Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky

Určení parametrů databáze telefonních čísel pro kontinuální CATI výzkum Radioprojekt na základě analýzy předchozích vln

Jiří Komárek

Bakalářská práce
2012

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: **2011/2012**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Jiří Komárek**
Osobní číslo:	**E08028**
Studijní program:	**B6209 Systémové inženýrství a informatika**
Studijní obor:	**Informatika ve veřejné správě**
Název tématu:	**Určení parametrů databáze telefonních čísel pro kontinuální CATI výzkum "Radioprojekt" na základě analýzy předchozích vln**
Zadávající katedra:	**Ústav systémového inženýrství a informatiky**

Zásady pro vypracování:

1. Představení a metodologie CATI projektu "Radioprojekt"
2. Popis databáze
3. Určení analytických metod
4. Teoretický popis vybraných metod
5. Aplikace vybraných metod
6. Interpretace výsledků

Rozsah grafických prací:

Rozsah pracovní zprávy:

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

Seznam odborné literatury:

[1] **BERKA, P. Dobývání znalostí z databází. Praha: Academia, 2003, 366 s. ISBN 80-200-1062-9.**

[2] **RUD, O.L. Data Mining: Praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM). Praha: Computer Press, 2001, 329 s. ISBN 80-7226-577-6.**

[3] **SPSS Inc. - CRoss Industry Standard Process for Data Mining [online]. Dostupné z WWW: (http://www.crisp-dm.org/)**

[4] **Interní materiály společnosti Median**

Vedoucí bakalářské práce: **Ing. Tomáš Kořínek**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **3. října 2011**

Termín odevzdání bakalářské práce: **30. dubna 2012**

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

L.S.

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

V Pardubicích dne 3. října 2011

PROHLÁŠENÍ

Prohlašuji, že jsem tuto práci vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.
Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako Školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně Univerzity Pardubice.

V Pardubicích dne 30. 6. 2012 Jiří Komárek

PODĚKOVÁNÍ:

Tímto bych rád poděkoval svému vedoucímu práce, Ing. Tomáši Kořínkovi, za jeho odbornou pomoc, cenné rady a poskytnuté materiály, které mi pomohly při zpracování bakalářské práce.

ANOTACE

Cílem bakalářské práce je nastavení parametrů databáze telefonních čísel, sloužící k provádění marketingového průzkumu Radioprojekt, což je národní mediální výzkum v České republice, do kterého jsou zařazeny všechny české celoplošné, regionální a lokální rozhlasové stanice. Tento průzkum je založen na kvartální bázi. Optimální nastavení databáze má za úkol minimalizovat mocnost výchozí databáze, a tím snížit režii tohoto transakčního systému.

KLÍČOVÁ SLOVA

CATI, časová řada, relační databáze

TITLE

Determination of parameters of phone numbers database for continuous CATI research project "Radioprojekt" based on previous waves analysis

ANNOTATION

Aim of this work is to set the parameters of phone numbers' database used to perform marketing research Radioprojekt, a national media research in the Czech Republic, in which are included all Czech nationwide, regional and local radio stations. This survey is based on a quarterly basis. Optimal setting up of the database is designed to minimize the thickness of the default database to reduce costs of the transaction system.

KEYWORDS

CATI, time series, relational database

OBSAH

SEZNAM TABULEK

SEZNAM OBRÁZKŮ

SEZNAM GRAFŮ

SEZNAM ZKRATEK

CATI	Computer Assisted Telephone Interviewing
CRISP-DM	CRoss-Industry Standard Proces for Data Mining
GUI	Graphical User Interface
RP	Radioprojekt
SQL	Structured Query Language
VMB	Velikost místa bydliště

ÚVOD

Úkolem bakalářské práce je provést analýzu databáze telefonních čísel sloužící k provádění marketingového průzkumu Radioprojekt a pomocí této analýzy navrhnout počáteční parametry této databáze.

Kvantitativní marketingový průzkum Radioprojekt je prováděn metodou CATI - COMPUTER AIDED TELEPHONE INTERVIEW. CATI metoda je, zjednodušeně řečeno, postavená na databázi telefonních čísel, nad kterou probíhá vlastní dotazování. Databáze telefonních čísel je reálně tabulka, která je součástí relační databáze, která slouží ke kompletní obsluze CATI výzkumných projektů. Nad touto databází běží aplikace CATISERVER, která řídí veškeré činnosti - jako například importy čísel, přiřazování tazatelů/počítačů k projektu a číselným řadám, přidělování telefonních čísel tazatelům (kteří pomocí aplikace CATICLIENT a elektronického dotazníku, provádějí vlastní dotazování), sběr výsledků od tazatelů, odkládání hovorů, filtrování databáze, uzavírání kvót, provádění statistik nad databázemi.

Před vlastním zahájením dotazování je třeba tuto databázi připravit. Kromě parametrů typu názvu databáze, číslo projektu, maximálního počtu odložených hovorů je nejdůležitější částí přípravy import telefonních čísel. Databáze telefonních čísel se skládá jak z čísel mobilních, tak z čísel na pevné linky. Metodika projektu jasně určuje, kolik má být navoláno kompletních interview na jednotlivé národní operátory. Zásadní otázkou je, kolik telefonních čísel je třeba připravit pro jednotlivé národní operátory. V případě prostého odhadu docházelo ke dvěma extrémům. Pokud byla databáze příliš malá, stávalo se, že čísla prostě došla. Pokud byla příliš velká, docházelo k pomalé reakční odezvě aplikace CATISERVER, a navíc rostla potřeba prostotu pro každodenní zálohování databáze.

Cílem této práce je provést analýzu konečných stavů databází jednotlivých vln a na základě této analýzy najít optimální vstupní parametry databáze pro jednotlivé národní operátory.

Vzhledem k tomu, že jde o problematiku reálného dobývání znalostí z databází, je nutné určit metodologii. Bylo rozhodnuto použít metodologii CRISP-DM. CRISP-DM se skládá z následujících kroků [3]:

- porozumění problematice (Business Understanding),
- porozumění datům (Data Understanding),
- příprava dat (Data Preparation),

- modelování (Modeling),
- vyhodnocení výsledků (Evaluation),
- využití výsledků - implementace vytvořeného modelu (Deployment).

Z těchto kroků bude celá práce složena. Bude představen Radioprojekt, budou vysvětleny potřebné datové struktury. Následně bude provedena příprava dat, na kterých bude postaven model. Tento model přinese výsledky, které budou vyhodnoceny a implementovány.

1. PŘEDSTAVENÍ A METODOLOGIE CATI PROJEKTU "RADIOPROJEKT"

1.1. Agentura Median

Firma Median je jedním ze dvou realizátorů Radioprojektu. Vzhledem k tomu, že oba realizátoři používají rozdílné technologie a my se budeme věnovat pouze části prováděnou agenturou Median, představme si tuto agenturu.

MEDIAN s. r. o. je nezávislá česká soukromá společnost s důležitým postavením na trhu v České i Slovenské republice. Představuje výzkumnou agenturu s plným servisem a vysokým standardem poskytovaných služeb. MEDIAN realizuje všechny typy kvalitativních i kvantitativních výzkumů trhu a výzkumů veřejného mínění.

MEDIAN se specializuje na realizaci výzkumných projektů pomocí tradičních i moderních technologií.

1.2. Marketingový výzkum

Radioprojekt, je marketingový výzkum. Na úvod si přibližme problematiku výzkumu trhu. Velmi často se průzkum trhu dává do souvislosti s průzkumem veřejného mínění. Vzhledem ke komerčnímu cíli se Radioprojekt týká pouze průzkumu trhu. Jaká je vlastně definice průzkumu trhu:

"Marketingový výzkum - systematická sbírka, analýza a interpretace informací relativních pro marketingová rozhodnutí." [4]

Marketingový výzkum se dělí na dvě základní kategorie, a to na průzkum kvalitativní a kvantitativní. Vzhledem k problematice se budeme dále zaobírat pouze kvantitativním průzkumem.

"Kvantitativní výzkum se zabývá měřením trhu a zahrnuje oblasti, jako například velikost trhu, velikost částí trhu, podíl značky, frekvence nákupu, míru povědomí o značce, úroveň prodeje atd. Kvantitativní údaje požadované pro určitou úroveň přesnosti (ačkoliv ne ve všech případech jde o úroveň vysokou) a užité metody musejí být takové, aby cíle bylo dosaženo. Alespoň, že ve spotřebních trzích je kvantitativní informace založena na výtahu ze vzorku průměrné populace či trhu a návrh výzkumu, zejména výběrové metody musí být dostatečně přesné, aby to umožnily." [4]

Vzhledem k úloze této práce budou představeny jednotlivé metody sběru dat.

1.2.1. Face-to-face (papírový dotazník, PAPI)

PAPI umožňuje podrobně mapovat vybrané cílové skupiny, jejich názory, myšlenky, potřeby. Pro realizaci výzkumu jsou připraveny papírové dotazové instrumenty, které slouží ke kladení otázek a zaznamenávání odpovědí respondentů. Navíc jsou zjišťovány identifikační znaky respondenta. Tazatel klade respondentovi otázky přesně podle dotazníku, řídí se instrukcemi a odpovědi zaznamenává do papírového dotazníku (tato metoda se nazývá pen & paper). Pomocí tohoto typu výzkumu je možné zjišťovat znalosti, využívání a hodnocení různých produktů, jejich penetraci trhu v regionech, informovanost vlastních klientů i běžné populace, získat popis vlastních klientů i klientů konkurence, zjišťovat spokojenost vlastních klientů i klientů konkurence apod. Častou součástí je hodnocení výroků na papírových kartách. [4]

1.2.2. Face-to-face - CAPI (Computer Assisted Personal Interviewing)

CAPI je technologicky vyspělejším typem kvantitativního výzkumu. Je to výzkum pomocí notebooků, netbooků, eventuelně PC. Tyto průzkumy lze provádět na celém území České republiky. Výhodou tohoto typu výzkumu je rychlost, přesnost a možnost použít různé multimediální ukázky (reklamní spoty, obrázky, videonahrávky, apod.). Při tomto typu výzkumu probíhá rozhovor podle předem odsouhlaseného dotazníku, který je poté naprogramován. Tento elektronický dotazník se zobrazuje na počítači přímo před respondentem, který tak může hodnotit multimediální materiály, jako obrázky, hudební spoty, videa a další materiály.

Elektronický dotazník umožňuje pokročilé filtry, rotace otázek a odpovědí, složité větvení dotazníku, což umožňuje dosáhnout přesnějších odpovědí. Zároveň elektronický přenos dat umožňuje velmi rychlé získání nasbíraných dat a následné zpracování. [4]

1.2.3. CATI dotazování (Computer Assisted Telephone Interviewing)

CATI dotazování, které se týká problematiky RP bude věnována kapitola 1.3.

1.2.4. CAWI (Computer Assisted Web Interviewing)

Mezi moderní kvantitativní výzkumné metody patří i technika sběru dat přes Internetové rozhraní. Základem této techniky sběru dat je vybudování panelu respondentů, kteří jsou ochotni vyplňovat elektronické dotazníky přes Internet. Dále je třena znát jejich základní sociodemografické charakteristiky.

Jako při jiných technikách kvantitativního výzkumu, i při použití techniky CAWI probíhá rozhovor podle předem odsouhlaseného elektronického dotazníku. Dotazník je umístěn na

chráněné internetové stránky. Stejně jako při CAPI je možno do dotazníku vkládat jakékoliv audiovizuální ukázky (videa, hudba, obrázky). Z panelu se provede náhodný stratifikovaný výběr respondentů, kteří odpovídají zadané cílové skupině. Těmto vybraným respondentům se rozesílá e-mail s jednoznačně identifikovaným linkem = přístupem na Internetovou stránku s dotazníkem, který zabraňuje tomu, aby jeden respondent vyplnil dotazník několikrát, a který rovněž identifikuje, který respondent dotazník vyplnil, aby k dotazníku mohly být připojeny sociodemografické znaky. Data jsou ukládána na zabezpečeném serveru, na konci výzkumu jsou exportována a zpracována standardními postupy kvantitativního výzkumu. [4]

1.2.5. AVL Studiové testy – CASI (Computer Assisted Studio Interviewing)

Poslední kvantitativní výzkumnou metodou jsou Studiové testy AVL, které navíc umožňují aplikaci některých postupů, známých spíše z metod kvalitativního výzkumu. Cílem studiových testů AVL je hodnocení zkoumaných objektů, ať již mají jakýkoliv charakter, navíc v kombinaci s posuzováním reálného objektu (chuťový vzorek, výrobek). Tyto testy jsou většinou poměrně obsáhle (např. délka hodina a půl oproti běžnému dvacetiminutovému CAPI dotazníku).

Studiové testy AVL se provádějí ve studiu vybaveném 10 až 40 počítači, kde respondenti vyplňují elektronický dotazník, vytvořeného technologií AVL (na rozdíl od CAPI, kde dotazník vyplňuje tazatel). Každý respondent pracuje svým individuálním tempem, což zvyšuje jeho komfort, a tím i přesnost odpovědí, kromě určitých specifických výzkumů, typu hromadné hodnocení reklam. Při hromadném hodnocení (za přítomnosti klienta), které se vyhodnocuje online, je naopak nutné, aby všichni hodnotili naráz. [2]

1.3. CATI sběr dat

Telefonické dotazování patří mezi hojně využívanou metodu dotazování. Je to dáno nejen vyšší vybaveností mobilními telefony, ale především spojením telefonického dotazování s počítači, tzv. CATI (Computer Assisted Telephone Interviewing), čímž došlo k výraznému zrychlení zpracování odpovědí a vyhodnocování výsledků. [6].

Základem metody CATI je dotazování pomocí telefonu. Telefonické dotazování patří mezi moderní metody, které se běžně používají při řadě výzkumů.

Speciálně vyškolený tazatel se telefonicky ptá respondenta a jeho odpovědi zaznamenává přímo do počítače. Pro realizaci výzkumu je připraven jednotný elektronický dotazník, tazatel klade po telefonu respondentovi otázky přesně podle obrazovek dotazníku.

Elektronický dotazník umožňuje flexibilní vedení rozhovoru, automaticky provádí tazatele i respondenta rozhovorem bez potřeby sledovat filtrační otázky, lze využít i náhodného generování pořadí otázek, sekcí či položek v dotazníku.

Tazatel může být během rozhovoru sledován supervizorem, který kontroluje kvalitu dotazování. Supervizor může za pomoci speciálního SW kontrolovat rovněž správnost zaznamenávání odpovědí do elektronického dotazníku a v případě chyby ji okamžitě opravit ještě během záznamu. Shrňme si výhody a úskalí CATI metody.

Výhody [2]:

- možnost náhodného výběru dotazovaných (telefonní seznam) včetně možnosti domluvy času, kdy se vybranému dotazovanému bude moci zavolat,
- průběžná automatická kontrola tazatelů (telefonuje se z agentury),
- zákazník může poslouchat (při zachování pravidel kodexu),
- daří se zastihnout i jinak těžko dosažitelné skupiny respondentů (opakovanými pokusy).

Úskalí [2]:

- infrastruktura,
- neochota, averze určitých skupin obyvatelstva k "technice",
- nemožnost při dotazování něco ukázat (např. obrázek ap.),
- vyšší nároky na dotazovaného, zejména na jeho paměť, soustředění,
- absence přímého kontaktu.

1.4. Radioprojekt - popis

Radioprojekt je národní mediální výzkum v České republice, do kterého jsou zařazeny všechny české celoplošné, regionální a lokální rozhlasové stanice.

Radioprojekt je realizován společnostmi STEM/MARK a MEDIAN; zadavateli jsou členové Radiové sekce Sdružení komunikačních a mediálních organizací České republiky (SKMO) a Sdružení reklamních agentur (ARA’s).

Výzkum je prováděn metodou telefonních rozhovorů. Při dotazování se používá CATI technologie, která umožňuje elektronický záznam odpovědí do CATI dotazníku a jejich logickou kontrolu. Výzkum probíhá kontinuálně v průběhu celého kalendářního roku s výjimkou posledních 14 dnů měsíce prosince a je rozdělen do 4 etap (čtvrtletí). Vzorek

rozhovorů provedených metodou CATI je doplněn o netelefonizované respondenty z výzkumu MML – TGI.

Díky náhodnému výběru a následnému převážení je zabezpečena reprezentativnost podle pohlaví, věku, vzdělání, kraje (včetně bývalého), (bývalého) okresu, velikosti místa bydliště a velikosti domácnosti.

Velikost vzorku je 28.000 uskutečněných kompletních telefonních interview. Z toho vyplývá 3.500 rozhovorů na agenturu a kvartál. Takže naším cílem je nalezení optimální původní databáze na oněch 3.500 rozhovorů.

