Metodologija za vrednovanje podataka dobijenih merenjem hidrotehničkih veličina

Doktorska disertacija

Mr Nemanja Branisavljević

Doktorska disertacija se bavi algoritmima za statističko vrednovanje merenih podataka na bazi relacija između pojedinih merenih veličina. Pretpostavlja se da su merene veličine sastavni deo nekog fizičkog (prirodnog ili izgrađenog) sistema, čime se potvrđuje postojanje relacija koje ih povezuju.

Predložena metodologija je posebno pogodna kod primene u hidrotehničkoj praksi zbog jasno definisanih odnosa između podataka. Metodologija se može primeniti i u drugim disciplinama ukoliko postoji zavisnost među informacijama koje nose mereni podaci.

Građevinski fakultet
Univerzitet u Beogradu

Sadržaj

Rezime

Razvoj metodologija za prikupljanje, transfer i čuvanje merenih podataka omogućio je jednostavno, pouzdano i povoljno formiranje velikih istorijskih baza, između ostalih, i hidro-meteoroloških podataka. Bez adekvatne analize kvaliteta čest je slučaj da se u istorijskim bazama podataka sačuvaju i podaci niže pouzdanosti, odnosno podaci sa potencijalnom greškom koji smanjuju pouzdanost i upotrebljivost baze podataka. Velika količina podataka u bazama često prevazilazi mogućnosti tradicionalnih pristupa provere kvaliteta podataka koji se uglavnom sastoje u vizuelnoj inspekciji podataka i zaključivanju od strane eksperta, zbog čega su neophodne metodologije provere podataka automatskim ili polu-automatskim putem, (dakle), pretežno uz pomoć računara.

U ovoj tezi opisuje se rezultat istraživanja na kompleksnom zadatku vrednovanja podataka, u kome je osmišljen i primenjen algoritam za statističko vrednovanje podataka na bazi relacija između merenih veličina. Prednosti predloženog algoritma u odnosu na postojeće algoritme predstavljene pretežno u naučnoj literaturi ogledaju se u mogućnosti da se primene sve postojeće matematičke interpretacije relacija između podataka, u lakom uvođenju novih veličina i relacija u sistem i u maloj osetljivosti na eventualno nedostajuće podatke. Predloženi algoritam je posebno pogodan za primenu u hidrotehničkoj praksi zbog mogućnosti da se jasno definišu relacije između podataka, mada se može primeniti i u drugim disciplinama kod kojih se podaci mogu dovesti u matematičku vezu.

Kao rezultat ustanovljenog sistema za vrednovanje, pored ocene kvaliteta svakog pojedinog izmerenog podatka, dobija se i veličina koja reprezentuje kvalitet samog sistema, odnosno njegovu osetljivost da registruje greške određenog intenziteta. Jedna od prednosti predloženog algoritma je i ta da se podaci posmatraju kao neodređene veličine, čime se posredno unosi informacija o oblasti u kojoj se tačna vrednost podatka nalazi. Značajna novina ogleda se i u uvođenju i modeliranju neizvesnosti samih relacija, što doprinosi pouzdanosti sistema za vrednovanje i smanjuje mogućnost greške pri vrednovanju.

Algoritam sistema za vrednovanje testiran je na primeru realnih merenih podataka o količini i kvalitetu vode na ispustu Beogradske kanalizacije. Rezultati predloženog algoritma na ovom primeru upoređeni su sa rezultatima drugih metoda za vrednovanje podataka objavljenih u stručnoj literaturi i sa rezultatima ručnog vrednovanja. Na osnovu nekoliko kriterijuma pokazan je značajan napredak u odnosu na postojeće metode vrednovanja, čime je potvrđena efektnost opisanog sistema za vrednovanje.

Abstract

Development of methodologies for sampling, transfer and storage of measured data made creation of large historical databases, including hydro-meteorological databases, simple, reliable and cheap. It is a common problem that, without an adequate quality analysis, historical databases also contain less reliable data, or data with possible errors, which reduce reliability and usefulness of the database. Large amounts of data in databases often exceed capacities of traditional approaches to data quality analysis, which mostly consist of visual data inspection and decision-making by an expert, and demand development of automatic or semi-automatic data analysis methodologies, mostly with help of a computer.

The thesis describes the result of a research on a complex task of data validation, where an algorithm for statistical data validation on the basis of relations between measured variables has been developed and applied. Advantages of the proposed algorithm in comparison to existing algorithms presented in literature are that it enables the application of all existing mathematical interpretations of data relations, that introducing new variables and relations into the system is easy, and that the system has little sensitivity to possibly missing data. The proposed algorithm is particularly convenient for application in hydraulic, hydrologic and environmental practice due to the possibility to clearly identify data relations, though it can also be applied in other disciplines where data relations can be presented mathematically.

The established data validation system, besides evaluating the quality of each measured data, provides variables which represent quality of the system itself, i.e. its sensitivity to errors of certain magnitude. One of the advantages of the proposed algorithm is that the data are presented as uncertain, thus providing indirectly the information on the interval in which the exact data value is to be found. A significant novelty is also that uncertainty of relations themselves is introduced and modeled, which contributes to the validation system reliability and reduces possibilities of mistakes in the validation process.

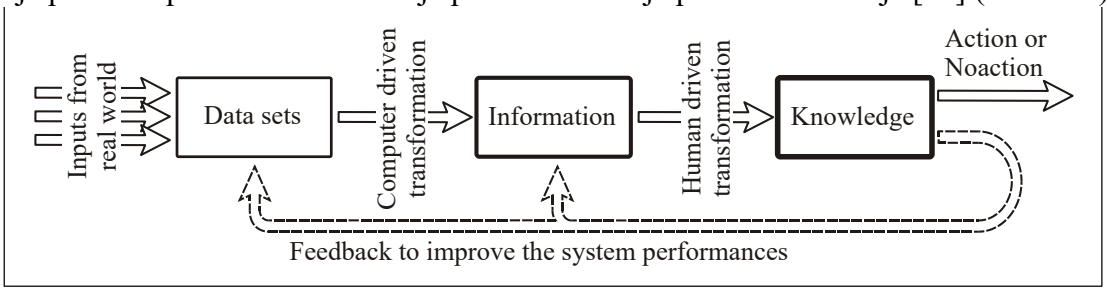
The validation system algorithm has been tested on measured data of the hydraulic and quality parameters in the Belgrade sewer outlet. The results of the proposed algorithm are compared to results obtained through other data validation methods presented in literature and with results obtained through manual validation. A significant progress in relation to the existing validation methods has been shown using several criteria, thus confirming the effectiveness of the described validation system.

Zahvalnost

TODO

1. Uvod

Opažanje predstavlja početak puta u procesu spoznaje karakteritika nekog hidrotehničkog procesa. Sledi tumačenje podataka i formiranje informacija koje predstavljaju protumačen i upotrebljiv podatak (ili grupu podataka). Informacija se dalje pretvara u znanje. Znanje se dalje može matematički ili lingvistički zabeležiti i upotrebiti kao osnov ili uzor u procesu upotrebe podatka. Na slici 1.1 je prikazan proces transformacije podatka u znanje preko informacije [36] (Slika 1.1)



Slika 1.1: Transfer od podatka do znanja
Pozajmljeno iz [36]

Podaci su numeričke ili opisne vrednosti koje se odnose na neku veličinu, dok je veličina u fizici, pa i u hidrotehnici "ono što se meri, tj. upravo ono što se može izmeriti" [45]. Zbog toga se poreklo podatka povezuje sa merenjem. Veličine se mogu podaliti u dve grupe: 1) osnovne i 2) izvedene. "Osnovne veličine su međusobno nezavisne i neuporedive i njih treba tako izabrati da se preko njih može izraziti i izmeriti sve ono što razmatrana problematika nameće" [45]. Na primer, podatak o brzini ne govori ništa o geometriji, dok se bez brzine i geometrije ne može izračunati protok. Brzina i geometrija (izražena dužinom) su osnovne veličine za jedan hidrotehnički problem, dok je protok izvedena.[1]

