SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Ing. Matej Kollár

Autoreferát dizertačnej práce

Aplikácia klasterizačných algoritmov na namerané dáta kalibrácií antén

na získanie akademického titulu philosophiae doctor, PhD.

v doktorandskom študijnom programe: 5.2.54 Meracia technika

Miesto a dátum: Bratislava, September 2012

Dizertačná práca bola vypracovaná v externej forme doktorandského štúdia

na ústave elektrotechniky FEI STU v Bratislave.

Predkladateľ: Ing. Matej Kollár
Ústav elektrotechniky FEI STU v Bratislave
Ilkovičova 3, 812 19 Bratislava

Školiteľ: doc. Ing. Karol Kováč, PhD.
Ústav elektrotechniky FEI STU v Bratislave
Ilkovičova 3, 812 19 Bratislava

Oponenti: Prof. Ing. Ivan Kneppo, DrSc. rektor
Trenčianska univerzita Alexandra Dubčeka
Študentská 2
911 50 Trenčín

Mgr. Svorad Štolc, PhD.
Ústav merania SAV
Dúbravská cesta 9
841 04 Bratislava 4

Autoreferát bol rozoslaný: .....

Obhajoba dizertačnej práce sa koná: ..... **o** ..... **h.**

na FEI STU v Bratislave, ….....,
FEI STU, Ilkovičova 3, 812 19 Bratislava.

Dekan FEI STU v Bratislave
prof. RNDr. Gabriel Juhás, PhD.

Obsah

1. ÚVOD

Kalibrácie antén sú dôležitým meraním, ktoré umožňuje súčasnému priemyslu uskutočňovať najrôznejšie emisné a imunitné testy elektronických zariadení. Účelom kalibračnej procedúry antény je získanie anténneho faktora slúžiaceho na korekciu prijímaného signálu. Pri kalibrácii dochádza k rôznym vplyvom, ktoré formujú výsledný anténny faktor, ako sú napríklad výrobné vlastnosti, chyby spôsobené meracím systémom, chyby konfigurácie, či chyby metódy.

Táto práca sa zameriava na analýzu a validáciu dát nameraných počas kalibrácie antén. Základnou myšlienkou je preskúmanie výsledkov približne 6000 kalibrácií uskutočnených v období posledných 14 rokov. Prostredníctvom tohoto jedinečného súboru dát je možné analyzovať rôzne modely antén, kalibračné metódy a ich parametre.

Hlavnou motiváciou projektu je potreba systematizácie daných dát a tiež vývoj procedúr, poprípade softvérového systému pre verifikáciu kalibračných meraní antén a tým zvýšenie kvality služieb zaoberajúcich sa kalibráciou antén. Zároveň, získané a vhodne usporiadané dáta je neskôr možné použiť pre budúce ciele.

Nevyhnutnou súčasťou spracovania kalibračných dát je ich získanie z rôznych distribouvaných zdrojov. Počas dlhého obdobia existencie kalibračných služieb antén na našom pracovisku, sa formát v ktorom boli dáta ukladané pozvoľna menil. Rovnako prechádzali vývojom aj použité kalibračné metódy (štandardy). Za týmto účelom je nutné vytvoriť softvérový systém, ktorý naše dáta vyextrahuje z kalibračných certifikátov uložených na disku počítača a následne ich uloží v databáze.

Pri zobrazení frekvenčných závislostí anténnych faktorov v grafe je možné pozorovať ich usporiadanie do prirodzených skupín tzv. klastrov. Rozlíšením týchto skupín je možné získať užitočné informácie. Napríklad je možné povedať, či sa špecifický kus antény líši od ostatných rovnakého modelu alebo či sú výsledky danej kalibrácie platné. Pre účely klasifikácie dát do podskupín preto použijeme klastrovú analýzu. Preštudujeme rôzne jej metódy a otestujeme ich aplikáciu na získaných kalibračných dátach.

Všetky kalibračné dáta boli poskytnuté rádio-frekvenčnou divíziou rakúskej spoločnosti Seibersdorf Labor GmbH, ktorá je od roku 1996 akreditovaným kalibračným laboratóriom ÖKD 13 pre antény a senzory poľa. Jej produkty a služby využívajú popredné svetové spoločnosti ako sú Philips, Rohde & Schwarz, Hewlett Packard, Liberty Labs, TÜV, Swisscom, Nokia, Samsung, LG, Hyundai, KIA, IBM, Cetecom, Motorola a mnoho ďalších.

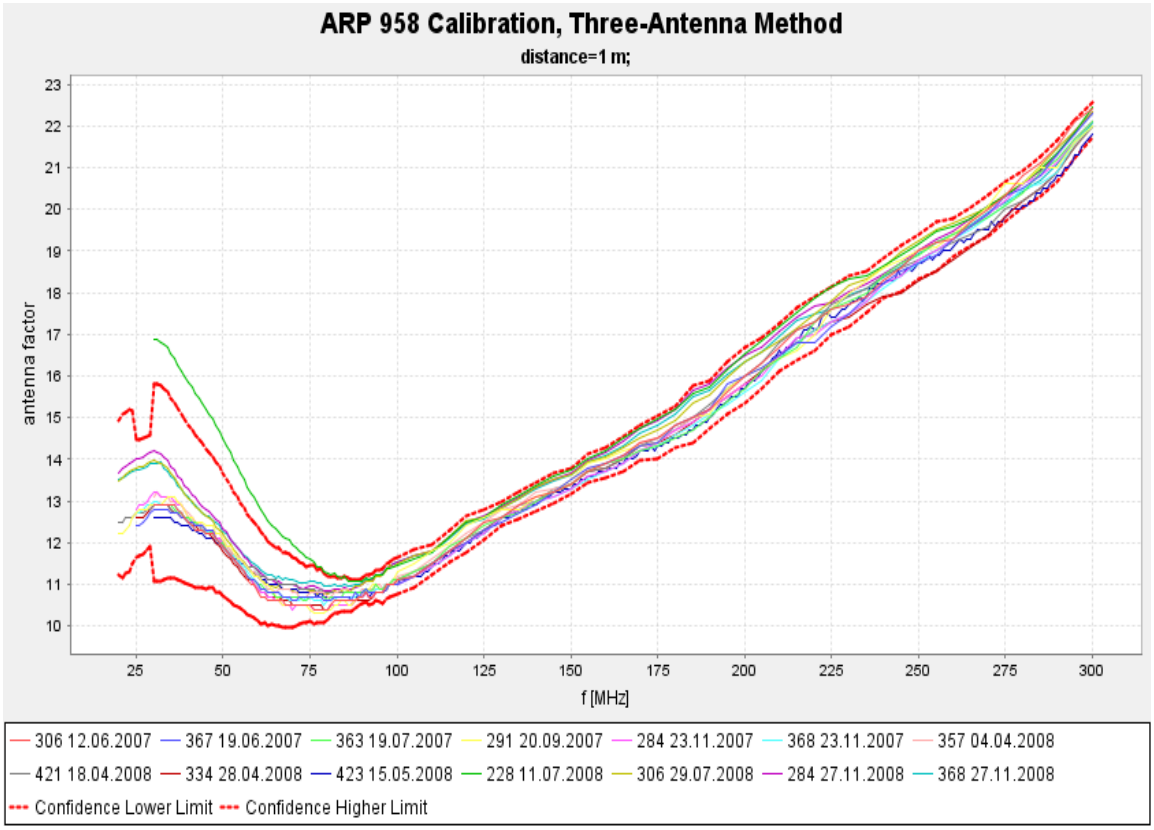
2. TEÓRIA A POUŽITÉ METÓDY

Pri posudzovaní správnosti výsledkov nameraných kalibračných dát môžeme vychádzať z predpokladu normálneho rozdelenia chyby merania. Na základe toho môžeme použiť nasledovnú štatistiku. Vzatím do úvahy počet vzoriek získaných pre danú frekvenciu, ak je ich počet menší ako 30 použijeme Studentone *t*-rozdelenie na odhad konfidenčných intervalov nameraných hodnôt pre danú frekvenciu. Toto rozdelenie sa používa pre odhad strednej hodnoty dát s normálnym rozdelením v situáciách, kde je počet dostupných vzoriek malý, čo je práve náš prípad keď v konkrétnej frekvencii je nameraný len obmedzený počet vzoriek. S narastajúcim počtov vzoriek sa Studentone *t*-rozdelenie blíži k normálnemu rozdeleniu.

Teoreticky, ak máme k dispozícii aspoň dve merania anténneho faktoru pre daný model antény, vieme týmto spôsobom vypočítať rozmedzie v ktorom sa s určitou pravdepodobnosťou budú nachádzať nasledujúce merania pre tento model antény. Použitím opísanej štatistiky dostávame mechanizmus na overovanie správnosti (validáciu) výsledkov kalibrácií antén.

Nanešťastie, ako je uvedené v obrázku Obr. 1 nie všetky kalibračné dáta pre špecifický model antény majú homogénny charakter. V takomto prípade nemôže byť výpočet konfidenčných intervalov aplikovaný priamo, pretože vzhľadom na ich šírku by to mohlo zamedziť detekcii neplatných kalibrácií. Preto, pred samotnou validáciou musia byť dáta najprv rozdelené do konzistentných podskupín, čo sa dá docieliť použitím klastrovej

analýzy [1]. Pre tento účel budeme naše krivky anténnych faktorov považovať za *n*-dimenzionálne vektory, kde *n* je počet nameraných bodov v danom frekvenčnom rozsahu. Mnohé klastrovacie algoritmy vyžadujú pred ich spustením ako vstupný parameter počet klastrov do ktorých majú byť dáta rozdelené. Pre určenie správnej hodnoty počtu klastrov je k dispozícii množstvo techník ako napríklad, metóda hľadania extrému funkcie podobnosti klastrov [2], delenie dendrogramu založené na entropii [3], prístup informačnej teórie [4], Silhouette [5], v-fold cross-validácia [6] a iné.



Obr. 1: Anténne faktory namerané pre model antény ARP 958. Demonštrácia potreby klasterizácie dát pre ich správne spracovanie.

