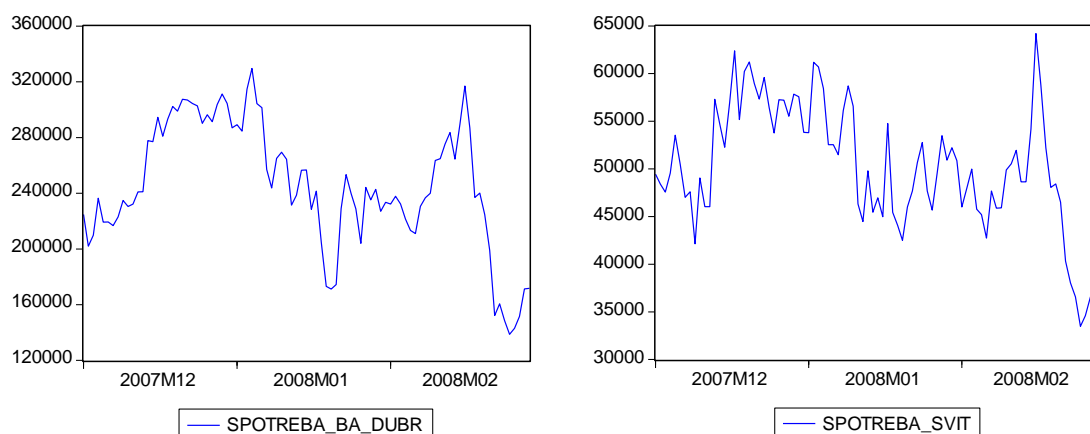
Obr. 07 – Denný priebeh spotreby plynu (Slovensko), zdroj: dáta SPP



Na ďalších grafoch si zobrazíme iba výseky z vyššie zobrazených grafov, pričom si zobrazíme priebeh spotrieb v období december 2007 – február 2008. Vo všetkých sú pomerne veľké skoky, nakoľko teplota sa v tomto období dosť výrazne menila. Avšak v prípade odb. z Považskej Bystrice sú dosť nezávislé od zmeny teploty.

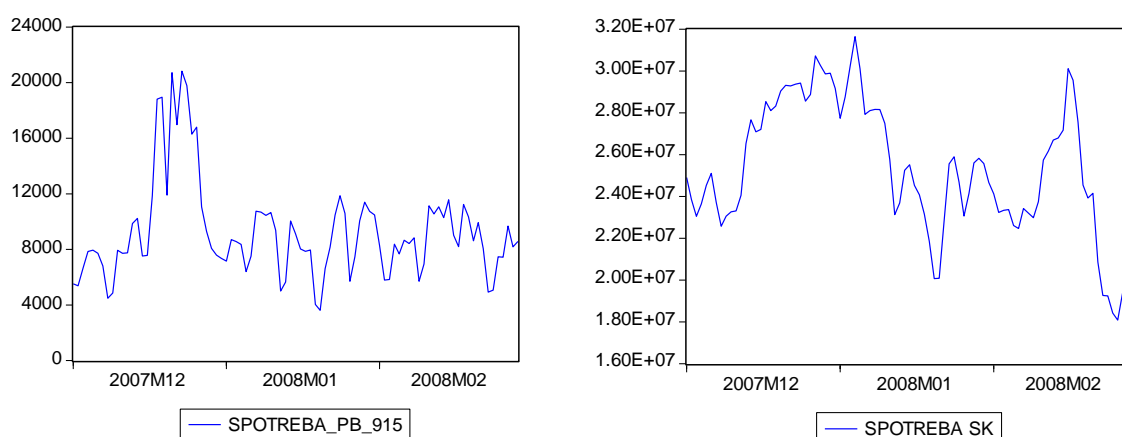
Obr. 08 (vľavo) – Denný priebeh spotreby plynu (BA - Dúbravka), zdroj: dáta SPP

Obr. 09 (vpravo) – Denný priebeh spotreby plynu (Svit), zdroj: dáta SPP



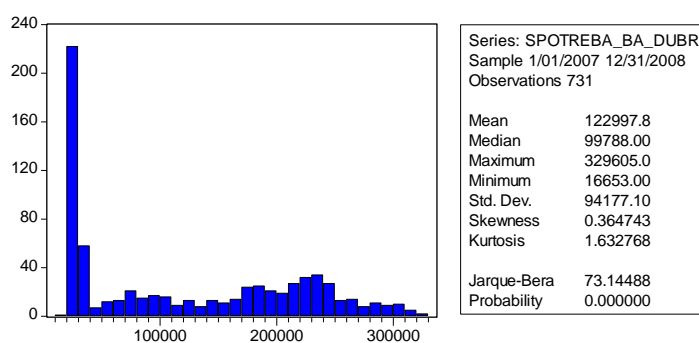
Obr. 10 (vľavo) – Denný priebeh spotreby plynu (BA – Považská Bystrica_915), zdroj: dáta SPP

Obr. 11 (vpravo) – Denný priebeh spotreby plynu (celé SK), zdroj: dáta SPP

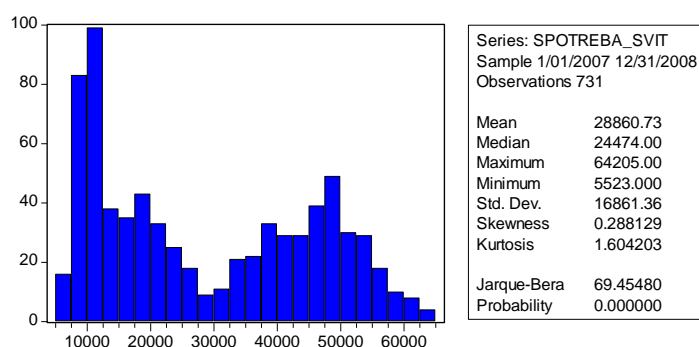


Ďalej si zobrazíme špecifické grafy, ide o tzv. histogramy, ktoré nám dávajú predstavu o rozdelení, z ktorého dáta pochádzajú. Vpravo sú vypísané základné deskriptívne štatistiky. Skewness znamená šikmosť hustoty a kurtosis jej špicatosť. Úplne dole je hodnota tzv. Jarque Berra štatistiky, ktorá na základe šikmosti a špicatosti rozhoduje o hypotéze, či dané dáta môžu pochádzať z normálneho rozdelenia. Ak je príslušná p-hodnota vyššia ako 5%, tak môžeme prijať nulovú hypotézu o normalite dát. Všetky tieto grafy sú konštruované v programe EViews 5.0. Obrázky 12-15 sa týkajú spotrieb za celé 2 roky (2007-2008) a obrázky 16-19 sa týkajú spotrieb za december 2007 až február 2008 (t.j. 3 zimné mesiace).

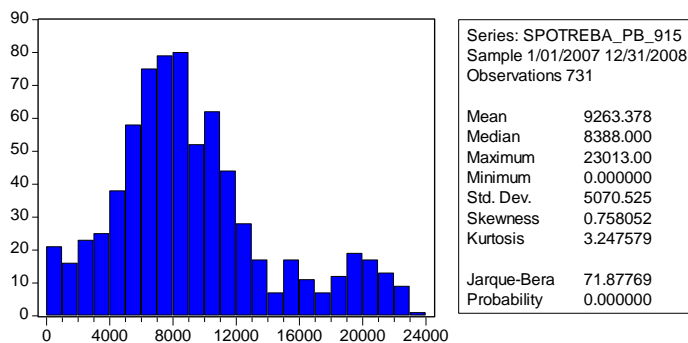
Obr. 12 – Histogram spotreby plynu (Bratislava - Dúbravka), zdroj: dáta SPP



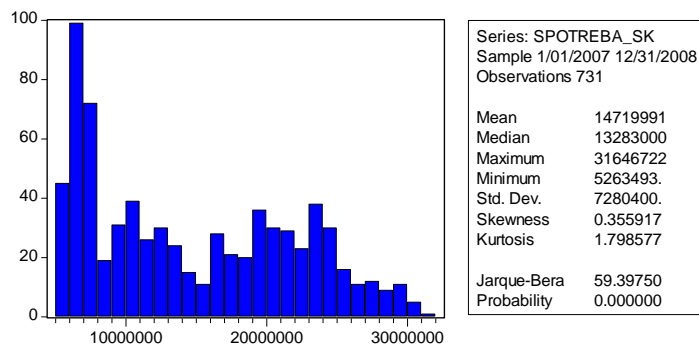
Obr. 13 – Histogram spotreby plynu (Svit), zdroj: dáta SPP



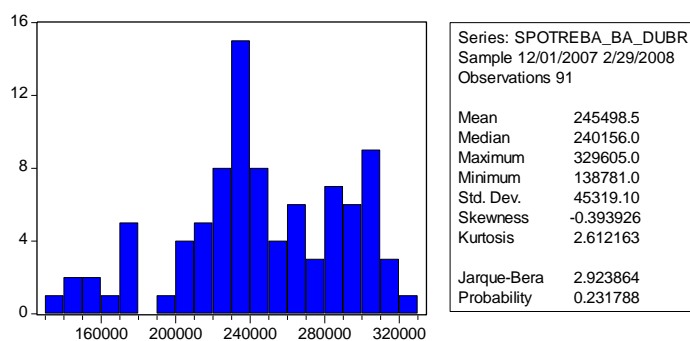
Obr. 14 – Histogram spotreby plynu (Považská Bystrica_915), zdroj: dáta SPP



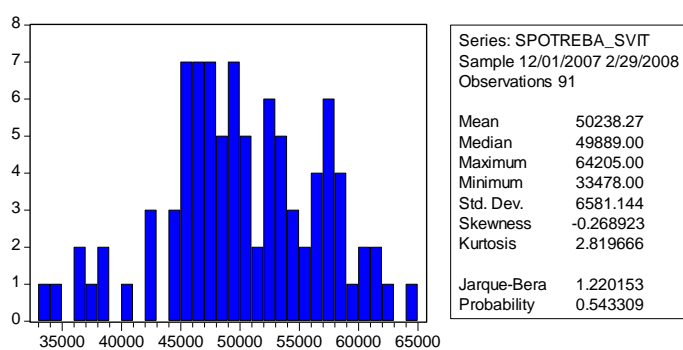
Obr. 15 – Histogram spotreby plynu (Považská Bystrica_915), zdroj: dáta SPP



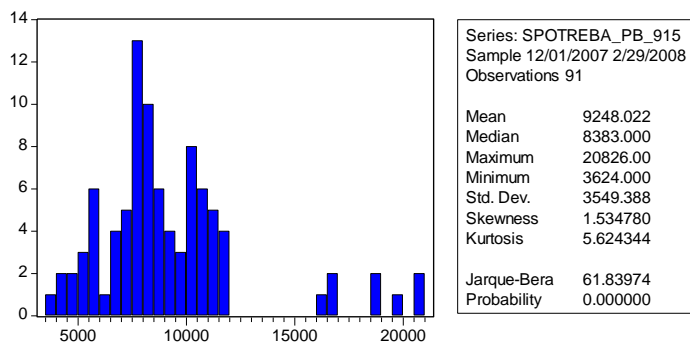
Obr. 16 – Histogram spotreby plynu (Bratislava - Dúbravka), zdroj: dáta SPP



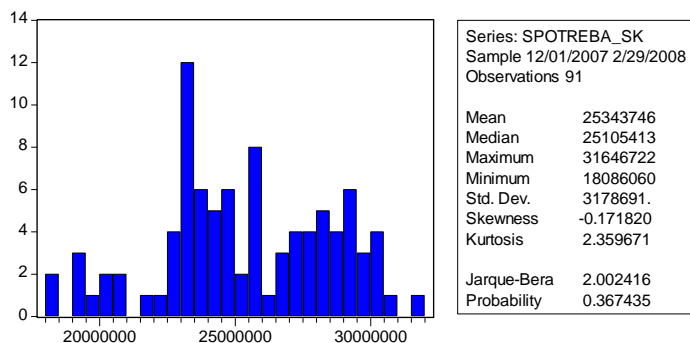
Obr. 17 – Histogram spotreby plynu (Svit), zdroj: dáta SPP



Obr. 18 – Histogram spotreby plynu (Považská Bystrica_915), zdroj: dáta SPP



Obr. 19 – Histogram spotreby plynu (Považská Bystrica_915), zdroj: dáta SPP



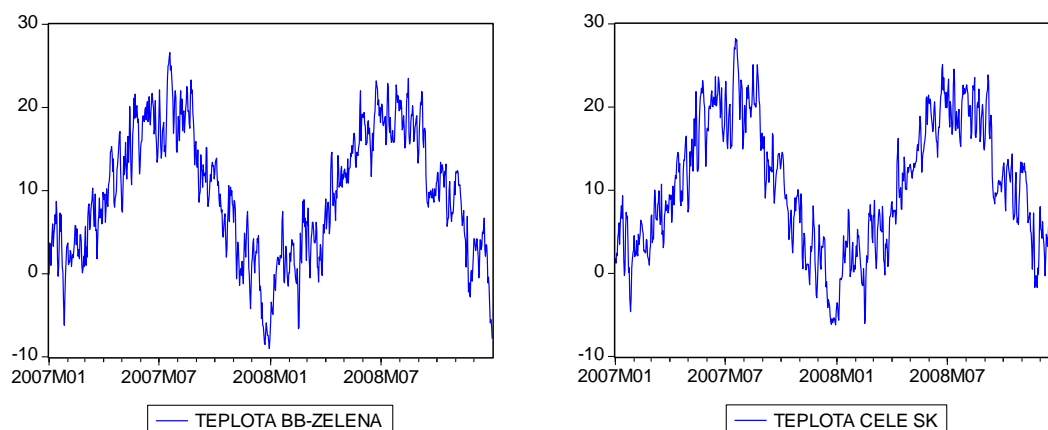
Z týchto grafov možno vyčítať napr. aj to, že kým spotreby za obdobie celých 2 rokov sa nesprávajú ako z normálneho rozdelenia (aj graficky aj podľa testu Jarque Berra) v žiadnom z našich príkladov, tak už spotreby v období 3 zimných mesiacov áno, okrem dát pre odberateľa z Považskej Bystrice, ktorého spotreba veľmi nezávisí od počasia. Ďalej z priemerov spotrieb týchto odberateľov vyplýva, že ich veľkosť sa výrazne líši. Preto ak budeme chcieť analyzovať vplyv počasia pre viacerých odberateľov naraz (napr. analýza panelových dát), budeme musieť na toto pamätať a vybrať si odberateľov s rádovo podobnou spotrebou.

2.2.2. Denné priemerné teploty

V tejto časti si zobrazíme priebehy teplôt pre vybranú lokalitu (Banská Bystrica - Zelená) a priemer pre celé územie SR, a to za obdobie 2 rokov (2007-2008) a zimné obdobie december 2007 – február 2008. Na osi „Y“ sú teploty v °C. Taktiež si zobrazíme grafy s príslušnými histogrammi, a vpravo budú opäť hodnoty deskriptívnej štatistiky.