1.5. CATI software

Vlastní výzkum je realizován pomocí CATI software. Toto programové vybavení se skládá z následujících komponent:

1. CATI Server,
2. CATI Klient,
3. elektronický dotazník,
4. generátor náhodných čísel.

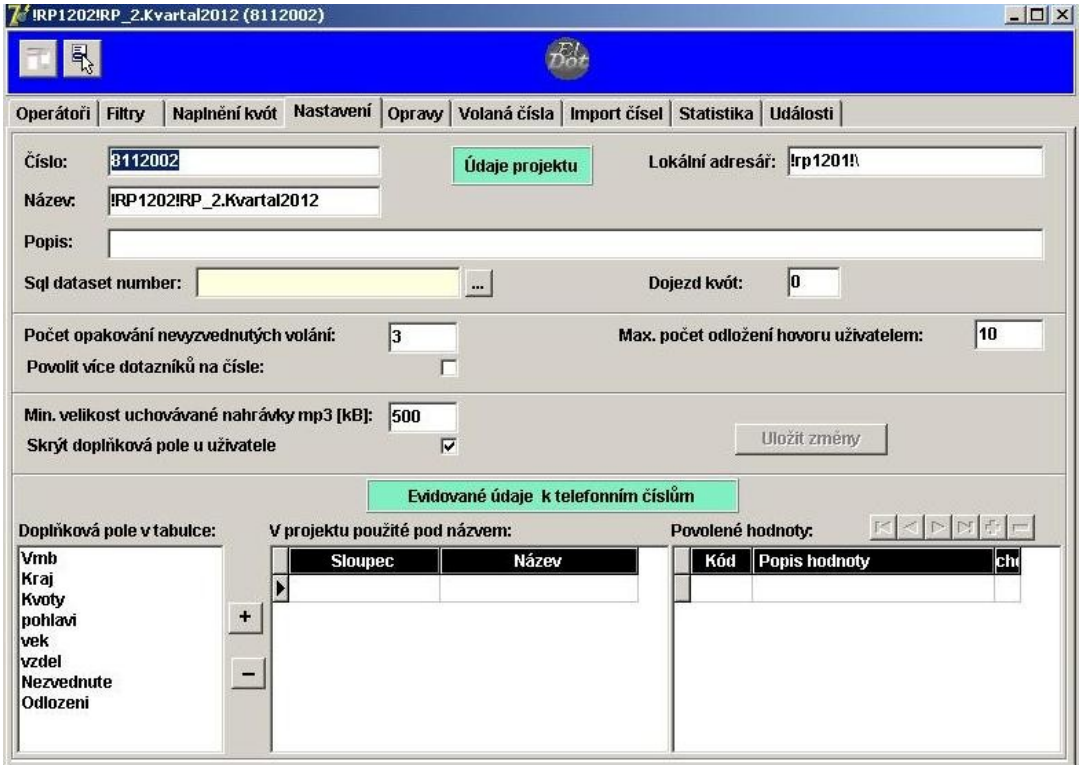
V rámci našeho zadání nás primárně zajímá CATI server a nepřímo generátor náhodných čísel.

1.5.1. CATI Server

CATI Server, jehož GUI je znázorněno na Obrázku 1, slouží k řízení běhu dotazování. Tato aplikace je naprogramována ve WIN32 prostředí (DELPHI 6). Tato aplikace je databázová a používá konektory na dva základní zdroje dat. Prvním zdrojem dat jsou projektové databáze, které se nacházejí v jedné složce ve formě MDB. Druhý datový konektor ukazuje na SQL server. SQL server je nainstalovaný na jiném počítači. V aplikaci je možné:

- definovat výzkumy (projekty),
- importovat telefonní čísla,
- k projektům přiřazovat telefonní operátory, pracovní stanice,
- hlídat kvóty,
- filtrovat databáze,
- generovat statistiky,

- zpravovat blacklist,
- atd.

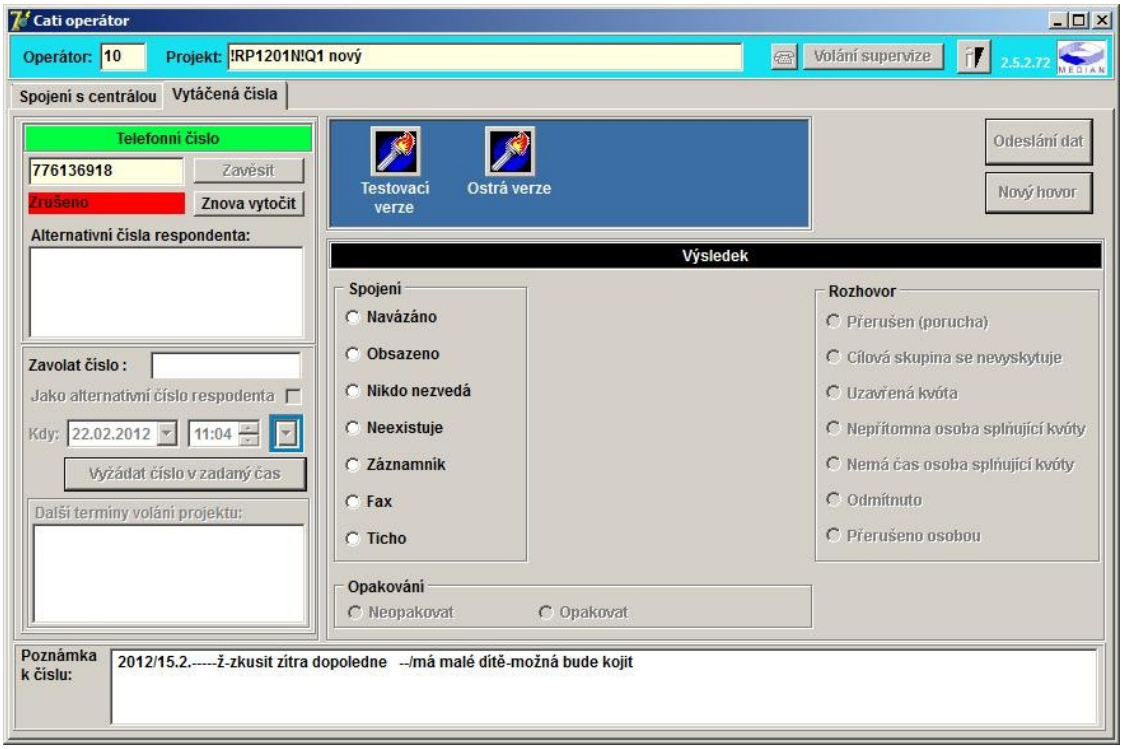


Obrázek 1: CATISW -CATISERVER- nastavení projektu

Zdroj:[9]

1.5.2. CATI Klient

CATI Klient, jehož GUI je zobrazeno na Obrázku 2, slouží k obsluze telefonie telefonním operátorem kromě vlastního sběru dat. Tazatel se musí přihlásit pod svým číslem, pod svým heslem, určit, jakou činnost bude vykonávat (telefonování, přestávka, školení, technické potíže…). Dále zde může komunikovat se supervizí, a to jak pomocí hovoru, tak pomocí chatovacího okna. Hlavní funkcí CATI Klienta je ale automatické vytočení serverem přiděleného čísla, spuštění elektronického dotazníku a následné zaznamenání výsledku telefonního hovoru telefonním operátorem. To jest práce nad tabulkou Telefony z projektové databáze. V okamžiku, kdy tabulka Telefony neobsahuje žádné řádky s odpovídajícím stavy volání, další dotazování již není možné a následně je upozorněn, jak tazatel, tak supervizor, že již nebylo přiřazeno žádné telefonní číslo. CATI Klient dále pořizuje nahrávku z dotazníku, pro případnou kontrolu a odesílá nasbíraná data na CATI Server.

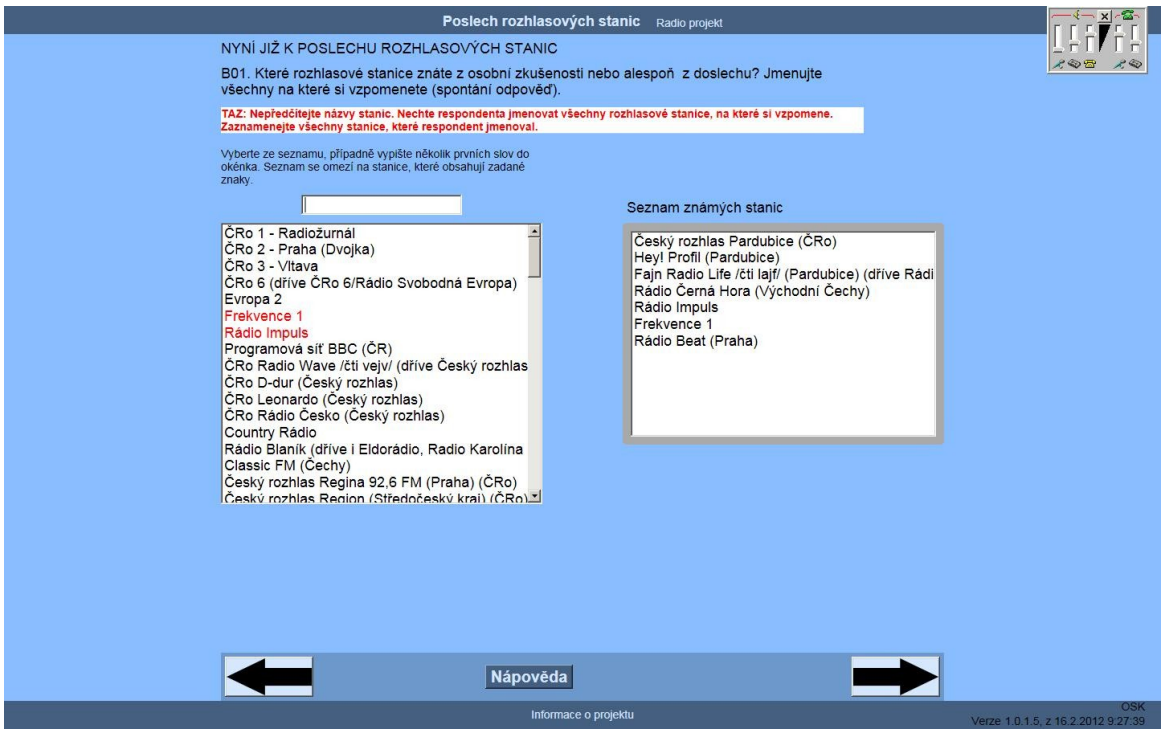


Obrázek 2: CATI Klient- práce operátora

Zdroj:[9]

1.5.3. Elektronický dotazník

Elektronický dotazník, slouží ke sběru dat s konkrétním respondentem. Operátor čte otázky, drží se instrukcí dotazníku a zaznamenává odpovědi. Pro ilustraci je na Obrázku 3 zobrazena otázka týkající se přímo poslechu rádií, místo klasických, například sociodemografických, otázek. Elektronický dotazník umožňuje zobrazit víc otázek na jedné obrazovce a obrazovky dynamicky měnit podle práce tazatele.

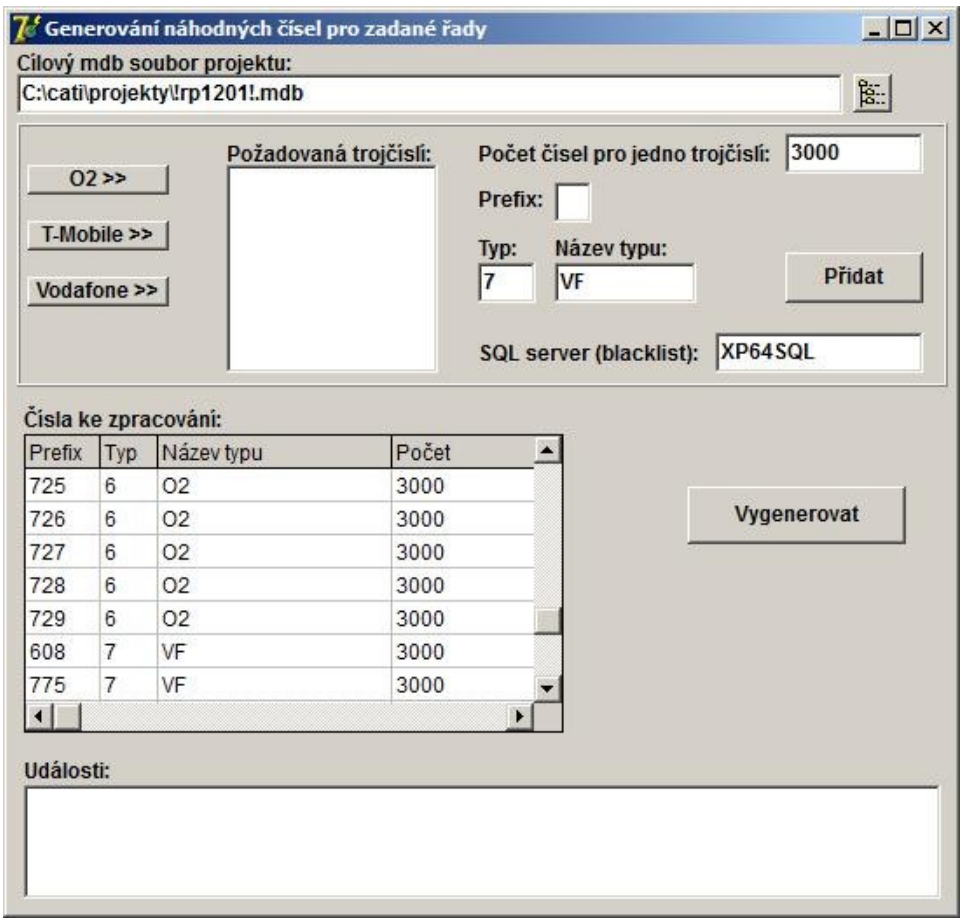


Obrázek 3: CATI - elektronický dotazník

Zdroj:[9]

1.5.4. Generátor náhodných čísel

Poslední aplikací je generátor náhodných čísel. Tato podpůrná aplikace je zobrazena na Obrázku 4. Tento program složí k získávání telefonních čísel. Díky znalosti prefixů národních operátorů se vygeneruje velké množství telefonních čísel, která se následně detekují. Toto množství bývá v řádech jednotek statisíců. Tato detekce, zda číslo opravdu existuje, má samozřejmě své limity, ale i tak je to velmi užitečný nástroj, který slouží jako zdroj pro tabulku Telefony.



Obrázek 4: Generátor náhodných čísel

Zdroj:[9]

1.6. Projektová databáze

CATI Server pracuje nad databází telefonních čísel. Realita je samozřejmě mnohem složitější. CATI Server pracuje nad projektovou databází. Pro každý výzkum (projekt), je tato databáze vytvořena z šablon. Typický návrh tabulek projektové databáze si lze prohlédnout na schématu v příloze A.

CATI Server ale pracuje v hybridním prostředí, protože kromě projektových databází využívá řadu tabulek na centrálním SQL serveru. Tyto databázové struktury nebudou přiblíženy, protože se našeho tématu dotýkají pouze okrajově.

Z celé této databáze je nejdůležitější tabulka Telefony, jejíž jméno padlo již mnohokrát. Tato tabulka bude popsána níže. Zobrazení návrhu tabulky Telefony obsahuje Obrázek 5.

	Název pole	Datový typ
🔑	IdTel	Číslo
	TelCislo	Text
	Vmb	Číslo
	Kraj	Číslo
	Typ	Číslo
	Stav	Číslo
	Vysledek	Číslo
	ZapisKdy	Datum a čas
	Kvoty	Text
	Poradi	Číslo
	Poznamka	Text
	pohlavi	Číslo
	vek	Číslo
	vzdel	Číslo
	Nezvednute	Číslo
	Odlozeni	Číslo

Obrázek 5: Návrh tabulky Telefony

Zdroj:[8]

Nyní bude následovat popis nejdůležitějších polí z této tabulky.

IdTel – automatický primární klíč.

TelCislo - vlastní telefonní číslo.

Typ – zakódovaný národní operátor. Hodnoty kódů lze vyčíst z Tabulky 1.

Tabulka 1: Telefony. Typ - kódy

kód	typ
0	N/A
1	pevné linky
6	O2 mobilní (ex Eurotel)
7	Vodafone (ex Oskar)
8	T-Mobile (ex Paegas)

Zdroj:[8]

Kraj – zakódovaný kraj ČR. U pevných linek je možno ve většině případů určit kraj přes veřejné databáze, které obsahují PSČ. To bohužel nelze u generovaných čísel pro mobilní operátory, takže hodnota proměnné kraj u těchto čísel nabývá hodnoty 0, což je znázorněno, včetně ostatních kódu, v Tabulce 2.