Da bi se podatak proizveo u informaciju potrebno je da on bude pre svega upotrebljiv, a zatim i transformisan (prilagođen) u formu koja ga čini "pitkijim" i korisnijim za korisnika. Upotreba merenih podataka u hidrotehnici se može sagledati iz više aspekata. U knjizi „Merenja u hidrotehnici" [75] svrha merenja u hidrotehnici je, između ostalog, opisana u odnosu na karakter istraživanja (fundamentalna istraživanja, primenjena istraživanja opšteg karaktera, primenjena istraživanja lokalnog karaktera i konkretna istraživanja), primenu u softverskom inženjerstvu (testiranje, validacija i verifikacija modela) i primenu u realizaciji složenog hidrotehničkog sistema. Time su navedene samo neke uloge informacija proisteklih iz podataka dobijenih merenjem hidrotehničkih veličina. Moderan pristup hidrotehnici, koji u hidrotehničku nauku i praksu kroz pojam hidroinformatike uvodi prof. Abbot [1], podrazumeva primenu informacionih tehnologija u hidrotehnici uzimajući u obzir sve veći rast količine podataka u bazama podataka i potrebu da se u proces njihovog tumačenja i upotrebe uključe napredni matematički alogoritmi i veštačka inteligencija [5].

Transformacija jednog izmerenog podatka u informaciju može se obaviti na više načina [36]: normiranjem i smeštanjem u organizovanu bazu podataka; statističkom analizom; simulacionim modelima; „data mining" algoritmima; itd. Kao što se vidi, prvi korak u transformaciji podatka i njegovoj kasnijoj upotrebi je smeštanje podatka na njegovo mesto. Kroz istoriju se način organizacije podataka menjao, pa se od tabela i kartoteka stiglo do objektnih relacionih baza za nestruktuirane podatke [128]. Statistička analiza predstavalja jedan od osnovnih vidova transformacija podataka koji datira još iz 17. veka. Srednja, minimalna ili maksimalna vrednost nekog skupa podataka

[1] Ili obrnuto. Ukoliko se, na primer, meri protok, onda je brzina izvedena veličina.

ponekad imaju veće značenje za korisnika nego ceo skup pojedinačnih vrednosti koji je teško protumačiti.

Transformacija izmerenog podatka u informaciju se može sprovesti i upotrebom simulacionih modela. Simulacioni modeli predstavljaju matematičku formulaciju relacija između veličina prilagođenu upotrebi pomoću računara. Pomoću simulacionih modela moguće je izračunati podatke koje iz nekog razloga nije bilo moguće izmeriti ili čak predvideti ponašanje nekog modeliranog dinamičkog sistema. Poslednjih decenija, razvojem tehnologije vezane za veštačku inteligenciju, sve više na popularnosti dobija nova grana nauke koja se odnosi na otkrivanje skrivenih informacija u velikom broju podataka. *Data mining* tehnike se mogu podeliti u četiri velike grupe: 1) klasifikaciju, 2) grupisanje, 3) regresiranje i 4) otkrivanje relacija između podataka [131].

Informacije dobijene transformacijom podataka predstavljaju osnovu za sticanje znanja o hidrotehničkom procesu. Ukoliko se informacije baziraju na lošim ili nedovoljno tačnim podacima, može se steći potpuno pogrešna slika i/ili izvesti pogrešni zaključci. Stiče se utisak da je bolje da informacija uopšte ne postoji nego da je loša. Zbog toga je neophodno utvrditi i proceniti kvalitet i reprezentativnost podataka koji se odnose na neku hidrotehničku pojavu, na osnovu čega bi mogla da se odredi i valjanost informacije. Postupak procene kvaliteta i reprezentativnosti podataka naziva se vrednovanje podataka.

Prikupljeni podaci se mogu koristiti ili jednokratno, najčešće u realnom vremenu, ili višestruko. Prvi način upotrebe pretpostavlja praćenje neke pojave preko podatka čija je vrednost najčešće značajna samo u trenutku osmatranja. Na ovaj način se, na primer, prate procesi u postrojenjima gde operater (ili automatika) na osnovu podataka koji daju informacije o stanju u postrojenju donosi odluke. Na primer, kod filterskih polja u postrojenjima za prečišćavanje vode na osnovu nivoa se određuje položaj ventila i tako održava konstantan nivo. Podaci se u tom slučaju ne čuvaju, pa se bilo kakva informacija o događanjima u prošlosti gubi. Operater, često nesvesno, na osnovu iskustva sprovodi vrednovanje podataka i na osnovu očekivanih vrednosti procenjuje da li je u odvijanju procesa sve u redu. S obzirom na to da često ne ostaje trag o istorijskim podacima u sistemu, ili se podaci čuvaju u vrlo kratkom intervalu vremena (na primer, radi rekonstrukcije događaja ako se dogodi nešto nepredviđeno), ne postoji potreba da se vrednovanje podataka sprovodi na sistematičan način i da se ostavlja trag o vrednosti izmerenog podatka. Osnovni razlog tome je činjenica da se operater može setiti koje su okolnosti vladale za vreme uzorkovanja u neposrednoj prošlosti.

Sa druge strane, ukoliko je predviđeno da se podaci koriste višestruko, neophodno ih je organizovati i čuvati zajedno sa dokumentacijom o tome kakve su bile okolnosti u vreme uzorkovanja (može se dogoditi da do upotrebe podatka prođe duži vremenski period). Istorijske baze podataka su neophodne u hidrotehničkoj praksi i ta činjenica izdvaja vrednovanje podataka u hidrotehnici kao disciplinu kojoj se mora pristupiti sistematično, vodeći računa o svim karakteristikama hidrotehničkih procesa, načinu osmatranja i merenja hidrotehničkih veličina, kao i o svrsi i načinima upotrebe.

Razvoj metodologija za prikupljanje, transfer i čuvanje merenih podataka omogućio je jednostavno, pouzdano i povoljno formiranje velikih istorijskih baza, između ostalih, i hidrometeoroloških podataka. Bez adekvatne analize kvaliteta čest je slučaj da se u istorijskim bazama podataka sačuvaju i podaci niže pouzdanosti,odnosno podaci sa potencijalnom greškom. Neprovereni podaci smanjuju pouzdanost i upotrebljivost baze podataka. Velika količina podataka u bazama često prevazilazi mogućnosti tradicionalnih pristupa provere kvaliteta podataka koji se uglavnom sastoje u vizuelnoj inspekciji podataka i zaključivanju od strane eksperta, zbog čega su neophodne metodologije provere podataka automatskim ili polu-automatskim putem, (dakle), pretežno uz pomoć računara.

U ovoj doktorskoj disertaciji prikazan je sistem za vrednovanje merenih podataka koji se koriste u hidrotehnici. Osnovne pretpostavke za vrednovanje podataka jedne ili više merenih veličina jesu da su vrednovani podaci mereni u okviru nekog fizičkog sistema i da se između njih mogu uspostaviti matematičke relacije. Relacije mogu biti bazirane na fizici, statistici ili algoritmima veštačke inteligencije. Iz merenih podataka i podataka koji predstavljaju predikciju merene veličine statističkim postupkom izdvajaju se podaci koji u sebi sadrže potencijalnu grešku. Ukoliko relacije ne postoje podaci se mogu smatrati nezavisnim, i u tom slučaju vrednovanje pomoću prikazanog sistema nije moguće.