Obr. 20 (vľavo) – Priebeh priemernej dennej teploty (BB-Zelená), obdobie 2007-2008, zdroj: dáta SHMÚ

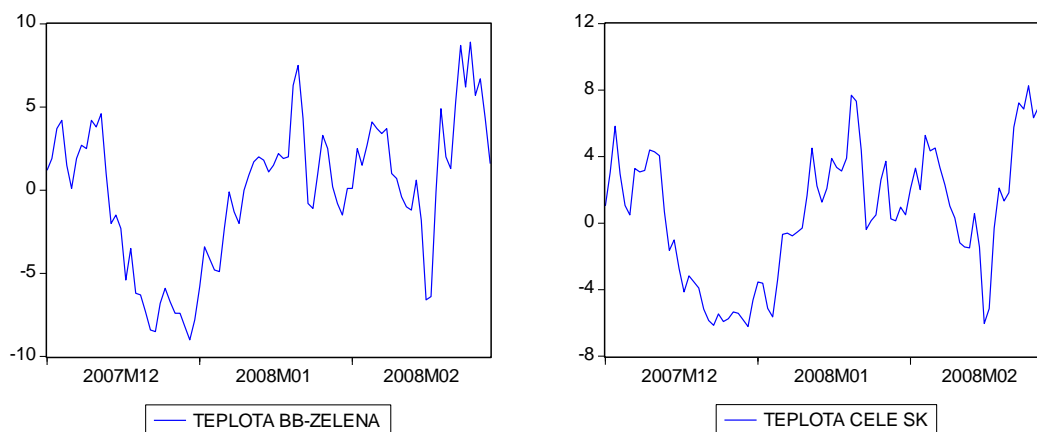
Obr. 21 (vpravo) – Priebeh priemernej dennej teploty (celé SK), obdobie 2007-2008, zdroj: dáta SHMÚ



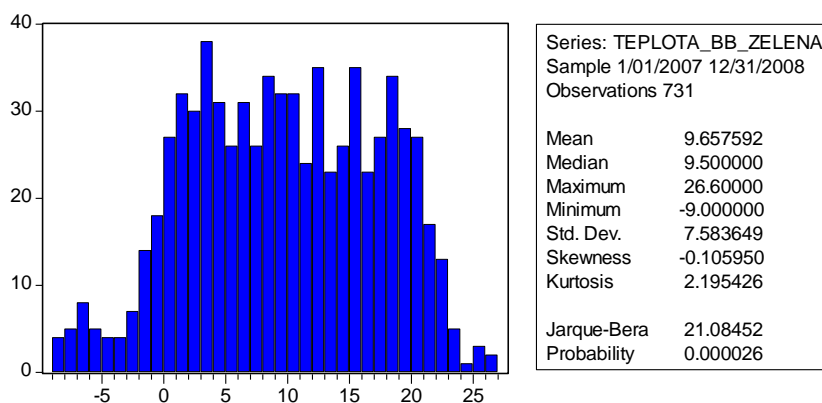
Z týchto priebehov vidíme, že majú relatívne rovnaký tvar. To sa aj dá očakávať, keďže Slovensko je malá krajina na to, aby tieto rozdiely boli extrémne a navyše Banská Bystrica leží v podstate v strede územia Slovenskej Republiky. Bližšie rozdiely uvidíme na nasledujúcich 2 grafoch, kde je aj na prvý pohľad viditeľné, že teploty v Banskej Bystrici sú o niečo nižšie, ako teploty za celoslovenský priemer.

Obr. 22 (vľavo) – Priebeh priemernej dennej teploty (BB-Zelená), obdobie december 2007- február 2008, zdroj: dáta SHMÚ

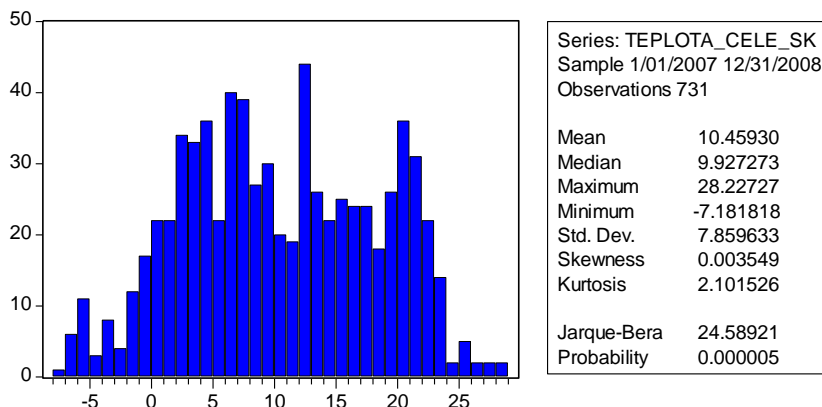
Obr. 23 (vpravo) – Priebeh priemernej dennej teploty (celé SK), obdobie december 2007- február 2008, zdroj: dáta SHMÚ



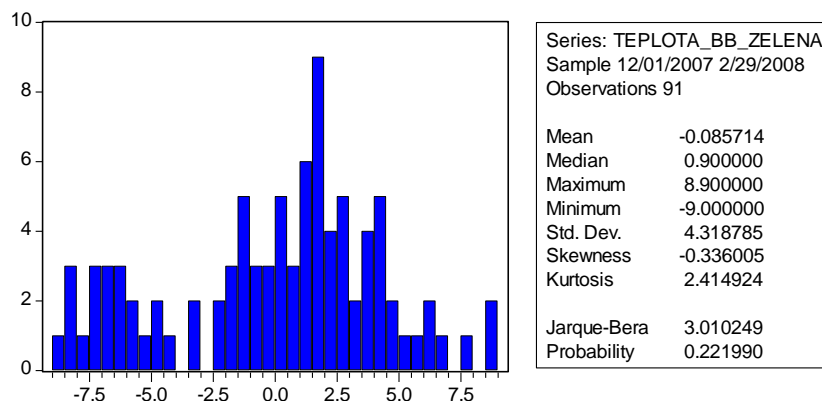
Obr. 24 – Histogram priemernej dennej teploty (BB-Zelená), obdobie 2007-2008, zdroj: dáta SHMÚ



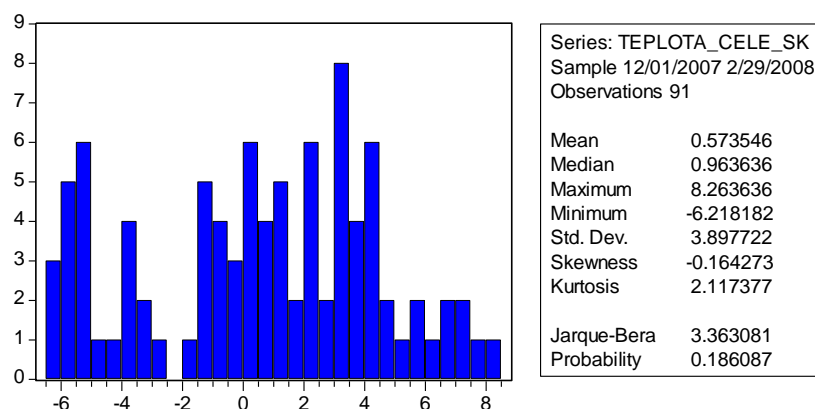
Obr. 25 – Histogram priemernej dennej teploty (celé SK), obdobie 2007-2008, zdroj: dáta SHMÚ



Obr. 26 – Histogram priemernej dennej teploty (BB-Zelená), obdobie december 2007-február 2008, zdroj: dáta SHMÚ



Obr. 27 – Histogram priemernej dennej teploty (celé SK), obdobie december 2007-február 2008, zdroj: dáta SHMÚ



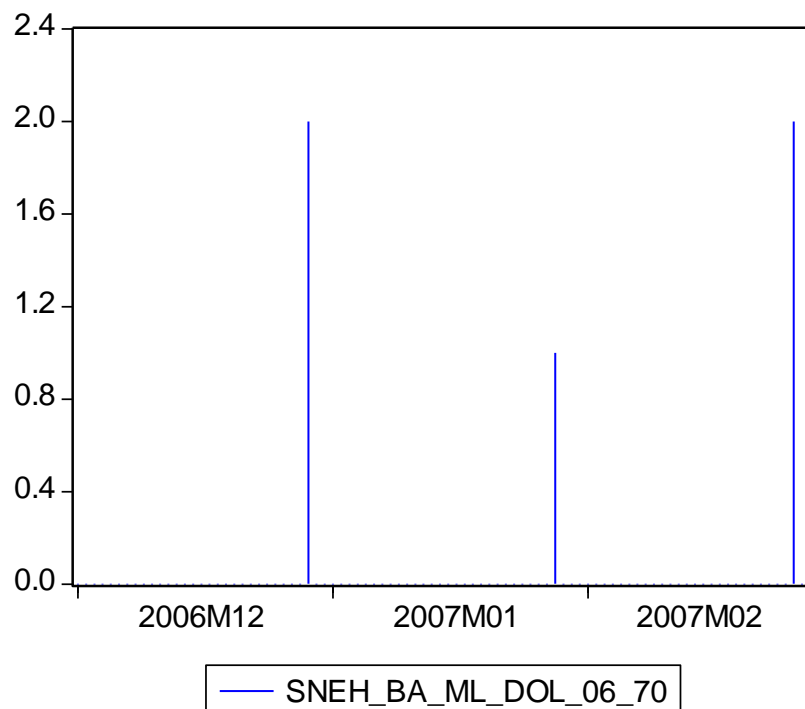
Z deskriptívnych štatistík vpravo pri histogramoch sa nám naozaj potvrdilo to, že priemerne sú v Banskej Bystrici teploty nižšie ako je celoslovenský priemer (viď. „mean“).

2.2.3. Celková snehová pokrývka

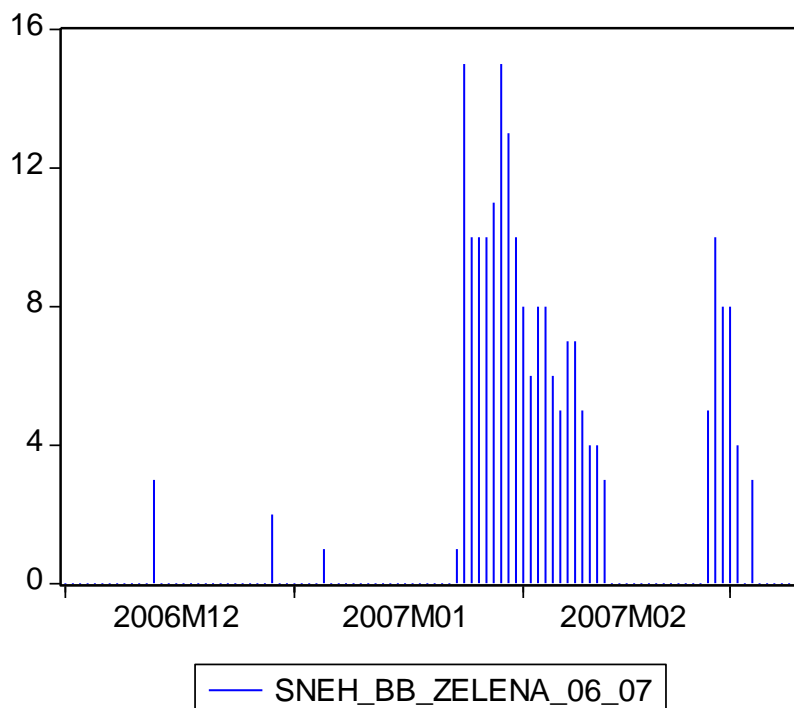
Teraz si zobrazíme priebehy výšky celkovej snehovej pokrývky a ich základné deskriptívne štatistiky za zimy 2006-2007 a 2007-2008 pre niekoľko vybraných lokalít. Ako lokality sme sa rozhodli ilustrovať dáta zo stanice Bratislava Mlynská Dolina (reprezentuje stanice na juhu a nížinné oblasti) a dáta zo stanice Banská Bystrica – Zelená (reprezentuje stanice z oblastí kde sú aj hory). Najprv si zobrazíme graf, ktorý sa volá stĺpcový diagram (tzv. Bar plot). Jeden stĺpec reprezentuje jeden deň, pričom dni sú na osi „X“ tak ako plynul čas. Výška každého stĺpca znamená výšku celkovej snehovej pokrývky

v daný deň a je meraná v „cm“. Najprv si zobrazíme tieto grafy pre zimné obdobie 2006/2007, pričom v prípade stanice Bratislava Mlynská Dolina to bude iba pre mesiace december 2006 až február 2007, keďže ďalej v marci už nebol žiadny sneh. V prípade stanice Banská Bystrica – Zelená to bude taktiež od decembra 2006, ale až po 10.marec 2007, lebo ešte začiatkom marca tam existovala snehová pokrývka, no neskôr už nebola.

Obr. 28 – Stĺpcový diagram celkovej snehovej pokrývky (Bratislava-Mlynská dolina), obdobie december 2006-február 2007, zdroj: dáta SHMÚ

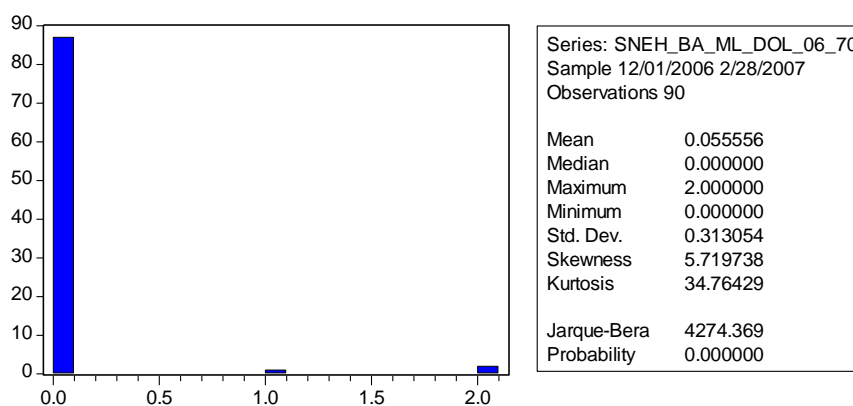


Obr. 29 – Stĺpcový diagram celkovej snehovej pokrývky (Banská Bystrica-Zelená), obdobie december 2006-10.marec 2007, zdroj: dáta SHMÚ