Tabulka 2: Telefony. Kraj - kódy

kód	kraj
0	N/A
1	Praha
2	Středočeský kraj
3	Jihočeský kraj
4	Plzeňský kraj
5	Karlovarský kraj
6	Ústecký kraj
7	Liberecký kraj
8	Královéhradecký kraj
9	Pardubický kraj
10	Vysočina
11	Jihomoravský kraj
12	Olomoucký kraj
13	Zlínský kraj
14	Moravskoslezský kraj

Zdroj:[8]

VMB – zakódovaná velikost místa bydliště. U pevných linek je ve většině případů možno určit VMB přes veřejné databáze, které obsahují PSČ. U mobilních čísel je VMB 0, což je patrné z Tabulky 3.

Tabulka 3: Telefony. Vmb - kódy

kód	VMB
0	N/A
1	do 1.000 obyvatel
2	1.000 - 4.999 obyvatel
3	5.000 - 19.999 obyvatel
4	20.000 - 49.999 obyvatel
5	50.000 - 99.999 obyvatel
6	100.000 obyvatel a více

Zdroj:[8]

Výsledek – Defaultní hodnota je 0 – nevoláno. Ostatní hodnoty se vyplňují podle výsledku telefonního hovoru, jak lze vidět v Tabulce 4.

Tabulka 4: Telefony. Výsledek - kódy

IdUdalosti	Nazev
-1	nevrácený výsledek
0	Nevoláno
1	Navázáno
2	Obsazeno
3	Nikdo nezvedá
4	Neexistuje
5	Záznamník
6	Ticho
7	Fax
101	Úplný
102	Přerušen (porucha)
103	Cílová skupina se nevyskytuje
104	Nepřítomna osoba splňující kvóty
105	Nemá čas osoba splňující kvóty
106	Odmítnuto
107	Přerušeno osobou
108	Uzavřená kvóta
1001	Rezervováno

Zdroj:[8]

1.7. Rekapitulace problematiky

V předchozích bodech bylo popsáno prostředí a smysl dotazování tohoto výzkumu. Jak je vidět, problematika je to velmi široká a obšírná, takže bude nutné zaměřit pozornost opět na hlavní cíl, kterým je stanovením parametrů výchozí databáze. Projekt RP běží jako dlouholetý projekt (trackový projekt), tudíž již existují výsledky předchozích vln. Zkušenost ukazuje, že projektová databáze je veliká řádově ve stovkách megabytů. Příliš velká databáze zatěžuje jak zálohovací proces, tak se může projevit i v transakční rychlosti. CATI server přistupuje k aplikaci jako MS Access. „Aplikace Microsoft Access si do souboru databáze zapisuje mnoho aktivit, které byly prováděny při práci s databázovými objekty. Tím jednak narůstá velikost databáze a navíc je možné, že dojde k částečné ztrátě integrity dat. Nyní bude uveden příklad. Byl smazán záznam v tabulce. Potom místo, které bylo původně pro tento záznam vyhrazeno, zůstane v databázi prázdné a nemůže být do budoucna použito (dokud nebude provedena komprimace). Toto prázdné místo pak vede pouze k problémům s nepravidelným rozmístěním dat a jejich poškození. Aby se předešlo problémům s databází, je vhodné pravidelně (dle četnosti používání databáze) provádět komprimaci a opravu databáze [7].“ Tato praxe ale řeší problematiku pouze částečně. Je nutné provést analýzu, jak zmenšit původní velikost databáze.

1.7.1. Hlavní úkol

Hlavním úkolem je zjistit, kolik je potřeba naimportovat telefonních čísel do databáze na začátku dotazování, když je známo, jaká je kvóta na jednotlivé národní telefonní operátory na vlnu dotazování pro jednotlivé typy 1,6,7,8, viz kapitola 1.6.

Jako zdroj analýzy budou sloužit všechny dostupné databáze předchozích vln, takže lze analyzovat od 1. kvartálu 2008 do 4. kvartálu 2011. To znamená, k dispozici je 12 vln tohoto projektu. Ze statistického úhlu pohledu není počet pozorování příliš veliký, ale vzhledem k časové posloupnosti není možné zajistit větší počet pozorování.

2. PŘEDZPRACOVÁNÍ DAT

„Příprava (předzpracování) dat je nejobtížnější a časově nejnáročnější krok celého procesu dobývání znalostí z databází. Současně je to ale krok, který má klíčový význam pro úspěch dané aplikace. Je to ta část práce, která (spolu s krokem porozumění problému) vyžaduje největší podíl spolupráce s expertem z dané aplikační oblasti. Skoro lze říci, že problémy se získáním „těch správných“ dat pro učící se systémy jsou podobné problémům se získáním znalostí u systémů expertních.

Cíl předzpracování je obvykle dvojí, a to [4]:

1. vybrat (nebo vytvořit) z dostupných dat ty údaje, které jsou relevantní pro zvolenou úlohu dobývání znalostí,
2. reprezentovat tyto údaje v podobě, která je vhodná pro zpracování zvoleným algoritmem.“

Jak bylo na konci předchozí kapitoly řečeno, je k dispozici dvanáct projektových databází nazvaných RP0801 až RP1201. Je patrné, že tyto MDB soubory obsahují příliš informací. Tyto soubory jsou pro další zpracování příliš veliké, obsahují příliš údajů. Je třeba se na tyto soubory podívat a rozhodnout, jaké relevantní informace lze z těchto databází získat.

Existují různé typy dat. Jsou data nestrukturovaná, jako například texty. Dále jsou data strukturovaná, jako například:

- prostorová data,
- časová data,
- atd.

„Časová data jsou obvykle tvořena hodnotami téže veličiny (nebo více veličin) zaznamenávaných v různých okamžicích. Veličinou (většinou numerickou) může být kurz akcií, spotřeba elektrické energie, výsledek laboratorního testu v nemocnici, transakce na účtu v bance nebo sekvence navštívených webových stránek (tzv. clickstream). Údaje mohou být zaznamenány jak v ekvidistantních časových intervalech (denní kurzy akcií), tak v libovolných okamžicích (návštěvy u lékaře) [3].“

Jak je patrné, cílem předzpracování je získat data časová. Tato časová data vytvoří časové řady, z kterých bude možné predikovat počáteční stav databáze pro jednotlivé typy telefonních čísel. Aby šlo vytvořit časovou řadu, bude nutné udělat čtyři kroky v rámci předzpracování:

1. separace dat,
2. čištění dat,
3. agregace dat,
4. vytvoření nových proměnných.

2.1. Separace dat

Separace dat byla provedena v následujících krocích:

1. v tabulce Telefony byla smazána telefonní čísla s výsledkem 0, tj. nepoužívaná,
2. z tabulky Telefony byla smazána některá, pro naši analýzu, nepotřebná pole, jako například pole „poznámka“, což byl 250 bytů dlouhý textový řetězec,
3. byly vyexportovány tabulky Telefony, které byly pojmenovány podle názvů projektových databází.

Výsledkem se staly tabulky, jejíž příklad je v Tabulce 5.

Tabulka 5: Tabulka separovaných dat

TelCislo	Vmb	Kraj	Typ	Stav	Vysledek
721569534	0	0	1	100	108
602851200	0	0	0	110	101
775142748	0	0	7	110	101
774186914	0	0	7	3	5
226517142	0	0	1	110	101
608781321	0	0	7	110	101
775435364	0	0	7	110	101
723554197	0	0	7	110	101
776799477	0	0	7	-100	-1

Zdroj: vlastní zpracování

V Tabulce 5 je pouze ukázka těchto dat. Z této ukázky je jasně patrné, že data je nutné vyčistit. Nekonzistenci v datech zobrazuje například první řádek, kde telefonní číslo začínající sedmičkou má přiřazeno typ 1 - pevná linka, což není reálné. Proto je ještě potřeba data vyčistit.

2.2. Čištění dat

Jak bylo zjištěno po exportu dat, data obsahují chyby v typech. Tyto chyby nastaly různým způsoby, od chyb při generování telefonních čísel, při importech čísel, neodbornými zásahy do databází. Díky znalosti, které počáteční trojčíslí patří ke kterému telefonnímu operátorovi,

lze tyto chyby opravit. Tato korekce není stoprocentní díky možnosti portování telefonního čísla mezi telefonními operátory. Tento fakt ale nelze žádným způsobem, vzhledem ke zdrojům dat, kvantifikovat, takže s možností portace nebude v této práci počítáno. Níže je výčet prefixů jednotlivých mobilních operátorů:

O2 (mobilní): 601,602,606,607,721,722,723,724,725,726,727,728,729

T-Mobile: 603,604,605,731,732,733,734,735,736,737,738,739

Vodafone: 608,773,774,775,776,777

Tato tabulka byla aplikována způsobem, že byl oříznut z volaného čísla třímístný prefix, porovnán s touto tabulkou a doplněn odpovídající typ. Pokud prefix v této tabulce nebyl nalezen, automaticky bylo toto číslo prohlášeno za pevnou linku, včetně prefixu 790 (U:fon).

Po aplikaci této tabulky bylo získáno 2 714 504 záznamů. Vzhledem ke znalosti kapacity většiny statistických softwarů, jako například SPSS, které je schopno pracovat pouze v řádech desetitisíců záznamů, je jasné, že analyzovat toto množství dat není jednoduché. Vzhledem k možnostem agregace dat by to bylo i zbytečné.

2.3. Agregace dat

K provádění analýz na 2 714 504 záznamů by bylo nutné mít k dispozici velmi výkonné vybavení a dostatek strojového času. Rozumným řešením je seskupení objektů se stejným výsledkem. Analýza se nebude provádět nad jednotlivými objekty, ale nad určitou statistikou vstupních objektů. Na jednotlivých kvartálech proto byly vytvořeny jednoduché kontingenční tabulky, jako je Tabulka 6.

Tabulka 6: Agregace dat jednoho kvartálu

Počet z Typ2	Popisky sloupců																
Popisky řádků	-1	1	2	3	4	5	6	7	101	102	103	104	105	106	107	108	(Prázdné)	Celkový součet
1	524	5	207	17664	29447	1652	10949	377	1423	94	4160	562	966	11646	2312	5205		87193
6	32	2	797	26651	1033	1024	3510	46	894	9	293	8	191	1169	198	161		36018
7	66	2	66	18258	19501	3201	1078	6	813	35	410	34	683	2727	435	1230		48545
8	13		25379	7202	40	408	2278	5	404	7	68	307	737	482	102	152		37584
(Prázdné)																		
Celkový součet	**635**	**9**	**26449**	**69775**	**50021**	**6285**	**17815**	**434**	**3534**	**145**	**4931**	**911**	**2577**	**16024**	**3047**	**6748**		**209340**

Zdroj: vlastní zpracování

Tabulky mají ve sloupcích výsledky a v řádcích typy, vlastním obsahem jsou počty z typů. Tyto výsledky byly označeny dle odpovídajících čísel kvartálů RP.

Poté, co bylo z dat za jednotlivé kvartály vytvořeny matice, bylo možné tyto matice sloučit a seřadit všechna data dle typů telefonních operátorů, což ilustruje Tabulka 7.

Tabulka 7: Agregovaná data kompletní

kvartál_typ	rok/kvartál	typ	-1	1	2	3	4	5	6	7	101	102	103	104	105	106	107	108	Celkem
RP0801_1	RP0801	1	524	5	207	17664	29447	1652	10949	377	1423	94	4160	562	966	11646	2312	5205	87193
RP0802_1	RP0802	1	691	5	460	16753	42333	1133	199	422	1411	71	1699	382	654	6036	1774	1798	75821
RP0803_1	RP0803	1	222	7	13369	50948	67016	2190	284	645	1394	67	5350	1027	732	12237	2314	4877	162679
RP0804_1	RP0804	1	311	2859	5205	27259	65668	1103	280	543	1390	68	2952	346	670	9785	2524	2558	123521
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
RP1101_8	RP1101	8	31	22	744	11828	16740	3859	69	43	813	19	924	301	387	2591	427	1222	40020
RP1102_8	RP1102	8	8	16	666	23374	9911	4094	81	74	817	23	1080	147	2373	3016	471	1219	47370
RP1103_8	RP1103	8	6	9	317	6338	889	995	68	18	814	8	823	12	2094	2457	442	1000	16290
RP1104_8	RP1104	8	0	4	339	5406	893	655	45	15	841	16	626	28	2024	2276	423	806	14397

Zdroj: vlastní zpracování

Náhle místo 2 714 504 objektů zůstala jedna tabulka o 64 záznamech (16 pro každý ze 4 typů). Nejspíše by bylo možné již zde ukončit předzpracování. Ale vzhledem k tomu, že je známo, že některá zadání projektu RP se můžou změnit, bude nutné ještě v předzpracování pokračovat.

2.4. Vytvoření nových proměnných

Vzhledem k dynamice trhu dochází ke změnám penetrace jednotlivých telefonních operátorů mezi obyvatelstvem. Z toho důvodu je nutné vytvořit nové odvozené (podílové) proměnné. Pomocí podílových proměnných je možné reagovat na změny kvótních předpisů. Dále se mění metody detekování existence náhodně generovaných čísel (vliv výsledků neexistuje). Proto bylo rozhodnuto o vytvoření několika nových proměnných:

Celkem-nonex – celkem bez neexistuje - Celkem - výsledek (4),

Aktivních – kdy se operátor dovolal **-** suma výsledků (101,102,103,104,105,106,107,108).

A nakonec se proměnné Celkem, Celkem-nonex a Aktivních vztáhnou k počtu kompletních dotazníků:

Celkem/hotový dotazník,

Celkem-nonex/hotový dotazník,

Aktivních/hotový dotazník.

Z dopočítání vyplývá Tabulka 8. Tato tabulka již představuje zdroj dat pro následující analýzu.

Tabulka 8: Dopočítané proměnné

kvartál_typ	rok/kvartál	Celkem/hotový dotazník	Celkem-nonex/hotový dotazník	Aktivních/hotový dotazník
RP0801_1	RP0801	61,27406887	40,58046381	18,52986648
RP0802_1	RP0802	53,73564848	23,73352232	9,798015592
RP0803_1	RP0803	116,6994261	68,62482066	20,08464849
RP0804_1	RP0804	88,86402878	41,62086331	14,59928058
RP0901_1	RP0901	70,08931186	34,22254758	12,58345534
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
RP1004_8	RP1004	36,02423469	22,62244898	7,354591837
RP1101_8	RP1101	49,22509225	28,63468635	8,221402214
RP1102_8	RP1102	57,98041616	45,8494492	11,19461444
RP1103_8	RP1103	20,01228501	18,92014742	9,398034398
RP1104_8	RP1104	17,11890606	16,05707491	8,37098692

Zdroj: vlastní zpracování

2.5. Rekapitulace předzpracování

Předpracování je jedna z nejdůležitějších částí analýzy. Na rozdíl od vlastních analýz, které provádí statické programy na základě většinou jednoduchého zadání, kdy je celé pozadí výpočtu uživateli skryto, předzpracování je dlouhá únavná ruční práce. Pokud by byla provedena špatně, mohly by být výsledky významně zkreslené. Dále předzpracování, obzvláště u velkých souborů dat, umožní provádět analýzy rychle a efektivně. Zde se ukázalo, že místo 2 714 504 objektů půjde do analýzy jen 64. A že místo 16 atributů absolutních vznikly tři podílové, z kterých nakonec, vzhledem k zadání, půjde do analýzy pouze jeden atribut. Tímto je základní předzpracování u konce. Poslední krok předzpracování, to jest existence extrémních hodnot, bude až součástí vlastní analýzy.

3. URČENÍ ANALYTICKÝCH METOD

Nyní je nutné určit analytické metody, které se použijí. Budou se analyzovat posloupnosti 16 podílových atributů v řadě. Na základě zkušeností a konzultací lze navrhnout následující postup:

3.1. Popisná statistika

Nejdříve bude provedena popisná statistiku, včetně grafického zobrazení. „V průzkumové (popisné) analýze dat se vyšetřují statistické zvláštnosti dat, jako je lokální koncentrace dat, tvarové zvláštnosti rozdělení dat a přítomnost podezřelých hodnot[10].“ Podezřelými hodnotami budou hodnoty chybějící, eventuelně extrémní.

3.2. Časová řada a její analýza

V čase srovnané hodnoty a predikce budoucích hodnot představují časovou řadu a její analýzu. Časová řada a její analýza mají následující definice:

„Časovou řadou budeme rozumět posloupnost věcně a prostorově srovnatelných pozorování (dat), která jsou jednoznačně uspořádána z hlediska času ve směru minulost-přítomnost. Analýzou (a podle potřeby i prognózou) časových řad se potom rozumí soubor metod, které slouží k popisu těchto dynamických systémů (a případně k předvídání jejich budoucího chování) [11].“

Analýza časových řad je velmi obtížná disciplína s mohutným matematickým aparátem. Naštěstí prudký rozvoj výpočetní techniky a statistického software tuto disciplínu zjednodušil.