S obzirom na to da se vrednovanje mora sprovesti upoređivanjem podataka sa odgovarajućim informacijama (npr. upoređivanjem izmerene vrednosti sa granicama mogućih vrednosti), uvođenjem u proceduru vrednovanja neodređenosti samih podataka i modeliranjem neodređenosti relacija između podataka omogućeno je i kvantifikovanje samog sistema za vrednovanje i određivanjem njegove osetljivosti da registruje greške u podacima određenog intenziteta.

Opisan sistem za vrednovanje testiran je na tri primera. Dva hipotetička primera imaju za cilj da pokažu da primenjeni algoritam daje smislene rezultate. Trećim primerom je sistem za vrednovanje testiran na realnim merenim podacima hidrotehničkog sistema. Rezultati dobijeni u trećem primeru nedvosmisleno pokazuju napredak u pogledu vrednovanja podataka u odnosu na metode objavljene u literaturi. Vrednovanje samog sistema za vrednovanje podataka sprovedeno je na osnovu nekoliko kriterijuma za ocenu kvaliteta rezultata vrednovanja.

Sam tekst doktorske disertacije podeljen je u 6 poglavlja. U ovom, *Uvodnom* poglavlju ukratko je opisana potreba za merenjem u hidrotehnici, proces transformacije merenog podatka u informaciju i znanje, kao i problemi sa kojima se susreće hidrotehnička praksa zbog sve većeg priliva podataka. Takođe je ukratko opisana potreba za vrednovanjem merenih podataka i osnovne karakteristike sistema predloženog u ovoj doktorskoj disertaciji.

U drugom poglavlju (*Pregled literature*) prikazani su najznačajniji naučni i stručni doprinosi u oblasti vrednovanja podataka poslednjih nekoliko decenija. Kritičkim osvrtom na mane i kvalitete predloženih metoda, postavljen je osnov za strategiju koja je korišćena u ovoj doktorskoj disertaciji.

U poglavlju *Problematika vrednovanja podataka* opisuju se karakteristike podataka koji se koriste u hidrotehnici, kao što su greške i neodređenost, navode se primeri grešaka u podacima iz hidrotehničke prakse i navode se definicije kvaliteta nekog podatka. Takođe se razmatraju načini uspostavljanja relacija između podataka i moguća rešenja problema neodređenosti relacija. Osnovni cilj ovog poglavlja je da se čitalac uvede u oblast vrednovanja podataka i upozna sa problemima koji se razmatraju dalje u disertaciji.

U četvrtom poglavlju, pod naslovom *Metodologija vrednovanja podataka*, detaljno su objašnjeni algoritmi za vrednovanje i tumačenje rezultata vrednovanja. Pored toga, u ovom poglavlju se razmatraju i kriterijumi za ocenu kvaliteta samog sistema za vrednovanje.

U pretposlednjem poglavlju – *Primeri*, navode se detalji i rezultati primera za vrednovanje podatka. U istom poglavlju se rezultati razvijenog sistema upoređuju sa rezultatima iz literature.

Na kraju, u šestom poglavlju 6 sumiraju se zaključci proizašli iz istraživanja opisanog u doktorskoj disertaciji i daje se kratak osvrt na moguća dalja istraživanja u oblasti vrednovanja podataka.

2. Pregled literature

Kvalitet podatka prepoznat je kao njegova važna osobina i to ne samo u hidrotehnici, već i na brojnim drugim naučnim i praktičnim poljima, kao što su: upadi u informacione sisteme – procena rizika od gubitka podataka u računarskim sistemima; detekcija kraĐa – na kreditnim karticama, mobilnim telefonima, osiguranju, itd.; podaci dobijeni pri kliničkim ispitivanjima – anomalije na EEG, EKG, itd. zapisima mogu biti indikator različitih patoloških stanja; detekcija kvarova – rano otkrivanje lošeg rada industrijskih i mehaničkih postrojenja ili detekcija štete na konstrukcijama; rad mreže senzora – otkrivanje poremećenog rada automatskih, a često i bežično umreženih senzora; itd.

Pošto je za problematiku odreĐivanjakvaliteta podatka izuzetno zainteresovana hidrotehnička praksa, na naučnom polju u oblasti hidrotehnike se metodologija neprestano poboljšava i unapreĐuje. Obimna literatura koja se bavi kvalitetom podataka može se klasifikovati prema brojnim kriterijumima. Najvažniji od njih su: 1) oblast primene, 2) primenjena metodologija i 3) način klasifikacije podataka prema kvalitetu.

Oblasti primene su brojne, od detekcije redudantnih zapisa u opštim bazama podataka do detekcije kvarova kod kompleksnih sistema. Metodologija odreĐivanja kvaliteta podataka obično je predstavljena preko matematičkih i logičkih alata (kao i načina njihovog kombinovanja) koji su primenjeni u procesu vrednovanja podataka. Njihov izbor i način primene se uglavnom značajno razlikuju od primera do primera koji su predstavljeni u literaturi.

Način klasifikacije podataka prema kvalitetu predstavlja poseban izazov u procesu vrednovanja podataka. U literaturi se susreću brojne varijante vrednovanja kao što su: detekcija anomalija (eng. anomaly detection), detekcija grešaka (eng. fault detection), detekcija odstupanja (eng. outlier detection), validacija podataka (eng. data validation), itd.

Svaka od navedenih varijanti vrednovanja podataka kao rezultat daje specifične klase u koje su podeljeni podaci. Kod detekcije anomalija i detekcije grešaka pretpostavljaju se dve klase – klasa regularnih podataka i klasa neregularnih podataka. Kod detekcije grešaka se u klasu neregularnih podataka svrstavaju i podaci koji se odnose na neuobičajena stanja i pojave iako su možda izmereni kako treba. Detekcija odstupanja predstavlja varijantu vrednovanja podataka kod koje se kao neregularni podaci registruju oni koji odstupaju od većine. Za razliku od detekcije grešaka i detekcije anomalija, kod detekcije odstupanja se može definisati i više klasa kvaliteta (prema nivou odstupanja) ili se čak i samo odstupanje, normiranjem, može prevesti u informaciju o kvalitetu kao kontinualna veličina. Validacija podataka pretpostavlja postojanje dve ili više diskretnih klasa kvaliteta, ili u opštem slučaju kontinualnu ocenu po kojoj se kvalitet podataka rangira. Kod svih navedenih varijanti vrednovanja se pretpostavlja generisanje i čuvanje dodatnih informacija (meta-podataka) koje imaju za cilj da pojasne karakteristike klasa kvaliteta i način na koji je neki podatak svrstan u odreĐenu klasu.

Pošto se način na koji se problematika vrednovanja podataka doživljava u praksi i nauci razlikuje, u ovoj tezi je literatura klasifikovana u dve klase:

1. preporuke hidro-meteoroloških organizacija i ostalih pružalaca podataka, i
2. naučna praksa

Ovakva podela napravljena je, pre svega, kako bi se prikazalo trenutno stanje u pogledu problematike vrednovanja podataka u praksi i kako bi se naglasio napredak koji se čini uplivom nauke u ovu oblast poslednjih decenija. Na taj način se pokušava da se istakne da ne postoji dovoljan

broj smernica koje postoje za provajdere i korisnike podatka i da se ukaže na moguć napredak koji bi se postigao uvoĐenjem neke od naprednijih metoda u praksu, kao što je metoda predložena u ovoj tezi.

Kao poseban odeljak u poglavlju izvdojen je pregled softverskih rešenja koji ili postoje na tržištu ili su u procesu razvoja i njihovo testiranje i primena se tek očekuju.