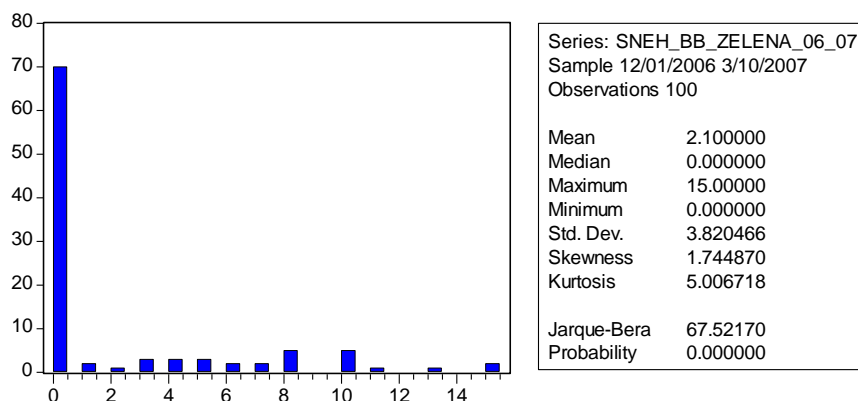


Z obr. 28 vidíme, že tu boli iba 3 dni, kedy existovala snehová pokrývka a to vo výške iba 1 a 2 cm. Je zrejmé, že v prípade takéhoto priebehu výšky snehovej pokrývky analýzu vplyvu snehu na spotrebu nebude možné urobiť. V prípade iného regiónu, obr. 29, to už bude možné, keďže jej výška dosahovala aj hodnoty 15 cm a jej trvanie bolo spojité v priebehu niekoľkých dní, kedy sa jej efekt už mohol prejaviť. K týmto priebehom ďalej prikladáme informácie o deskriptívnych štatistikách a príslušných histogramoch (obr. 30 a 31)

Obr. 30 – Histogram celkovej snehovej pokrývky (Bratislava-Mlynská dolina), obdobie december 2006-február 2007, zdroj: dáta SHMÚ

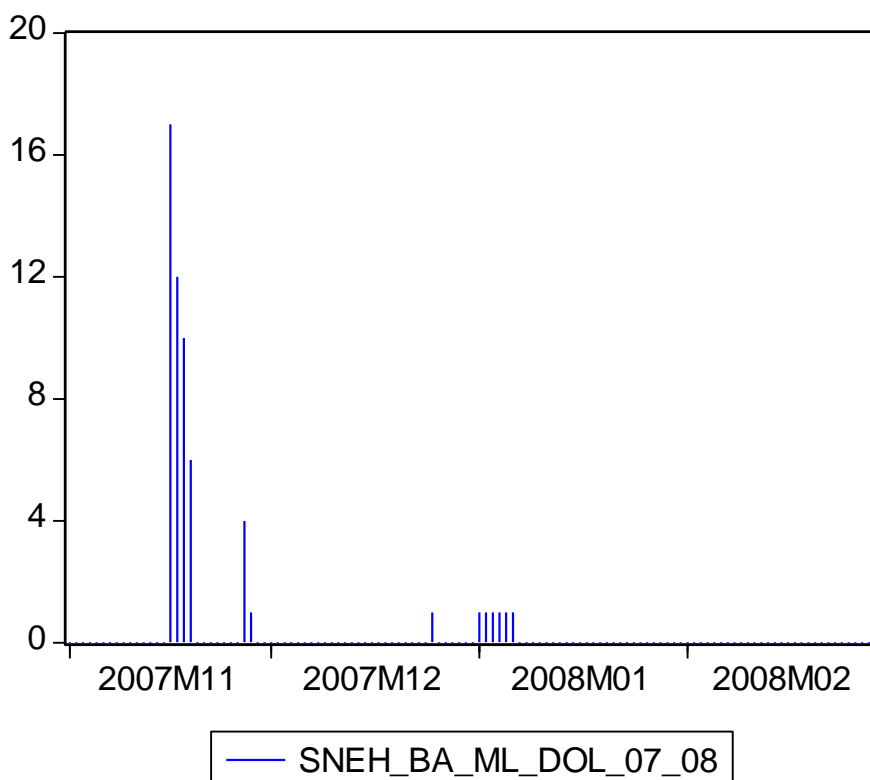


Obr. 31 – Histogram celkovej snehovej pokrývky (Banská Bystrica-Zelená), obdobie december 2006-10.marec 2007, zdroj: dáta SHMÚ

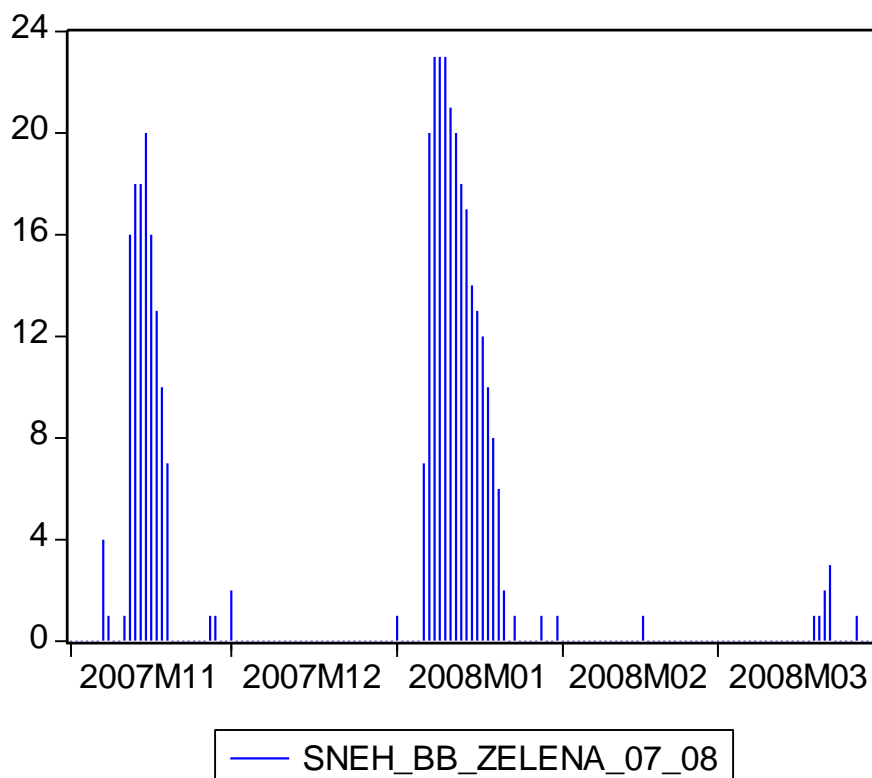


Ďalej si zobrazíme podobne ako na predchádzajúcich obrázkoch stĺpcové diagramy pre tie isté stanice, ale tento krát za zimné obdobie 2007/2008. Avšak keďže už v novembri 2007 napadol prvý sneh, obdobie na jeho analýzu sa nám teda predĺžilo o jeden mesiac. Navyše v prípade stanice v Banskej Bystrici bol sneh ešte aj v marci 2008, takže tu máme celkové obdobie až 5 mesiacov.

Obr. 32 – Stĺpcový diagram celkovej snehovej pokrývky (Bratislava-Mlynská dolina), obdobie október 2007-február 2008, zdroj: dáta SHMÚ



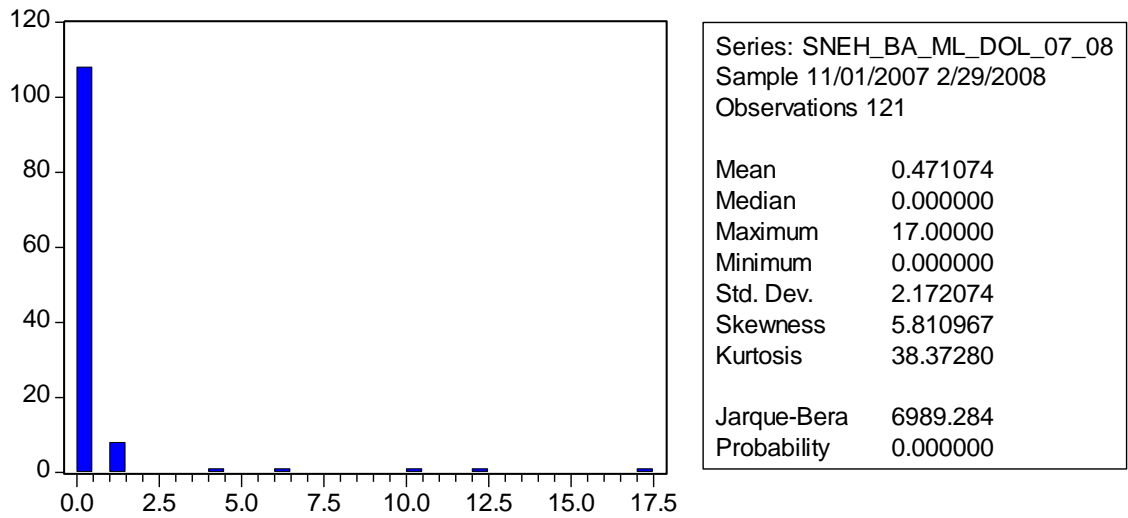
Obr. 33 – Stĺpcový diagram celkovej snehovej pokrývky (Banská Bystrica-Zelená), obdobie október 2007-marec 2008, zdroj: dáta SHMÚ



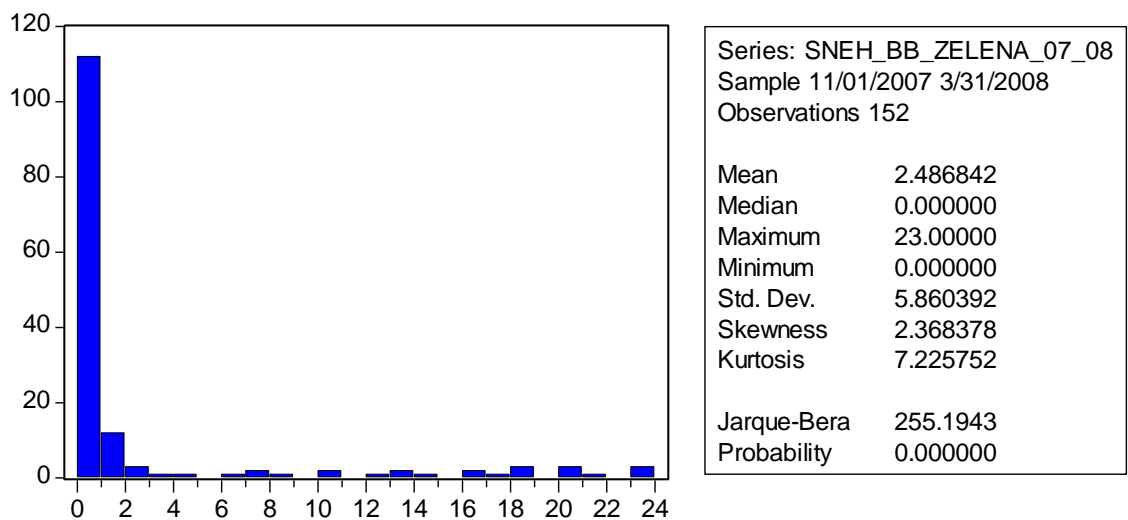
Z obr. 32 je jasne vidieť, že ani táto zima nebola v prípade stanice v Bratislave čo do snehu veľmi štedrá. V podstate až paradoxne bolo v novembri viac sneh než potom v ďalších typicky zimných mesiacoch. V prípade regiónov strednej a hlavne severnej oblasti Slovenska to bolo o niečo lepšie, ale ani táto zima nebola veľmi štedrá na snehovú pokrývku. Bohužiaľ na tomto mieste musíme podotknúť, že dáta ktoré máme k dispozícii rozhodne nie sú ideálne pre analýzu vplyvu snehovej pokrývky na spotrebu. Ideálne by bolo disponovať s dátami, kde by nie len že bola súvislá vrstva snehovej pokrývky v priebehu aspoň jedného týždňa spojito, ale aby v danom období prišli aj väčšie mrazy, kedy by sme očakávali, že sa izolačný efekt môže dostaviť. Tieto naše zimy ani veľmi nespĺňajú požiadavku pre dosť výrazné mrazy.

Nakoniec ešte pripájame základné deskriptívne štatistiky s histogrammi dopĺňujúce obrázky 32 a 33.

Obr. 34 – Histogram celkovej snehovej pokrývky (Bratislava-Mlynská dolina), obdobie október 2007-február 2008, zdroj: dáta SHMÚ



Obr. 35 – Histogram celkovej snehovej pokrývky (Banská Bystrica-Zelená), obdobie október 2007-marec 2008, zdroj: dáta SHMÚ



3. Teória a doterajšie poznatky

Táto kapitola bude o matematickom základe, na ktorom budú naše modely postavené. Samozrejme, že nie je možné detailné rozpísanie všetkých jej častí a odvodení, nakoľko by to rozsahom niekoľkonásobne prekročilo rozsah tejto práce a navyše to ani nie je jej účelom. Pre samotných odborníkov z oblasti štatistiky, regresnej analýzy alebo analýzy časových radov sú tieto poznatky aj tak veľmi dobre známe.

3.1. Lineárna regresia

V tejto podkapitole si zdefinujeme niekoľko matematických formúl o ktoré sa opierajú ekonometrické analýzy. Čerpali sme pritom so zdrojov [4], [5] a [6].

Medzi základné pojmy patrí:

- Vysvetľovaná premenná (závislá) – je tá, ktorej zmeny sa snažíme vysvetliť pomocou iných premenných (regresorov), ozn. spravidla ako Y .
- Vysvetľujúca premenná (nezávislá) – je tá, ktorá pôsobí na inú premennú (závislú) a teda spôsobuje jej zmenu. V regresných modeloch ich býva spravidla niekoľko. Sú to tzv. regresory, ozn. spravidla ako X .

Formulácia modelu:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

kde:

α ... predstavuje konštantu alebo tzv. konštantný člen, často ozn. ako β_0

β_1, \dots, β_k predstavujú parametre regresnej nadroviny (priamky ak máme iba β_1), pričom určujúce jej sklon

ε_i ... predstavuje náhodný (disturbančný) člen, tzv. rezíduum

i ... predstavuje index označujúci jednotlivé pozorovania, pričom $i = 1, \dots, N$, pričom v našom prípade sú tieto pozorovania idúce v čase a preto index „ i “ môžeme nahradiť indexom „ t “, pričom $t = 1, \dots, T$

Dôvody existencie disturbančného člena:

- Vynechanie vplyvu niektorých regresorov.
- Chyby merania (problémy pri zbere dát).
- Individuálne preferencie.

Často sa pracuje s maticovým zápisom modelu, kde „X“ voláme ako matica plánu, ktorej prvý stĺpec pozostáva z „1“. Ostatné zložky sú vektory.

$$y = X\beta + \varepsilon$$

Základným krokom v regresnej analýze je odhad neznámych parametrov modelu. V praxi sa často používa tzv. metóda najmenších štvorcov. Je postavená na princípe minimalizácie súčtu štvorcov rezíduí. Rezíduá predstavujú zvislú vzdialenosť medzi skutočnými (pozorovanými) hodnotami vysvetľovanej premennej a fitovanými hodnotami určujúcich regresnú „priamku“. Ďalej platí, že ich suma je rovná nule.