3.2.1. Druhy časových řad

Základní druhy časových řad se rozlišují [11]:

1. podle rozhodného časového hlediska na časové řady **intervalové** (časové řady intervalových ukazatelů) a na časové řady okamžikové (časové řady okamžikových ukazatelů),
2. podle periodicity, s jakou jsou údaje v řadách sledovány, na časové řady **roční** (někdy též dlouhodobé) a na časové řady **krátkodobé**, kde jsou údaje zaznamenávány ve čtvrtletních, měsíčních, týdenních aj. periodách,
3. podle druhu sledovaných ukazatelů na časové řady **absolutních** ukazatelů a na časové řady **odvozených** charakteristik,

4. podle způsobu vyjádření údajů na časové řady **naturálních** ukazatelů (hodnoty ukazatele jsou vyjadřovány v naturálních jednotkách) a na časové řady **peněžních** ukazatelů.

Při pohledu na analyzovanou časovou řadu je jasné, že se bude analyzovat následující časová řada:

1. intervalová (potřeba čísel na hovor během kvartálů),
2. krátkodobá (délka intervalu je kvartál),
3. odvozená (analyzována bude dopočítaná podílová proměnná),
4. naturální (údaje nejsou v peněžních jednotkách).

3.2.2. Způsoby modelování časových řad

V rámci této práce se bude uvažovat pouze o jednorozměrném modelu. K tomuto modelu lze přistupovat následujícími způsoby [11]:

1. pomocí **klasického (formálního) modelu**, kdy jde pouze o popis forem pohybu (a nikoliv o poznání věcných příčin dynamiky časové řady). Tento model vychází z dekompozice na čtyři formy (složky) časového pohybu. Souběžná existence všech těchto forem však není nutná a je podmíněna věcným charakterem zkoumaného ukazatele. Časovou řad řadu lze dekomponovat na složku trendovou, sezónní, cyklickou a náhodnou,
2. pomocí **Box-Jenkinsovy metodologie,** která považuje za základní prvek konstrukce modelu časové řady náhodnou složku, jež může být tvořena korelovanými náhodnými veličinami. Těžiště postupu spočívá v korelační analýze více či méně závislých pozorování uspořádaných do tvaru časové řady. Předpokladem aplikace tohoto postupu je obvykle požadavek disponovat delší časovou řadou, řádově alespoň o cca 50 pozorováních,
3. pomocí **spektrální analýzy**, kdy časovou řadu považujeme za směs sinusovek a kosinusovek o rozličných amplitudách a frekvencích. Tato koncepce pak umožní provést explicitní popis periodického chování časové řady a především vystopovat ty významné složky periodicity, které se podílejí na věcných vlastnostech zkoumaného procesu. V této koncepci tedy není stěžejním faktorem časová proměnná, ale právě faktor frekvenční.

Na analyzovanou řadu bude použit klasický (formální) model.

4. TEORETICKÝ POPIS VYBRANÝCH METOD

Pokud je třeba analyzovat nějakou situaci, je třeba mít k tomu připravené matematické a statistické metody.

4.1. Popisná statistika

U jednotlivých typů budou spočteny následující popisné statistiky: střední hodnota, chyba střední hodnoty, medián, směrodatná odchylka, rozptyl výběru, špičatost, šikmost, variační rozpětí, minimum, maximum, součet, počet, hladina spolehlivosti (95,0%), useknutý průměr, horní kvartil, dolní kvartil 75, mezikvartilové rozpětí a variační koeficient.

Dalším krokem popisné statistiky bude grafické znázornění. Z grafického znázornění lze vyčíst, zda soubor obsahuje chybějící nebo extrémní hodnoty. Ty by pak bylo eventuelně nutné vypustit, nebo dopočítat (například pomocí rovnice trendu nebo pomocí průměru).

4.2. Analýza časové řady pomocí klasického (formálního) modelu

Jako odrazový můstek slouží grafické znázornění z popisné statistiky. Již z grafu lze odhadnout, zda časová řada má trend, sezónnost a cyklus. Náhodná složka představuje ostatní vlivy, které často nelze vůbec posoudit.

Analýzu bude provedena jako dekompozice časové řady. Vzhledem ke krátkosti časové řady se neprojeví cyklická složka. Úkolem bude proto objevit funkční zápis jednotlivých složek, aby bylo možné provést predikci časové řady. Postupovat se bude následujícím způsobem.

4.2.1. Sezonní čištění

V prvé řadě je nutné identifikovat a odstranit sezónnost. Sezonností se rozumí periodické kolísání v časové řadě, které má systematický charakter. Toto kolísání se odehrává během jednoho kalendářního roku a každý rok se ve stejné nebo modifikované podobě opakuje. Periodické změny jsou způsobeny především střídáním ročních období a různými institucionalizovanými lidskými zvyky [1]."

K odstranění sezónnosti lze přistoupit několika způsoby:

- klouzavé průměry,
- sezónní indexy,
- exponenciální vyrovnání.

V případě nejistoty lze použít vícero metod a podle grafu vybrat tu, která přinese lepší výsledky.

Podstata vyrovnání pomocí klouzavých průměrů spočívá v tom, že posloupnost empirických pozorování se nahradí řadou průměrů vypočítaných z těchto pozorování. Každý z těchto průměrů tedy reprezentuje určitou skupinu pozorování. Velmi důležitou otázkou, kterou je nutné při tomto způsobu vyrovnání řešit, je stanovení počtu pozorování, z nichž jsou jednotlivé klouzavé průměry počítány. Tento počet pozorování se bude nazývat klouzavá část období interpolace a značit symbolem m=2p+1 pro m<n, kde n je celkový počet pozorování analyzované řady. Klouzavou částí období interpolace se bude tedy rozumět časový interval určité délky, který se posunuje po časové ose vždy o jednotku [5].

Při provádění těchto výpočtů se vyplatí i provést výpočty absolutních a relativních přírůstků.

Dále lze spočítat sezonní indexy podle vzorce (1).

$$S_j = \frac{m \cdot \sum_{i=1}^{n} X_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{m} X_{i,j}} \qquad (1)$$

Legenda ke vzorci (1) je následující:

n se rovná počtu let (period),

m se rovná počtu období v rámci periody/roku (počet kvartálů, měsíců…),

Sj se rovná hodnotě sezónního indexu pro j-té období,

Xi,j se rovná hodnotě pozorování časové řady v i-té periodě a v j-tém období v rámci příslušného roku/periody.

Výhodou sezonních indexů je, na rozdíl od klouzavých průměrů, to, že zůstane původní počet hodnot, při použití klouzavých průměrů se počet hodnot do dalšího zpracování snižuje.

Pomocí tohoto kroku bude získána časová řada očištěná o sezónnost.

4.2.2. Modelování trendu

„Dalším krokem dekompozice bude určení trendu. Trend odráží dlouhodobé změny v průměrném chování časové řady, resp. obecnou tendenci vývoje zkoumaného jevu

za dlouhé období. Je výsledkem faktorů, které dlouhodobě působí ve stejném směru, jako např. technologie výroby, demografické podmínky či podmínky trhu v dané oblasti. Trend může mít různý charakter, může být rostoucí, klesající, strmý, mírný, v průběhu času se může měnit, takže jej lze pokládat spíše za cyklus. Může být hladší, než je vlastní časová řada, nebo také variabilnější [1].“

Trendová funkce může být různých typů, jako například:

- lineární trend,
- exponenciální trend,
- parabolický trend,
- logistický trend,
- atd.

V této práci se budou využívat pouze trendová funkce lineární a exponenciální. Lineární je základní trendová funkce a exponenciální by měla nejvěrněji odpovídat odhadu ze zadání. Lineární trend má rovnici zobrazenou ve vzorci (2).

$$y = mx + b \qquad (2)$$

Exponenciální trend má rovnici podle vzorce (3).

$$y = ce^{bx} \qquad (3)$$

V obou rovnicích x zastupuje čas a m, b, c jsou parametry přímek, které bude hledat statistický software.

Znalost těchto rovnic umožňuje dopočítat budoucí hodnotu bez sezónního vlivu na libovolný počet budoucích období.

4.2.3. Náhodná složka

Nyní je časová řada očištěna o sezónní a trendovou složku. Zobrazená náhodná složka by neměla mít žádné pravidelnosti obvyklé pro sezónní a trendovou složku, o čemž se lze přesvědčit z grafu. Za předpokladu použití lineárního trendu lze pomocí grafického znázornění provést vizuální kontrolu, zda byly výpočty provedeny správně. Tato kontrola se dělá pomocí proložení reziduí přímkou. V případě správných výpočtu tato přímka odpovídá ose X.

4.2.4. Predikce budoucích hodnot časové řady

Nyní lze provést předpověď budoucího pokračování časové řady. Tuto predikci lze provést díky znalosti sezónních indexů (nemění se) a rovnice trendové přímky. Co se týká náhodné složky, tu predikovat opravdu nelze, takže s ní se v předpovědi nepočítá. To znamená, že časová řada se protáhne o požadovaný počet období a dopočítá se pomocí rovnice trendu a sezonních indexů (opak dekompozice), přičemž náhodná složka bude rovna nule.

Dopočítání splní hlavní cíl – zjištění odhadu vstupních parametrů databáze pro jednotlivé národní operátory.

5. APLIKACE VYBRANÝCH METOD

V následujících kapitolách bude provedena vlastní aplikace vybraných metod na datech, jež již byla předzpracována. Ve všech grafech a tabulkách je pořadové číslo kvartálu prezentováno písmenem *t*.

5.1. Typ 1 – pevné linky

Jako první bude provedena analýza číselné řady týkající se typu 1, tj. pevných linek. Vzhledem k tomu, že půjde o první analýzu, bude provedena podrobněji. Protože analýzy dalších řad budou analogické, budou proto provedeny stručněji.

5.1.1. Zobrazení dat a popisná statistika

Díky poměrně malé velikosti lze zobrazit i analyzovaná data v Tabulce 9.

Tabulka 9: Typ 1 - data

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
čís/dot	61,27	53,74	116,70	88,86	70,09	54,49	81,55	36,79	27,44	31,46	40,25	31,33	31,50	37,50	23,03	16,71

Zdroj: vlastní zpracování

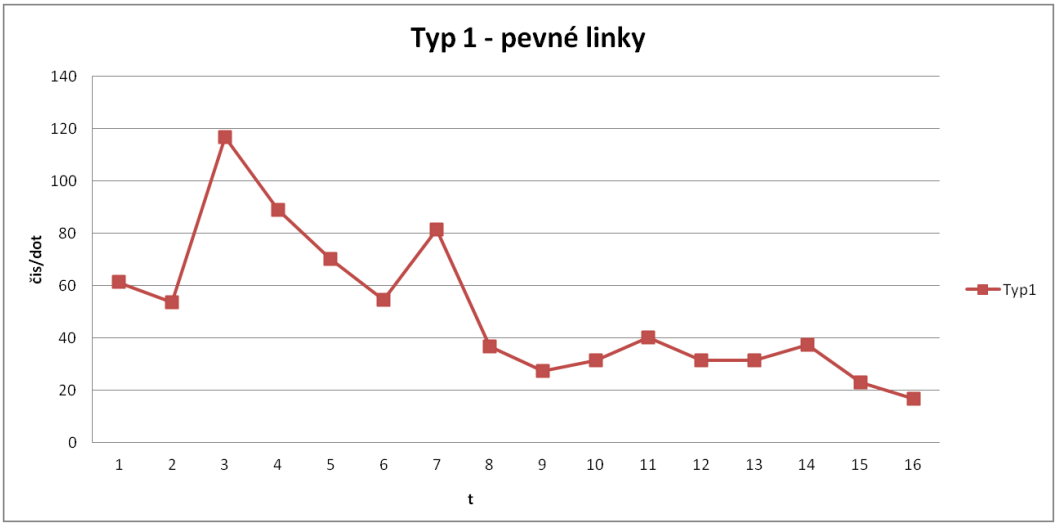
Z těchto dat byla poté spočítána popisná statistika, která je zobrazena v Tabulce 10.

Tabulka 10: Typ 1 - popisná statistika

Stř. hodnota	50,17
Chyba stř. hodnoty	6,87
Medián	38,88
Směr. odchylka	27,48
Rozptyl výběru	754,92
Špičatost	0,75
Šikmost	1,10
Variační rozpětí	99,99
Minimum	16,71
Maximum	116,70
Součet	802,71
Počet	16,00
Hladina spolehlivosti (95,0%)	13,46
Useknutý průměr	47,81
Kvartily 25	31,43
Kvartily 75	63,48
Mezikvartilové rozpětí	32,05
Variační koeficient	54,77%

Zdroj: vlastní zpracování

A nakonec znázornění vstupních dat je zobrazeno v Grafu 1.



Graf 1: Typ 1 - znázornění dat

Zdroj: vlastní zpracování

Z grafického znázornění je patrné, že trend je sestupný. Dále je vidět, že třetí kvartály obsahují mimořádné vrcholky – řada obsahuje sezonní složku. Tato složka je dokonce interpretovatelná. Do třetího kvartálu spadají letní prázdniny a dovolené, což je období, kdy se populace všeobecně hůře dotazuje. Variační koeficient není příliš veliký, takže data ani nejsou extrémně vychýlená.

5.1.2. Sezonní čištění

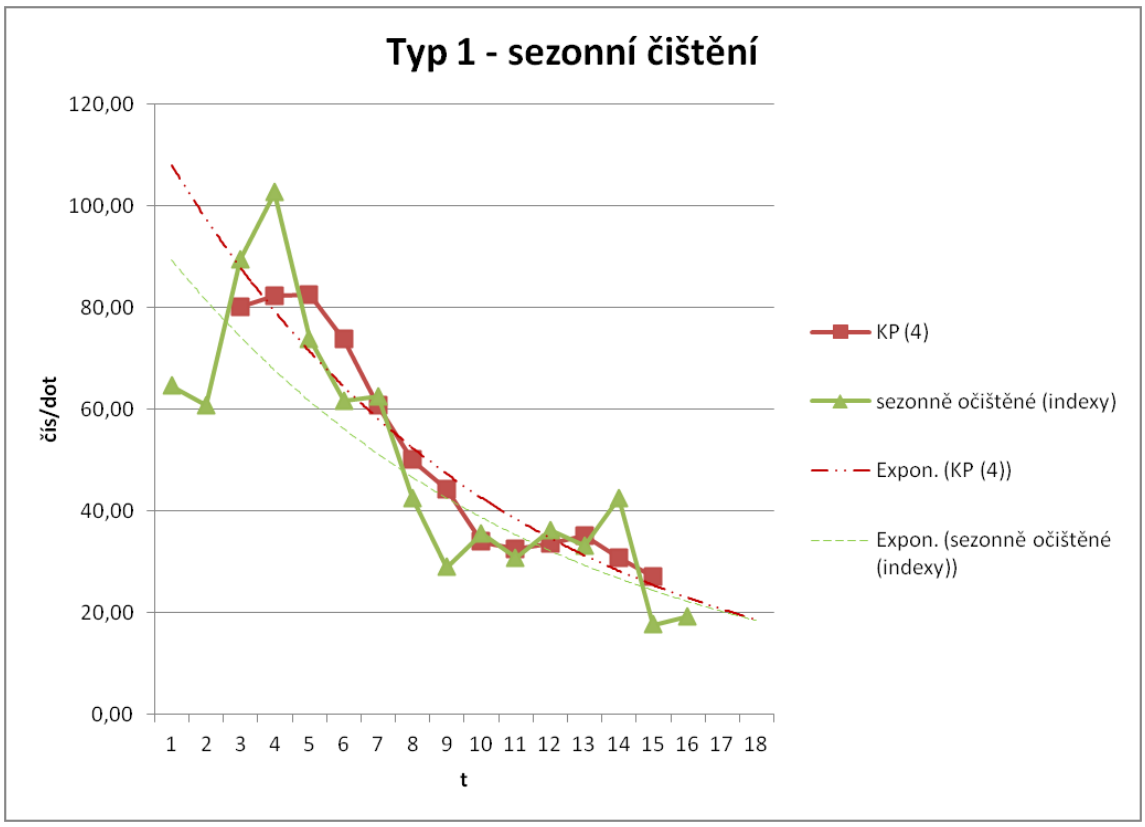
K sezonnímu čištění byla použita metoda sezonních indexů a pro ilustraci i klouzavé průměry tříčlenné a čtyřčlenné. Zároveň byly spočítány absolutní a relativní přírůstky. Vše shrnuje Tabulka 11.