2.1.Preporuke hidro-meteoroloških i ostalih orgnizacija koje isporučuju podatke

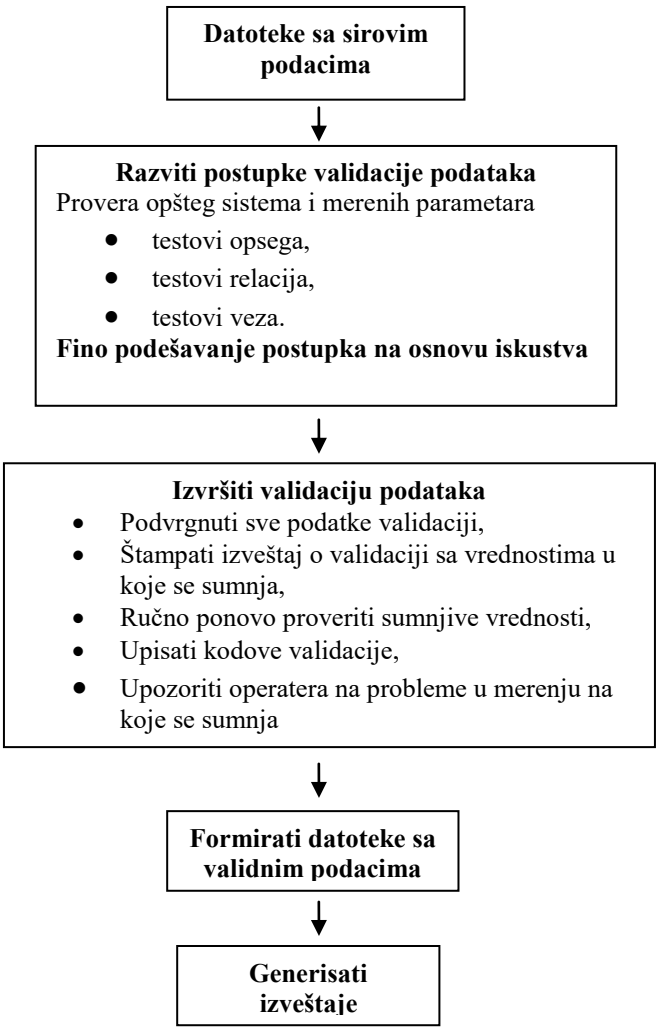
Standardi kvaliteta podataka koje hidrometeorološke službe zemalja u svetu pružaju klijentima uglavnom ne postoje. Svetska meteorološka organizacija (WMO) u svom uputstvu [133] daje informacije kao što su brzina vetra, temperatura, vlažnost i sl. neekspertima kojima je potrebno da shvate osnovne pojmove i važnost kvaliteta podataka u oblasti merenja meteoroloških veličina. Ovo uputstvo daje teorijske okvire rada mernih uređaja i njihove fizičke povezanosti sa veličinom koju mere, principe kalibracije i osnovne i napredne tehnike obrade signala. TakoĐe, precizirani su faktori koji utiču na kvalitet merenih podataka:

1. Odnos zahteva za kvalitetnim podacima i mogućnosti sistema da te zahteve ispune. Ova stavka daje ocenu da li je sistem uopšte sposoban da proizvede dovoljno kvalitetne podatke;
2. Tehničke i funkcionalne mogućnosti. Ova stavka obuhvata širok spektar uslova koji su neophodni da bi se obavilo kvalitetno merenje – od stručnog kadra do pravne regulative, od tehničkih uputstava do standarda. Nedostaci kod činioca ovog tipa mogu dovesti do značajnog gubitka kvaliteta podataka;
3. Izbor senzora i instrumenata. Ukoliko je izabran neadekvatan merni instrument, ne može se očekivati pouzdan podatak;
4. Pravilna ugradnja i adekvatni uslovi rada senzora;
5. Kompatibilnost mernih instrumenata;
6. Mikrolokacija i adekvatna pozicija mernog isnstrumenta;
7. Pravilno upravljanje greškama senzora. Osrednjavanje ima značajnu ulogu;
8. Prikupljanje podatataka (*data acquisition*);
9. Obrada podataka (*data processing*);
10. Kontrola kvaliteta u realnom vremenu (*real time quality control*);
11. Posmatranje celokupnog procesa (*performance monitoring*);
12. Testiranje i kalibracija (*test and calibration*);
13. Održavanje (*maintenance*);
14. Obuka kadrova (*training and education*);
15. Kvalitet metadata podataka (*metadata*).

Svaki od navedenih 15 faktora može značajno uticati na kvalitet izmerenog podatka. Registrovanjem anomalije u vremenskoj seriji potrebno je, ukoliko je moguće, proveriti i ispitati koji od navedenih faktora utiče na pojavu greške i otkloniti problem.

Pored ovog, specijalizovanog uputstva za održavanje kvaliteta podataka, postoje i opšti sistemi za održavanje kvaliteta proizvoda i usluga. Internacionalna organizacija za standardizaciju (ISO) proizvela je niz standarda od kojih se oni sa oznakom ISO 9000 odnose na kvalitet. Pored ISO vredi navesti i niz procedura nazvanih *Total Quality Management* (TQM) čiji je glavni cilj ispravljanje lošeg kvaliteta tako što se popravlja sistem u celini, a ne samo neki njegovi delovi, pri čemu se informacije o lošem kvalitetu dobijaju na osnovu indikatoraa kvaliteta koji se računaju statističkim putem.

Pored opštih pravila i uputstava, postoje i individualne studije koje se bave kvalitetom pre svega meteoroloških podataka. Jedna od njih sprovedena je od strane grupe autora u vezi s merenjem podataka vezanih za vetar [4]. U ovoj studiji se, pored detaljno opisanih sistema za merenje brzine i pravca vetra, nalaze i algoritmi za validaciju i određivanje kvaliteta izmerenih podataka. Proces validacije dobijenih podataka je predstavljen dijagramom na slici 2.1.



Slika 2.1: Procedura definisana u priručniku [4]

Ukoliko se raspolaže dodatnim potrebnim podacima, jednostavne procedure predstavljene u ovom dokumentu može sprovesti ekspert ručno ili se mogu sprovesti iautomatski. Procedure koje je potrebno sprovesti podeljene su u dve grupe: pretraživanje podataka (eng. *data screening*), gde se vremenske serije pretražuju u potrazi za sumnjivim podacima, i verifikacija podataka (eng. *data verification*), gde je potrebno odlučiti da li je podatak zadovoljavajući i da li ga treba zadržati, odbaciti ili zameniti redudantnim podatkom ukoliko on postoji. U fazi verifikacije kvalitet podatka označava se stepenom pouzdanosti (u dokumentu nije definisana metodologija kvantifikacije pouzdanosti merenja). U ovom koraku ključnu ulogu ima stručno lice – ekspert, kome je poznat proces monitoringa meteoroloških podataka. Takođe, u ovom dokumentu predlažu se specifični algoritmi validacije:

1. Sveobuhvatni pregled podataka
    a. Provera da li broj podataka odgovara očekivanom broju u vremenskoj seriji;

   b. Provera da li postoje praznine u vremenskoj skali vremenske serije u kojoj nedostaju i podaci;
2. Pojedinačni pregled i validacija podataka
   a. Provera opsega merenja;
   b. Provera relacija između izmerenih veličina;
   c. Provera trenda.

Poražavajući izveštaj WMO iz 2003. godine [112] koji se odnosi na zvanične hidrometeorološke ustanove u pojedinim zemljama sveta govori o slaboj kontroli kvaliteta podataka koji se isporučuju klijentima. U najvećem broju slučajeva ne postoji kontrola kvaliteta podataka, iako je većina evropskih i severnoameričkih zemalja uvela ISO 9000 standard u svoje organizacije.