Keď je spomínaný model odhadnutý, kontroluje sa jeho kvalita a splnenie istých predpokladov. Pre posúdenie kvality modelu, t.j. ako dobre nám naša odhadnutá regresná „priamka“ preložila dáta sa často používa tzv. koeficient determinácie, ozn. R^2 . Už aj pohľad na samotné rezíduá nám dáva dosť dobrý prehľad o kvalite modelu, lebo vidíme, že čím bližšie sú skutočné hodnoty „Y“ k regresnej „priamke“, tým je chyba v modeli menšia a teda rezíduá sú menšie. Keďže v praxi je týchto hodnôt rezíduí toľko, koľko je samotných pozorovaní, pri ich veľkých počtoch už nevieme jednoznačne z obrázku rozhodnúť o kvalite modelu. Preto je užitočné vypočítať si koeficient determinácie, lebo ten zohľadňuje všetky rezíduá naraz.

$$R^2 = 1 - \frac{RSS}{TSS}$$

RSS ... reziduálny súčet štvorcov (residual sum of squares)

TSS ... celkový súčet štvorcov (total sum of squares)

$$RSS = \sum_{t=1}^T (y_t - \hat{y}_t)^2$$

$$TSS = \sum_{t=1}^T (y_t - \bar{y})^2$$

\hat{y}_t ... fitované hodnoty regresnou „priamkou“

\bar{y} ... aritmetický priemer z Y

Ďalším kritériom je tzv. upravené (adjusted) R^2 , ktoré zohľadňuje aj počet parametrov v modeli. Pointa je v tom, že čím viac vysvetľujúcich premenných v modeli máme, tým je model komplikovanejší a môžu vznikáť nežiaduce problémy ako multikolinearita, alebo sa môže viac prejaviť chyba pri meraní dát,... a preto tento ukazovateľ penalizuje pridávanie ďalších premenných. V praxi sa preto väčšinou preferujú čo najjednoduchšie modely.

$$R_A^2 = 1 - \frac{n-1}{n-k} (1 - R^2)$$

n ... počet pozorovaní (ozn. ako N resp. pri časových priebehoch T)

k ... počet koeficientov

Ďalšie často používané kritériá sú Akaikeho informačné kritérium (AIC) a Schwarzovo (SC) informačné kritérium, ktoré sa používajú pre porovnanie modelov s rôznym počtom parametrov.

$$AIC = \ln \frac{\hat{e}e}{n} + 2 \frac{k}{n}$$

$$SC = \ln \frac{\hat{e}e}{n} + \frac{k}{n} \ln n$$

e ... vektor odhadov skutočných rezíduí

Predpoklady lineárneho regresného modelu:

- $E[\varepsilon_i] = 0$
- $\text{Var}[\varepsilon_i] = \sigma^2$... homoskedasticita
- $\text{Kov}[\varepsilon_i, \varepsilon_j] = 0$ pre $i \neq j$... nekorelovanosť rezíduí
- ε_i má normálne rozdelenie
- X je pevná matica (nenáhodná)

Pri splnení týchto predpokladov je odhad parametrov cez MNŠ (OLS) najlepší lineárny nevychýlený odhad, tzv. BLUE.

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

3.1.1. Problém heteroskedasticity rezíduí

Jeden z predpokladov pre prácu s našim modelom je konštantnosť rozptylu náhodnej zložky. Ak je to splnené, nazývame to homoskedasticita náhodnej zložky. Teda znamená to, že náhodné zložky sú z rozdelenia s rovnakým parametrom disperzie. To či je v modeli prítomná heteroskedasticita sa dá testovať. Najčastejšie používané sú testy Goldfeldov-Quandtov test, Breuschov-Paganov test a obľúbený Whiteov test. Ak nám tieto testy potvrdia heteroskedasticitu, bude nutné model transformovať, logaritmováť jednu alebo viac premenných, predeliť ich nejakou premennou,... Jedná sa o tzv. zovšeobecnený lineárny regresný model.

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

Ω je $n \times n$ matica, $E[\varepsilon\varepsilon'] = \sigma^2\Omega$

3.1.2. Problém autokorelácie rezíduí

Na to aby bol OLS odhad najlepším (BLUE) a mohli sme s ním ďalej pracovať (testy o parametroch, predikcie,...), musí byť splnená aj táto podmienka, teda nekorelovanosť rezíduí. Problém s korelovanosťou rezíduí je typický pre údaje, ktoré tvoria časové rady a teda sú veľmi korelované. To je presne aj náš prípad, kedy spotreba dnes bude určite veľmi podobná včerajšej spotrebe plynu. Na to aby bol náš odhad BLUE, musíme urobiť transformáciu modelu. Ideálny prípad je tzv. kvázi-diferenciácia časového radu, kedy sa vypočíta sila korelácie v danom časovom rade a ňou sa potom robia „diferencie“. Bežne v praxi sa používajú aj obyčajné diferencie a autokorelovanosť rezíduí sa takmer úplne vytratí. Pre testovanie autokorelácie rezíduí sa používajú tieto testy: Breuschov-Godfreyov test, Boxov-Piercov test, Ljungov test a veľmi populárny Durbinov-Watsonov test (DW). Pri DW teste sa kontroluje, či je hodnota testovacej štatistiky čo najbližšie k 2, pričom môže padnúť do intervalu $< 0 ; 4 >$. Za uspokojivý výsledok sa berie hodnota $2 \pm 0,2$. Výpočet testovacej štatistiky je nasledovný:

$$DW = \frac{\sum_{t=2}^T (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^T \hat{\varepsilon}_t^2}$$

3.1.3. Normalita rezíduí

Jedným z predpokladov pre náš model je aj normalita rezíduí, teda že rezíduá pochádzajú z normálneho rozdelenia, pričom sa predpokladá ich nulová stredná hodnota a konštantná disperzia. Všeobecne môže konštatovať, že ak tento predpoklad nie je splnený, tak testovaniu hypotéz o parametroch, predikciám, ... by sme nemali veľmi veriť, lebo oni sú postavené práve na tejto normalite. Avšak podľa asymptotickej teórie (nekonečná vzorka pozorovaní) sú aj napriek nesplneniu normality rezíduí výsledky rovnaké, ako keby normalita bola splnená. V praxi to znamená, že ak máme dosť veľa pozorovaní, tak nesplnená normalita rezíduí nie je až takým veľkým problémom. Jedným z najpoužívanejších testov pre overenie normality je Jarque-Berra test, ktorý sme už spomínali pri deskriptívnej štatistike našich dát.

3.1.4. Problém multikolinearity

Multikolinearita nastáva vtedy, keď sú vysvetľujúce premenné navzájom silne korelované. Teda vlastne nesú v sebe veľmi podobné informácie a teda to modelu neprinesú výrazne zlepšenie (tá ďalšia z nich). Prečo je to problémom je fakt, že multikolinearita robí značné problémy pri odhade metódou najmenších štvorcov. Deje sa to, že determinant matice $X'X$ je blízky nule. Jej dôsledok môže byť aj to, že sa nám môžu zdať odhady parametrov ako nesignifikantné. Multikolinearita nemá vplyv na koeficient determinácie. Častým riešením tohto problému je vypustenie jednej z korelovaných premenných z modelu, čím sa navyše model zjednoduší. Pre testovanie multikolinearity sa používa obyčajný výpočet korelácie daných premenných. Obvykle ak je tento odhad korelácie väčší ako 0,9, vtedy celkom určite hovoríme o silnej multikolinearite.

3.1.5. Testovanie hypotéz o parametri

Ak už máme odhadnuté parametre modelu, často sa testuje, či nie je niektorý z parametrov nulový a teda by sme príslušný regresor z modelu mohli odstrániť. Samozrejme, že sa dá testovať aj iná než nulová hodnota daného parametra, ale nás bude predovšetkým zaujímať test na nulovosť. V našom prípade to bude znamenať, že ak by pri premennej reprezentujúcej výšku snehu dopadol test tak, že nezamietame hypotézu o nulovosti parametra, vplyv snehu by teda nehral v modeli významnú rolu a z modelu by sme ho mohli odstrániť. Pre toto testovanie budeme používať tzv. kontrasty, kedy

správnou voľbou vektora „a“ testujeme daný parameter. Táto testovacia štatistika sa riadi studentovým t-rozdelením, pričom počet stupňov voľnosti je rovný „n-k“, kde „n“ je počet pozorovaní a „k“ je počet parametrov modelu. Testovacia štatistika má tvar:

$$T = \frac{a^t \hat{\beta} - a^t \beta}{S \sqrt{a^t (X^T X)^{-1} a}}$$

S ... je odhad pre σ , odmocninu z disperzie rezíduí

Pre hodnotu testovacej štatistiky sa zráta príslušná p-hodnota (p-value, skratene p-v) väčšinou na hladine významnosti 95% a porovná sa s príslušnou kritickou hodnotou studentovho rozdelenia.

Špeciálnym prístupom k testovaniu hypotéz o nulovosti parametrov je ich súčasná nulovosť, pričom sa vynechá iba parameter reprezentujúci konštantný člen. Tieto testy sa zvyknú všeobecne nazývať ako reštrikcie, kedy testujeme hypotézu o viacerých parametroch súčasne. Pričom ak testujeme hypotézu o nulovosti všetkých (okrem konšt. člena) parametrov naraz, voláme to test významnosti regresie. Hodnota testovacej štatistiky (F) sa porovnáva s kritickou hodnotou Fisherovho rozdelenia s príslušnými stupňami voľnosti ($F_{k-1, n-k}$).

$$F = \frac{\frac{TSS - RSS}{k - 1}}{\frac{RSS}{n - k}}$$

3.2. Časové rady a dynamické regresné modely

V tejto podkapitole sa oboznámime so základnou teóriou časových radov a modelov dynamickej regresie. Keď hovoríme o časovom rade, myslíme tým modelovanie nejakej premennej pomocou jej predchádzajúcich hodnôt, keďže jednotlivé pozorovania sú po sebe v čase idúce a preto sú výrazne korelované. Túto koreláciu nazývame autokorelácia, keďže sa jedná o ten istý časový rad. Svojím spôsobom už aj takýto model je dynamickým modelom, keďže regresory sú iba časovo posunuté hodnoty vysvetľovanej premennej. Ale keďže my sa budeme snažiť vysvetliť hodnoty spotreby plynu v závislosti od meteorologických veličín, takýto model obohatený o predchádzajúce spotreby budeme chápať ako dynamický. Ďalšiu dynamiku modelu by sme mohli získať pridaním časovo posunutých premenných samotných regresorov. Túto teóriu budeme čerpať zo zdrojov [8], [9], [10].

Autokorelovanosť rezíduí je veľkým problémom v klasickom prístupe ako je regresná analýza z predchádzajúcej podkapitoly. Ale Box a Jenkins vymysleli také modely, kde je táto autokorelácia žiaduca, lebo ju využívajú. Preto sa táto metodológia volá Boxova-Jenkinsonova. Existujú tu 2 základné prístupy, a to autoregresné modely (ozn. AR, z angl. auto-regresive) a modely s kľzavými priermi (ozn. MA z angl. moving average). Tieto dva prístupy sa zvyčajne kombinujú do tzv. ARMA modelov. Keďže sa pre prácu s nimi vyžaduje tzv. stacionarita, často je nevyhnutné najprv časový rad zdiferencovať. Vtedy sa tieto modely označujú ako ARIMA. V prípade, že sa jedná o priebeh sa dlhšie obdobie, kde sú výrazné sezónne vplyvy, je nutné daný časový rad ešte aj sezónne zdiferencovať. Vtedy sa tieto modely označujú ako SARIMA.

AR(p) proces môžeme zapísať v rovnicovom tvare:

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \dots + \Phi_p Y_{t-p} + \varepsilon_t$$

kde ε_t je tzv. biely šum a Φ_i sú AR-koeficienty.

AR(p) proces vieme zapísať pomocou operátora spätného posunu. Aby bol tento proces stacionárny, korene nasledovnej rovnice musia ležať mimo jednotkového kruhu:

$$(1 - \Phi_1 B - \Phi_2 B^2 - \dots - \Phi_p B^p) = \Phi_p(B) = 0$$

Biely šum je stacionárny proces, ktorý spĺňa nasledujúce predpoklady:

- $E[\varepsilon_t] = 0$ pre všetky „t“
- $\text{Var}[\varepsilon_t] = \sigma^2$ pre všetky „t“
- $\text{Kov}[\varepsilon_t, \varepsilon_s] = 0$ pre všetky „t“ \neq „s“

Pod pojmom stacionarita budeme myslieť tzv. slabú stacionaritu, ktorá je definovaná nasledovne:

- $E[x_t] = \mu$ pre všetky „t“
- $\text{Kov}[x_t, x_s] = \gamma(|t-s|)$ pre všetky „t“
- $\text{Var}[x_t] = \text{konšt.}$ pre všetky „t“, čo vyplýva z predch. predp.

MA(q) proces môžeme zapísať v rovnicovom tvare:

$$Y_t = \mu + \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \Theta_2 \varepsilon_{t-2} + \dots + \Theta_q \varepsilon_{t-q}$$

Tieto procesy sú vždy stacionárne a preto overujeme ich invertovateľnosť. Tá platí ak korene rovnice $\Phi_q(B) = 0$ ležia mimo jednotkového kruhu. Samozrejme aj tento proces vieme prepísať pomocou operátora spätného posunu:

$$Y_t = (1 + \Theta_1 B + \Theta_2 B^2 + \dots + \Theta_q B^q) \varepsilon_{t-p}$$

Keď tieto dva procesy spojíme, dostaneme ARMA(p,q) proces, ktorý vieme zapísať nasledovnou rovnicou:

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \dots + \Phi_p Y_{t-p} + \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \Theta_2 \varepsilon_{t-2} + \dots + \Theta_q \varepsilon_{t-q}$$

A pomocou operátora spätného posunu:

$$(1 - \Phi_1 B - \Phi_2 B^2 - \dots - \Phi_p B^p) Y_t = (1 - \Theta_1 B - \Theta_2 B^2 - \dots - \Theta_q B^q) \varepsilon_t$$

Overovanie stacionarity a invertovateľnosti sa robí podľa príslušných častí modelu.

Pri konštrukcii týchto modelov nám pomáha priebeh autokorelačnej (ACF) a parciálnej autokorelačnej funkcie (PACF), ktorý si vieme zobrazit' aj v programe Eviews 5.0, cez tzv. korelogram. Ak ACF rýchlo klesá k nule (monotónne alebo oscilujúco), a PACF sa po „p“ hodnotách rovná nule, toto signalizuje AR(p) proces. Pre MA(q) procesy to je presne opačne, teda PACF je nenulová a ACF je nulová po niekoľkých členoch (q).

Pri navrhovaní modelu pomocou korelogramu nie je niekedy celkom zrejmé, že koľko AR alebo MA členov by sme mali do modelu pridať. Všeobecne platí, že sa preferujú čo najjednoduchšie modely. Pre porovnanie modelov kde sa okrem veľkosti σ_ε^2 , čo je rozptyl bieleho šumu (my ho odhadne a podľa počtov „p“ a „q“ členov, označených ako „k“ a „l“) zohľadňuje aj penalizácia za pridávanie ďalších členov používa tzv. Akaikeho informačné kritérium (AIC).

$$AIC(k, l) = \ln \hat{\sigma}_{k,l}^2 + \frac{2(k+l)}{n}$$

Jeho modifikáciou získame tzv. Bayesovo informačné kritérium (BIC).

$$BIC(k, l) = \ln \hat{\sigma}_{k,l}^2 + \frac{\ln n (k+l)}{n}$$

Overenie modelu je veľmi dôležité. Overujeme taktiež prítomnosť jednotkového koreňa. Toto sa dá testovať pomocou tzv. Dickey-Fullerovho testu.