Tabulka 11: Typ 1 - sezonní čištění

kvartál_typ	t	čís/dot	KP (3)	KP (4)	absolutní přírustky	relativní přírustky	index	S indexy	sezonně očištěné
RP0801_1	1	61,27					S1	94,83%	64,62
RP0802_1	2	53,74	77,24		-7,54	-12,30%	S2	88,30%	60,86
RP0803_1	3	116,70	86,43	80,14	62,96	117,17%	S3	130,32%	89,55
RP0804_1	4	88,86	91,88	82,35	-27,84	-23,85%	S4	86,55%	102,67
RP0901_1	5	70,09	71,15	82,53	-18,77	-21,13%	S1	94,83%	73,91
RP0902_1	6	54,49	68,71	73,75	-15,60	-22,26%	S2	88,30%	61,71
RP0903_1	7	81,55	57,61	60,73	27,06	49,67%	S3	130,32%	62,57
RP0904_1	8	36,79	48,59	50,07	-44,76	-54,89%	S4	86,55%	42,51
RP1001_1	9	27,44	31,90	44,31	-9,35	-25,42%	S1	94,83%	28,93
RP1002_1	10	31,46	33,05	33,99	4,03	14,68%	S2	88,30%	35,64
RP1003_1	11	40,25	34,35	32,62	8,78	27,92%	S3	130,32%	30,88
RP1004_1	12	31,33	34,36	33,63	-8,92	-22,17%	S4	86,55%	36,19
RP1101_1	13	31,50	33,44	35,14	0,17	0,54%	S1	94,83%	33,21
RP1102_1	14	37,50	30,68	30,84	6,01	19,08%	S2	88,30%	42,48
RP1103_1	15	23,03	25,75	27,19	-14,47	-38,58%	S3	130,32%	17,67
RP1104_1	16	16,71			-6,32	-27,45%	S4	86,55%	19,31

Zdroj: vlastní zpracování

Nyní se porovnají výsledky získané pomocí klouzavých průměrů a řada očištěná pomocí sezonních indexů, a to včetně exponenciální predikce na dvě období dopředu, což je zobrazeno v Grafu 2.



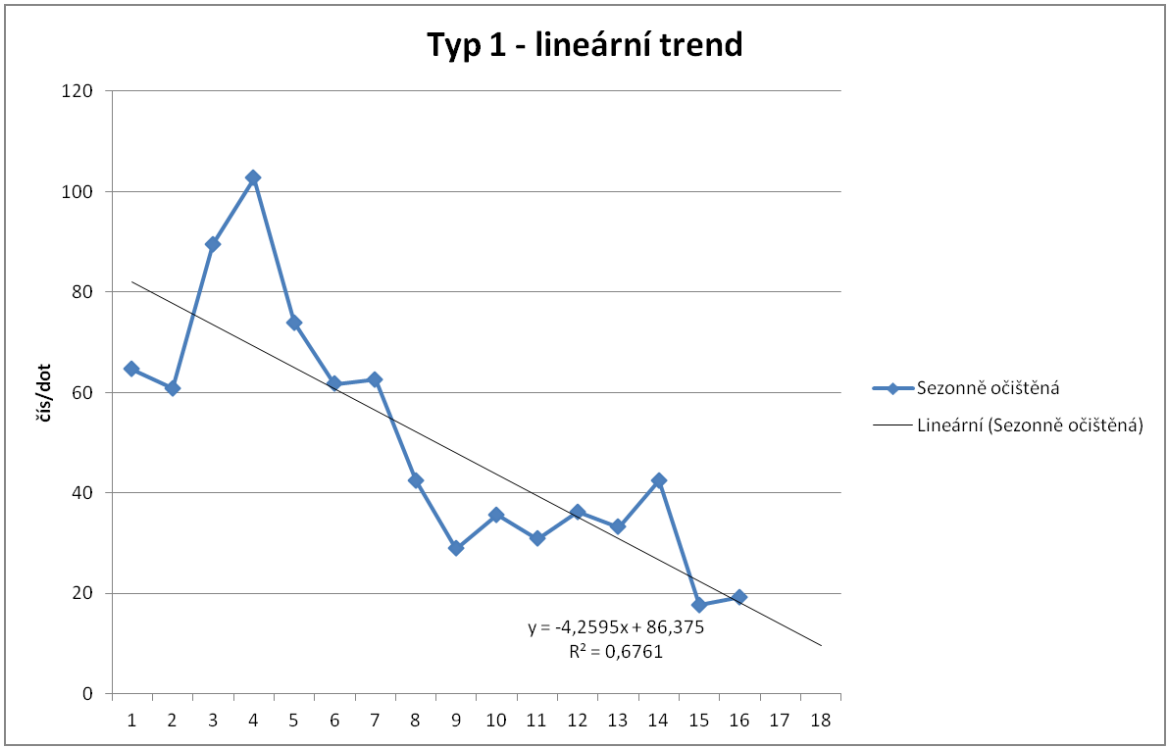
Graf 2: Typ 1 - sezonní čištění

Zdroj: vlastní zpracování

Z Grafu 2 je vidět, že výsledky obou metod mají finálně podobný výsledek. Z důvodu zachování všech hodnot bude zvolena metoda sezonních indexů.

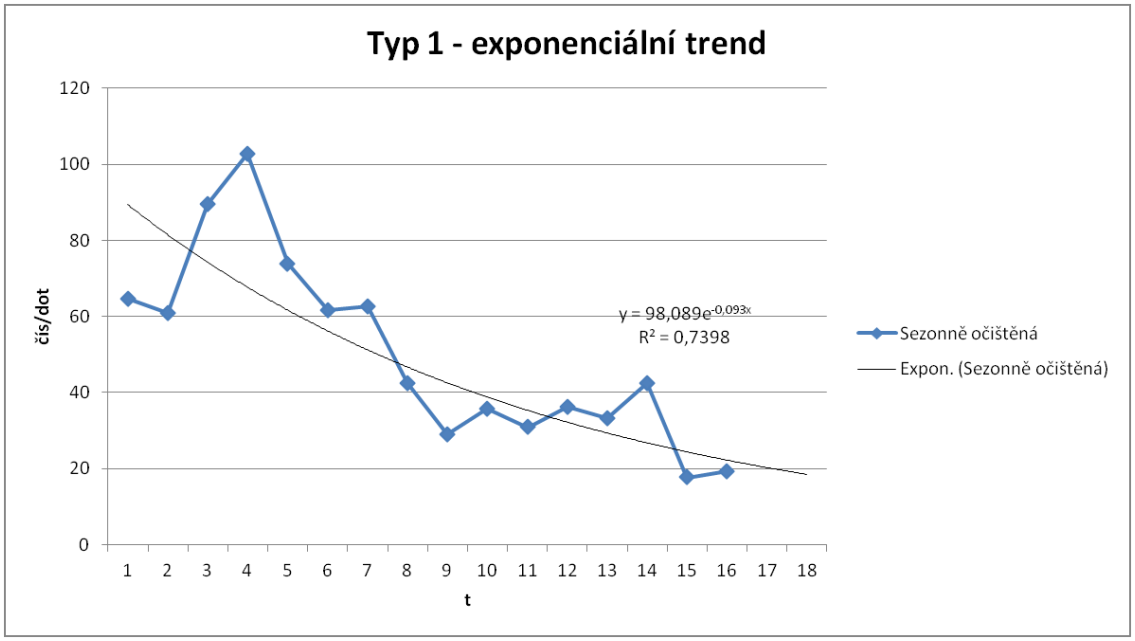
5.1.3. Modelování trendu

K modelování trendu se dá přistoupit různými způsoby. První možnost je spočítat parametry trendové funkce, dopočítat budoucí hodnoty a následně vytvořit graf. Druhá možnost je nechat vykreslit trend, včetně predikce MS Excel, který zároveň spočítá parametry rovnice, a následně pouze dopočítat predikované hodnoty. Druhá možnost je efektivnější, a proto bude využita. Výsledky zobrazené v Grafu 3 a Grafu 4. Tyto grafy slouží k porovnání obou modelů.



Graf 3: Typ 1 - lineární trend

Zdroj: vlastní zpracování



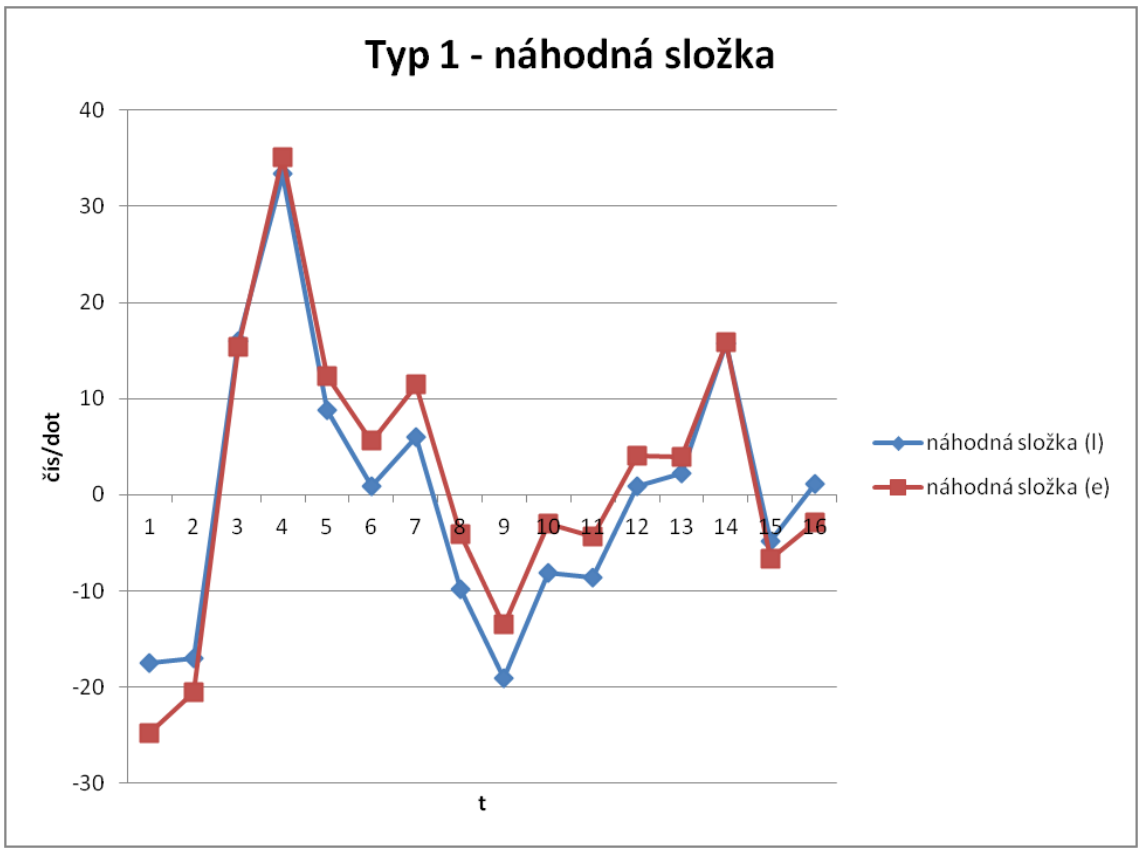
Graf 4: Typ 1 - exponenciální trend

Zdroj: vlastní zpracování

MS Excel vykreslil trendovou křivku, spočítal parametry trendových rovnic a spočetl hodnotu spolehlivosti. Je vidět, že exponenciála přináší lepší výsledek nejen dle odhadu, ale i podle hodnoty spolehlivosti R^2.

5.1.4. Náhodná složka

Dalším krokem zpracování je náhodná složka (residua). Ta je znázorněna v Grafu 5, zda neobsahuje nějaký trend, či cyklus. V legendě (l) znamená lineární a (e) znamená exponenciální trend.



Graf 5: Typ 1 - náhodná složka

Zdroj: vlastní zpracování

Náhodná složka neobsahuje žádný trend u lineárního trendu. (proložená křivka je rovna ose X) a zároveň z ní nevyčnívá žádný cyklus. U exponenciálního trendu lehce roste, což je dáno tvarem exponenciály.

5.1.5. Predikce budoucích hodnot časové řady

Díky znalosti rovnice trendu se snadno dopočítají hodnoty pro t=17 a t=18. Výsledky jsou zobrazeny v Tabulce 12.

Tabulka 12: Typ 1 - predikce

kvartál_typ	t	čís/dot (l)	čís/dot (e)
RP1201_1	17	13,24	19,14
RP1202_1	18	8,57	16,24

Zdroj: vlastní zpracování

Na základě exponenciálního modelu trendu vyšlo, že v 1. kvartále roku 2012 bude potřeba pro pevné linky 19,14 telefonních čísel na dotazník, což znamená, že celkem bude potřeba, při kvótě 1400 dotazníků, 26796 telefonních čísel.

5.2. Typ 6 - O2 mobilní

Jako druhá bude provedena analýza číselné řady týkající se typu 6, tj. O2 mobilních telefonů. Vzhledem k tomu, že půjde o druhou analýzu, bude provedena stručněji.

5.2.1. Zobrazení dat a popisná statistika

Díky poměrně malé velikosti lze zobrazit i analyzovaná data v Tabulce 13.

Tabulka 13: Typ 6 - data

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
čís/dot	40,29	62,34	125,71	85,50	42,40	15,40	74,41	40,35	12,77	42,59	62,19	29,73	54,10	61,35	21,16	21,08

Zdroj: vlastní zpracování

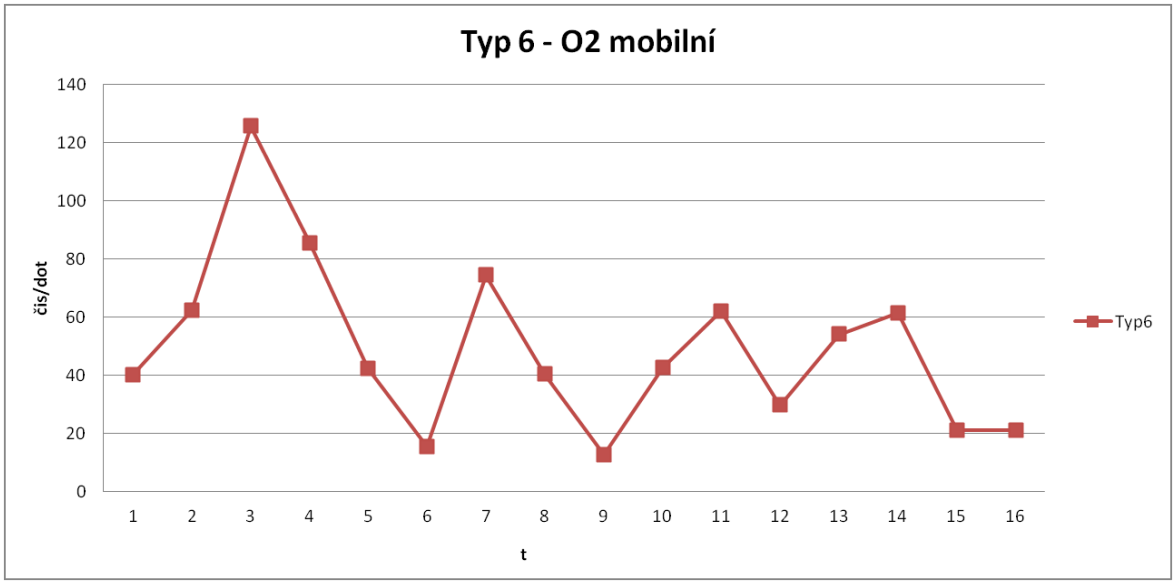
Z těchto dat byla poté spočítána popisná statistika, která je zobrazena v Tabulce 14.

Tabulka 14: Typ 6 - popisná statistika

Stř. hodnota	49,46
Chyba stř. hodnoty	7,36
Medián	42,49
Směr. odchylka	29,43
Rozptyl výběru	866,11
Špičatost	1,68
Šikmost	1,10
Variační rozpětí	112,94
Minimum	12,77
Maximum	125,71
Součet	791,36
Počet	16,00
Hladina spolehlivosti (95	14,42
Useknutý průměr	46,63
Kvartily 25	27,59
Kvartily 75	62,22
Mezikvartilové rozpětí	34,63
Variační koeficient	59,50%

Zdroj: vlastní zpracování

A nakonec znázornění vstupních dat je zobrazeno v Grafu 6.



Graf 6: Typ 6 - znázornění dat

Zdroj: vlastní zpracování

Z Grafu 6 je patrné, že trend je sestupný. Oproti typu 1 ale v menší míře. Dále je vidět, že 3. kvartály obsahují mimořádné vrcholky – řada obsahuje sezonní složku, letní prázdniny, což je období, kdy se populace všeobecně hůře dotazuje. Variační koeficient je větší než u předchozího typu, takže data jsou více vychýlená.

5.2.2. Sezonní čištění

K sezonnímu čištění byla použita metoda sezonních indexů a pro ilustraci i klouzavé průměry tříčlenné a čtyřčlenné. Zároveň byly spočítány absolutní a relativní přírůstky. Vše shrnuje Tabulka 15.