Na *IAHS* (International Association of Hzdrological Sciences) *Workshop*-u o sistemu kvaliteta koji je potrebno uspostaviti u hidrologiji (*Quality Assurance in Hydrology Measurement*), održanom u julu 1995, raspravljalo se o primeni internacionalnog standarda ISO 9002 koji se odnosi na kvalitet proizvoda (pri čemu je hidrološki podatak proizvod) [54]. Iako se na skupu nije raspravljalo o metrologiji već se diskusija vodila oko transfera podataka od mernih nivoa do protoka (preko QH krivih), koje inače i unose najveću neizvesnost u procesu formiranja informacije o protoku, zaključeno je da se, između ostalog, moraju uspostaviti procedure koje će nedvosmisleno ukazivati na sve promene koje su na podacima izvršene, uključujući i korekcije u toku pripreme podataka. Na taj način je skrenuta pažnja učesnicima i javnosti da se kvalitet podataka mora ostvariti pre svega praćenjem promena u sistemu i pravovremenim reagovanjem, pre nego što podaci dođu u javnost.

Svaki mereni podatak mora u bazi podataka imati relaciju ka odgovarajućem meta-podatku. Meta-podaci oslikavaju uslove i način na koji je neki podatak dobijen. U dokumentu Svetske meteorološke organizacije (WMO) [132] se nalazi i predlog šta sve treba da sadrže prateći podaci merenih vremenskih serija (meta-podaci):

1. Informacije o mernoj stanici:
   a. Administrativne informacije;
   b. Lokacija: geografske koordinate, kota nivoa;
   c. Opis mikrolokacije i ograničenja;
   d. Izgled stanice i podešavanja;
   e. Mogućnosti: prenos podataka, napajanje, povezivanje;
   f. Klima na mestu merenja;
2. Informacije o pojedinim instrumentima:
   a. Tip: proizvođač, model, serijski broj, način rada;
   b. Karakteristike;
   c. Kalibracioni podaci i vreme kalibracije;
   d. Mesto i izloženost pojavi: lokacija, zaštita, kota;
   e. Program merenja i setovanje;
   f. Setovano vreme;
   g. Osoba koja je zadužena za podatke;

h. Prikupljanje podataka: semplovanje, osrednjavanje;
i. Obrada podataka, metode i algoritmi;
j. Preventivno i korektivno održavanje;
k. Kvalitet podataka prema stanju mernog instrumenta.

Poslednja stavka, koja označava kvalitet podatka, možda je i najvažniji pratilac podatka na putu ka daljoj upotrebi. Istovremeno, taj meta-podatak je i najteže odrediti, jer treba na osnovu raspoloživog znanja, drugih meta-podataka, kao i rezultata merenja na određenom mestu, ali i rezultata merenja na drugim, srodnim mestima, proceniti kvalitet podatka i to kvantifikovati. U nastavku teksta će se tom problemu posvetiti posebna pažnja.

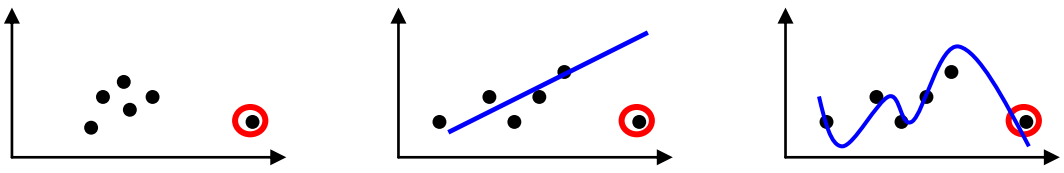
2.2 Naučna praksa

2.2.1 Vrednovanje podataka – uvod

Osnovni aspekt prema kom se mora posmatrati način vrednovanja podataka je priroda podatka koji se vrednuje [77]. Podatak se u najopštijem smislu može shvatiti kao kolekcija različitih objekata (*object*, *record*, *point*, *vector*, *pattern*, *event*, *case*, *sample*, *observation*, *entity*) [44]. Svaki od objekata može biti predstavljen nizom osobina ili atributa (*variable*, *characteristic*, *feature*, *field*, *dimension*) [44]. Atributi mogu biti različitog tipa, na primer binarni, kontinualni ili u vidu neke kategorije. Svaki objekat može biti predstavljen samo jednim atributom (jednodimenzionalni) ili preko više atributa (višedimenzionalni). U slučaju višedimenzionalnih objekata, atributi mogu biti istog ili različitih tipova. Na osnovu svega toga, može se zaključiti da se metode vrednovanja moraju odabrati na osnovu prirode podataka kojima se raspolaže. Na primer, ukoliko se podaci vrednuju na osnovu udaljenosti jednih od drugih (u odnosu na neki koordinatni sistem i izabranu transformaciju),i ukoliko su podaci u obliku kategorija, moraju se definisati posebni principi za određivanjerastojanja jednih podataka od drugih.

Takođe, metode vrednovanja se moraju odabirati i na osnovu raspoloživih relacija između podataka [77]. Relacije između podataka pružaju redudantne informacije o odnosu između podataka. Tako, na primer, postoje metode prilagođene podacima koji su u prostornim ili vremenskim relacijama ili metode koje su razvijene isključivo za podatke čije se relacije mogu izraziti nekom vrstom grafa.

Proces vrednovanja mora biti oslonjen na neku bazu u odnosu na koju se proverava regularnost podataka. Bazu mogu činiti sami podaci (uglavnom gustina podataka – slika 2.2A), relacije između podataka (slika 2.2B) ili kombinacija relacija i istorijskih podataka (slika 2.2C).



Slika 2.2: Vrednovanje koje se oslanja na gustinu podataka (A), linearnu relaciju između podataka (B) i kalibrisan model pomoću istorijskih podataka (C)

Anomalije se mogu podeleiti na sledeća tri tipa [77]:

1. Tačkaste anomalije – anomalije koje postoje samo u izolovanim podacima. Ovo je najčešći tip anomalija i najveći broj metoda za vrednovanje podataka se odnosi upravo na njih.

2. Kontekstualne anomalije – anomalije koje se moraju posmatrati u skladu sa nekim kontekstom koji mora biti naveden u formulaciji problema [46]. Na primer, neki podatak može biti regularan pod nekim uslovima, dok pod nekim drugim uslovima može predstavljati anomaliju. Veliki problem kod metoda koje se koriste za detektovanje ovog tipa anomalija je nepostojanje dodatnih podataka koji bi omogućili klasifikaciju u predviĎene kontekste. Ponekad je ta klasifikacija jednostavna (na primer prema vremenu ili prostoru), a ponekad njena složenost prevazilazi i složenost same metode za vrednovanje.
3. Grupne anomalije – anomalije koje se mogu uočiti ako se posmatra čitava grupa podataka. Ove anomalije obično se manifestuju kao nemogućnost da se podatak uklopi u predefinisan obrazac javljanja. Ovaj tip anomalija obično se javlja kod vremenskih serija ili prostornih podataka.

Anomalije u podacima ne moraju biti isključivo jednog tipa. Naime, i tačkaste i grupne anomalije mogu se posmatrati u skladu sa kontekstom u kom se nalaze.