Pri úlohách analýzy nameraných dát je dôležité byť schopný rozlíšiť vzorky, ktoré nezapadajú do celkového modelu, tzv. outliery. Hoci sú outliery často považované za chyby alebo šum, môžu byť nositeľom dôležitej informácie. Vo väčšine prípadov sú však kandidátmi na odchýlené dáta, ktoré by mohli viesť k nesprávnym výsledkom. Preto je dôležité ich odhaliť ešte pred tým ako dôjde k samotnej analýze dát [8], [9]. Metódy detekcie outlierov boli navrhnuté pre širokú škálu aplikácií, ako napríklad detekcia zneužitia kreditných kariet, klinické testy, čistenie dát, detekciu narušenia počítačových sietí, predpovedanie počasia, geografické informačné systémy, analýzu výkonnosti atlétov a iné úlohy z oblasti data-miningu ([11], [12], [13], [14], [15], [16]). V našom prípade sú metódy detekcie outlierov využité pri určovaní klastrov v analyzovaných dátach.

3. CIELE DIZERTAČNEJ PRÁCE

Východiskom práce je potreba automatickej validácie kalibračných meraní antén na pracovisku autora. Základná idea dizertačnej práce vychádza z unikátnej dostupnosti veľkého množstva kalibračných údajov antén, ktoré pochádzajú z mnohoročnej kalibračnej praxe. V rámci riešenia je potrebné zozbierať a systematicky usporiadať kalibračné dáta antén a pripraviť ich pre ďalšie spracovanie.

Naším cieľom je stanoviť neurčitosti anténnych faktorov a navrhnúť metódy validácie kalibračných meraní pre rôzne modely antén. Predbežná analýza kalibračných dát ukázala, že anténne faktory sa vplyvom rôznych faktorov zhlukujú do skupín. Preto je potrebné na ich rozlíšenie využiť klastrovú analýzu. Za týmto účelom preskúmame rôzne klasterizačné metódy a ich použiteľnosť na náš súbor dát. S rozdeľovaním kalibračných

kriviek do klastrov je spojený dodatočný problém, a to určenie konkrétneho počtu klastrov v dátach. V tejto oblasti máme v úmysle navrhnúť inovatívny prístup založený na metódach detekcie outlierov.

Na základe uvedeného môžeme definovať ciel dizertačnej práce nasledovne:

- implementácia algoritmov na získanie kalibračných dát a vytvorenie databázy kde tieto dáta budú usporiadané na základe rôznych kritérií ako model a sériové číslo antény, kalibračná metóda, jej parametre a podmienky prostredia, čím sa vytvorí základ pre nasledujúce úlohy,
- vypracovať vhodný spôsob aplikácie klastrovej analýzy na kalibračné dáta antén. Návrh a otestovanie klasterizačných metód s automatickou detekciou počtu klastrov založenou na princípoch detekcie outlierov,
- po splnení predchádzajúcich dvoch cieľov, navrhneme postup pre výpočet konfidenčných intervalov a tým mechanizmus na overovanie správnosti výsledkov meraní kalibrácií antén. Nakoniec stanovíme počet klastrov v dátach pre jednotlivé typy antén a následne vypočítame varianciu hodnôt ich anténnych faktorov.

4. POPIS NAVRHNUTÝCH METÓD

4.1. Vytvorenie databázy kalibračných dát a nástroja na ich analýzu

Ako bolo spomenuté v úvode, kalibračné dáta antén sú súčasťou kalibračných certifikátov uložených na disku počítača v súborovom formáte programu Microsoft Word. Z tohto dôvodu, pred tým ako ich bude možné analyzovať, musíme vytvoriť softvérový nástroj na ich nájdenie v súborovom systéme a následné vyextrahovanie číselných údajov frekvenčných závislostí anténnych faktorov. Až potom bude možné vytvorenie databázy umožňujúcej ďalší výskum.

Východiskovým bodom riešenia tejto úlohy je tabuľka vo formáte Microsoft Excel s firemnými záznamami týkajúcich sa kalibrácií antén. Každý záznam tvorený riadkom tabuľky, obsahuje informácie o identifikátore objednávky a prislúchajúceho certifikátu, meno zákazníka, typ antény, jej sériové číslo, kalibračnú metódu a dátum kalibrácie. Tieto údaje tvoria kalibračné meta-dáta. Kvôli vnútornej organizácii firemných dát, identifikátor objednávky odkazuje na určitý adresár súborového systému kde je uložený prislúchajúci kalibračný certifikát, ktorý sa dá nájsť na základe jeho identifikátora. Certifikát samotný už obsahuje namerané hodnoty frekvenčných závislostí anténneho faktora, tvz. kalibračnú krivku.

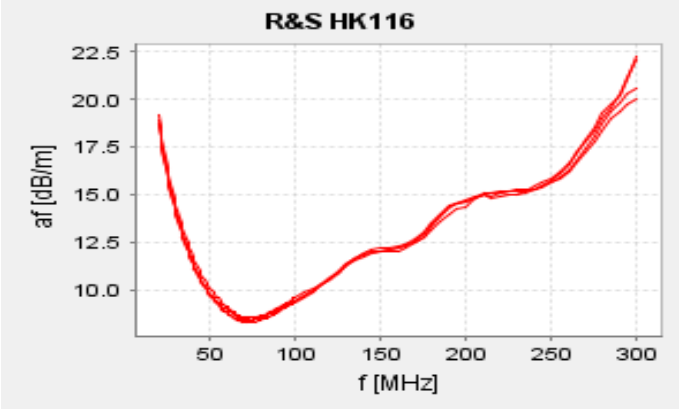
Vzhľadom na množstvo dát (približne 6000 kalibračných certifikátov), zber surových dát a prislúchajúcich meta-dát nemôže byť uskutočnený manuálne. Je potrebná istá forma automatizácie. Z tohto dôvodu sme vytvorili softvér, ktorý uskutoční všetky manuálne náročné úlohy.

Softvér implementuje nasledujúcu funkcionalitu:

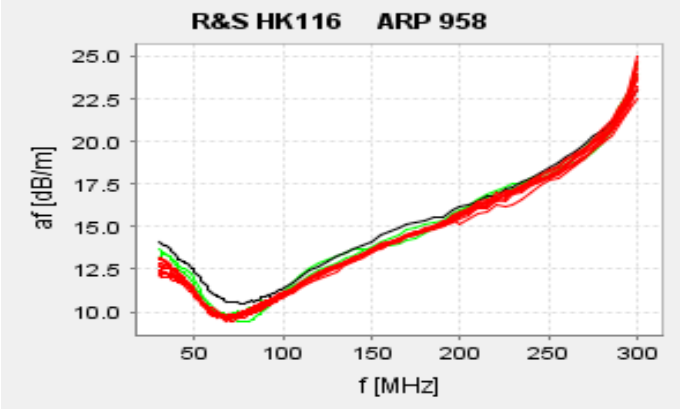
1. prechádzanie tabuľky Microsoft Excel, kvôli získaniu kalibračných meta-dát,
2. určenie počtu kalibrácií pre špecifický model antény,
3. nájdenie súboru obsahujúceho kalibračný certifikát v súborovom systéme disku počítača,
4. extrakcia detailov kalibrácie vrátane frekvenčných závislostí anténnych faktorov zo súboru vo formáte Microsoft Word,
5. separácia dát na základe kalibračnej metódy a typu výsledných dát merania a ich uloženie v databáze,
6. zobrazenie klasifikovaných dát v grafe,
7. implementácia analytických modulov (klastering, detekcia outlierov, výpočet konfidenčných intervalov)

4.2. Aplikácia klastrovej analýzy na kalibračné krivky

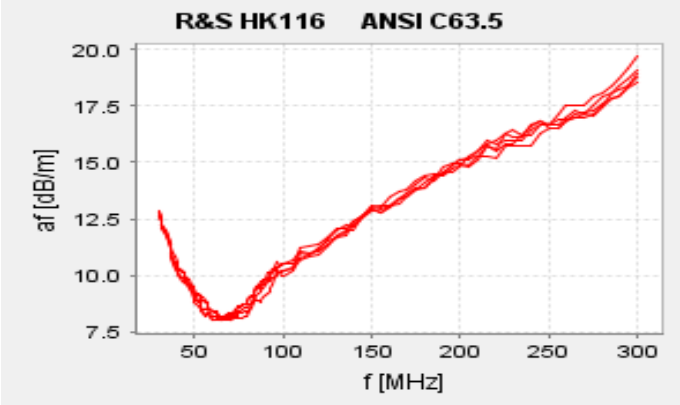
Predtým ako môžu byť použité klasterizačné metódy, kalibračné dáta musia byť najskôr predspracované. Ako prvé musíme kalibračné krivky orezať na jednotný frekvenčný rozsah a rovnaký počet nameraných bodov. Toto dosiahneme implementáciou algoritmu, ktorého vstupným parameterom bude kolekcia kalibračných kriviek pre daný model antény. Každá krivka je reprezentovaná ako séria hodnôt X,Y kde X je frekvencia a Y je amplitúda. Pre každú frekvenciu potom zrátame počet dostupných hodnôt Y. Orezaná séria bude potom obsahovať iba frekvencie, pre ktoré je počet amplitúd rovný počtu kriviek v kolekcii. Týmto spôsobom získame, všetky hodnoty pre frekvencie kde sa krivky prekrývajú.