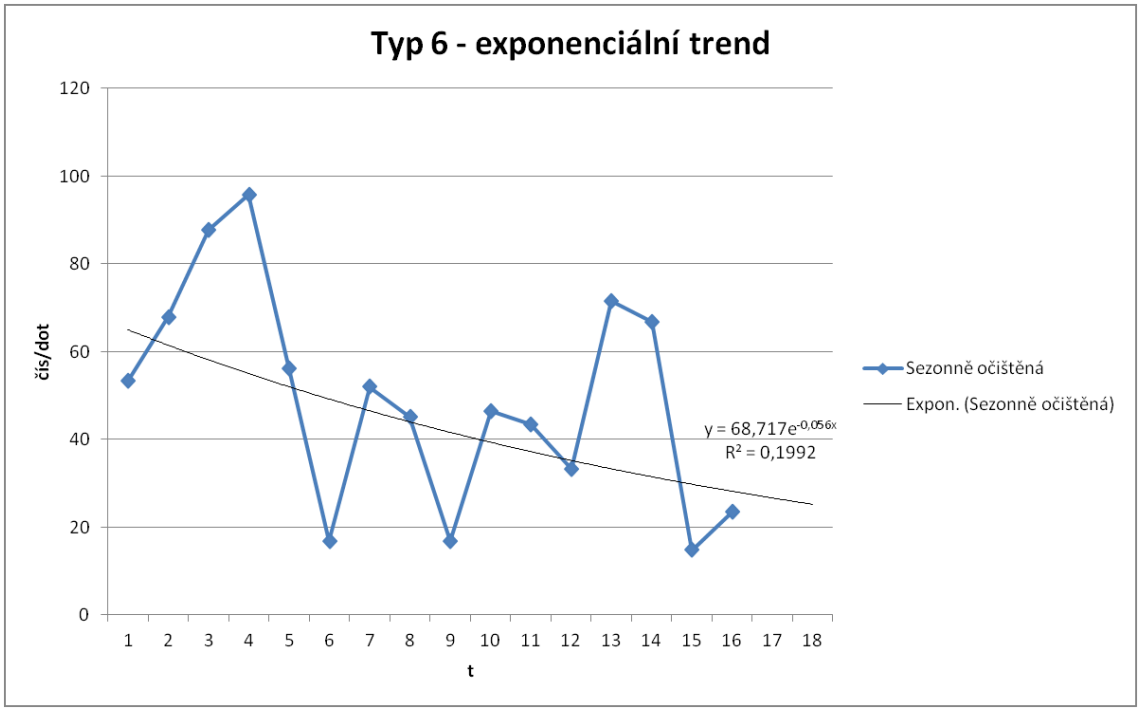
Tabulka 15: Typ 6 - sezonní čištění

kvartál_typ	t	čís/dot (l)	KP (3)	KP (4)	absolutní přírustky	relativní přírustky	index	S indexy	sezonně očištěné
RP0801_6	1	40,29					S1	75,59%	53,30
RP0802_6	2	62,34	76,11		22,05	54,73%	S2	91,83%	67,88
RP0803_6	3	125,71	91,18	78,46	63,37	101,65%	S3	143,28%	87,74
RP0804_6	4	85,50	84,53	78,99	-40,20	-31,98%	S4	89,30%	95,75
RP0901_6	5	42,40	47,77	67,25	-43,10	-50,41%	S1	75,59%	56,09
RP0902_6	6	15,40	44,07	54,43	-26,99	-63,67%	S2	91,83%	16,77
RP0903_6	7	74,41	43,39	43,14	59,00	383,02%	S3	143,28%	51,93
RP0904_6	8	40,35	42,51	35,73	-34,05	-45,77%	S4	89,30%	45,19
RP1001_6	9	12,77	31,90	42,53	-27,59	-68,36%	S1	75,59%	16,89
RP1002_6	10	42,59	39,18	39,47	29,82	233,59%	S2	91,83%	46,38
RP1003_6	11	62,19	44,84	36,82	19,60	46,01%	S3	143,28%	43,40
RP1004_6	12	29,73	48,67	47,15	-32,45	-52,19%	S4	89,30%	33,30
RP1101_6	13	54,10	48,40	51,84	24,37	81,96%	S1	75,59%	71,57
RP1102_6	14	61,35	45,54	41,59	7,25	13,40%	S2	91,83%	66,81
RP1103_6	15	21,16	34,53	39,42	-40,20	-65,52%	S3	143,28%	14,77
RP1104_6	16	21,08			-0,08	-0,38%	S4	89,30%	23,60

Zdroj: vlastní zpracování

5.2.3. Modelování trendu

Co se týká modelování trendu, bude již uvažována pouze exponenciální rovnice trendu. Graf vytvoří a parametry rovnice opět spočítá MS Excel. Výsledky jsou zobrazené v Grafu 7.



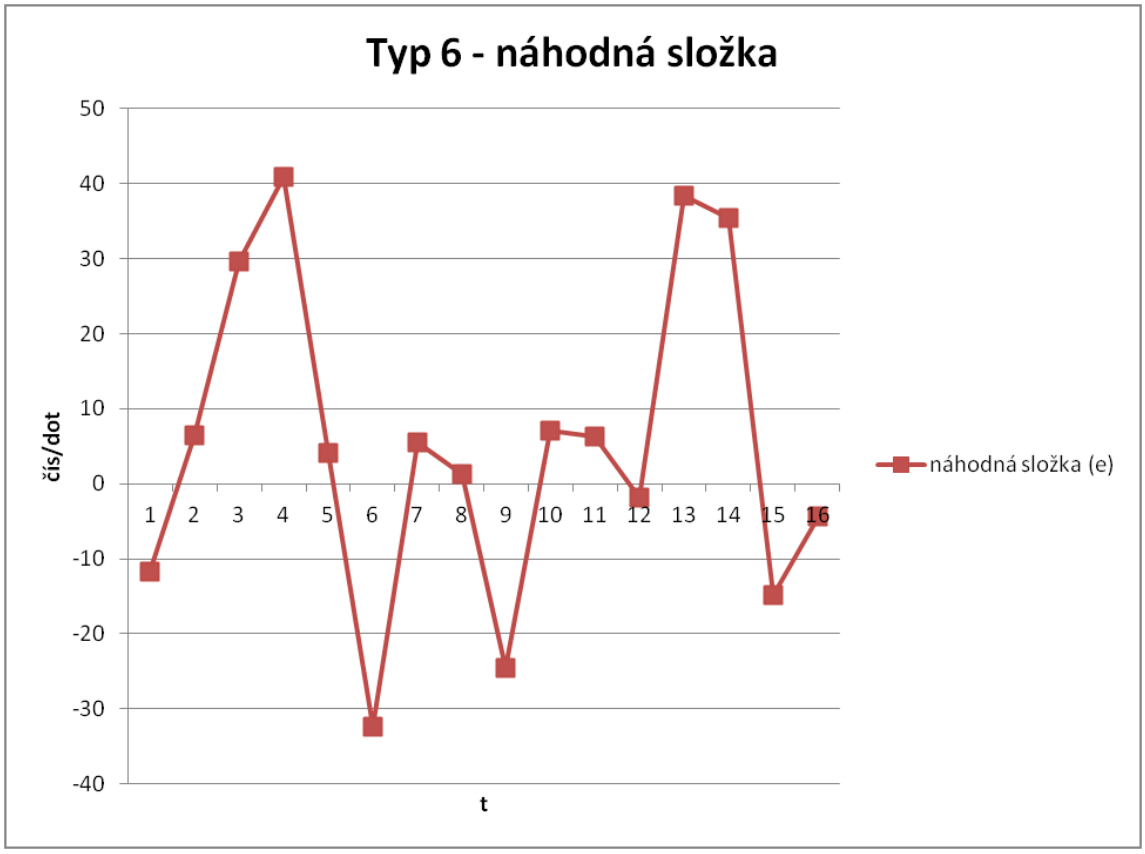
Graf 7: Typ 6 - exponenciální trend

Zdroj: vlastní zpracování

Z Grafu 7 je vidět malá hodnota spolehlivosti R^2. V ideálním případě se blíží k jedné. To jasně ukazuje velmi silnou náhodnou složku. Hodnota spolehlivosti by šla zlepšit vysledováním odlehlých hodnot a jejich vypuštěním. Bohužel by bylo potřeba vypustit větší množství hodnot, což by při tomto množství hodnotu zlepšilo, ale zároveň by to zkreslilo výsledky.

5.2.4. Náhodná složka

Dalším krokem zpracování je náhodná složka (residua). Ta je znázorněna pouze v Grafu 8, zda neobsahuje nějaký trend, či cyklus.



Graf 8: Typ 6 - náhodná složka

Zdroj: vlastní zpracování

Opět je vidět značný vliv náhodné složky, je vidět řada poměrně zásadních zlomů.

5.2.5. Predikce budoucích hodnot časové řady

Díky znalosti rovnice trendu se snadno dopočítají hodnoty pro t=17 a t=18. Výsledky jsou zobrazeny v Tabulce 16.

Tabulka 16: Typ 6 - predikce

kvartál_typ	t	čís/dot (e)
RP1201_6	17	20,05
RP1202_6	18	23,03

Zdroj: vlastní zpracování

Na základě exponenciálního modelu trendu vyšlo, že v 1. kvartálu roku 2012 bude potřeba pro mobilního operátora O2 20,05 telefonních čísel na dotazník, což znamená, že celkem bude potřeba, při kvótě 798 dotazníků, 16000 telefonních čísel.

5.3. Typ 7 - Vodafone

Jako třetí bude provedena analýza číselné řady týkající se Typu 7, tj. mobilních telefonů národního operátora Vodafone. Tato analýza bude opět provedena stručněji.

5.3.1. Zobrazení dat a popisná statistika

Díky poměrně malé velikosti lze zobrazit i analyzovaná data v Tabulce 17.

Tabulka 17: Typ 7 - data

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
čís/dot	59,71	32,50	66,95	53,82	25,77	24,96	66,57	29,09	26,84	36,12	64,85	42,16	47,15	50,30	20,82	21,55

Zdroj: vlastní zpracování

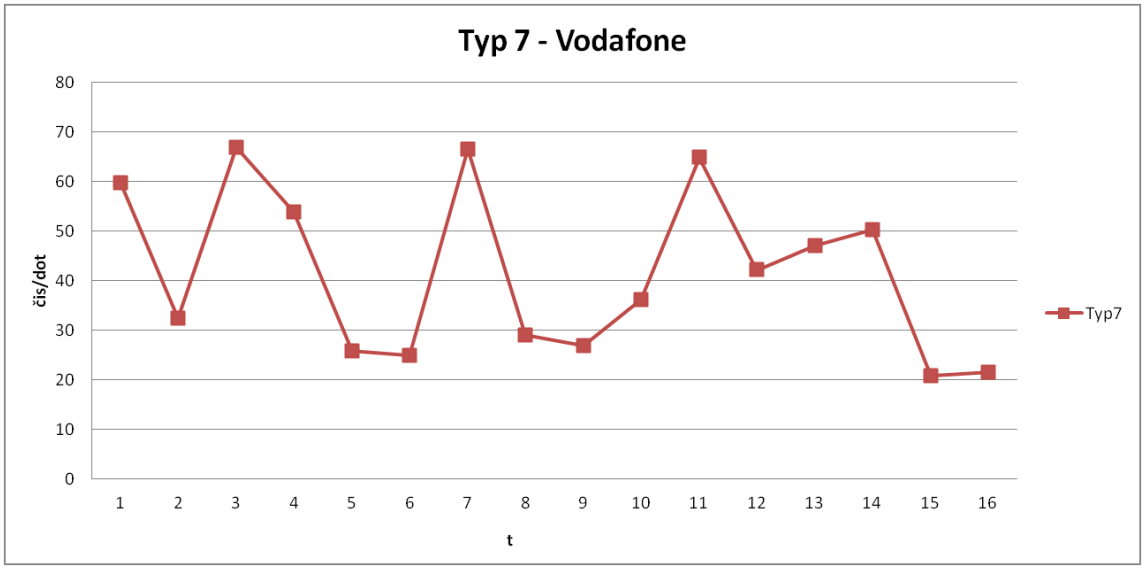
Z těchto dat byla poté spočítána popisná statistika, která je zobrazena v Tabulce 18.

Tabulka 18: Typ 7 - popisná statistika

Stř. hodnota	41,82
Chyba stř. hodnoty	4,20
Medián	39,14
Směr. odchylka	16,81
Rozptyl výběru	282,57
Špičatost	-1,49
Šikmost	0,30
Variační rozpětí	46,13
Minimum	20,82
Maximum	66,95
Součet	669,17
Počet	16,00
Hladina spolehlivosti (95	8,24
Useknutý průměr	41,53
Kvartily 25	26,57
Kvartily 75	55,29
Mezikvartilové rozpětí	28,72
Variační koeficient	40,19%

Zdroj: vlastní zpracování

A nakonec znázornění vstupních dat je zobrazeno v Grafu 9.



Graf 9: Typ 7 - znázornění dat

Zdroj: vlastní zpracování

Z Grafu 9 lze odhadnout, že trend je sestupný, a to celkem mírně. Dále je vidět, že 3. kvartály obsahují mimořádné vrcholky – řada obsahuje sezonní složku, letní prázdniny, což je období, kdy se populace všeobecně hůře dotazuje.

5.3.2. Sezonní čištění

K sezonnímu čištění byla použita metoda sezonních indexů a pro ilustraci i klouzavé průměry tříčlenné a čtyřčlenné. Zároveň byly spočítány absolutní a relativní přírůstky. Vše shrnuje Tabulka 19.

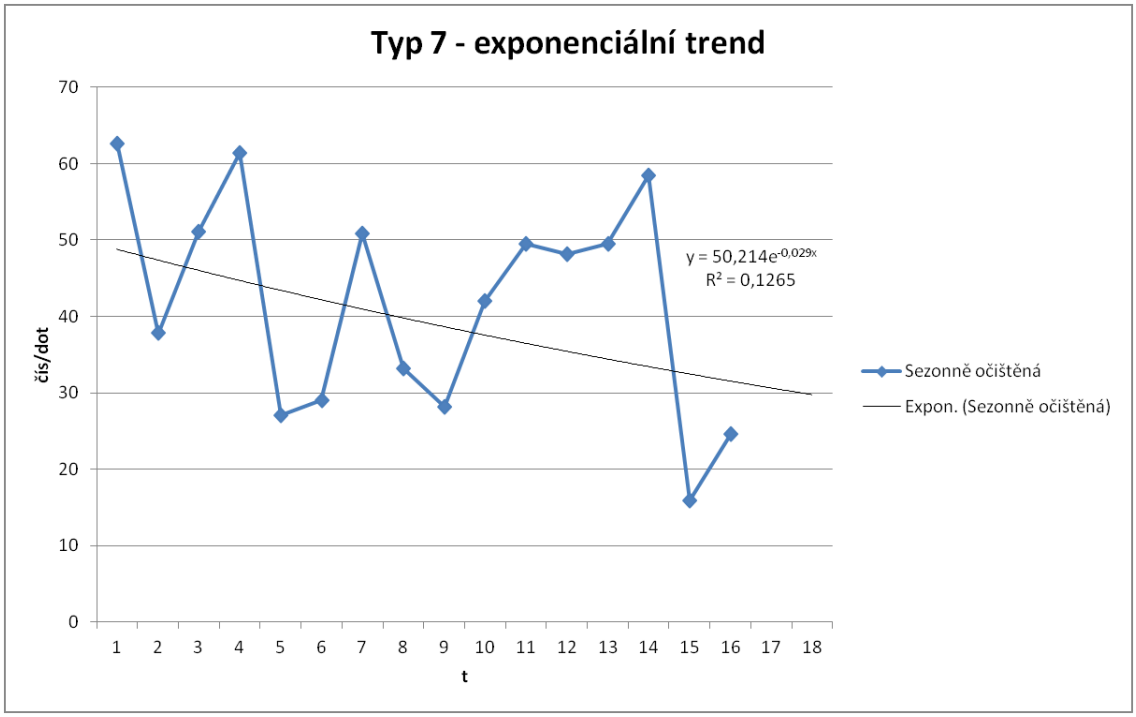
Tabulka 19: Typ 7 - sezonní čištění

kvartál_typ	t	čís/dot (l)	čís/dot (e)	KP (3)	KP (4)	absolutní přírustky	relativní přírustky	index	S indexy	sezonně očištěné
RP0801_7	1	59,71	59,71					S1	95,32%	62,64
RP0802_7	2	32,50	32,50	53,05		-27,21	-45,57%	S2	86,01%	37,79
RP0803_7	3	66,95	66,95	51,09	53,25	34,45	106,01%	S3	131,02%	51,10
RP0804_7	4	53,82	53,82	48,85	44,76	-13,13	-19,61%	S4	87,65%	61,40
RP0901_7	5	25,77	25,77	34,85	42,88	-28,05	-52,12%	S1	95,32%	27,03
RP0902_7	6	24,96	24,96	39,10	42,78	-0,81	-3,14%	S2	86,01%	29,02
RP0903_7	7	66,57	66,57	40,21	36,60	41,61	166,71%	S3	131,02%	50,81
RP0904_7	8	29,09	29,09	40,83	36,87	-37,48	-56,30%	S4	87,65%	33,19
RP1001_7	9	26,84	26,84	30,68	39,66	-2,25	-7,75%	S1	95,32%	28,15
RP1002_7	10	36,12	36,12	42,60	39,22	9,28	34,59%	S2	86,01%	42,00
RP1003_7	11	64,85	64,85	47,71	42,49	28,73	79,53%	S3	131,02%	49,49
RP1004_7	12	42,16	42,16	51,39	47,57	-22,68	-34,98%	S4	87,65%	48,10
RP1101_7	13	47,15	47,15	46,54	51,12	4,99	11,83%	S1	95,32%	49,46
RP1102_7	14	50,30	50,30	39,42	40,11	3,15	6,68%	S2	86,01%	58,49
RP1103_7	15	20,82	20,82	30,89	34,96	-29,48	-58,61%	S3	131,02%	15,89
RP1104_7	16	21,55	21,55			0,73	3,52%	S4	87,65%	24,59

Zdroj: vlastní zpracování

5.3.3. Modelování trendu

Co se týká modelování trendu, bude již uvažována pouze exponenciální rovnice trendu. Graf vytvoří a parametry rovnice spočítá MS Excel. Výsledky jsou zobrazené v Grafu 10.