Hlavnou podstatou celej analýzy časových radov je navrhnutie modelu, ktorým by sme vedeli predikovať budúci vývoj. Pre overenie kvality týchto predikcií sa používa viacero kritérií. My si ich tu teraz uvedieme.

Prvým porovnaním kvality predikcií je tzv. odmocnina z priemernej štvorcovej chyby (root mean squared error).

$$RMSE = \sqrt{\frac{1}{h} \sum_{t=T+1}^{T+h} (C_t^f - C_t)^2}$$

Druhým porovnaním kvality predikcií je tzv. priemerná absolútna chyba (mean absolute error).

$$MAE = \frac{1}{h} \sum_{t=T+1}^{T+h} |C_t^f - C_t|$$

Tretím porovnaním kvality predikcií je tzv. priemerná absolútna percentuálna chyba (mean absolute percentage error)

$$MAPE = \frac{100}{h} \sum_{t=T+1}^{T+h} \left| \frac{C_t^f - C_t}{C_t} \right|$$

Štvrtým porovnaním kvality predikcií je tzv. Koefficient Theilovej nerovnosti (Theil inequality coefficient)

$$TIC = \frac{RMSE}{\sqrt{\frac{1}{h} \sum_{t=T+1}^{T+h} (C_t^f)^2 + \frac{1}{h} \sum_{t=T+1}^{T+h} C_t^2}}$$

Dynamické regresné modely alebo častejšie používaný ich anglický názov Autoregressive distributed lag models (ARDL) sú akýmsi spojením prístupov obyčajnej regresie, kde vysvetľovanú premennú vysvetľujeme iba pomocou vysvetľujúcich premenných, a prístupu ARMA modelov, kde sa zase snažíme vysvetliť priebeh jednej premennej pomocou jej minulých hodnôt.

Všeobecne môžeme zapísať takúto rovnicu ako:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \dots + \beta_p y_{t-p} + \delta_{1q} x_{1t-1} + \dots + \delta_{1q} x_{1t-q} + \dots + \delta_{k1} x_{kt-1} + \dots + \delta_{kq} x_{kt-q} + u_t$$

ARDL predpokladá nasledujúce podmienky:

- Podmienená stredná hodnota náhodnej premennej u_t vzhľadom ku všetkým lagom exogénnych premenných je rovná nule, t.j.

$$E[u_t | y_{t-1}, y_{t-2}, \dots, x_{1t-1}, x_{1t-2}, \dots, x_{kt-1}, \dots, x_{kt-2}] = 0$$

- Všetky premenné y, x_1, x_2, \dots, x_k sú stacionárne, ich rozdelenie má definované prvé štyri momenty, a pozorovania v čase „t“ a „t-j“ sú nezávislé, pre „j“ dosť vysoké (slabá závislosť)

Tieto modely sú vhodné, ak premenné „x“ nie sú ovplyvňované endogénnou premennou „y“

3.3. Panelová analýza

Používa sa pre dáta, ktoré majú časovú a aj priestorovú dimenziu. Z pohľadu dát sa delia na základné 2 skupiny. V prvej skupine máme relatívne málo jednotiek, ale relatívne veľa pozorovaní pre každú jednotku. Často sa označujú ako makro-panely, lebo sú pre ne typické makroekonomické dáta. Druhou skupinou sú také panely, ktoré majú relatívne veľa jednotiek, ale zase relatívne krátku časovú dimenziu. Tieto sú skôr typické pre firmy a preto sa zvyknú označovať ako mikro-panely. V našom prípade ide síce o dáta zo súkromného sektora, lenže povaha týchto dát je skôr podľa skupiny prvej, teda že máme menej jednotiek a relatívne veľkú dĺžku časovej dimenzie. Panel sa zvykne označovať ako tzv. balanced panel vtedy, ak pre každú jednotku máme k dispozícii rovnaké množstvo pozorovaní. Pre analýzu dát nášho typu je veľmi vhodným softvérom práve EViews, v ktorom sme robili doterajšie modely regresnej analýzy a časových radov.

Výhody tejto metodológie:

- Veľa pozorovaní
- Ohľad na nepozorovateľné faktory (omitted variable bias), ktoré sú konštantné buď pre jednotky alebo pre časové obdobie
- Lepšie štatistické vlastnosti odhadu

Nevýhody:

- Porovnateľnosť výsledkov nie je vždy jasná (poolability)
- Výsledky môžu byť skreslené koreláciou dát medzi prierezovými jednotkami
- Nie pri všetkých odhadoch je možné zahrnúť časovo konštantné premenné

3.3.1. Model s fixnými efektmi

Model s fixnými efektmi je veľmi často používaným modelom panelovej analýzy, a aj my ho aplikujeme na naše dáta. Využíva sa tu OLS-regresia s fixovanými efektmi:

$$y_{it} = \beta_1 x_i + \alpha_i + u_{it}$$

pričom všetky dáta sú v logaritmoch. Parameter α zahŕňa všetky efekty, ktoré sa pre prierezové jednotky v čase nezmenili. Člen $\alpha_i + u_{it}$ môže byť chápaný ako spoločný chybový člen. Podľa toho sa tento model tiež označuje ako „error component model“. Fixné efekty môžu predstavovať rôzne jednotky: krajiny, dvojice krajín, ale aj všeobecné efekty ako skupina krajín. Tieto informácie som čerpal zo študijných materiálov z predmetu seminár z ekonómie III, na FMFI UK v Bratislave, z roku 2012.

4. Praktická časť

V tejto kapitole si popíšeme konkrétne modely, ktoré sme zvolili pre analýzu vplyvu snehovej pokrývky na spotrebu plynu. Ako prvú sme urobili regresnú analýzu, kedy sa budeme snažiť vysvetliť priebeh spotreby pomocou regresorov počasia. Potom si ukážeme sofistikovanejší prístup cez analýzu časových radov, kde sa spotreba plynu bude modelovať aj pomocou predchádzajúcich hodnôt spotrieb. Napokon si ukážeme ešte analýzu cez metódy pre panelové dáta, tzv. pool modely.

4.1. Regresná analýza

Tento podkapitole sa budeme venovať o niečo menej ako nasledujúcej o časových radoch, nakoľko jej výsledky nie sú v praxi veľmi dobré v porovnaní s inými metódami pre analýzu našich dát. Regresné modely sa totižto veľmi neprispôsobujú novým údajom. Navyše spotreba plynu vykazuje dosť značnú mieru zotrvačnosti a preto bude lepšie pracovať s modelmi, ktoré zohľadňujú aj minulé spotreby a nie len regresory počasia. To bude v ďalšej podkapitole, tzv. dynamické modely.

V prvom rade si musíme vybrať dáta pre analýzu. Keďže máme viacero odberateľov, vyberieme si hlavne tých, ktorí sú z oblasti najmä severného Slovenska, pretože v týchto oblastiach sa nachádzalo najviac snehu (podľa intuície a aj dát). Pre ukážku výstupov z modelovania sme si vybrali odberateľa z Nižnej (Orava), ku ktorému sme priradili meteorologickú stanicu Oravský Podzámok. Ako obdobie sme si zvolili december 2006 – február 2007.

Pre začiatok uvažujem jednoduchý model (obr. 36), kde spotrebu modelujeme iba pomocou konštanty, teploty a vplyvu snehovej pokrývky. Snehovú pokrývku berieme tú, kde sa nesúvislá vrstva snehu berie ako žiadna snehová pokrývka. V tomto modeli sa nám potvrdil problém s rezíduami, nakoľko najmä spotreby plynu vykazujú veľkú mieru závislosti. Normalitu rezíduí síce nezamietame, ale ich vzájomné korelácie sú dosť veľké (obr. 37). Pre dokumentáciu uvádzame výstup z programu Eviews 5.0, ktorý budeme aj v ďalších modeloch používať. Pri jeho používaní sme sa opierali aj o materiál [11].

Obr. 36

Sample: 12/01/2006 2/28/2007
Included observations: 90

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	21265.12	215.3924	98.72738	0.0000
TEPL	-625.7564	51.50707	-12.14894	0.0000
SNEH1	38.77539	19.36549	2.002294	0.0484

R-squared	0.699782	Mean dependent var	20519.63
Adjusted R-squared	0.692880	S.D. dependent var	2755.651
S.E. of regression	1527.138	Akaike info criterion	17.53294
Sum squared resid	2.03E+08	Schwarz criterion	17.61627
Log likelihood	-785.9824	F-statistic	101.3946
Durbin-Watson stat	1.165888	Prob(F-statistic)	0.000000

Z tohto výstupu sa zamerajme najmä na odhady koeficientov rovnice modelu. Konštantný člen vyšiel 21265. To znamená, že ak by bolo práve 0 °C a žiaden sneh, spotreba by mala byť práve 21265 m³. V prípade, že sa teplota zvýši o 1 °C, spotreba by mala klesnúť o 625 m³ plynu. Ako vidíme, znamienko tohto koeficientu je záporné, čo je v súlade s našimi očakávaniami, lebo kladná zmena teploty vyvolá zápornú zmenu spotreby a naopak. Koeficient pri výške snehovej pokrývky je kladný a má hodnotu 38,7. To znamená, že ak existuje snehová pokrývka, spotreba plynu ešte narastie, a to s každým „cm“ o 38 m³. Toto je v rozpore s našimi očakávaniami, lebo sme práve naopak očakávali nižšiu spotrebu kôly izolačným vlastnostiam výšky snehovej pokrývky. Hodnota tohto koeficientu v porovnaní s koeficientom pri teplote je viac než 10 násobne nižšia, teda vplyv snehu nie je veľmi výrazný. Navyše p-hodnota pri koeficiente snehu je takmer na hranici 5%, teda takmer by sme ho mohli z modelu vyhodiť. Nie je veľmi signifikantný.

Obr. 37

Sample: 12/01/2006 2/28/2007
Included observations: 90

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
		1	0.402	0.402	15.071	0.000
		2	0.081	-0.097	15.683	0.000
		3	-0.214	-0.254	20.048	0.000
		4	-0.316	-0.162	29.648	0.000
		5	-0.136	0.088	31.446	0.000
		6	0.044	0.064	31.640	0.000
		7	0.278	0.185	39.328	0.000
		8	0.167	-0.098	42.155	0.000
		9	0.035	-0.035	42.278	0.000
		10	-0.139	-0.060	44.265	0.000
		11	-0.245	-0.073	50.581	0.000
		12	-0.139	0.016	52.633	0.000
		13	0.013	0.045	52.650	0.000

Na obrázku č. 37 máme v stĺpci AC vypísané hodnoty autokorelácií rezíduí. Vidíme, že tieto korelácie sú dosť silné, prvá z nich je až 0,40. Teda nie je prekvapením, že v poslednom stĺpci sme dostali p-hodnoty blízke nule, čo znamená, že zamietame nulovú hypotézu o tom, že prísušné korelácie sú nulové. O tom, že tieto autokorelácie sú dosť veľké hovorí aj Durbin-Watson štatistika, ktorá vyšla 1,16, čo je dosť ďaleko od 2. V praxi to znamená, že predikcie na základe tohto modelu by neboli veľmi dôveryhodné.

Ako ďalšie vylepšenie modelu sme vyskúšali zlogaritmováť spotreby plynu. (obr. 38) Tento model navrhol pán Jenkins[7] (pôvodne pre modelovanie spotreby elektrickej energie v závislosti od teploty), pričom my do neho zatiaľ nezavedieme žiadne autoregresné členy.

Obr. 38

Sample: 12/01/2006 2/28/2007				
Included observations: 90				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.960892	0.010917	912.4151	0.0000
TEPL	-0.032367	0.002611	-12.39815	0.0000
SNEH1	0.001483	0.000982	1.510608	0.1345
R-squared	0.699721	Mean dependent var	9.919793	
Adjusted R-squared	0.692819	S.D. dependent var	0.139655	
S.E. of regression	0.077402	Akaike info criterion	-2.246836	
Sum squared resid	0.521227	Schwarz criterion	-2.163509	
Log likelihood	104.1076	F-statistic	101.3655	
Durbin-Watson stat	1.150945	Prob(F-statistic)	0.000000	

S týmto modelom sme si taktiež zatiaľ veľmi nepolepšili z pohľadu rezíduí, ktoré sú síce opäť normálne, ale ich autokorelácie sú dosť podobné tomu modelu bez logaritmu spotreby. V tomto modeli sa dá veľmi pekne interpretovať odhad koeficientov pri teplote a snehu. Ak sa nám totižto zvýši teplota o 1°C, spotreba plynu by mala klesnúť o 3,23%. Podobne ak nám narastie výška snehovej pokrývky o 1cm, spotreba plynu by mala stúpnuť o 0,148%. Opäť sa potvrdilo, že vplyv snehu je takmer zanedbateľný faktor, a to nie len podľa hodnoty príslušného koeficientu, ale aj podľa p-hodnoty, podľa ktorej je táto premenná nesignifikantná. Musíme však opäť konštatovať, že smer vplyvu je opačný, než sme očakávali.

Ďalej sme sa rozhodli, že sa pokúsime tieto premenné diferencovať, čím by sme chceli odstrániť alebo aspoň znížiť autokorelovanosť rezíduí. To nám aj v celku podarilo,

o čom svedčí aj hodnota Durbin-Watson štatistiky, ktorá je už relatívne blízka hodnote 2. Tento výstup je na nasledujúcom obrázku (obr. 39)

Obr.39

Sample (adjusted): 12/02/2006 2/28/2007
Included observations: 89 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001130	0.008557	0.132012	0.8953
D(TEPL)	-0.023568	0.003585	-6.574572	0.0000
D(SNEH1)	0.000984	0.002501	0.393563	0.6949
R-squared	0.335409	Mean dependent var	0.001973	
Adjusted R-squared	0.319954	S.D. dependent var	0.097880	
S.E. of regression	0.080717	Akaike info criterion	-2.162618	
Sum squared resid	0.560305	Schwarz criterion	-2.078732	
Log likelihood	99.23652	F-statistic	21.70149	
Durbin-Watson stat	2.300615	Prob(F-statistic)	0.000000	

Z výstupu je pre nás podstatné to, že znamienka koeficientov sa nezmenili a opäť sa potvrdilo, že vplyv snehu je zanedbateľný a ešte aj nesignifikantný.

Ďalej sme sa pokúsili model vylepšiť o pridané regresory, a to konkrétne vietor a dĺžku slnečného svitu. Všetko nechávame v diferenciách a spotreba je ako predtým navyše logaritmovaná. Dostali sme nasledovný výstup (obr. 40).