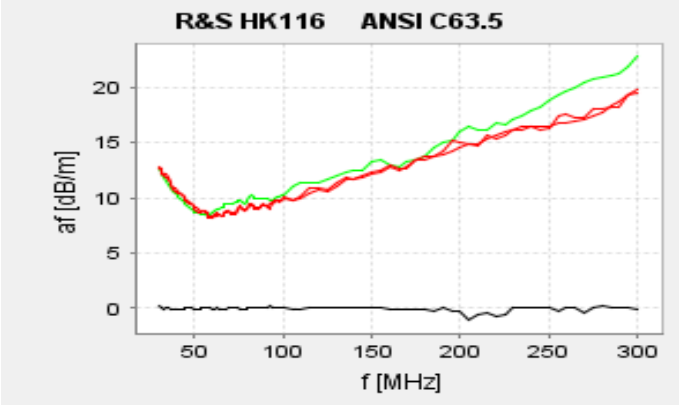
1 klaster



3 klaster



1 klaster



3 klaster

Obr. 2: Výsledky aplikácie klastrovej analýzy na namerané anténne faktory. Požadovaný počet klastrov bol určený manuálne. Jednotlivé klastre sú zobrazené rôznymi farbami. Dolný pravý obrázok obsahuje chybné dáta (čiernou) spôsobené nesprávnym spracovaním súboru obsahujúceho kalibračný certifikát.

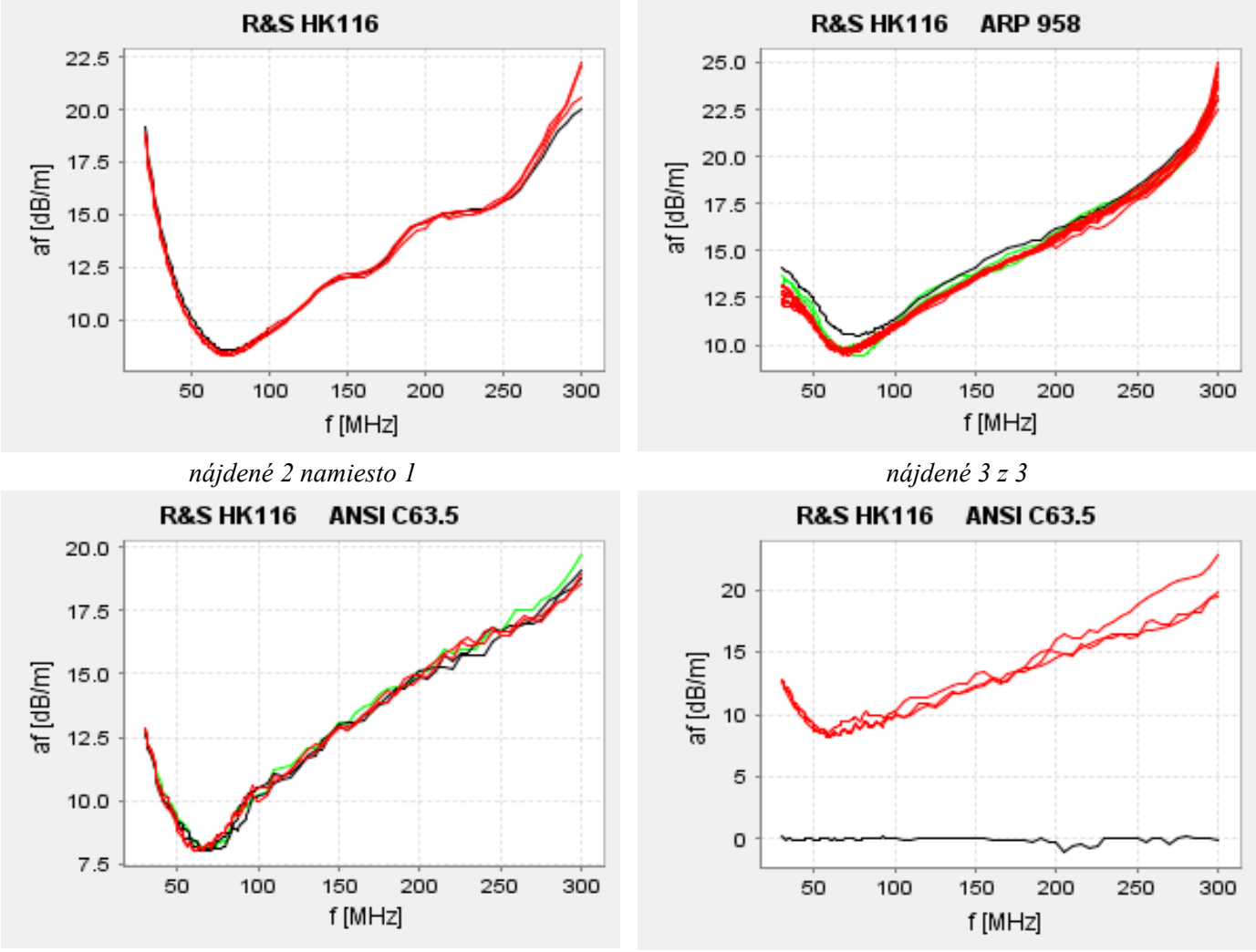
V našich experimentoch uvažujeme o krivkách ako o vektoroch. Ako miery vzdialenosti sme testovali Euklidovskú vzdialenosť a vzdialenosť city-block. Euklidovská vzdialenosť progresívne kladie väčší dôraz na body, ktoré sú ďalej od seba. Na druhej strane vzdialenosť city-block je viac tolerantná k vzdialenejším bodom. Testy uskutočnené na väčšej vzorke našich dát nakoniec ukázali, že v prípade Euklidovskej vzdialenosti klasterizačné algoritmy pracú lepšie.

Ďalej sme na kalibračné krivky aplikovali dve klasterizačné metódy, aglomeratívne hierarchické klastrovanie a k-means++ [10] klastrovanie. Ukázalo sa, že každá z metód je dobre použiteľná pre naše potreby.

Klastrované dáta boli vybrané zo súboru kalibrácií uskutočnených na základe noriem ANSI C63.5 a ARP 958 z rôznymi parametrami merania. Na obrázku Obr. 2 sú zobrazené výsledky klastrovania. Uvedené dáta sú modelovými prípadmi, ktoré možno pozorovať v úplnom súbore. Použitím Euklidovskej vzdialenosti nebol pozorovaný rozdiel medzi výsledkami dosiahnutými oboma klasterizačnými metódami.

Kritickou vlastnosťou oboch klastrovacích metód je nutnosť zadať počet klastrov na vstupe algoritmu. Ak explicitne nie je známy počet klastrov, je potrebné ho empiricky odhadnúť z dát. Toto je samostatný problém,

na ktorý sa dá aplikovať množstvo techník riešenia. Niektoré z nich sú: metóda hľadania extrému funkcie podobnosti klastrov [2], delenie dendrogramu založené na entropii [3], prístup informačnej teórie [4], Silhouette [5], v-fold cross-validácia [6]. Okrem prvej spomenutej metódy, všetky ostatné majú jeden spoločný problém, a to, že sú určené pre súbory dát obsahujúce veľké množstvo prvkov. Kdežto nami analyzované dáta obsahujú v priemere 3 – 10 kalibračných kriviek pre špecifický model antény. Navyše niektoré zo spomínaných metód zahŕňajú algebraické maticové operácie, ktoré zlyhávajú kvôli singularitám a nesplneným predpokladom v prípadoch keď sú aplikované na iné dáta pre aké boli navrhnuté.



Obr. 3: Výsledok klastrovacieho algoritmu využívajúceho kritérium celkovej variancie klastrov pre odhad počtu klastrov.

Jedným z postupov určovania počtu klastrov v ľubovoľnom súbore dát je definovať kriteriálnu funkciu, ktorá meră kvalitu klastrovania akéhokoľvek rozkladu dát. Problémom je nájsť také rozdelenie dát, ktoré minimalizuje túto kriteriálnu funkciu [2]. Ak súbor dát obsahuje rozumný počet prvkov, je možné spúšťať klastrovací algoritmus iteratívne pre rôzne počty klastrov, postupujúc od jedna až do počtu prvkov v súbore, pričom vyberieme práve taký počet klastrov, ktorý minimalizuje kriteriálnu funkciu.

Pri skúmaní rôznych kriteriálnych funkcií, kritérium celkovej variancie klastrov (total cluster variance) S_T prinieslo najlepšie výsledky [7]. V našom prípade je celková variancia definovaná ako suma medzi klastrovej variancie S_B a vnútro klastrovej variancie S_W.

$$S_T = S_B + S_W$$

Pri určovaní počtu klastrov v dátach sme použili kriteriálnu funkciu celkovej variancie v kombinácii s aglomeratívnym hierarchickým klastrovaním. Ako vidieť z obrázku Obr. 3, výsledky nie sú optimálne vzhľadom na manuálne určenie klastrov z obrázka Obr. 2.

Hoci v publikácii [7] sme uviedli, že kritérium celkovej variancie klastrov, prináša akceptovateľné výsledky, ďalší výskum na väčšom súbore dát ukázal obmedzenia tohto prístupu. Pravdepodobne najväčšie z nich je neschopnosť detekcie prípadu keď je v dátach iba jeden klaster. Z vykonaných experimentov vyplýva, že dva klastre budú vždy mať celkovú varianciu menšiu ako v prípade jedného klastra. Pre presnejší odhad počtu klastrov v našich dátach je preto nutné vyvinúť iné kritérium, poprípade celkom nový prístup.

4.3. Outlier detection in measurement data

V našej situácii súbor dát pozostáva z prvkov (kalibračných kriviek), ktoré sa formujú do klastrov, a preto môžeme na prvky jedného klastra nazerať ako na outliery voči prvkom iného klastra. Pri aplikácii existujúcich metód na detekciu outlierov sme opäť narazili na problém, že nami klastrované dáta obsahujú len malý počet prvkov. Táto skutočnosť má za následok zlyhávanie viacrozmerných metód na detekciu outlierov. Z tohto dôvodu sme sa v ďalšom výskume sústredili na jednorozmerné metódy detekcie.