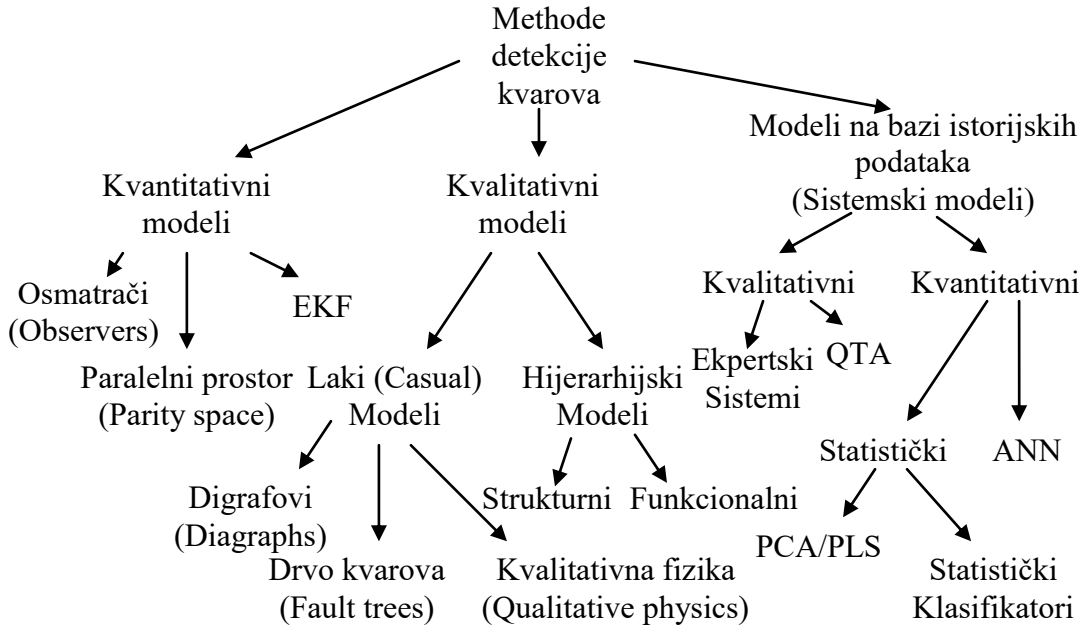
2.2.2 Pregledni radovi

Detekcija kvarova detaljno je opisana u preglednim člancima [37], [41], [56], [63], [64], [123], [124] i [125], detekcija izuzetaka u preglednim člancima [52], detekcija anomalija u [19] i [87], a kvalitet podataka u člancima [44] i [62].

Analizom literature prikazane u navedenim preglednim radovima može se uočiti da postoji tendencija da se u jednom radu uglavnom primenjuje samo jedna metoda sa svim ograničenjima koje nosi. TakoĎe, stiče se i utisak da se neke očigledne relacije izmeĎupodataka (npr. postojanje fizičkih zavisnosti) svesno zanemaruju ukoliko njihova implementacija kod primenjenih metoda nije trivijalna. Isto tako, često se za potrebe elegantnije primene odabrane metode uvode neadekvatna ograničenja i pretpostavke, čime se svesno povećava neizvesnost relacije izmeĎu podataka i smanjuje mogućnost odreĎivanjaanomalija u podacima.

2.2.3 Detekcija kvarova (*fault detection*)

Postoje brojni radovi u kojima se opisuje detekcija kvarova pretežno industrijskih sistema (*plants*). U preglednim radovima [123], [124] i [125] metode iz literature podeljene su prema tipovima matematičkih modela koji su korišćeni. Metode koje su prikazane u detekciji kvarova dele se na metode za koje je potrebno posedovati model redovnog stanja sistema (kada nema kvarova) i metode za koje je potrebno model formirati na osnovu istorijskih podataka. Prva grupa modela se deli na kvantitativne i kvalitatine modele. Hijerarhija metoda prikazana je dijagramom na slici 2.3.



Slika 2.3: Klasifikacija metoda koje se mogu koristiti pri detekciji kvarova u sistemaima (*fault detection/diagnosis*) [123], [124] i [125]

Kod kvantitativnih modela procesi su opisani matematičkim zavisnostima i funkcionalnim vezama, dok se kod kvalitativnih modela veze između promenljivih ostvaruju kvalitativnim funkcijama. Pri modeliranju osmatranog sistema na osnovu istorijskih podataka o sistemu pretpostavlja se samo postojanje velikog broja istorijskih podataka, dok se modeli svode na *data mining* algoritme koji se ne oslanjaju na prethodno znanje i iskustvo o sistemu koji se osmatra.

2.2.3.1 Kvantitativni modeli

Kod detekcije kvarova, tradicionalno se najviše članaka i knjiga odnosi na metode bazirane na upotrebi linearnih modela stanja dobijenih iz jednačina održanja ili nekom od metoda identifikacije sistema. Metode koje su obrađene u literaturi mogu se svrstati u tri grupe: 1) metode bazirane na osmatračima (*observer*), 2) metode bazirane na rezidualima (*parity space*) i 3) metode bazirane na frekventnim karakteristikama signala. U prvu grupu metoda spadaju metode u kojima se merene vrednosti upoređuju sa rezultatima modela, a razlika se tretira deterministički ili stohastički. Razlike merenih i izračunatih vrednosti se dalje predstavljaju linearnim modelima, linearizovanim nelinearnim, determinističkim ili stohastičkim modelima (sa upotrebom Kalmanovih filtera).

Metode bazirane na formiranju prostora reziduala imaju za cilj da transformišu matricu transformacije ulaza izlaza i stanja sistema, tako da u modelu figurišu isključivo ulazi i izlazi iz kojih se formiraju reziduali [88]. Postoje dva opšta postupka u literaturi koji se odnose na metode bazirane na rezidualima: izračunavanje matrice transformacije za statičke modele i za dinamičke modele [130]. Identifikaciju varijable koja je u procesu u stanju različitom od redovnog moguće je obaviti ukoliko se prostor reziduala tako formira da oni budu ortogonalni jedni u odnosu na druge. U [28] prikazan je i objašnjen proces detekcije kvarova na osnovu transformacija transfer funkcije osmatranog sistema.

U skorijoj literaturi istraživanje je nastavljeno u pogledu obezbeđivanjarobustnosti postupka, dok se veza sa ostala dva pravca detekcije kvarova navedena u ovom odeljku može naći u [20].

Osnovni nedostatak primene navedenih metoda kod hidrotehničkih veličina su nelinearne veze koje postoje između njih. Naime, ukoliko bi se veze između hidrotehničkih veličina linearizovale, porasla bi neizvesnost takvih relacija, čime bi se smanjila mogućnost da se detektuju anomalije u podacima.

2.2.3.2 Kvalitativni modeli

Kvalitativni modeli koji se u literaturi koriste u procesu detekcije kvarova mogu se klasifikovati u tri grupe: 1) označeni orijentisani direkcioni grafovi (*digraphs*), 2) stabla kvarova (*fault trees*) i 3) metode kvalitativne fizike. Označeni orijentisani direkcioni grafovi (OODG) predstavljaju grafove sa orijentisanim vezama između čvorova sa pozitivnim ili negativnim oznakama. Oni označavaju pravac od uzroka do posledice. Pošto mogu veoma jednostavno i slikovito da opišu uzročno posledične veze između varijabli procesa koji se osmatra, izuzetno su popularni i postoji brojna literatura o njima. U [55] prvi put se OODG opisuju kao moguć put kako da se detektuju kvarovi u sistemu. Iz OODG autori su osmislili takozvane uzrok-posledica grafove, koji su se dalje razvijali u nizu kasnijih radova. U [114] je pokazano kako se OODG mogu izvesti iz diferencijalnih jednačina koje opisuju osmatrani sistem, što predstavlja izuzetan doprinos za slučajeve kada se osmatrani proces može opisati jednačinama, ali za njihovo rešavanje ne postoje pouzdani podaci. OODG su dalje prošireni uvođenjem rasplinutih skupova od strane [48].

Stabla kvarova prvi put su predstavili *Bell Telephone Laboratories* 1961. godine. Ona predstavljaju niz na logičan način poređanih događaja povezanih uzročno posledičnim vezama. Osnovna odlika stabla kvarova je da se jedan događaj može propagirati od mesta nastanka kvara do njegovog uzroka. Stabla kvarova su obično organizovana u formu grafa sa čvorovima opremljenim AND ili OR operatorima koji služe u procesu propagacije događaja [40].