Obr. 40

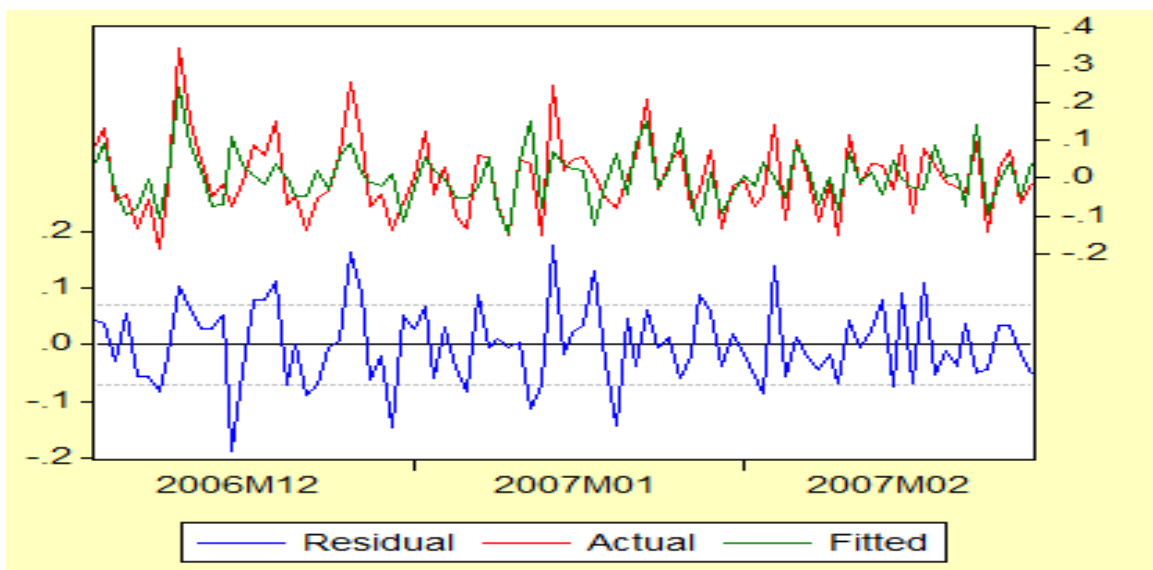
Sample (adjusted): 12/02/2006 2/28/2007
Included observations: 89 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-5.73E-05	0.007471	-0.007663	0.9939
D(TEPL)	-0.029817	0.003387	-8.803741	0.0000
D(SNEH1)	-0.000577	0.002204	-0.261655	0.7942
D(VIET)	0.018051	0.018287	0.987102	0.3264
D(SVIT)	-0.015961	0.002973	-5.368514	0.0000
R-squared	0.505622	Mean dependent var	0.001973	
Adjusted R-squared	0.482080	S.D. dependent var	0.097880	
S.E. of regression	0.070441	Akaike info criterion	-2.413545	
Sum squared resid	0.416801	Schwarz criterion	-2.273734	
Log likelihood	112.4028	F-statistic	21.47762	
Durbin-Watson stat	2.172948	Prob(F-statistic)	0.000000	

Forma modelu ostala rovnaká, iba sme pridali nové vysvetľujúce premenné. To spôsobilo nárast fitu modelu, čo vidíme na zvýšenej hodnote R-squared. Pridaním nových premenných sa navyše znížila hodnota koeficientu pre sneh, čo ešte viac potvrdzuje nevýrazný vplyv snehu na spotrebu. Takisto aj príslušná p-hodnota sa zväčšila, takže o to viac sme presvedčený, že tento koeficient je nesignifikantný. Hodnota autokorelácií rezíduí

sa taktiež o niečo zlepšila, o čom svedčí znížená hodnota Durbin-Watson štatistiky. Keď sa pozrieme na hodnoty koeficientov pri premenných vietor a dĺžka slnečného svitu, ich veľkosť je výrazne väčšia než pri snehu. To znamená, že na spotrebu vplyvajú oveľa výraznejšie ako napríklad nami analyzovaný sneh. Čo sa týka ich zmanienok, je to v súlade s našou intuíciou, nakoľko koeficient pri vetre je kladný (čím silnejší vietor fúka, tým je spotreba väčšia) a koeficient pri dĺžke slnečného svitu je záporný (čím dlhšie svieti slnko, tým je spotreba plynu nižšia). Podľa p-hodnoty je síce koeficient pri vetre nesignifikantný, to však môže byť spôsobené tým, že v danej lokalite fúkal počas tohto obdobia veľmi slabí vietor (väčšinou 0,3 alebo 1 m/s), čo neumožnilo ukázať jeho vplyv na spotrebu. K tomuto modelu ešte pripájam výstup (obr. 41) o priebehu rezíduí (nižšie) a priebeh premennej $d(\log(\text{spot}))$ s jej fitovanými hodnotami našim modelom, pričom „d“ znamená diferencovanie, „log“ je v skutočnosti prirodzený logaritmu premennej „spot“, čo je spotreba plynu v m^3 pre daného odberateľa.

Obr. 41



Ďalej sme sa rozhodli skúsiť analyzovať vplyv snehovej pokrývky pomocou tzv. dummy premenných. Premenná označená ako D2 nadobúda hodnotu „1“ práve vtedy, keď je v daný deň snehová pokrývka o výške 1, 2 alebo 3 cm. Inak nadobúda hodnotu „0“. Premenná D3 nadobúda hodnotu „1“ práve vtedy, keď je výška snehovej pokrývky 4 a viac cm, inak nadobúda hodnotu „0“. Premennú D1, ktorá by popisovala nulovú výšku snehovej pokrývky sme do modelu z dôvodu multikolinearity nemohli dať. Tento model

definujeme ako $\log(\text{spot})$, ktorú chceme vysvetliť cez teplotu, D2, D3, vietor a svit. V programe sme dostali nasledovný výstup (obr. 42).

Obr. 42

Sample: 12/01/2006 2/28/2007
Included observations: 90

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.970223	0.020604	483.9028	0.0000
TEPL	-0.035355	0.002511	-14.08185	0.0000
D2	0.017535	0.018974	0.924133	0.3581
D3	0.012308	0.020516	0.599908	0.5502
VIET	0.028822	0.018855	1.528631	0.1301
SVIT	-0.016945	0.003569	-4.748470	0.0000
R-squared	0.776032	Mean dependent var	9.919793	
Adjusted R-squared	0.762700	S.D. dependent var	0.139655	
S.E. of regression	0.068031	Akaike info criterion	-2.473375	
Sum squared resid	0.388767	Schwarz criterion	-2.306721	
Log likelihood	117.3019	F-statistic	58.21055	
Durbin-Watson stat	1.114056	Prob(F-statistic)	0.000000	

V tomto modeli máme síce autokorelované rezíduá, ale keď sa pozrieme na odhady koeficientov a ich p-hodnoty, tak dostaneme dosť dobrý prehľad o vplyve jednotlivých faktorov. Premennú D2 by sme teraz mohli interpretovať tak, že ak sa zmení výška snehovej pokrývky z „0 cm“ na „1, 2 alebo 3 cm“, tak by mala narásť spotreba plynu o 1,7%. Ak by sme z nuly skočili rovno na „4 a viac cm“ snehu, tak by spotreba mala narásť iba o 1,2%. Ale keďže príslušné p-hodnoty vyšli vyššie ako 5%, tieto koeficienty sú v modeli prakticky bezvýznamné a bolo ich možné z modelu vyhodiť.

Fakt, že hodnoty koeficientov pri premenných snehu nám vyšli opačné v porovnaní s očakávaniami by sa dal napr. vysvetliť tak, že izolačné účinky snehu by sa mohli prejavovať pri nižších teplotách (vyšších záporných, napr. -15°C), ktoré v našich dátach bohužiaľ nemáme. Ďalej by bolo zaujímavé preskúmať nejaké špeciálne efekty, ako by „dobré“ a „zlé“ počasie. To je však veľmi subjektívne, i keď dalo by sa očakávať že väčšine ľudí z populácie by vyhovovalo slnečné počasie, ktoré by bolo pre nich dobré. Ak by bol stav počasia pre nich „dobrý“, mohli by sa rozhodnúť nesediť doma, ale radšej svoj čas tráviť v prírode (lyžovačka,...) a vtedy by spotreba plynu mohla klesnúť. Týmto by sa dalo relatívne odôvodniť to, že ak je málo snehu (premenná D2), ľudia radšej ostanú doma, čím spotreba plynu ešte narastie. O tomto by sa dalo ešte dlho polemizovať, takže to radšej nechajme pre iných, ktorý budú mať kvalitnejšie dáta a môžu tieto súvislosti preskúmať. Pre nás je podstatné to, že sneh podľa tejto analýzy nehrá významnú úlohu.

Ďalej sme sa v modeloch pokúsili jemne pozmeniť priebeh premennej o výške snehu. Bolo to urobené dvomi rôznymi spôsobmi. Prvý z nich bol ten, že sme nesúvislú vrstvu snehu charakterizovali nie nulovou hodnotou ako doteraz, ale ako výšku „1 cm“. Tým sme získali o niekoľko pozitívnych pozorovaní čo do snehu navyše. Výsledky týchto odhadov sa takmer vôbec nezmenili. Je to vidieť z nasledujúceho obr. 43, ktorý je možné porovnať s obr. 36.

Obr. 43

Sample: 12/01/2006 2/28/2007
Included observations: 90

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	21250.05	218.0103	97.47268	0.0000
TEPL	-624.2534	51.58388	-12.10171	0.0000
SNEH2	40.04449	19.57177	2.046033	0.0438

Druhou možnou zmenou bola idea časového posunu premennej sneh. Dôvodom je časová nekonzistentnosť plynúca zo spôsobu zberu týchto dát (viac v kapitole 2). Urobili sme teda priemer dnešnej hodnoty snehovej pokrývky so zajtrajšou hodnotou a priradili sme ju ku dnešnej hodnote o výške snehovej pokrývky. Potom sme odhadli rovnaký model ako v predchádzajúco prípade a dostali sme nasledovný výstup (obr. 44).

Obr. 44

Sample: 12/01/2006 2/28/2007
Included observations: 90

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	21230.07	218.5838	97.12553	0.0000
TEPL	-617.5099	52.24918	-11.81856	0.0000
SNEH3	43.46130	20.00211	2.172835	0.0325

Vidíme, že táto zmena sa už prejavila výraznejšie pri odhade parametrov, než len zmena pohľadu na nesúvislú vrstvu snehovej pokrývky.

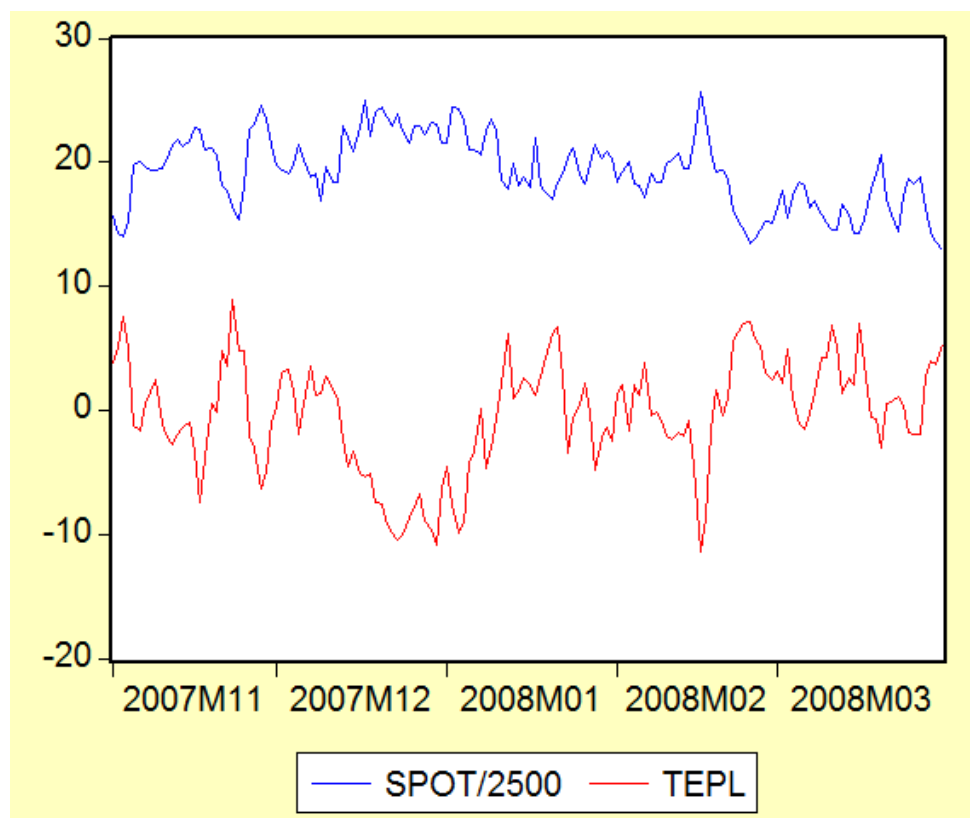
Ďalej sme sa pozreli aj na iné lokality, kde boli iní odberatelia a iné meteorologické priebehy premenných, no avšak ani v týchto prípadoch sa nám nepotvrdil výrazný vplyv výšky snehovej pokrývky na spotrebu plynu. Avšak v niektorých prípadoch sme už dostali negatívne znamienka koeficientov pri snehu, čo je v zhode s našimi očakávaniami. Ďalej sme zistili, že efekt vetra hrá dosť významnú a signifikantnú úlohu.

4.2. Časové rady a dynamické regresné modely

V tejto podkapitole sa pozrieme na praktickú aplikáciu teórie časových radov pre analýzu našich dát. V zásade platí, že čím dlhší časový rad berieme do úvahy, tým sú naše modely lepšie. Preto sme sa pokúsili nájsť takého odberateľa, ktorý sa správal v zmysle zimného modelu čo najdlhšie, t.j. že teploty v danej lokalite boli čo najdlhšie najnižšie, aby sme mohli pridať vplyv snehovej pokrývky do modelu. I napriek nie moc vydareným zimám, či už čo sa týka extrémne nízkych teplôt alebo nízkej a krátkodobej snehovej pokrývky sme mali šťastie aspoň v tom, že zima 2007/2008 sa začala už v novembri, kedy prvý sneh spadol už 7.11.2007 a 12.11.2007 napadlo ďalších 8cm (na stanici Poprad). Vybrali sme si teda na ukážku odberateľa zo Svitú. Pri tomto výbere sme taktiež prihliadali na hodnotu korelačného koeficientu uvedeného v tab. 05 z druhej kapitoly.

Najprv sa pozriem na priebeh časového radu spotreby plynu. Ten si zobrzíme na obr. 45. Na osi „X“ je čas idúci od začiatku novembra 2007 po koniec marca 2008, t.j. spolu 5 mesiacov denných údajov. Na osi „Y“ máme hodnoty spotrieb v m³ predelené 2500 (kvôli ukážke vplyvu teploty) pre daného odberateľa (hore, modrou farbou) a taktiež aj hodnoty teplôt (dole, červenou farbou) v danom čase. Z obr. 45 je krásne vidieť opačné zmeny teplôt a spotrieb.