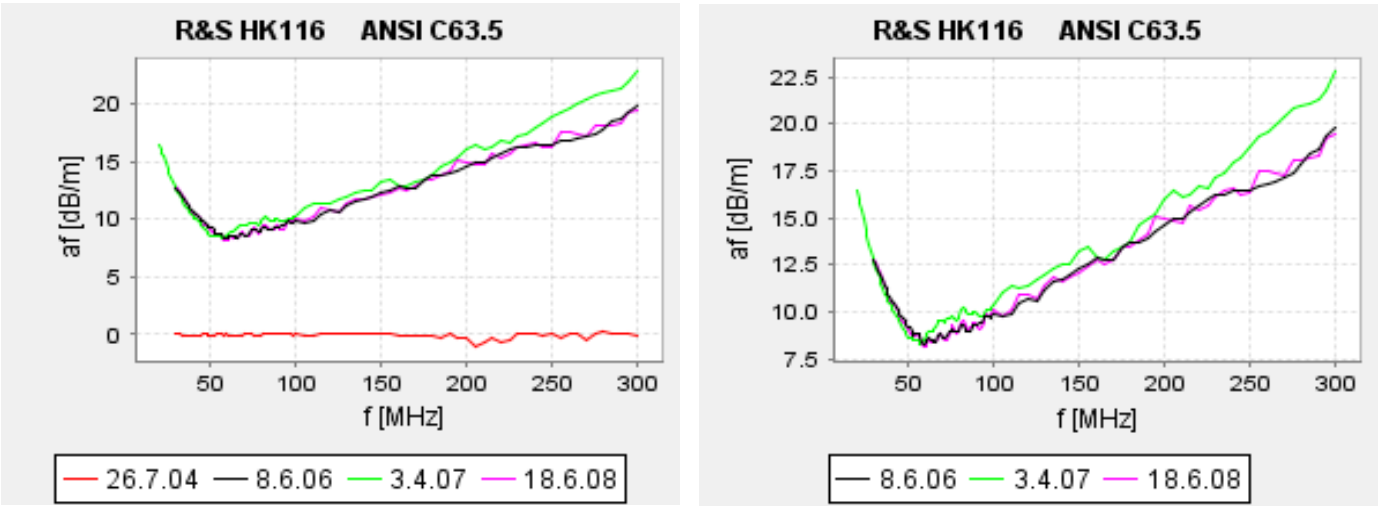
4.3.1. Klasický jednorozmerný prístup k detekcii outlierov

V tejto kapitole sa zaoberáme preskúmaním aplikovateľnosti jednorozmerného prístupu k detekcii outlierov založenom na štatistike χ^2 (*Chi-kvadrát)*.

Z pohľadu štatistickej teórie, ak X_i je k nezávislých, normálne rozdelených náhodných premenných s priemermi μ_i a štandardnými odchýlkami σ_i, potom štatistika

$$Y = \sum_{i=1}^{k} \left(\frac{(X_i - \mu_i)}{\sigma_i} \right)^2$$

má rozdelenie Chi-kvadrát, kde parameter k zodpovedá počtu stupňov voľnosti, čo zodpovedá počtu náhodných premenných X_i a teda rozmeru našich vektorov. Hodnoty vypočítané na základe tejto štatistiky, potom môžeme testovať oproti χ^2 konfidenčným intervalom pre daný stupeň voľnosti, ktorý je rovný počtu bodov v kalibračných krivkách. Týmto spôsobom vieme vyradiť krivky, ktorých vzdialenosť od priemeru kriviek je za akceptovateľným limitom.



pred detekciou *nájdený 1 outlier*

Obr. 4: Výsledok chi-kvadrát outlier detektora.

Na Obr. 4 môžeme vidieť výsledky detekcie outlierov za použitia tohto prístupu. Krivka 26.7.04 nezapadá do 95% konfidenčného intervalu vypočítaného na základe χ^2 štatistiky a preto bola algoritmom vyradená. Avšak na druhej strane, hoci krivka 3.4.07 je z nášho pohľadu outlierom, táto nebola vyradená. Tento jav bol spôsobený vplyvom krivky 26.7.04 na výpočet priemeru kriviek, tzv. maskovací efekt.

V snahe zvýšiť robustnosť uvedeného algoritmu, sme použili tzv. robustnú štatistiku mediánu a mediánovej absolútnej odchýlky, navrhnutú v Hampel [17], [18]. Táto robustná štatistika poskytuje alternatívu ku klasickým štatistickým

metódam priemeru a smerodajnej odchýlky, kde jej motiváciou bolo vytvorenie odhadov neovplyvnených malými odchýlkami od predpokladaného modelu.

Medián je robustnou mierou centrálnej tendencie, na rozdiel od priemeru. Pre príklad medián má bod zlomu 50%, pričom priemer má bod zlomu 0% (aj jedna veľká hodnota ho môže vychýliť).

Mediánová absolútna odchýlka (MAD) je robustnou mierou štatistického rozptylu, a je používaná ako odhad smerodajnej odchýlky. Pre jednorozmerné dáta X_1, X_2, ..., X_n, je MAD definovaný ako medián absolútnych odchýliek od dátového mediánu:

$$MAD = median_i\left(\left|X_i - median_j\left(X_j\right)\right|\right) \quad ,$$

uvažujúc odchýlky od dátového mediánu, MAD je mediánom ich absolútnych hodnôt. Aby sme MAD mohli použiť ako konzistentný odhad smerodajnej odchýlky σ, musíme uvažovať

$$\hat{\sigma} = k.MAD$$

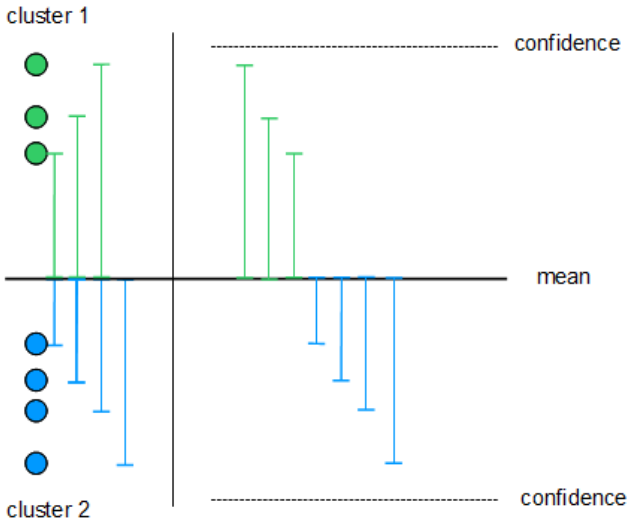
kde k je škálovacia konštanta, ktorá závisí od rozdelenia. Pre normálne rozdelené dáta k = 1.4826. Inými slovami, očakávame že smerodajná odchýlka bude 1.4826 násobkom MAD pre veľké súbory normálne rozdelených vzoriek X_i. Na druhej strane, pre súbor vzoriek z rovnomerného rozdelenia je táto konštanta rovná 1.1547.

V ďalších experimentoch sme aplikáciou χ^2 štatistiky za použitia klasickej i robustnej verzie dokázali, že tento prístup k detekcii outlierov nebude vhodný pre náš typ dát. Hlavnou príčinou zlyhania tohto prístupu bude malé množstvo analyzovaných dát, čím sa odhad priemeru stáva nestabilný vo vzťahu k outlierom.

Hoci aplikácia χ^2 štatistiky nebola úspešná pri riešení problému detekcie outlierov, stále ju môžeme využiť pri validácii kalibrácií antén. Naším pôvodným zámerom bolo počítať konfidenčné intervaly na úrovni jednotlivých frekvencií, pre zistenie správnosti kalibračnej krivky. Tento prístup by vyžadoval, dodatočný mechanizmus na rozlíšenie situácií, kedy iba časť posudzovanej krivky je mimo povolený rozsah. Použitím χ^2 štatistiky posudzovaná krivka bude reprezentovaná jednou hodnotou, testovanou vzhľadom na konfidenčné intervaly, čo robí validáciu antén jednoznačnou. Vypočítaním χ^2 konfidenčných intervalov zo skupiny kalibračných kriviek považovaných za správne, môže byť validovaná krivka testovaná na základe rovnakej štatistiky.

4.3.2. Inovatívny prístup k detekcii outlierov

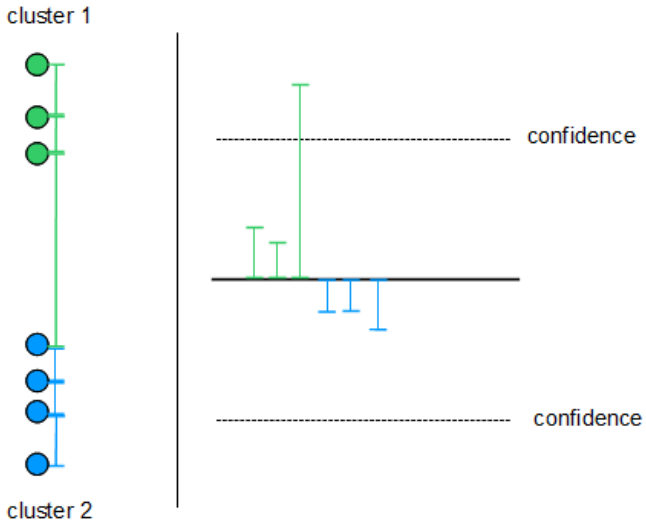
Analýza problematiky detekcie outlierov, ukázala, že odhad priemeru analyzovaných dát nie je vhodný pre ich detekciu, pretože je silne ovplyvniteľný prítomnosťou outlierov samotných, ktoré môžu v našom prípade byť prvkami iného klastra (viď Obr. 5).



Obr. 5: Výpočet konfidenčných intervalov z pozícií prvkov môže viesť k zlyhaniu detekcie outlierov.

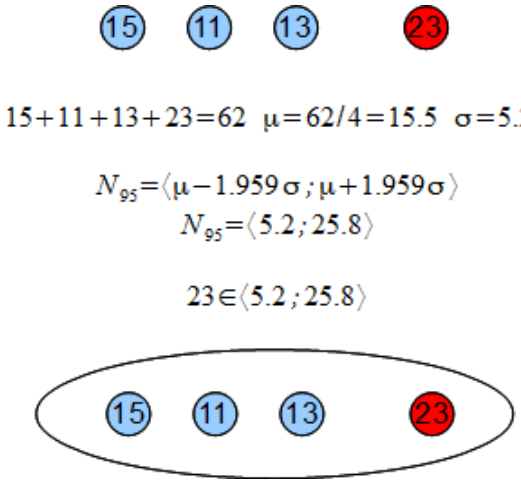
V tejto kapitole preskúmame myšlienku výpočtu štatistiky použitím vzájomných vzdialeností medzi krivkami namiesto použitia ich vzdialenosti od ich priemeru.