Kvalitativna fizika je disciplina u kojoj se fizičke zakonitosti opisuju tzv. jednačinama kvalitativne fizike. Postoje dva glavna pravca razvoja kvalitativne fizike. Prvi pravac se odnosi na definisanje jednačina kvalitativne fizike iz diferencijalnih jednačina, dok se drugi pravac odnosi na određivanje ponašanja sistema.Izvedene jednačine i pravila mogu zatim biti upotrebljeni kao baza znanja za potrebe detekcije kvarova u sistemima.

Detekcija anomalija u podacima na bazi kvalitativnih modela može se primeniti u hidrotehnici kada se iz nekog razloga ne mogu uspostaviti matematičke relacije između veličina. To je obično slučaj kada postoji samo načelno znanje o pojavama koje merene veličine reprezentuju. Matematička interpretacija relacija između veličina uglavnom, ako ne i uvek, sadrži u sebi i informaciju o odnosu veličina.

2.2.3.3 Kvalitativni sistemski (Data Driven Models, DDM) modeli

Kod primene kvalitativnih sistemskih DDM modela dva najvažnija pristupa su ekspertski sistemi i kvalitativna analiza trenda podataka. Ekspertski sistemi predstavljaju sisteme pravila koja se koriste za rešavanje nekog dobro definisanog problema. Prvi pokušaji da se primeni ekspertski sistem u detekciji kvarova može se naći u [49] i [83]. Ekspertski sistemi su ostali izuzetno popularni u domenima kod kojih se ne poznaju pravila funkcionisanja sistema. Tako se u [92] daju se primeri primene ekspertskih sistema u hemijskom inženjerstvu. Kombinovanje ekspertskih sistema sa drugim metodama za modeliranje može poboljšati rezultate detekcije kvarova. Na primer, u [9] razmatra se kombinacijaveštačkih neuralnih mreža (Artificial Neural Network, ANN) i ekspertskih sistema, u [110] se kombinuju ekspertski sistemi sa OODG i rasplinutim skupovima, a u [136] se uvodi kombinacija ekspertskih sistema, ANN i *wavelet* transformacije podataka.

Analiza trendova predstavlja izuzetno važnu grupu metoda koja se koristi kako za detekciju kvarova, tako i za predviđanje stanja sistema, osmatranje i kontrolu procesa. U [21] se definiše platforma za

predstavljanje procesa trednovima. U [58] su definisani tipovi trendova. U [121] se predlaže identifikacija tipova trendova pomoću *wavelet* transformacije vremenskih serija, dok se u [122] za istu svrhu razvija primena *spline* tehnika.

Metode bazirane na kvalitativnim DDM modelima se, kao i one bazirane na kvantitativnim, oslanjaju se isključivo na istorijske podatke. Pošto je redak slučaj da se kod hidrotehničkih procesa ne poznaju granični uslovi i pravila funkcionisanja sistema, kao nedostatak navedeih metoda može se uzeti upravo to što se njihovo postojanje zanemaruje.

2.2.3.4 Kvantitativni sistemski (DDM) modeli

Kvantitativni sistemski modeli grubo se mogu klasifikovati u statističke i nestatističke. Na primer, *principal component analysis* (PCA) i višestruka regresija (*partial least squares*, PLS) ubrajaju se u statističke metode, dok veštačke neuralne mreže (ANN) predstavljaju primer nestatističkih metoda.

U grupi metoda koje se mogu koristiti kada su podaci predstavljeni u više dimenzija sigurno je najuspešnija PCA metoda. PCA metoda je detaljno opisana u [69]. Osnovna dva nedostatka ove tehike su što ne uzima u obzir vreme (*time invariant*) i što se odnosi samo na linearne veze između podataka. Unapređenja PCA metode se uglavnom odnose na rešavanje ova dva problema. Prvi problem se rešava uvođenjem pokretnog prozora, dok se drugi problem rešava ili primenom PCA metode parcijalno nad delovima podataka [84] ili se uvodi nelinearni model u formi neuralne mreže [101]. Metode bazirane na PCA se stalno razvijaju jer je njihova primena elegantna, a izvršenje metoda na bazi PCA je uglavnom izuzetno brzo. Jedno od unapređenja predstavlja i metoda koja kombinuje PCA i *wavelet* analizu podataka [7].

Mnogi autori prikazali su mogućnosti primene veštačkih neuralnih mreža. Istraživanja na primeni ANN mogu se podeliti u dva pravca: razvoj arhitekture ANN i razvoj strategije učenja (treninga) koje može biti nadgledano ili nenadgledano (*supervised and unsupervised*). Najpopularniji vid učenja ANN je svakako *back-propagation* algoritam. Detaljno objašnjenje i način upotrebe ovog tipa ANN je opisano u članku [125]. U [73], u kome se predlaže upotreba *radial basis* funkcije u ANN za potrebe detekcije kvarova.

Različite arhitekture ANN su predlagane za rešavanje problema detekcije kvarova. Na primer, u [6] predlaže se upotreba tzv. *wavenet* ANN – određene kombinacije *wavelet* transformacije i *back-propagation* ANN. Takođe, samoorganizujuće ANN, poznatije kao samoorganizujuće mape (*Self Organizing Maps*, SOM) [68] uspešno su primenjene na problemu detekcije kvarova.

Primena PCA metode i metode bazirane na ANN testirana je na podacima hidrotehničkih veličina merenih na ispustima Beogradske kanalizacije [11]. Dobijeni rezultati upoređeni su sa rezultatima dobijenim sličnim metodama u kojima su figurisale fizičke relacije između podataka. Fizičke relacije između podataka, pored toga što unose i dodatne, redudantne informacije u korišćene metode, omogućavaju i primenu u situacijama ekstrapolacije, na primer, pri oticajima koji nisu zabeleženi u istorijskim podacima korišćenim za kalibraciju matematičkog modela.

2.2.4 Detekcija izuzetaka (*outlier detection*)

Dok su istraživanja u oblasti detekcije kvarova (Poglavlje 2.2.3) orijentisana isključivo na sisteme koji mogu da se pokvare (tehnički sistemi, a ponekad i prirodni sistemi usled kvara mernog uređaja) i pokrivaju uglavnom neregularne pojave u samom sistemu, detekcija izuzetaka predstavlja širi pojam i odnosi se na sve podatke koji nisu u skladu sa predefinisanim šablonom u kom se podaci pojavljuju ili šablonom u kom se javlja većina podataka.

Definicija izuzetka može se naći u [8]: Podatak koji nije u skladu sa ostalim merenim podacima sa kojima se upoređuje[2].

Navedena definicija izuzetno je široka, pa se zbog toga može primeniti u mnogim disciplinama koje zbog svoje složenosti ili specifičnosti ne mogu biti opisane kao sistemi sa ulazima, izlazima i stanjima. Neki od primera takvih disciplina su otkrivanje prevara kreditnim karticama, detekcija upada internet mrežom, analiza satelitskih snimaka i otkrivanje novih detalja, detekcija unosa podataka u baze, itd.

Postoje tri pristupa problemu detekcije izuzetaka:

1. Određivanje izuzetaka bez prethodnog znanja o podacima. Ovaj problem se može definisati preko problema nenadgledanog grupisanja podataka. U [98] se ovaj pristup deli na dve podgrupe. Kod prve, kada se detektuje izuzetak, on se odbaci, dok se kod druge on uključuje u model smanjujući osetljivost same metode koja se koristi.
2. Modeliranje i regularnog i neregularnog stanja podataka. Ovaj pristup odgovara nadgledanoj klasifikaciji podataka i zahteva podatke za učenje modela sa predefinisanim oznakama o tome da li su regularni ili ne. Da bi se postigle dobre performanse algoritma za klasifikaciju potrebno je da regularni i neregularni podaci budu jednako raspoređeni po prostoru.
3. Modeliranje samo regularnog stanja podataka, sa mogućnošću da se neregularno stanje prepozna. Na ovaj način se mogu otkriti i novi podaci koji bi se uključili u novi model [59], pa autori ovaj vid detekcije izuzetaka nazivaju još i detekcija novog (*novelty detection*). Ovaj pravac se koristi pre svega kada je lako modelirati granice regularnosti podataka.