Obr. 45



Priebeh spotreby sa nám nezdá byť veľmi nestacionárny. Zobrazme si však priebeh korelogramu pre spotrebu. Dostaneme nasledovný výstup (obr. 46)

Obr. 46

Sample: 11/01/2007 3/31/2008
Included observations: 152

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob	
		1	0.802	0.802	99.763	0.000
		2	0.614	-0.084	158.54	0.000
		3	0.503	0.103	198.28	0.000
		4	0.430	0.036	227.58	0.000
		5	0.375	0.029	250.00	0.000
		6	0.365	0.121	271.40	0.000
		7	0.335	-0.040	289.47	0.000
		8	0.252	-0.114	299.79	0.000
		9	0.199	0.044	306.25	0.000
		10	0.167	-0.010	310.85	0.000
		11	0.156	0.047	314.87	0.000
		12	0.165	0.052	319.40	0.000
		13	0.206	0.103	326.52	0.000
		14	0.237	0.061	336.07	0.000
		15	0.235	-0.004	345.51	0.000
		16	0.215	-0.020	353.43	0.000
		17	0.218	0.072	361.71	0.000
		18	0.206	-0.048	369.09	0.000
		19	0.182	-0.027	374.94	0.000
		20	0.167	-0.014	379.88	0.000

Vidíme, že hodnoty autokorelačnej funkcie na začiatku klesajú dosť rýchlo, ale potom sa to výrazne spomalilo a prestali klesať k nule. Podľa všetkého budeme musieť urobiť ešte prvé diferencie. No podľa PACF by to mohol byť AR(1) proces. Skúsime ho odhadnúť (obr. 47) a skontrolovať, či je splnená podmienka stacionarity (prevrátené hodnoty koreňov príslušného polynómu musia ležať vnútri jednotkového kruhu) a či sú rezíduá nekorelované (obr. 48).

Obr. 47

Sample (adjusted): 11/02/2007 3/31/2008
Included observations: 151 after adjustments
Convergence achieved after 4 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	47955.18	2212.825	21.67148	0.0000
AR(1)	0.857530	0.043880	19.54243	0.0000
R-squared	0.719348	Mean dependent var		48228.35
Adjusted R-squared	0.717464	S.D. dependent var		7281.070
S.E. of regression	3870.186	Akaike info criterion		19.37315
Sum squared resid	2.23E+09	Schwarz criterion		19.41311
Log likelihood	-1460.673	F-statistic		381.9064
Durbin-Watson stat	1.794688	Prob(F-statistic)		0.000000

Inverted AR Roots .86

Z tohto výstupu vidíme, že koeficienty modelu sú signifikantné a že model je stacionárny, nakoľko v spodnej časti výstupu máme hodnotu invertovaného koreňa, ktorý leží vnútri jednotkového kruhu, lebo jeho hodnota je menšia ako 1. Ešte sme sa pozreli na normalitu rezíduí a tá bola tesne splnená.

Obr. 48

Sample: 11/02/2007 3/31/2008
 Included observations: 151
 Q-statistic probabilities adjusted for 1 ARMA term(s)

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.088	0.088	1.1904	
		2 -0.121	-0.130	3.4528	0.063
		3 -0.080	-0.058	4.4459	0.108
		4 0.016	0.013	4.4841	0.214
		5 -0.096	-0.119	5.9441	0.203
		6 0.090	0.113	7.2275	0.204
		7 0.151	0.113	10.875	0.092
		8 -0.087	-0.112	12.089	0.098
		9 -0.021	0.055	12.160	0.144
		10 -0.037	-0.063	12.380	0.193
		11 -0.010	-0.002	12.396	0.259
		12 -0.091	-0.074	13.781	0.245
		13 0.040	0.000	14.047	0.298
		14 0.105	0.097	15.911	0.254
		15 0.064	0.053	16.596	0.278
		16 -0.068	-0.062	17.390	0.296
		17 0.055	0.107	17.911	0.329
		18 0.084	0.071	19.143	0.320
		19 0.025	0.051	19.251	0.377
		20 0.041	0.046	19.541	0.423

Z tohto výstupu je podstatná autokorelovanosť rezíduí, ktorá ako vidíme je naozaj veľmi nízka, čo nám potvrdzujú aj p-hodnoty, ktoré sú vyššie ako 5%, teda nezamietame nulovú hypotézu o nulovosti autokorelácií, teda ich môžeme považovať za nekorelované. Teda zatiaľ môžeme potvrdiť prítomnosť bieleho šumu v modeli. Ale ešte sa pozrime aj na druhé mocniny týchto rezíduí (obr. 49).

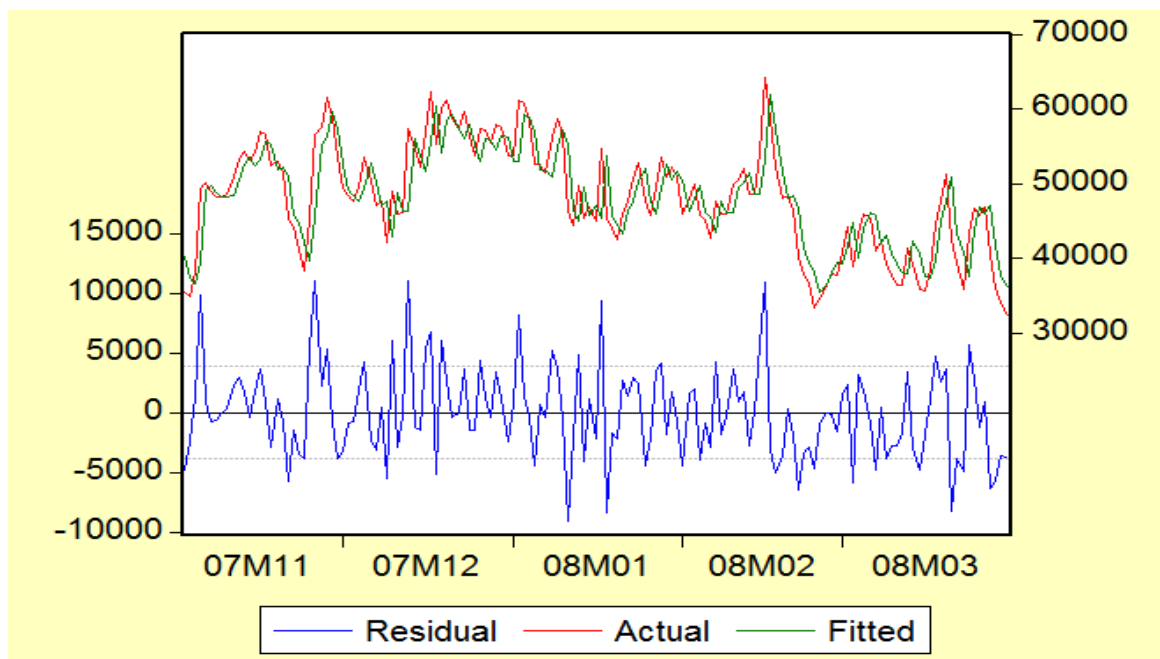
Obr. 49

Sample: 11/02/2007 3/31/2008
 Included observations: 151
 Q-statistic probabilities adjusted for 1 ARMA term(s)

Autocorrelation	Partial Correlation	AC	PAC	Q-Stat	Prob
		1 0.017	0.017	0.0421	
		2 -0.056	-0.056	0.5214	0.470
		3 0.073	0.075	1.3577	0.507
		4 -0.022	-0.029	1.4365	0.697
		5 -0.089	-0.080	2.6906	0.611
		6 0.100	0.097	4.2723	0.511
		7 -0.008	-0.019	4.2817	0.639
		8 -0.097	-0.077	5.8040	0.563
		9 0.022	0.009	5.8807	0.661
		10 -0.058	-0.071	6.4315	0.696
		11 -0.131	-0.103	9.2683	0.507
		12 -0.120	-0.143	11.662	0.390
		13 -0.068	-0.085	12.427	0.412
		14 -0.051	-0.039	12.871	0.458
		15 0.000	-0.017	12.871	0.537
		16 0.030	0.012	13.023	0.601
		17 0.053	0.057	13.508	0.635
		18 -0.099	-0.108	15.200	0.581
		19 -0.083	-0.102	16.400	0.565
		20 0.060	0.024	17.024	0.588

Vidíme, že tento model je veľmi jednoduchý, a splňa všetky potrebné požiadavky. Na nasledujúcom grafe (obr. 50) si demonštrujeme fit tohto modelu s priebehom rezíduí v dolnej časti. V hornej časti sú pri sebe 2 priebehy dát, pričom ten vykreslený červenou farbou je označený ako Actual, teda skutočné hodnoty, no a ten druhý priebeh, vykreslený zelenou farbou sú nami odhadnuté hodnoty modelu, tzv. fitted values. Na to, aký je tento model zatiaľ jednoduchý, je jeho fit veľmi dobrý.

Obr. 50



Teraz si skúsme model vylepšiť o vysvetľujúce premenné. Ako prvú skúsme pridať teplotu. Z programu sme dostali nasledujúci výstup (obr. 51).

Obr. 51

Sample (adjusted): 11/02/2007 3/31/2008
 Included observations: 151 after adjustments
 Convergence achieved after 8 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	47656.55	838.7985	56.81525	0.0000
TEPL	-1108.794	90.60606	-12.23753	0.0000
AR(1)	0.729772	0.059383	12.28932	0.0000
R-squared	0.855954	Mean dependent var	48228.35	
Adjusted R-squared	0.854007	S.D. dependent var	7281.070	
S.E. of regression	2782.026	Akaike info criterion	18.71941	
Sum squared resid	1.15E+09	Schwarz criterion	18.77936	
Log likelihood	-1410.316	F-statistic	439.7232	
Durbin-Watson stat	2.318783	Prob(F-statistic)	0.000000	
Inverted AR Roots	.73			

Vidíme že hodnoty parametrov sú zrozumiteľné a ľahko interpretovateľné. Napríklad ak sa teplota zvýši o 1°C, tak spotreba plynu by sa mala znížiť o 1108 m³. P-hodnoty vyšli pekne signifikantné a model je stacionárny, čo vidieť aj z toho, že invertovaný koreň leží vnútri jednotkového koreňa. Koeficient determinácie sa taktiež zvýšil, čo svedčí o lepšom fite modelu. Skontroloval som aj normalitu rezíduí, bola splnená ale čo nebolo splnené je nekorelovanosť rezíduí. I keď samotné hodnoty ACF vyšli pomerne malé, príslušné p-hodnotu pre Q-štatistiku už boli menšie ako 5%, teda nulovú hypotézu o nekorelovanosti zamietame a teda rezíduá nie sú bielym šumom. Pri pohľade na fit modelu je jasne vidieť, že sa zlepšil, keďže fitované hodnoty na mnohých miestach úplne splynuli so skutočnými hodnotami. Teraz sa skúsme pozrieť na to, ako by sa dala korelovanosť rezíduí odstrániť.

Navrhujem zdiferencovať pridaný časový rad teploty, lebo pravdepodobne on je na vine a spôsobil nekorelovanosť rezíduí. Dostali sme nasledovný výstup (obr. 52) a pri kontrole rezíduí sa potvrdila nekorelovanosť rezíduí.

Obr. 52

Sample (adjusted): 11/03/2007 3/31/2008
 Included observations: 150 after adjustments
 Convergence achieved after 9 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	48168.12	2285.962	21.07127	0.0000
D(TEPL)	-454.8022	83.95693	-5.417090	0.0000
AR(1)	0.873718	0.041614	20.99587	0.0000
R-squared	0.764245	Mean dependent var		48312.97
Adjusted R-squared	0.761037	S.D. dependent var		7230.575
S.E. of regression	3534.581	Akaike info criterion		19.19837
Sum squared resid	1.84E+09	Schwarz criterion		19.25859
Log likelihood	-1436.878	F-statistic		238.2640
Durbin-Watson stat	1.780950	Prob(F-statistic)		0.000000
Inverted AR Roots	.87			

Teraz poďme tento model vylepšiť o premennú sneh4 (obr. 53), čo je mierne upravený priebeh dát tak, že nesúvislú vrstvu snehovej kvantifikujeme ako výšku snehu „1 cm“ a potom ešte urobíme priemer dnešných hodnôt so zajtrajšími a toto tvorí dnešné upravené hodnoty priebehu výšky celkovej snehovej pokrývky.

Obr. 53

Sample (adjusted): 11/03/2007 3/31/2008
 Included observations: 150 after adjustments
 Convergence achieved after 9 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	47412.59	2149.952	22.05286	0.0000
D(TEPL)	-452.3083	83.42060	-5.422022	0.0000
SNEH4	480.8005	233.0534	2.063049	0.0409
AR(1)	0.865161	0.043028	20.10683	0.0000
R-squared	0.770858	Mean dependent var	48312.97	
Adjusted R-squared	0.766149	S.D. dependent var	7230.575	
S.E. of regression	3496.568	Akaike info criterion	19.18326	
Sum squared resid	1.78E+09	Schwarz criterion	19.26354	
Log likelihood	-1434.744	F-statistic	163.7197	
Durbin-Watson stat	1.820795	Prob(F-statistic)	0.000000	
Inverted AR Roots	.87			

Model je stacionárny, rezíduá sú normálne, nekorelované a aj v druhých mocninách. Fit modelu sa málo zlepšil, keďže sa zvýšila hodnota R-squared. Pozrime sa však na koeficienty modelu. Diferencia teploty má záporný odhad koeficientu, čo je v súlade s logikou, keďže nárast teploty musí spôsobiť pokles spotreby plynu. Koeficient pri snehu je tiež kladný, čo je zhodný výsledok s modelmi z klasickej regresnej analýzy. I keď sme očakávali skôr záporné znamienko, podľa teórie o izolačných vplyvoch, stále máme podozrenie, že by sa táto premenná mohla z modelu úplne vylúčiť. Jej p-hodnota vyšla síce pod 5%, teda jej nesignifikantnosť sa zamietá, ale na hladine 99% by už bol nesignifikantný.

Podme si tento model ešte trochu vylepšiť o ďalšie vysvetľujúce premenné, a to vietor a slnečný svit (obr. 54). V klasickej regresnej analýze nám signifikantne vylepšili model a efekt snehu sa ešte znížil.

Obr. 54

Sample (adjusted): 11/03/2007 3/31/2008
 Included observations: 150 after adjustments
 Convergence achieved after 10 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	47740.44	2217.823	21.52581	0.0000
D(TEPL)	-433.7084	88.40385	-4.905989	0.0000
SNEH4	491.9068	234.7598	2.095362	0.0379
VIET	-107.5316	125.6081	-0.856088	0.3934
SVIT	15.07387	80.22510	0.187895	0.8512
AR(1)	0.863905	0.044020	19.62550	0.0000
R-squared	0.772078	Mean dependent var	48312.97	
Adjusted R-squared	0.764164	S.D. dependent var	7230.575	
S.E. of regression	3511.382	Akaike info criterion	19.20459	
Sum squared resid	1.78E+09	Schwarz criterion	19.32501	
Log likelihood	-1434.344	F-statistic	97.55887	
Durbin-Watson stat	1.816348	Prob(F-statistic)	0.000000	
Inverted AR Roots	.86			

Ešte predtým než sa vyjadríme k odhadnutým koeficientom, musíme konštatovať, že model je stacionárny, rezíduá sú normálne, a taktiež sú nekorelované, ako aj ich druhé mocniny. Keď sa pozrieme na znamienka odhadnutých koeficientov, musíme konštatovať, že sú presne opačné, ako by sme očakávali (vietor, svit, sneh). Taktiež nás prekvapil fakt, že vplyv snehu sa zdá byť ešte viac signifikantný, kým vplyv vetra a slnečného svitu nie.