Ľubovoľná dvojrozmerná krivka môže byť prezentovaná ako vektor, ktorý je zároveň bodom v multi-dimenzionálnom priestore. V snahe vyriešiť horeuvedený problém týkajúci sa priemeru kriviek, sme navrhli použiť vzájomné vzdialenosti kriviek namiesto ich pozície, pre výpočet priemeru a smerodajnej odchýlky. Myšlienka je graficky znázornená na Obr. 6.



Obr. 6: Výpočet konfidenčných intervalov zo vzájomných vzdialeností prvkov.

Pri tomto prístupe sú prvky reprezentované vzdialenosťami medzi nimi, čím je možné získať informáciu o ich vzájomných pozíciách vrámci/mimo klastra. Vzdialenosti medzi bodmi vrámci klastra sú výrazne menšie ako vzdialenosti medzi klastrami. V skutočnosti je ľudské vnímanie veľmi blízke tomuto prístupu, pretože mi sa pri posudzovaní klastrov pozeráme na vzdialenosti medzi prvkami a nie na ich vzdialenosti od nejakej vypočítanej priemernej hodnoty.



Obr. 7: Priemerná hodnota ovplyvnená outlierom ústiaca do zlyhania jeho detekcie.

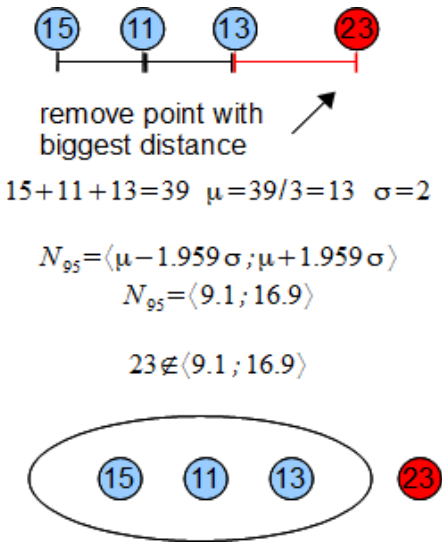
Použitie medzi-prvkových vzdialeností pre výpočet štatistiky, má za následok, že smerodajná odchýlka je relatívne menšia a tým je konfidenčný interval užší, čo umožňuje lepšiu detekciu outlierov. Potom môžeme definovať outlier ako prvok, ktorého vzdialenosť k ostatným je nezvyčajne veľká.

Ako už bolo spomenuté, je treba mať na zreteli problém robustnosti detekcie outlierov. Stačí zopár prvkov, ktoré ovplyvnia konfidenčný interval do tej mieri, že detekcia zlyhá (viď Obr. 7).

V snahe riešiť tento problém sme navrhli vylúčiť prvok s najväčšou vzdialenosťou z výpočtu konfidenčných intervalov. Princíp je načrtnutý na Obr. 8. Následkom tejto operácie konfidenčný interval pokrýva iba platné prvky súboru, čím sme schopný odhaliť outlier, ktorého vzdialenosť do neho nespadá.

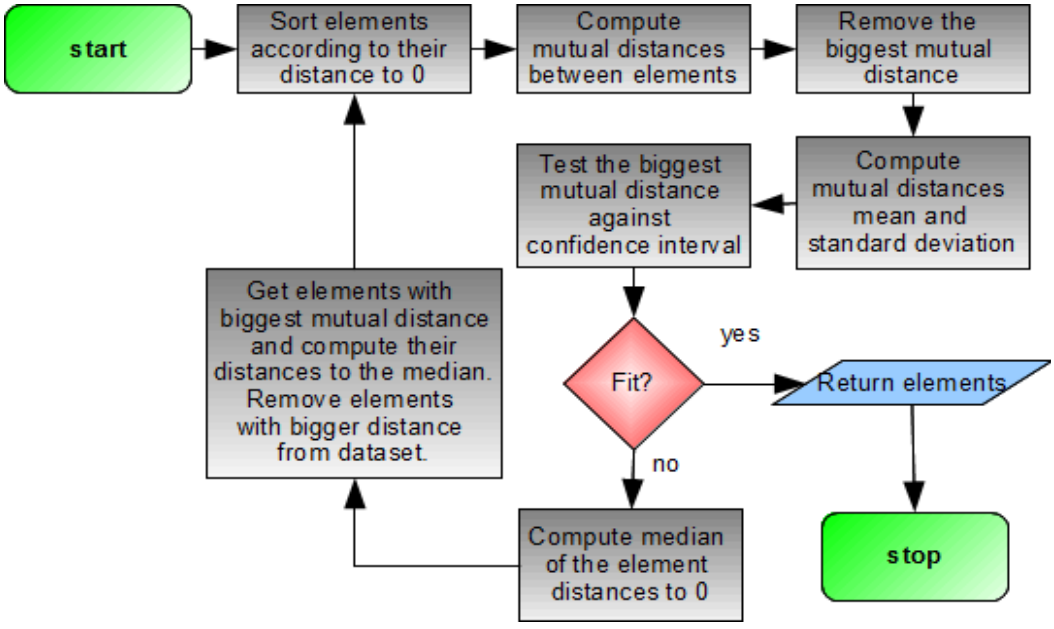
Ak by sme pri výpočte použili vzdialenosti prvkov od ich priemeru namiesto ich vzájomných vzdialeností, navrhnutú prístup by nefungoval, pretože daný priemer by už bol ovplyvnený outlierom. V situácii keď je prítomných viacero outlierov, ich náš algoritmus postupne odstráni v po sebe nasledujúcich iteráciách. Hodnoty

priemeru a smerodajnej odchýlky sú po odstránení outliera prepočítané a test najväčšej vzdialenosti sa uskutočňuje nad novým konfidenčným intervalom.



Obr. 8: Výpočet konfidenčného intervalu po vylúčení prvku s najväčšou vzdialenosťou.

Nami navrhnutý algoritmus pre detekciu outlierov (viď Obr. 9) vo všeobecnosti začína usporiadaním prvkov v zostupnom poradí. Potom vypočítame vzájomné vzdialenosti každého prvku od prvku pod ním. Ak máme *n* prvkov, výsledkom bude *n-1* hodnôt vzdialeností. V ďalšom kroku vylúčime najväčšiu vzájomnú vzdialenosť *d* a vypočítame priemer μ a smerodajnú odchýlku σ pre zostávajúce vzdialenosti. Na základe toho určíme konfidenčné intervaly $<\mu - z\sigma; \mu + z\sigma>$ kde *z* je napríklad 1.959 pre 95% interval. Ak najväčšia vzdialenosť spadá do intervalu ($d \leq \mu + z\sigma$) potom neprišlo k detekcii outliera a algoritmus skončí. V opačnom prípade vylúčime príslušný element ako outlier a pokračujeme v ďalšej iterácii. Opísaný postup je možné použiť ako alternatívu k robustnému estimátoru navrhnutému v Tukey [19], založenom na tzv. Box plote.



Obr. 9: Robustný algoritmus pre detekciu outlierov.

V prípade klastrov sa procedúra počítania vzájomných vzdialeností prvkov na základe poradia nahrádza počítaním najmenších vzdialeností medzi prvkami v rámci klastra, pretože usporiadanie prvkov podľa poradia v mnohorozmernom priestore nie je uskutočniteľné. Potom zistíme najkratšie vzdialenosti medzi klastrami a najväčšiu z nich testujeme, ako potenciálny outlier, pretože predpokladáme, že vzdialenosti medzi klastrami nie sú z rovnakého rozdelenia ako vzdialenosti prvkov v rámci klastrov.

Stále je tu však situácia, ktorú sme zatiaľ neanalyzovali. A to prípad kedy je k dispozícii súbor obsahujúci iba tri prvky. Na základe opísaného algoritmu, počítaním vzdialeností medzi prvkami by sme dostali dve hodnoty, s ktorých väčšiu by sme vylúčili a výpočet priemeru a smerodajnej odchýlky z jednej hodnoty nie je možné rovnako ako výpočet konfidenčných intervalov. Pre tento prípad sme preto navrhli rozdielnu metódu založenú na porovnávaní priemerov dvoch populácií.

T-test dvoch priemerov je metóda založená na testovaní štatistických hypotéz, kde je nulová hypotéza formulovaná nasledovne:

$$H_0: \mu_1 = \mu_2 \quad \text{proti} \quad H_a: \mu_1 \neq \mu_2 \quad \text{t.j.}$$

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{proti} \quad H_a: \mu_1 - \mu_2 \neq 0.$$

Na základe toho, sa počíta konfidenčný interval pre rozdiel priemerov dvoch populácií (μ_1 - μ_2), ktorý označuje rozsah hodnôt, v ktorom sa tento rozdiel môže pohybovať pre nezamietnutie nulovej hypotézy.

Vo všeobecnosti nepoznáme smerodajné odchýlky populácií, a preto počítame ich odhady s_1 and s_2 . V tomto prípade je dvoj vzorková *t* štatistika definovaná nasledovne:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} - \frac{s_2^2}{n_2}}} \quad .$$

Hoci táto štatistika nemá striktne *t* rozdelenie (čo je spôsobené použití odhadov smerodajných odchýlok), *p-hodnoty* môžu byť získané pre použitím *t(k)* distribúcie kde *k* reprezentuje menŠiu z dvoch hodnôt n_1 - *1* a n_2 - *1*. Inou možnosť by bolo odhadnúť počet stupňov voľnosti z dát.