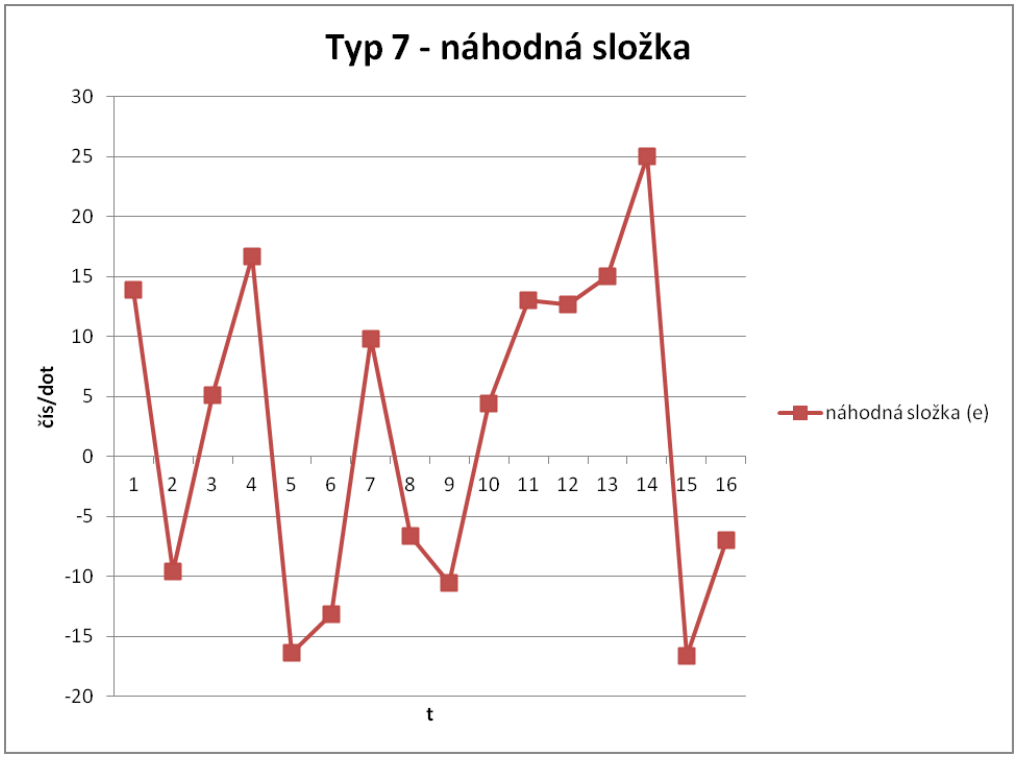
Graf 10: Typ 7 - exponenciální trend

Zdroj: vlastní zpracování

Graf 10 má velmi neuspořádaný průběh. Malá spolehlivosti R^2 je daná obrovským skokem z bodu, kde t nabývá hodnotu 14 na hodnotu 15, který byl způsoben zlepšenou technologii detekce existence telefonních čísel. Lze říct, že po cyklickém průběhu prvních dvou let nabyla na intenzitě náhodná složka. Tím pádem je hodnota R^2 od jedné hodně vzdálená. Dá se pouze spekulovat, čím to bylo způsobeno, ale k odpovědi nejsou dostupné žádné podklady.

5.3.4. Náhodná složka

Dalším krokem zpracování je náhodná složka (residua). Ta je znázorněna Grafu 11, zda neobsahuje nějaký trend, či cyklus.



Graf 11: Typ 7 - náhodná složka

Zdroj: vlastní zpracování

Náhodná složka neobsahuje žádný trend a zároveň z ní nevyčnívá žádný jasný cyklus. Zároveň jsou vidět poměrně vysoké hodnoty náhodné složky.

5.3.5. Predikce budoucích hodnot časové řady

Díky znalosti rovnice trendu se již snadno dopočítají hodnoty pro t=17 a t=18. Výsledky jsou zobrazeny v Tabulce 20.

Tabulka 20: Typ 7 - predikce

kvartál_typ	t	čís/dot (e)
RP1201_7	17	29,24
RP1202_7	18	25,62

Zdroj: vlastní zpracování

Na základě exponenciálního modelu trendu vyšlo, že v 1. kvartále roku 2012 bude potřeba pro mobilního operátora Vodafone 29,24 telefonních čísel na dotazník, což znamená, že celkem bude potřeba, při kvótě 462 dotazníků, 13508 telefonních čísel.

5.4. Typ 8 T-Mobile

Jako poslední bude provedena analýza číselné řady týkající se typu 8, tj. národního operátora T-Mobile. Opět půjde o stručnější analýzu.

5.4.1. Zobrazení dat a popisná statistika

Díky poměrně malé velikosti lze zobrazit i analyzovaná data v Tabulce 21.

Tabulka 21: Typ 8 - data

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
čís/dot	93,03	39,94	104,04	69,79	30,36	39,57	69,28	41,88	42,74	40,94	68,82	36,02	49,23	57,98	20,01	17,12

Zdroj: vlastní zpracování

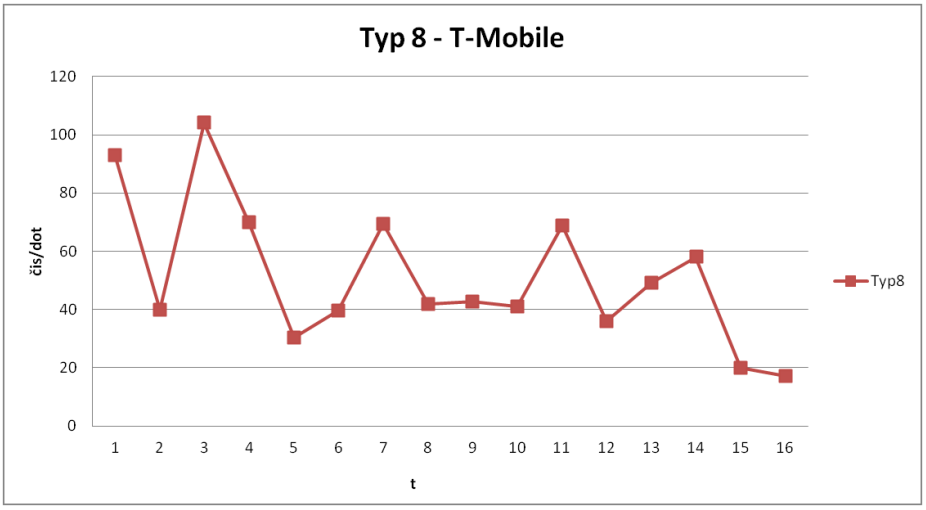
Z těchto dat byla poté spočítána popisná statistika, která je zobrazena v Tabulce 22.

Tabulka 22: Typ 8 - popisná statistika

Stř. hodnota	51,30
Chyba stř. hodnoty	6,09
Medián	42,31
Směr. odchylka	24,36
Rozptyl výběru	593,28
Špičatost	0,17
Šikmost	0,78
Variační rozpětí	86,92
Minimum	17,12
Maximum	104,04
Součet	820,75
Počet	16,00
Hladina spolehlivosti (95	11,93
Useknutý průměr	49,97
Kvartily 25	38,69
Kvartily 75	68,93
Mezikvartilové rozpětí	30,25
Variační koeficient	47,48%

Zdroj: vlastní zpracování

A nakonec jsou vstupní data zobrazena v Grafu 12.



Graf 12: Typ 8 - znázornění dat

Zdroj: vlastní zpracování

Z Grafu 12 je vidět, že trend je mírně sestupný. Dále je vidět, že 3. kvartály obsahují mimořádné vrcholky – řada obsahuje sezonní složku, letní prázdniny, což je období, kdy se populace všeobecně hůře dotazuje.

5.4.2. Sezonní čištění

K sezonnímu čištění byla použita metoda sezonních indexů a pro ilustraci i klouzavé průměry tříčlenné a čtyřčlenné. Zároveň byly spočítány absolutní a relativní přírůstky. Vše shrnuje Tabulka 23.

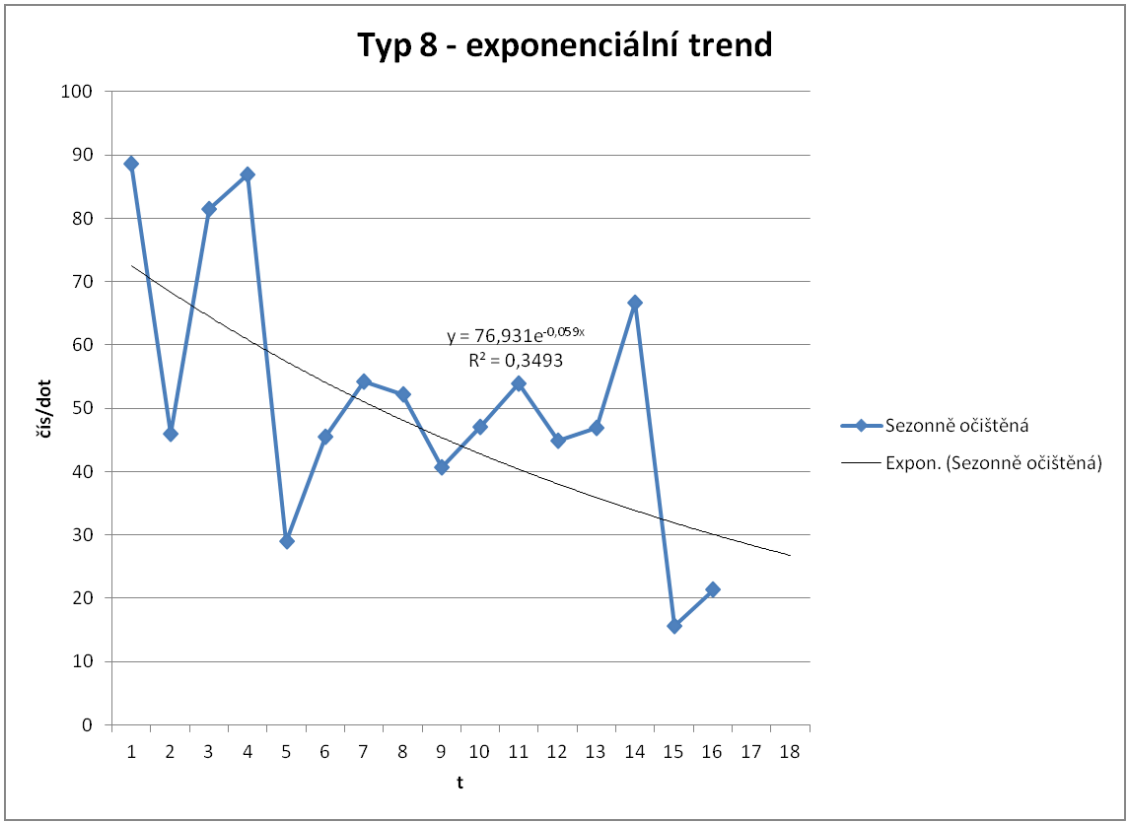
Tabulka 23: Typ 8 - sezonní čištění

kvartál_typ	t	čís/dot (l)	čís/dot (e)	KP (3)	KP (4)	absolutní přírustky	relativní přírustky	index	S indexy	sezonně očištěné
RP0801_8	1	93,03	93,03					S1	1,05	88,64
RP0802_8	2	39,94	39,94	79,00		-53,09	-0,57	S2	0,87	45,93
RP0803_8	3	104,04	104,04	71,26	76,70	64,10	1,61	S3	1,28	81,43
RP0804_8	4	69,79	69,79	68,06	61,03	-34,25	-0,33	S4	0,80	86,89
RP0901_8	5	30,36	30,36	46,57	60,94	-39,44	-0,57	S1	1,05	28,92
RP0902_8	6	39,57	39,57	46,40	52,25	9,22	0,30	S2	0,87	45,51
RP0903_8	7	69,28	69,28	50,24	45,27	29,71	0,75	S3	1,28	54,23
RP0904_8	8	41,88	41,88	51,30	48,37	-27,40	-0,40	S4	0,80	52,14
RP1001_8	9	42,74	42,74	41,85	48,71	0,86	0,02	S1	1,05	40,72
RP1002_8	10	40,94	40,94	50,83	48,60	-1,80	-0,04	S2	0,87	47,08
RP1003_8	11	68,82	68,82	48,59	47,13	27,87	0,68	S3	1,28	53,86
RP1004_8	12	36,02	36,02	51,36	48,75	-32,79	-0,48	S4	0,80	44,85
RP1101_8	13	49,23	49,23	47,74	53,01	13,20	0,37	S1	1,05	46,90
RP1102_8	14	57,98	57,98	42,41	40,81	8,76	0,18	S2	0,87	66,67
RP1103_8	15	20,01	20,01	31,70	36,08	-37,97	-0,65	S3	1,28	15,66
RP1104_8	16	17,12	17,12			-2,89	-0,14	S4	0,80	21,31

Zdroj: vlastní zpracování

5.4.3. Modelování trendu

Co se týká modelování trendu, bude již uvažována pouze exponenciální rovnice trendu. Graf vytvoří a parametry rovnice spočítá MS Excel. Výsledky jsou zobrazené v Grafu 13.



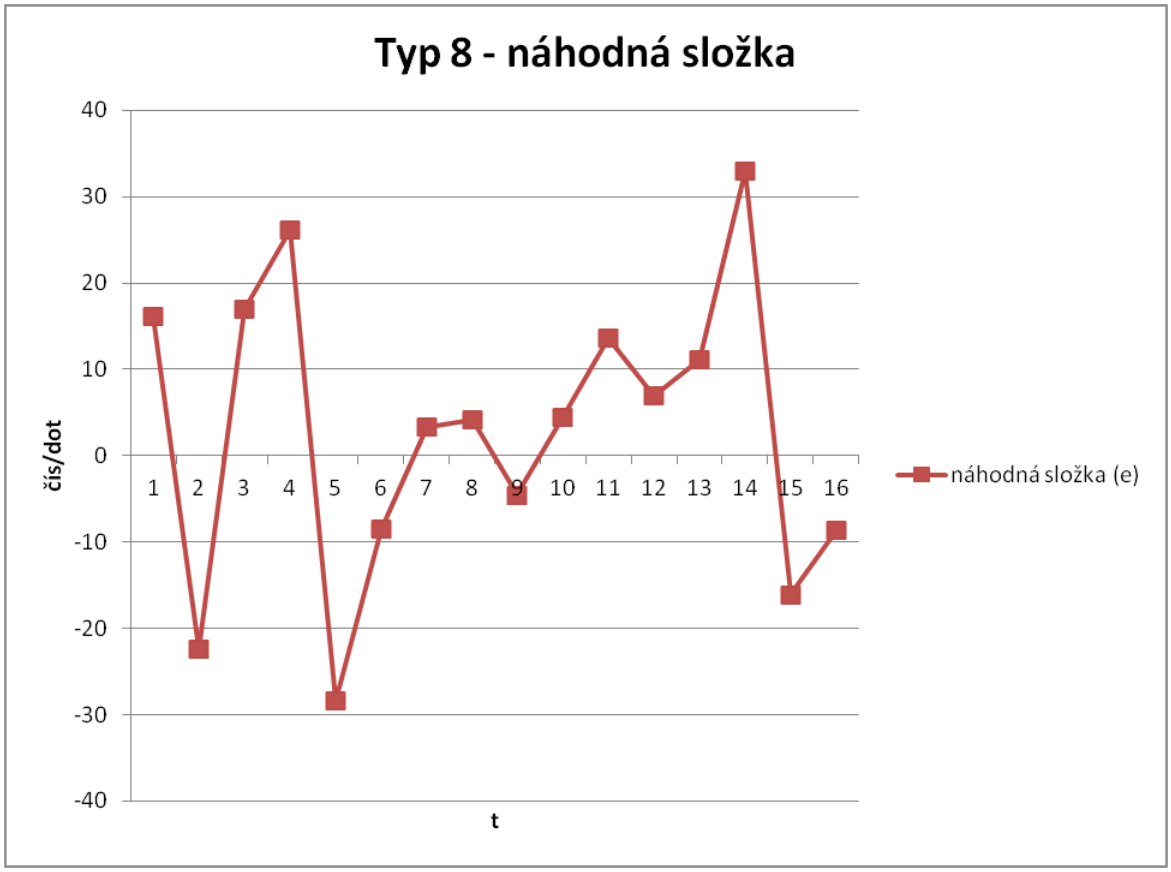
Graf 13: Typ 8 - exponenciální trend

Zdroj: vlastní zpracování

V Grafu 13 je opět vidět malá hodnota spolehlivosti R^2. Interpretace je identická, jako u Graf 10., viz kapitola 5.3.3.