Potom sme ešte skúsili niekoľko ďalších modelov, ale vplyv vetra a svitu sa nám nepodarilo potvrdiť, kým vplyv snehu áno, ale s tým, že ak je sneh, tak spotreba ešte narastá. Preto sa ešte pozrime na posledný spôsob analýzy a to panelovú analýzu s fixnými efektmi.

4.3. Panelová analýza s fixnými efektmi

Do tejto analýzy sme vybrali 8 odberateľov, ktorých spotreba je dosť podobná. Týmto sme chcelo doceliť čo najmenšie rozdiely vo fixných efektoch jednotlivých odberateľov. Tým pádom bude mať interpretácie koeficientov jednotlivých premenných presnejší fit pre jednotlivého odberateľa. Kritérium voľby týchto 8 odberateľov bolo okrem relatívne podobnej spotreby aj rozdielnosť lokality, čím sme zabezpečili pre každého odberateľa jedinečné údaje o počasi. Taktiež sme sa museli pozrieť aj na to, či sa tento odberateľ správa podľa počasia, alebo na jeho zmeny nereaguje. Posledným faktorom pre výber odberateľa bola dostupnosť všetkých meteorologických charakteristík, ktoré sme v práci zmieňovali, lebo napr. pri meteorologickej stanici Beluša neboli dostupné informácie o dĺžke slnečného svitu. Takže ako prienik týchto kritérií sme dostali týchto 8 odberateľov:

_badv ...Bratislava Devínska Nová Ves

_bap1 ...Bratislava Petržalka

_skal ...Skalica

_nitr ... Nitra

_stur ...Štúrovo

_nizn ...Nižná

_rsob ... Rimavská Sobota

_po1 ... Prešov

Ďalej budeme pracovať so snehovými údajmi, kde nesúvislá vrstva snehu je charakterizovaná nulovou výškou snehovej pokrývky. Tieto hodnoty nie sú nijako

priemerované s budúcimi hodnotami snehovej pokrývky. Výšku snehu sme si rozdelili podľa tzv. dummy premenných a to tak, že D2 nadobúda hodnoty „1“ práve vtedy, keď je snehová pokrývka o výške 1, 2 alebo 3 cm. Premenná D3 nadobúda hodnoty „1“ práve vtedy, keď je 4 a viac cm snehu.

Prvým modelom je veľmi jednoduchí prípad, kedy spotrebu modelujeme priamo pomocou vysvetľujúcich premenných nijako neupravených. Po spustení modelu sme dostali nasledovný výstup (obr. 55)

Obr. 55

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	22998.08	126.9936	181.0964	0.0000
TEPL?	-980.5574	15.94364	-61.50147	0.0000
VIET?	236.8852	34.53368	6.859540	0.0000
SVIT?	-155.8510	19.99968	-7.792675	0.0000
D2?	125.4975	198.2378	0.633065	0.5268
D3?	563.5024	264.3617	2.131558	0.0332
Fixed Effects (Cross)				
_BADV--C	4530.796			
_BAP1--C	2933.234			
_SKAL--C	-4745.916			
_NITR--C	-2693.859			
_STUR--C	-2707.097			
_NIZN--C	475.7033			
_RSOB--C	9259.745			
_PO1--C	-7052.606			
Effects Specification				
Cross-section fixed (dummy variables)				
R-squared	0.913866	Mean dependent var	21157.54	
Adjusted R-squared	0.913007	S.D. dependent var	6862.678	
S.E. of regression	2024.122	Akaike info criterion	18.07429	
Sum squared resid	4.93E+09	Schwarz criterion	18.12885	
Log likelihood	-10976.17	F-statistic	1063.632	
Durbin-Watson stat	0.720114	Prob(F-statistic)	0.000000	

Vidíme, že tento odhad má veľmi dobrý fit. Konštantný člen je interpretuje teraz pre každého odberateľa individuálne a to tak, že napr. pre prvého v poradí odberateľa _BADV sa hodnota spoločnej konštanty (úplne hore) sčíta s hodnotou konštanty príslušného odberateľa, teda u nás 22998 + 4530. Analogicky sa vypočíta hodnota konštantného člena

pre každého odberateľa. Keď sa pozrieme na znamienka jednotlivých koeficientov, tak musíme skonštatovať v podstate iba jediné prekvapenie a to v prípade snehu. Opäť nám odhady pre dummy premenné vyšli kladné, čo sa znamená kladný vplyv na spotrebu plynu. Teda ak máme sneh, spotreba plynu bude ešte väčšia. Musíme sa však taktiež pozrieť na príslušné p-hodnoty a vidíme, že vplyv snehu by sme podľa nich mohli v modeli zanedbať podľa D2.

Ďalej sme sa rozhodli logaritmovať (prirodzeným logaritmom) všetky premenné okrem dummy premenných. Navyše pri logaritmovaní teploty sme k týmto hodnotám ešte pripočítali +20°C, aby sme sa vyhli logaritmovaniu záporných hodnôt. Dostali sme nasledovný výstup (obr. 56).

Obr. 56

Sample: 11/01/2007 3/31/2008

Included observations: 139

Cross-sections included: 8

Total pool (unbalanced) observations: 784

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.00123	0.063727	204.0137	0.0000
LOG(TEPL?+20)	-1.009439	0.020416	-49.44301	0.0000
LOG(VIET?)	0.036376	0.005572	6.528459	0.0000
LOG(SVIT?)	-0.017502	0.003029	-5.777991	0.0000
D2?	-0.024705	0.012688	-1.947184	0.0519
D3?	0.008466	0.018347	0.461451	0.6446
Fixed Effects (Cross)				
_BADV--C	0.226749			
_BAP1--C	0.153697			
_SKAL--C	-0.223264			
_NITR--C	-0.119683			
_STUR--C	-0.125830			
_NIZN--C	0.008737			
_RSOB--C	0.372740			
_PO1--C	-0.339407			
Effects Specification				
Cross-section fixed (dummy variables)				
R-squared	0.911622	Mean dependent var	9.867882	
Adjusted R-squared	0.910247	S.D. dependent var	0.307446	
S.E. of regression	0.092107	Akaike info criterion	-1.915284	
Sum squared resid	6.540964	Schwarz criterion	-1.837941	
Log likelihood	763.7914	F-statistic	662.7431	
Durbin-Watson stat	1.158670	Prob(F-statistic)	0.000000	

Tieto logaritmované premenné sa dajú vysvetliť cez percentuálne zmeny, t.j. napr. aj sa zvýši rýchlosť vetra o 1%, spotreba by sa mala zvýšiť 0,03%. Opäť sa pozrieme na znamienka pri jednotlivých koeficientoch. Zistíme, že okrem snehu nám vyšli v súlade

s očakávaniami. Teraz však nastala malá zmena aj pri premennej D2, kde vyšiel koeficient už záporný, teda aj je málo snehu (1, 2 alebo 3 cm), tak by spotreba plynu mala klesnúť. To je síce v rozpore s našou domnienkou, že izolačné účinky by sa mohli skôr prejaviť až pri vyššej snehovej pokrývke, ale vzhľadom na príslušné p-hodnoty musíme konštatovať, že vplyv snehu je opäť nesignifikantný. I keď pri D2 sa signifikancia výrazne zlepšila.

Teraz dodajme tomuto modelu ešte jedno vylepšenie. Skúsme modelovať vplyv teploty nelineárne, pridaním vysvetľujúcej premennej $(\log(\text{tepl?}+20))^2$. Potom dostaneme nasledovný výstup (obr.57).

Obr. 57

Sample: 11/01/2007 3/31/2008
 Included observations: 139
 Cross-sections included: 8
 Total pool (unbalanced) observations: 784

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.639825	0.449152	14.78301	0.0000
LOG(TEPL?+20)	3.208703	0.296001	10.84019	0.0000
(LOG(TEPL?+20))^2	-0.696048	0.048752	-14.27736	0.0000
LOG(VIET?)	0.035595	0.004958	7.179106	0.0000
LOG(SVIT?)	-0.017098	0.002695	-6.343651	0.0000
D2?	-0.002732	0.011394	-0.239752	0.8106
D3?	0.027665	0.016380	1.688918	0.0916
Fixed Effects (Cross)				
_BADV--C	0.225763			
_BAP1--C	0.158994			
_SKAL--C	-0.222056			
_NITR--C	-0.122408			
_STUR--C	-0.126887			
_NIZN--C	0.016496			
_RSOB--C	0.370548			
_PO1--C	-0.347634			
Effects Specification				
Cross-section fixed (dummy variables)				
R-squared	0.930121	Mean dependent var	9.867882	
Adjusted R-squared	0.928942	S.D. dependent var	0.307446	
S.E. of regression	0.081955	Akaike info criterion	-2.147593	
Sum squared resid	5.171822	Schwarz criterion	-2.064300	
Log likelihood	855.8563	F-statistic	788.3922	
Durbin-Watson stat	1.369855	Prob(F-statistic)	0.000000	

Tento model má opäť lepší koeficient determinácie. Ak sa pozrieme na koeficient pri novom člene, vidíme, že je signifikantný a preto je naozaj dobré modelovať vplyv teploty nelineárne. Má to aj logické vysvetlenie, konkrétne ak budeme uvažovať vplyv zmeny teploty o 1°C pri 10°C, túto zmenu už nepocítíme tak výrazne ako keby bolo napr. -10°C. Keď sa však pozrieme na koeficienty pri dummy premenných, opäť vidíme zápornosť pri D2, pričom tento koeficient je podľa p-hodnoty nesignifikantný, ale koeficient pri D3 je

kladný, a ten síce nie je takisto signifikantný, ale je k tomu bližšie. Ale z týchto zmien znamienok pri dummy premenných môžeme takisto usudzovať relatívnu nestabilitu týchto odhadov. Môže to byť spôsobené relatívne nie vhodnými dátami na analýzu snehovej pokrývky, keďže vieme, že tieto zimy boli čo do snehu a silného mrazu veľmi slabé. Škoda sme nemohli získať dáta zo zimy, kedy táto práca vznikala, keďže táto zima 2012/2013 bola veľmi dlhá a s množstvom snehových zrážok počas dlhého obdobia.

Záver

Cieľom tejto práce bolo preskúmať vplyv snehových zrážok na spotrebu zemného plynu. Pred samotným začiatkom analýzy dát rôznymi ekonometrickými metódami sme sa sústredili na čo najlepší výber dát ako aj ich vlastnosti a špecifiká. Pri tom sme detailne vysvetlili základy meteorológie, ako i zber týchto dát. Potom sme sa pokúsili bližšie preskúmať dáta, pričom sme v nich objavili časovú nekonzistenciu. Vysvetlili sme si niekoľko zaujímavých zistení, ako bolo napríklad odhalenie niektorých odberateľov plynu, ktorý sa nesprávali veľmi korelované s prislúchajúcim počasím. Po príprave dát a ich charakterizácii pomocou deskriptívnej štatistiky sme si popísali v teoretickej rovine tri základné postupy, ktorými sme sa rozhodli analyzovať vplyv snehových zrážok na spotrebu plynu.

Prvým prístupom bola regresná analýza, kde sme modelovali spotrebu plynu iba pomocou regresorov počasia. Tu sme narazili na problém, že spotreba plynu je silne autokorelovaná, čo spôsobovalo autokorelovanosť rezíduí. Tento problém sme ale eliminovali pomocou diferencovania spotreby plynu a aj niektorých regresorov. V takmer všetkých prípadoch sa vplyv snehových zrážok preukázal ako nesignifikantný, ale skôr nás prekvapil jeho väčšinou pozitívny vplyv na spotrebu plynu. Navyše sme zistili, že vplyv ďalších vysvetľujúcich premenných, vietor a slnečný svit, zohrali v modeloch dôležitejšiu úlohu, ako snehové zrážky.

Druhým prístupom bola analýza samotného časového radu spotrieb plynu, pričom sme tento model ešte obohatili o vysvetľujúce premenné, teda jednotlivé charakteristiky počasia. V týchto modeloch nás prekvapil signifikantný vplyv snehových zrážok, pričom ďalšie prídavné charakteristiky, vietor a slnečný svit boli nesignifikantné.

Posledným prístupom k vplyvu snehových zrážok na spotrebu plynu bola panelová analýza s fixnými efektmi. Jej výsledky nám potvrdili prakticky bezvýznamný vplyv snehovej pokrývky na spotrebu plynu. Naopak nám potvrdili signifikantný vplyv ďalších meteorologických charakteristík, a to vplyv vetra a dĺžky slnečného svitu, ktoré by podľa nás stálo za to, v budúcnosti preskúmať. Na záver však musíme poznamenať, že priebeh zimných období 2 rokov, ktoré sme mali k dispozícii, nebol k analyzovaniu tejto problematiky veľmi vhodný, nakoľko v týchto zimách bolo veľmi málo snehu a taktiež neboli ani silné mrazy. To mohlo rozhodne prispieť k nesignifikantnosti premenných týkajúcich sa snehu vo väčšine modelov. Preto by sme tieto analýzy odporúčali zopakovať napr. na dátach z roku 2012/2013.

Zoznam použitej literatúry a internetových zdrojov

- [1] – Slabá Natálie, Návod pre pozorovateľov meteorologických staníc ČSSR, Hydrometeorologický ústav Bratislava, 1976
- [2] - Zoznam klimatologických staníc na Slovensku k 1.10.2011, http://www.shmu.sk/File/Klima/zoznam_klimatologickych_stanic_SK.pdf
- [3] - Dostupnosť meteorologických/klimatologických údajov, http://www.shmu.sk/File/Klima/Dostupnost_klimatologickych_udajov_1.pdf
- [4] –Zvára Karel, Regrese, matfyzpress, Praha, 2008, ISBN 978-80-7378-041-8
- [5] –Green Willam H., Econometric analysis, 5th edition, Prentice Hall, 2003, ISBN 0-13-066189-9
- [6] –Koop G., Introduction to Econometrics, John Wiley & Sons, 2008, ISBN 978-80-7378-041-8
- [7]– Jenkins G.M. , Practical experiences with modelling and forecasting time series, Jersey, GJP Publication 1979
- [8] – Cipra Tomáš, Analýza časových řad s aplikacemi v ekonomii, SNTL/Alfa, Praha 1986
- [9] - Cipra Tomáš, Finanční ekonometrie, Ekopress, 2008, ISBN 978-80-86929-43-9
- [10] – Enders Walter, Applied Econometric Time Series, John Wiley & Sons, 1995, ISBN 0-471-03941-1
- [11] – Marius Ooms, Introduction EViews for Orientation course Econometrics, Amsterdam, 2004