Pre prípad 3 kriviek, prvá populácia je reprezentovaná rozdielmi súradníc (bodov) dvoch najbližších kriviek, pričom druhá populácia obsahuje rozdiely súradníc dvoch vzdialenejších kriviek. Potom na základe opísanej štatistiky testujeme priemery daných populácií. Tu predpokladáme, že v prípade krivky-outliera sú vzdialenosti medzi jej bodmi a bodmi k nej bližšej krivky z dvoch zostávajúcich kriviek distribuované rozdielne. Tiež predpokladáme normálne rozdelenie vzdialeností medzi bodmi kriviek.

Teraz môžeme formulovať predpoklady pre detekciu outlierov navrhnutým algoritmom.

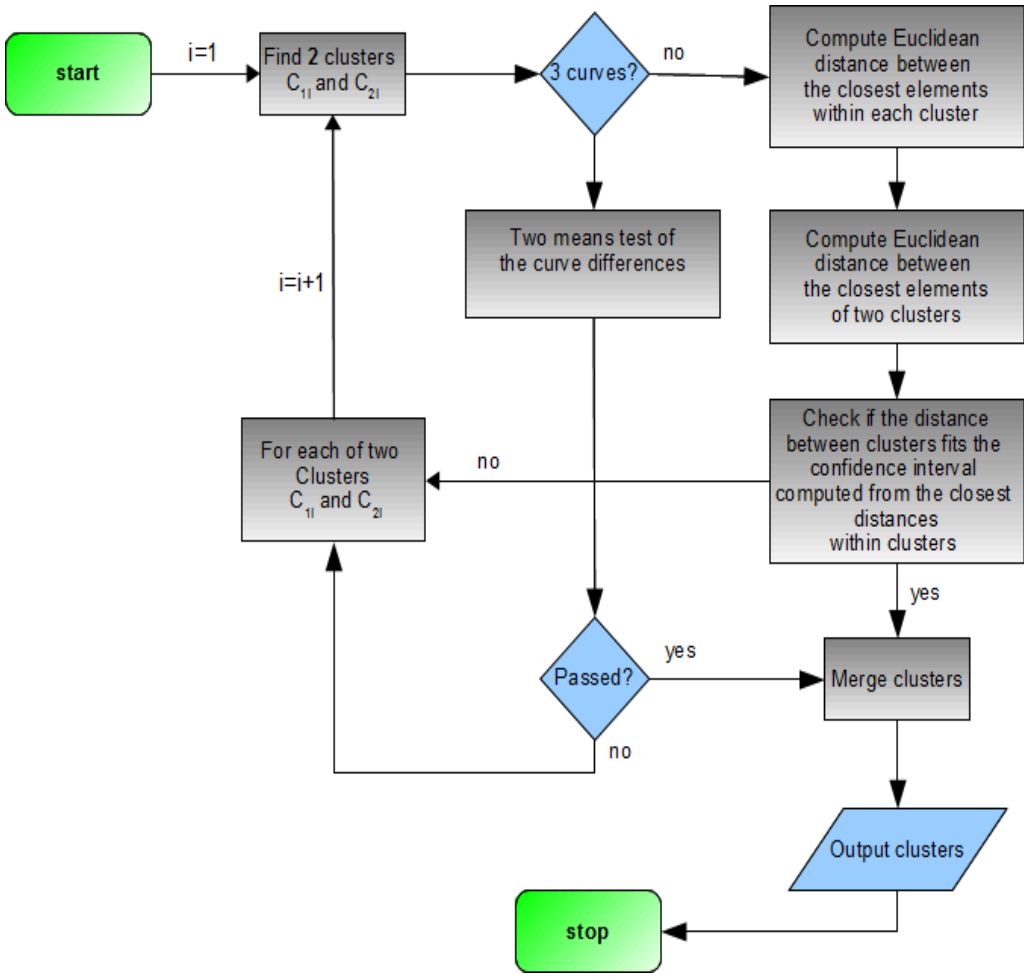
1. Predpokladáme, že vstupné dáta obsahujú klaster správnych dát a tiež môžu obsahovať jeden alebo viac outlierov alebo klaster/e outlierov.
2. Pravdepodobnostné rozdelenie vzdialeností medzi prvkami v rámci klastrov je rozdielne ako rozdelenie medzi klastrových vzdialeností.

Navrhnutý algoritmus je svojou podstatou efektívny pri prekonávaní tzv. *swamping* a *masking* efektov opísaných v [11]. *Masking* efekt nastáva keď outlier zdeformuje odhad priemeru a kovariancie smerom k sebe, čo má za následok malú vzdialenosť outliera od priemeru a tým jeho neodhalenie. V našom prístupe sa priemer počíta zo vzdialeností medzi prvkami a tým sa minimalizuje vplyv potenciálneho outliera. *Swamping* efekt, kedy prítomnosť outliera spôsobí, že sa platný prvok začne javiť ako outlier, je eliminovaných na rovnakom princípe.

Ak je v analyzovanom súbore dát viacej správnych prvkov ako outlierov, algoritmus eliminuje outliery na základe ich vzdialeností k správnym prvkom. Z tejto úvahy vieme určiť Hampelov bod zlomu pre náš algoritmus. Ten sa stanovuje ako najmenšie percento ľubovoľne veľkých outlierov v dátach, ktoré povedie k zlyhaniu algoritmu. V našom prípade je to 2/n, čo je limit kedy nastane *masking* alebo *swamping* efekt.

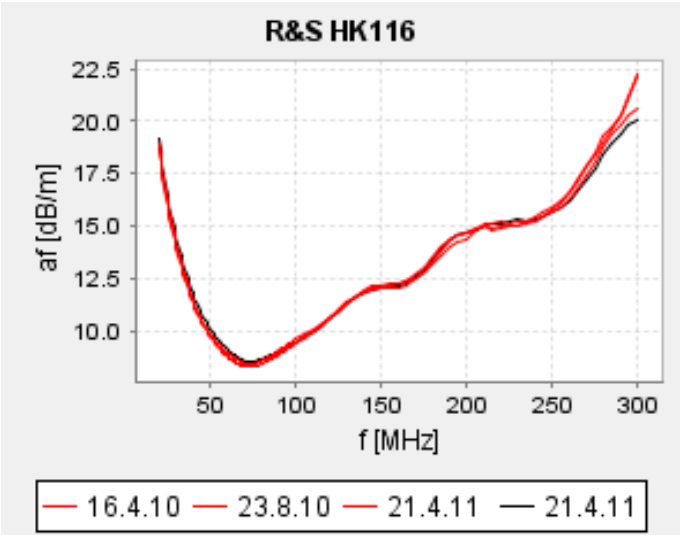
4.4. Detekcia outlierov a stanovenie počtu klastrov

Ako bolo už skôr spomenuté, pri klastrovacích algoritmoch je potrebné dopredu vedieť počet klastrov v spracovávaných dátach. V rámci nášho výskumu sme skúmali možnosti určenia tohto parametru algoritmickou cestou. Keďže existujúce metódy sa ukázali ako nevhodné pre našu aplikáciu, rozhodli sme sa venovať vývoju celkom novej metódy na automatické určovanie počtu klastrov v dátach, založenej na princípoch detekcie outlierov. Po mnohých experimentoch sme dospeli k vhodnej kombinácii klasterizačnej metódy a nami vyvinutej metódy na detekciu outlierov. Myšlienka spočíva v skúmaní klastrov ako potenciálnych outlierov navzájom.

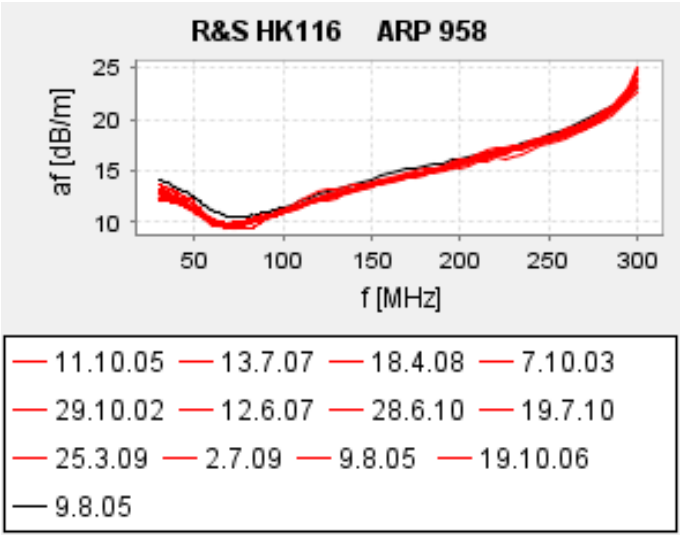


Obr. 10: Automatická klastrovacia metóda kombinovaná s detekciou outlierov.

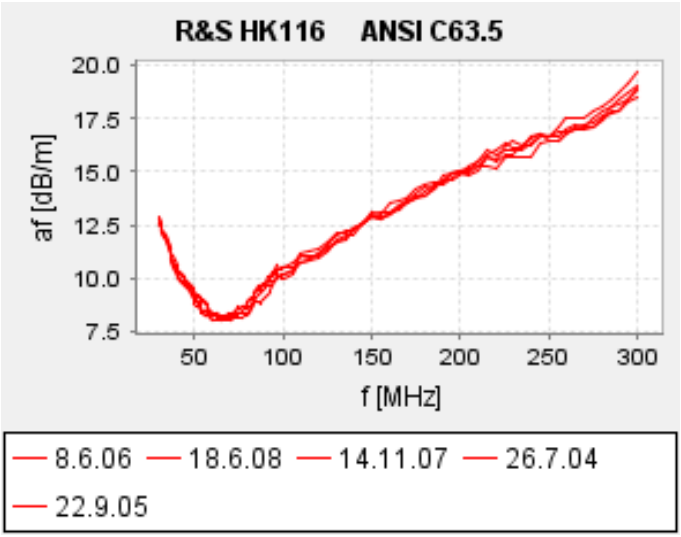
Bloková schéma výsledného algoritmu je vyobrazená na Obr. 10. Proces detekcie môže byť vysvetlený nasledovne. V prvej itérácii začíname určením dvoch klastrov vo vstupných dátach použitím metódy aglomeratívneho hierarchického klastrovania. Následne vypočítame Euklidovské vzdialenosti medzi najbližšími krivkami v rámci klastrov a tiež ich priemer a smerodajnú odchýlku. Z nich potom stanovíme 98% konfidenčný interval. Vypočítame Euklidovskú vzdialenosť medzi klastrami ako vzdialenosť ich najbližších kriviek, ktorú otestujeme voči danému konfidenčnému intervalu. Ak je test úspešný klastre sú zlúčené a algoritmus skončí. Ak test nie je úspešný, spustíme algoritmus rekurzívne pre každý z dvoch klastrov zvlášť. Po skončení všetkých vetiev dostávame výsledné klastre. V prípade 3 krivek, je detekcia outlierov uskutočnená použitím štatistického testu dvoch priemerov. Obr. 11 zobrazuje výsledky algoritmu pre vybrané skupiny kriviek. Hoci algoritmus nefunguje dokonale, zaručuje konzistentné výsledky klasterizácie.