5.4.4. Náhodná složka

Dalším krokem zpracování je náhodná složka (residua). Ta je pouze znázorněna v Grafu 14, zda neobsahuje nějaký trend, či cyklus.



Graf 14: Typ 8 - náhodná složka

Zdroj: vlastní zpracování

Náhodná složka neobsahuje žádný jasný trend a zároveň z ní nevyčnívá žádný cyklus. Chování náhodné složka je odpovídající teorii.

5.4.5. Predikce budoucích hodnot časové řady

Díky znalosti rovnice trendu se snadno dopočítají hodnoty pro t=17 a t=18. Výsledky jsou zobrazeny v Tabulce 24.

Tabulka 24: Typ 8 - predikce

kvartál_typ	t	čís/dot (e)
RP1201_8	17	29,61
RP1202_8	18	23,13

Zdroj: vlastní zpracování

Na základě exponenciálního modelu trendu vyšlo, že v 1. kvartále roku 2012 bude potřeba pro mobilního operátora T-Mobile 29,61 telefonních čísel na dotazník, což znamená, že celkem bude potřeba, při kvótě 840 dotazníků, 24872 telefonních čísel.

6. SHRNUTÍ VÝSLEDKŮ

V Tabulce 25 si lze prohlédnou shrnutí výsledků.

Tabulka 25: Finální výsledky

	typ1	typ6	typ7	typ8	Celkem
čís/dot	19,14	20,05	29,24	29,61	23,19
kvóta	1400	798	462	840	3500
čísel	26795	15999	13507	24876	81178

Zdroj: vlastní zpracování

Z toho shrnutí vyplývá jak počet telefonních čísel pro jednotlivé typy, tak i celková mocnost databáze. Průběh dat v časové ose pro všechny národní operátory je zobrazen v souhrnném grafu v příloze B.

U typu 1 je při importu potřeba dodržet, v rozumné míře, rozložení na VMB a kraje. Zdálo by se, že pokud by se čísla z databáze pevných linek náhodně vybírala, pak by podle statistických zákonitostí mělo rozložení vybraných čísel ze vzorku odpovídat rozložení ČR. Vzhledem k tomu, že se provádějí průzkumy cílené na jednotlivé regiony, a čísla, která se použijí, se nesmí jeden rok používat, je opravdu nutné ohlídat rozložení vzorku. U ostatních typů, vzhledem k náhodnému generování, není třeba při výběru věnovat nutnou péči dalším parametrům.

Dosavadní praxe, připravit okolo 50.000 čísel pro jednotlivé typy tím pádem může být opuštěna. Výsledek znamená, že výchozí databáze bude mít čtyřicetiprocentní velikost oproti původní, což přenáší následující výhody:

- rychlejší otvírání projektu,
- rychlejší zálohování,
- méně diskového prostoru potřebného pro zálohy,
- rychlejší rekční doba při práci nad databází (statistiky, změny filtrů).

ZÁVĚR

Cílem této práce bylo stanovení počátečních parametrů databáze telefonních čísel sloužící k provádění marketingového průzkumu Radioprojekt, a to pomocí analýzy předchozích vln za využití metodologie CRISP-DM.

V rámci porozumění problematice byl v úvodu představen průzkum trhu Radioprojekt, a to v základních rysech včetně realizátora. Finální výsledek tohoto kontinuálního marketingového průzkumu je veřejně přístupný ve formě tiskové zprávy a prezentace. Dále bylo prezentováno programové vybavení sloužící ke sběru dat pro tento kvantitativní CATI průzkum.

Co se týká porozumění datům, tak byly představeny potřebné datové struktury, zobrazené v příloze A, včetně vysvětlení kódování nejdůležitějších tabulek.

Poté bylo v rámci přípravy dat provedeno předzpracování dat, které poskytlo po separaci, čištění, agregaci a vytvoření nových proměnných zdroj dat pro vlastní analýzu. Tímto zdrojem se staly čtyři časové řady po šestnácti časových obdobích - kvartálech. Tato etapa byla časově nejnáročnější.

Následovalo vlastní modelování postavené na dekompozici časové řady na jednotlivé složky, v tomto případě na sezonní složku, trendovou a složku náhodnou. Díky této dekompozici byla možnost predikovat vývoj v časové ose dopředu. Vzhledem k zadání úlohy bylo rozhodnuto používat exponenciální trend, což potvrdilo jeho srovnání s trendem lineárním.

Model poskytl výsledky, to jest predikci budoucích hodnot časových řad pro jednotlivé telefonní operátory. Zároveň model ukázal, že dlouhodobější předpověď nedává přílišný smysl, a že bude nutné, po každém kvartále, model aktualizovat. Vzhledem k předpokladu menšího vlivu budoucích technologických a organizačních změn by se model měl časem stabilizovat.

Využití výsledků modelu do praxe přineslo zmenšení počtu záznamů tabulky telefonních čísel z 200 000 na 82 000, to jest na dvě pětiny původního stavu. Toto zmenšení přineslo jasnou úsporu režie správy této databáze a zlepšilo reakční doby běhu aplikačního software běžícího nad touto databází.

Dalším cílem bude, aby se udržela podoba trendu odpovídající pokročilé části exponenciály, eventuálně lehce rostoucí přímce. Lidé projevují čím dál menší ochotu odpovídat na výzkumy trhu, takže po vyčerpání všech zlepšení, tj. když křivka v modelu bude prakticky rovnoběžná s osou X, tato neochota trend otočí na mírně rostoucí.

POUŽITÁ LITERATURA

[1] ARLT Josef, ARLTOVÁ Markéta. *Ekonomické časové řady*. 1. vydání. Praha: Grada Publishing 2007. 285 s. ISBN 978-80-247-1319-9

[2] BÁRTOVÁ Hilda, BÁRTA Vladimír, KOUDELKA Jan. *Spotřebitel (chování spotřebitele a jeho výzkum).* 1. vydání. Praha: Vysoká škola ekonomická v Praze, Nakladatelství Oeconomica 2007. 254 s. ISBN 978-80-245-1275-4

[3] BERKA Pavel. *Dobývání znalostí z databází.* 1. vydání. Praha: Academia, nakladatelství Akademie věd České republiky 2003. 366 s. ISBN 80-200-1062-9

[4] HAUGE Paul. *Průzkum trhu.* 3. vydání. Brno: Computer Press 2003. 234s. ISBN 80-7226-917-8

[5] HINDLS Richard, HRONOVÁ Stanislava, SEGER Jan. *Statistika pro ekonomy*. 3. vydání. Praha: Professional Publishing 2003. 415 s. ISBN 80-86419-34-7

[6] KOZEL Roman, MYNÁŘOVÁ Lenka, SVOBODOVÁ Hana. *Moderní metody a techniky marketingového průzkumu.* 1. vydání. Praha: Grada 2011. 304 s. ISBN 978-80-247-3527-6

[7] KRUCZEK Aleš. *Microsoft Office Access 2007 Podrobná uživatelská příručka*. 1. vydání. Brno: Computer Press 2007. 364 s. ISBN 978-80-251-1608-1

[8] Median Projektová databáze RP. Praha: Median 2012. RP1201.MDB

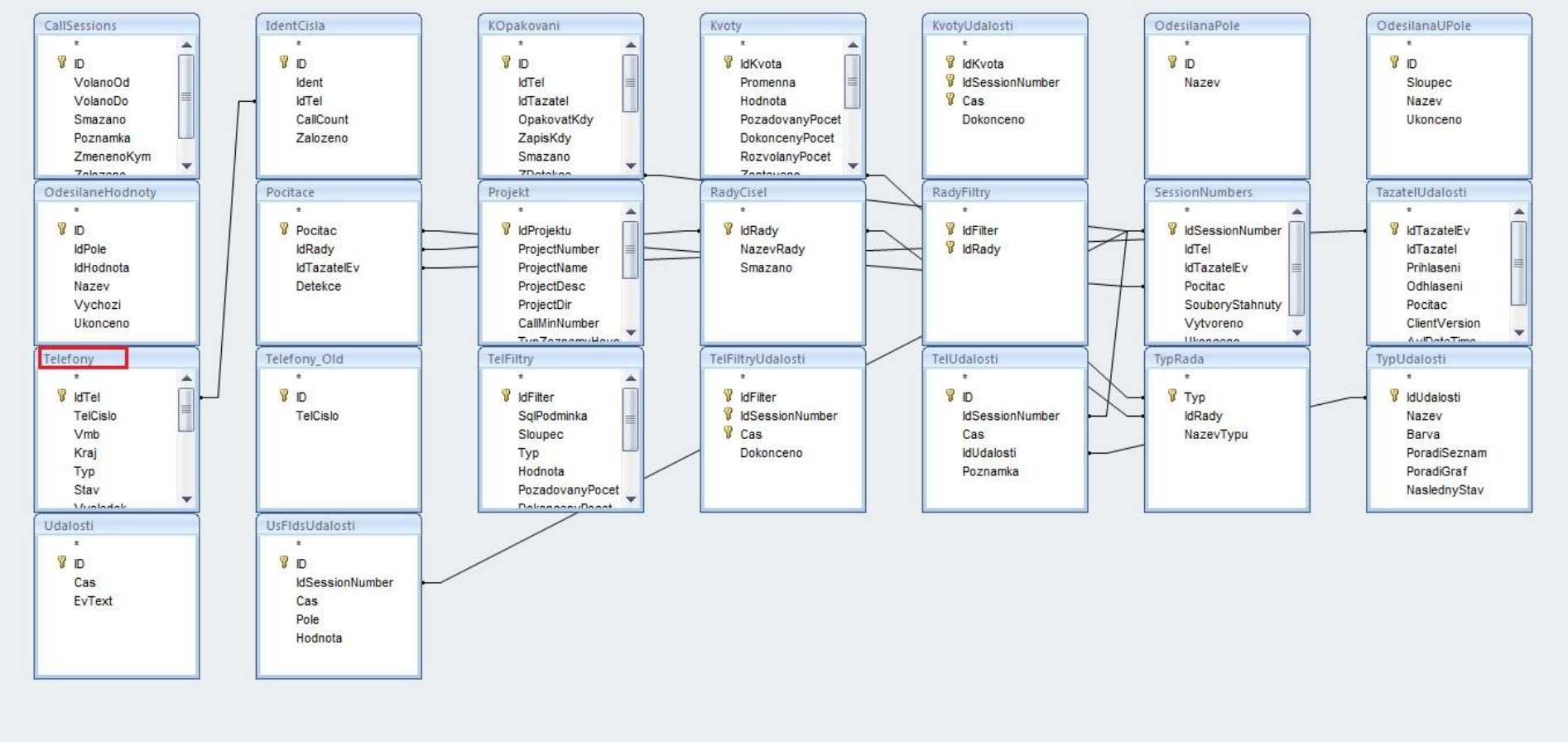
[9] Median. CATISW [online]. Praha: Median 2005. CATISW.

[10] MELOUN Milan, MILITKÝ Jiří. *Kompendium statistického zpracování dat.* 1. vydání. Praha: Academia, nakladatelství Akademie věd České republiky 2002. 764 s. ISBN 80-200-1008-4

[11] SEGER Jan, HINDLS Richard. *Statistické metody v tržním hospodářství.* 1. vydání. Praha: Victoria Publishing 1995. 435 s. ISBN 80-7187-058-7

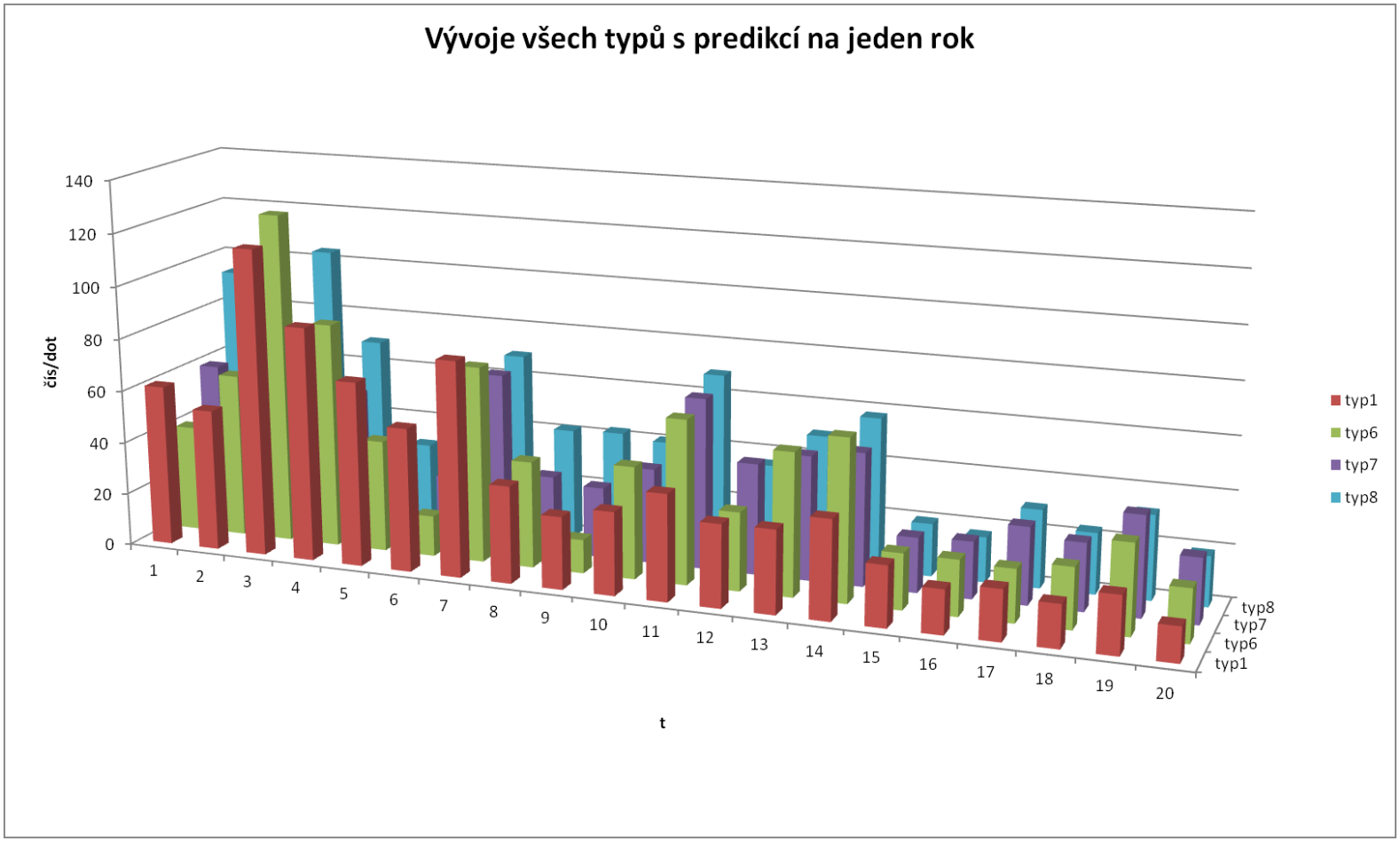
SEZNAM PŘÍLOH

Příloha A: Schéma projektové databáze



Zdroj:[9]

Příloha B: Grafické znázornění všech typů s predikcí na jeden rok



Zdroj: vlastní zpracování