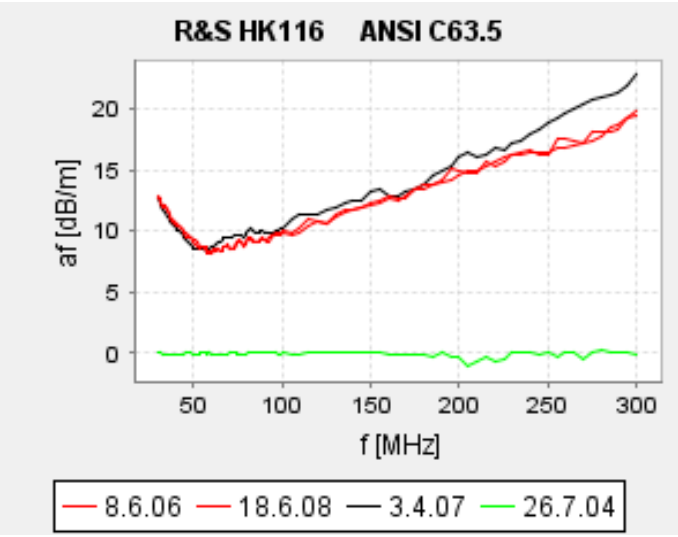
nájdené 2 namiesto 1 klastra



nájdené 2 z 3 klastrov



správne nájdený 1 klaster



nájdené 3 z 3 klastrov

Obr. 11: Výsledok kombinovaného algoritmu.

4.5. Validácia kalibračných meraní

Vychádzajúc zo záverov výskumu detekcie outlierov, použitím χ^2 štatistiky môžeme definovať validáciu meraní anténneho faktora v nasledovných krokoch:

1. Výpočet χ^2 štatistiky a stanovenie konfidenčného intervalu pre krivky, ktoré sú považované za správne.
2. Výpočet χ^2 štatistiky pre testovanú krivku, za požitia priemeru a smerodajnej odchýlky vypočítanej v kroku 1.
3. Test hodnoty štatistiky vypočítanej pre testovanú krivku v bode 2 vzhľadom na konfidenčný interval z bodu 1. V prípade úspešnosti testu považujeme krivku za správnu.

5. SUMARIZÁCIA VÝSLEDKOV

Pre overenie navrhnutého algoritmu na určovanie počtu klastrov, sme uskutočnili test na celom súbore 170 skupín obsahujúcich 1008 kalibračných kriviek. Jednotlivé skupiny boli tvorené krivkami frekvenčných závislostí anténnych faktorov pre rôzne modely antén.

Kvalita automatickej klasterizácie bola meraná ako penalizácia stanovená spočítaním sumy absolútnych rozdielov medzi požadovaným počtom klastrov v jednotlivých skupinách určeným expertom a počtom klastrov určeným algoritmom. Navyše sme vypočítali celkovú správnosť klasterizácie, spočítaním prípadov kedy sa výsledok automatickej metódy zhodoval s názorom experta. Počet správnych detekcií sme podelili celkovým počtom skupín aby sme dostali percentuálne vyjadrenie zhody algoritmu a experta.

Rovnaký postup vyhodnotenia sme použili pre kombináciu aglomeratívneho hierarchického klastrovania a kritéria celovej variancie klastrov pre určovanie počtu klastrov. Porovnanie výsledkov oboch metód je uvedený v Tab. 1.

Algoritmus	**Úspešnosť**	**Penalizácia**	**Analyzované skupiny**
Naša metóda	80.59%	38	170
Metóda celkovej variancie klastrov	37.06%	154	170

Tab. 1: Vyhodnotenie automatickej detekcie klastrov.

Nami navrhnutá metóda na automatickú detekciu klastrov bola viac ako 2-krát úspešnejšia ako metóda založená na kritériu celkovej variancie klastrov. Tu treba podotknúť, že klasifikácia kriviek do jednotlivých klastrov, je v niektorých prípadoch zložitá, pretože daná krivka môže byť považovaná za outlier vzhľadom na jej vzdialenosť od ostatných kriviek a to čiste z dôvodu malého počtu analyzovaných kriviek. Napríklad ak máme skupinu 10 kriviek, ktoré sa dajú považovať za jeden klaster, vybratím jednej krivky z hornej časti klastra a dvoch kriviek z dolnej časti klastra dostaneme situáciu kedy sa horná krivka bude javiť ako outlier voči dvom spodným.

Pre kompletizáciu výsledkov nášho výskumu, sme využili výsledky kastrovej analýzy pre výpočet neurčitosti merania anténnych faktorov pre jednotlivé modely antén. Hodnoty neurčitosti boli stanovené na základe 95% konfidenčného intervalu pre amplitúdy na celom frekvenčnom rozsahu. Fragment získaných výsledkov je uvedený v Tab. 2. Stĺpce tabuľky označujú model antény, kalibračnú metódu, typ výsledku metódy, počet kalibrácií (#), počet nájdených klastrov (DC), skutočný počet klastrov (RC) a rozšírenú neurčitosť (2Uc).

Antenna model	**Method**	**Result**	**#**	**DC**	**RC**	**2Uc [dB]**
EMCO 3115	ANSI C63.5	af	5	1	1	0.72
R&S HK116		af	4	2	1	0.35
R&S HK116	ARP 958	af	13	2	3	0.36
R&S HK116	ARP 958	af	4	1	1	0.38
R&S HK116	ANSI C63.5	af	5	1	1	0.38

Tab. 2: Výsledky klasteriácie a výpočtu neurčitostí pre jednotlivé modely antén. Stĺpec # obsahuje počet analyzovaných kalibračných kriviek, DC označuje počet automaticky detekovaných klastrov, RC je počet skutočných klastrov, 2Uc označuje rozšírenú neurčitosť pre danú skupinu kriviek.

6. ZÁVER

V tejto práci sme sa zamerali na tri hlavné témy:

- zber kalibračných dát antén a ich usporiadanie do podoby vhodnej pre výskumnú prácu,
- klasterová analýza zozbieraných dát s cieľom získania konzistentných skupín kalibračných kriviek,
- hľadanie metódy validácie meraní frekvenčných závislostí anténnych faktorov.

Ako prvé sme vytvorili databázu obsahujúcu tisíce kalibrácií rôznych modelov antén, uskutočnených na základe rozdielnych kalibračných metód. Z celkového množstva realizovaných kalibrácií sa nám podarilo automaticky nájsť 75% všetkých kalibračných certifikátov uložených v spletitej štruktúre adresárov súborového systému. Z osemdesiatich percent nájdených certifikátov bolo možné vyextrahovať kalibračné dáta v číselnej podobe. Pre splnenie tejto úlohy bolo nutné vyvinúť robustný algoritmus na prehľadávanie a spracovanie súborov kalibračných certifikátov uložených na disku počítača. V ďalšom kroku sme vytvorili iný druh počítačového programu s cieľom usporiadať kalibračné dáta podľa modelov antén, parametrov meraní a typu ich výstupov. Nakoniec bolo vytvorené grafické užívateľské rozhranie, ktoré umožnilo prechádzanie a vizualizáciu dát uložených vo vytvorenej databáze. Takto pripravené dáta nám umožnili ďalší výskum.

Predbežná analýza usporiadaných kalibračných dát ukázala, že krivky frekvenčných závislostí anténnych faktorov (kalibračné krivky) získaných v praxi vytvárajú zhluky (klastre). Táto ich charakteristická črta znemožňuje priamu aplikáciu štatistických metód pre výpočet konfidenčných intervalov počas validácie meraní. Takto vypočítané hranice intervalov by s veľkou pravdepodobnosťou zamedzovali detekcii chybných kalibráčných kriviek. Z toho vyplýva, že jednotlivé krivky musia byť rozdelené do podskupín na základe ich podobnosti, ešte predtým ako sa začne vyhodnocovať ich validita. Na účel klasifikácie kriviek do podskupín sme použili klastrovú analýzu.

Vytvorením si prehľadu o existujúcich klasterizačných metódach a ich aplikáciách sme dospeli k zisteniu, že klastrová analýza je komplexným, aplikačne špecifickým problémom. Aglomeratívne hierarchické klastrovanie a k-means klastrovanie sa ukázali byť dobre použiteľné pre naše potreby. Nakoniec sme pristúpili k uprednostneniu aglomeratívneho hierarchického klastrovania pred k-means klastrovaním, pretože jeho charakter je viac deterministický. Skúmaním dostupných mier vzdialeností, ako jedného z parametrov klastrovacích algoritmov, sme sa zamerali na *city-block* a Euklidovskú vzdialenosť. Ukázalo sa, že Euklidovská vzdialenosť je vhodnejšia pre párovanie vektorov počas klasterizačného procesu ako *city-block* vzdialenosť. Kladením väčšej váhy na body, ktoré sú ďalej od seba, produkuje lepšie výsledky.

Byť schopný nájsť klastre v kalibračných dátach je však len časťou riešenia problému klasterizácie. Viac kritickým problémom je zistenie množstva klastrov, ktoré dáta obsahujú. Pri detekcii počtu klastrov vyvstáva otázka: „Čo je to skutočný počet klastrov?“. Zvyčajne je táto otázka zodpovedaná expertom, ktorý explicitne priradí určité prvky dát do predefinovaných skupín. My však hľadáme spôsob ako túto úlohu vyriešiť automaticky použitím algoritmu. Za týmto účelom sme študovali existujúce riešenia v dostupnej literatúre, ktoré sme následne tiež testovali. Väčšina nájdených metód zlyhala v dôsledku malého počtu prvkov v nami spracovávanom súbore dát. Obvyklé množstvo kalibračných kriviek v analyzovaných skupinách je od 3 do 10. „Total cluster variance“ bolo jediné použiteľné kritérium pre hľadanie počtu klastrov v dátach pre náš prípad. Hoci prvé výsledky jeho aplikácie vyzerali sľubne, po jeho použití na rozsiahlejší súbor dát, sme dospeli k záveru, že táto metóda principiálne nie je schopná detekovať situáciu, keď sa v dátach nachádza iba jeden klaster. Taktiež v prípadoch prítomnosti viac ako jedného klastra, výsledky tejto metódy neboli v zhode s názorom experta. Na základe tohoto zistenia sme začali s vývojom novej metódy automatického určovania počtu klastrov v dátach.

Pri pohľade na skupiny kalibračných kriviek sa dá konštatovať, že krivky z individuálnych klastrov sú tzv. outliermi vzhľadom na ostatné klastre. Tento uhol pohľadu nás priviedol na myšlienku použitia metód detekcie

outlierov. Následne sme preskúmali existujúce postupy v tejto oblasti a ich aplikáciu na naše dáta. Opäť sme dospeli k záveru, že súčasné metódy nie sú úplne vhodné pre riešenie našej situácie a to kôly riedkosti dát v danom súbore, teda malého množstva kalibračných kriviek v analyzovaných skupinách. V rámci nášho riešenia sme preto navrhli použitie hodnôt vzdialeností medzi krivkami pre výpočet konfidenčných intervalov a následne sme vyvinuli , ako aj robustnú metódu na detekciu outlierov. Tieto navrhnuté postupy sa ukázali byť vhodným riešením daného problému.

Testovaním či je hodnota medzi klastrovej vzdialenosti outlierom voči vzdialenostiam medzi krivkami v rámci klastrov, sme našli vhodnú kombináciu klastrovania a detekcie outlierov. Tento princíp bol implementovaným do finálnej verzie algoritmu na automatickú detekciu počtu klastrov v dátach. Overovaním výkonnosti algoritmu na 170 skupinách kalibračných dát bola dosiahnutá 80% zhoda s manuálnym určovaním počtu klastrov v dátach expertom. Pre porovnanie, „total cluster variance“ kritérium dosiahlo iba 37% zhodu s expertom.

Poslednou adresovanou témou bola validácia meraní frekvenčných závislostí anténnych faktorov. Pôvodná myšlienka bola rátať konfidenčné intervaly na báze jednotlivých frekvencií na získanie rozsahu kde je daná závislosť považovaná za správnu. Tento prístup by však vyžadoval dodatočný mechanizmus pre rozlišovanie situácií, kde iba časť krivky je mimo platného rozsahu. Požitím χ^2 štatistiky validovaná krivka je reprezentovaná jednou hodnotou testovanou vzhľadom na konfidenčný interval, čo robí validáciu kalibrácie antén priamočiarou. Neznáma krivka sa testuje vzhľadom na konfidenčný interval založený na χ^2 štatistike skupiny kalibračných kriviek, ktoré sú považované za správne.

Na konci nášho výskumu sa nám podarilo dosiahnuť všetky stanovené ciele. Vytvorená databáza kalibračných údajov bude ďalej využitá na analýzu kalibračných procedúr a zariadení v rádio-frekvenčnej divízii spoločnosti Seibersdorf Labor GmbH. Vyvinuté klastrovacie metódy spolu s metódou na validáciu meraní je v pláne integrovať do existujúceho softvéru Calstan, ktorý sa používa na kalibrácie antén. Z hľadiska softvérového inžinierstva, bolo v rámci tohoto projektu napísaných 14 198 riadkov kódu v programovacom jazyku Java. Tento kód bude ďalej použitý v budúcich projektoch.

ZOZNAM POUŽITEJ LITERATÚRY

[1] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

[2] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification,. New York: John Wiley & Sons, 2001, Unsupervised Learning and Clustering, pp. 29-37a

[3] Casado et al. (1997). “An objective method for partitioning dendrograms based on entropy parameters”. Plant Ecology Volume 131, Number 2, Springer Netherlands, pp. 193-197.

[4] Catherine A. Sugar and Gareth M. James (2003). "Finding the number of clusters in a data set: An information theoretic approach". Journal of the American Statistical Association 98 (January), pp. 750–763.

[5] Peter J. Rousseuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65.

[6] Geisser, Seymour (1993). Predictive Inference. New York: Chapman and Hall.

[7] M. Kollár, K. Kováč. “Application of Cluster Analysis on Antenna Factor Measurements Data”; International Conference Measurement 2011, Smolenice, Slovakia

[8] Liu H., Shah S., Jiang W., ”On-line outlier detection and data cleaning,” Computers and Chemical Engineering, 28, 1635–1647, 2004.

[9] Williams G. J., Baxter R. A., He H. X., Hawkins S., Gu L., ”A Comparative Study of RNN for Outlier Detection in Data Mining,” IEEE International Conference on Data-mining (ICDM’02), Maebashi City, Japan, CSIRO Technical Report CMIS-02/102, 2002.

[10] Arthur, D. and Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding". Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms: 1027–1035.

[11] Acuna E., Rodriguez C. A., ”Meta analysis study of outlier detection methods in classification,” Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, Retrived from academic.uprm.edu/ eacuna/ paperout.pdf. In proceedings IPSI 2004, Venice, 2004.

[12] Fawcett T., Provost F., ”Adaptive fraud detection,” Data-mining andKnowledge Discovery, 1(3), 291–316, 1997.

[13] Johnson T., Kwok I., Ng R., ”Fast Computation of 2-Dimensional Depth Contours,” In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, 224-228. AAAI Press, 1998.

[14] Penny K. I., Jolliffe I. T., ”A comparison of multivariate outlier detection methods for clinical laboratory safety data,” The Statistician 50(3), 295-308, 2001.

[15] Lu C., Chen D., Kou Y., ”Algorithms for spatial outlier detection,” In Proceedings of the 3rd IEEE International Conference on Data-mining (ICDM’03), Melbourne, FL, 2003.

[16] Ruts I., Rousseeuw P., ”Computing Depth Contours of Bivariate Point Clouds,” In Computational Statistics and Data Analysis, 23,153-168, 1996.

[17] Hampel F. R., ”A general qualitative definition of robustness,” Annals of Mathematics Statistics, 42, 1887–1896, 1971.

[18] Hampel F. R., ”The influence curve and its role in robust estimation,” Journal of the American Statistical Association, 69, 382–393, 1974.

[19] Tukey J.W., Exploratory Data Analysis. Addison-Wesley, 1977.

[20] Martin R. D., Thomson D. J., ”Robust-resistant spectrum estimation,” In Proceeding of the IEEE, 70, 1097-1115, 1982.

[21] Iglewics B., Martinez J., Outlier Detection using robust measures of scale, Journal of Sattistical Computation and Simulation, 15, 285-293, 1982.

[22] Rousseeuw P., ”Multivariate estimation with high breakdown point,” In: W. Grossmann et al., editors, Mathematical Statistics and Applications, Vol. B, 283-297, Akademiai Kiado: Budapest, 1985.

[23] Knorr E. M., Ng R. T., Zamar R. H., ”Robust space transformations fordistancebased operations,” In Proceedings of the 7th International Conference on Knowledge Discovery and Data-mining (KDD01), 26-135, San Francisco, CA, 2001.

[24] Oliver J. J., Baxter R. A.,Wallace C. S., ”Unsupervised Learning using MML,” In Proceedings of the Thirteenth International Conference (ICML96), pages 364-372, Morgan Kaufmann Publishers, San Francisco, CA, 1996.

[25] Shekhar S., Lu C. T., Zhang P., ”A Unified Approach to Spatial Outliers Detection,” GeoInformatica, an International Journal on Advances of Computer Science for Geographic Information System, 7(2), 2003.

[26] Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers, 2005.

PUBLIKÁCIE AUTORA

1. BAJLA, I. – Holländer, I. – Fluch, S. – Burg, K. – Kollár, M.: An alternative method for electrophoretic gel image analysis in the GelMaster software. In: *Computer Methods and Programs in Biomedicine*. 2005, Elsevier, Vol. 77, p. 209-231
2. KOLLÁR, M. – KOVÁČ, K.: Application of Cluster Analysis on Antenna Factor Measurements Data. In: *Measurement 2011 : Proceedings. 8th International Conference on Measurement. Smolenice, Slovak Republic, 27.-30.4.2011.* Bratislava: Institute of Measurement Science Slovak Academy of Sciences, 2011, s. 84--87. ISBN 978-80-969-672-4-7.
3. KOLLÁR, M. – KOVÁČ, K.: Continuous vs. Discrete Height Scan Method in Normalized Site Attenuation Measurement for EMC Test Site Evaluation. In: *Measurement 2009 : Proceedings. 7th International Conference on Measurement. Smolenice, Slovak Republic, 20.-23.5.2009.* Bratislava: Institute of Measurement Science Slovak Academy of Sciences, 2009, s. 348--351. ISBN 978-80-969672-1-6.
4. KOLLÁR, M.: Calstan - A Software for Automation of Radio Frequency Measurements and Calibrations. In FROLLO, I. -- MAŇKA, J. -- JURÁŠ, V. *Measurement 2007 : Proceedings. 6th International Conference on Measurement. Smolenice, Slovak Republic, 20.-24.5.2007.* Bratislava: Institute of Measurement Science Slovak Academy of Sciences, 2007, s. 255--258. ISBN 978-80-969672-0-9.