Veliki broj metoda za detekciju izuzetaka se izvode iz ili se oslanjaju na statističke metode, ANN metode i metode mašinskog učenja (*machine learning*, ML). Stoga se metode za detekciju izuzetaka, prema vrsti algoritma koji koriste, mogu klasifikovati na:

1. metode bazirane na razdaljini između podataka;
2. metode bazirane na statističkim zavisnostima;
3. metode bazirane na algoritmima klasifikacije;
4. metode bazirane na algoritmima grupisanja;
5. metode bazirane na algoritmima mašinskog učenja.

2.2.4.1 Metode bazirane na razdaljini između podataka

Metode bazirane na razdaljini između podataka predstavljaju tehnike koje je lako implementirati i koje ne zahtevaju nikakvo prethodno znanje o rasporedu podataka u prostoru. Ipak, velika mana im je to što se kod podataka sa više dimenzija eksponencijalno povećava broj operacija koje su potrebne za izračunavanje rastojanja između podataka (često svakog od svakog). Broj operacija je proporcionalan i sa dimenzijom podataka i sa brojem podataka koji se ispituju. Jedna od najpoznatijih tehnika ovog tipa je svakako „k-najbližih suseda" (*k-nearest neighbour*, k-NN). Kod ove tehnike se rastojanje računa najčešće kao Euklidovo, ali i kao Mahalanobis rastojanje koje uzima u obzir i unutrašnji raspored podataka preko matrice kovarijansi. Za podatke sa velikim brojem dimenzija računanje Mahalanobis rastojanja je izuzetno zahtevno jer je potrebno odrediti matricu kovarijansi za ceo razmatrani skup podataka.

[2] An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.

Na bazi k-NN tehnike moguće je sprovesti detekciju anomalija na različite načine. Rezultat same tehnike je mera razdaljine između k najbližih suseda razmatranog podatka. Mera može biti broj najbližih suseda koji nisu dalje od nekog predefinisanog graničnog rastojanja d [67], ili ukoliko je ulazna veličina broj suseda, mera može biti najveće ili kumulativno rastojanje razmatranog podatka od njih.

S obzirom na to da je k-NN izuzetno popularna tehnika, u literature se može sresti niz modifikacija kojima se unose poboljšanja u osnovnu verziju. Mana osnovne metode je što je potrebno izračunati udaljenosti svake vrednosti u razmatranom skupu od ostalih vrednosti. U [94] predstavljen je optimizovan k-NN algoritam koji kao rezultat daje rangirane potencijalne izuzetke. Po ovom algoritmu, vrednost p je izuzetak ako manje od $n-1$ drugih vrednosti u skupu posmatranih veličina ima veću udaljenost od D_m (udaljenost od m-tog suseda), pri čemu je m ulazni parametar metode. U [94] je takođe predložena tehnika za ubrzavanje ovog algoritma, tako što se prostor na kome se nalaze razmatrane vrednosti podeli na ćelije. Ukoliko u jednoj ćeliji postoji više od k vrednosti pretpostavlja se da je raspored vrednosti u razmatranom regionu zadovoljavajuće gustine, dok ukoliko to nije ispunjeno, cela ćelija se proglašava autlajerom. Autori su takođe predložili način indeksiranja podataka tako da se potrebno vreme i računarska snaga još više smanje.

U [67] predloženo je poboljšanje osnovnog algoritma na sledeći način: ukoliko m od k najbližih suseda (za $m < k$) leži na udaljenosti manjoj od neke granične, podatak se klasifikuje kao regularan. Ukoliko to nije ispunjeno, podatak se označava kao izuzetak. Slična logika je primenjena u [18], gde je algoritam primenjen na problemu pronalaženja mina pomoću satelitskih fotografija terena.

Ukoliko postoji skup podataka za formiranje algoritma sa oznakama da li su vrednosti razmatranog skupa podataka regularni podaci ili izuzeci, može se primeniti postupak prikazan u [129]. U njemu se nova vrednost klasifikuje u jednu od dve kategorije podataka (regularnih i izuzetaka) po sistemu udaljenosti od najbližih suseda. Najveća mana ovog algoritma je što se, da bi dao dobre rezultate, zahteva da podaci obe klase približno podjednako pokrivaju prostor u kom se nalaze. Ovaj nedostatak efikasno je prevaziđen algoritmom prikazanom u [109], gde se kao mera udaljenosti ne pretpostavlja prost zbir udaljenosti pojedinačnih vrednosti kao u [129], već se računaju težinski osrednjena kumulativna rastojanja između razmatrane tačke i k najbližih suseda.

Veliki problem kod algoritama koji su bazirani na udaljenosti podataka jednih od drugih predstavlja i količina razmatranih podataka. U [25] predlaže se formiranje tzv. prototipova podataka koji bi predstavljali podatke. Klasifikacijom podataka u predefinisane prototipove dolazi se do izuzetaka na osnovu podataka koji nisu klasifikovani. Jedan od najpoznatijih metoda u kojima se koristi ova logika je *k-means* algoritam za klasifikaciju, gde se grupa podataka predstavlja srednjom vrednosti jedne klase podataka. U [3] ovaj postupak je upotrebljen za automatsko otkrivanje novih vesti u elektronskim novinama preko klasifikacije naslova. Grupisanjem sličnih vesti smanjen je broj upoređivanja (rastojanja) između novih vesti (naslova) i ranije postojećih, tako što se nove vesti upoređuju isključivo sa predstavnikom cele grupe (srednjom vrednosti). Ukoliko nije moguće svrstati novu vest u već postojeću grupu (granično rastojanje se zadaje kao ulazna veličina) u pitanju je nova vest.

U [103] prikazan je jedan od algoritama baziranih na rastojanju između podataka, u kom se rastojanje računa isključivo preko veza između čvorova grafa (rastojanje preko povezanih čvorova). Na taj način je analizirano postojanje izuzetka kod merenja intenziteta saobraćaja u određenim tačkama saobraćajne mreže. Podaci su označeni kao izuzeci ukoliko se vrednost izmerena na nekoj mernoj stanici razlikovala od srednje vrednosti susednih mernih stanica više nego što iznosi granična

vrednost. Granična vrednost je određena kao procenat srednje vrednosti razlika svih čvorova u kojima se meri intenzitet saobraćaja i njihovih topoloških suseda. Važno je napomenuti da se susedi određuju na osnovu grafa povezanosti, pa broj suseda varira od čvora do čvora (nije potrebno zadati broj suseda *k*).

Već je napomenuto da je najveća mana opisanih metoda povećanje kompleksnosti zbog povećanja dimenzionalnosti podataka i zbog povećanja broja podataka u razmatranom skupu. Ovaj problem se može prevazići tako što bi se na osnovu raspoloživih podataka formirali modeli podataka čija bi složenost ostala ista sa pojavom novih dimenzija ili novih podataka. Veliki nedostatak ovakvog pristupa je pretpostavka o vrsti modela koji bi se primenio da opiše podatke. Ukoliko je on poznat i odgovara podacima, može se očekivati visoka tačnost metoda ovog tipa, ali ne može se, nažalost, očekivati da postoji odgovarajući model za sve skupove podataka.

Dve slične metode koje pripadaju ovoj grupi predstavljene su u [98] i [113]. U njima se formira model koji odgovara minimalnom elipsoidu koji se može opisati oko većine podataka (koliko iznosi većina podataka je ulazni parametar). Model se može formirati na dva načina: predefinisanjem procenta podataka koji se smatra regularnim, ili iterativnim odbacivanjem podataka koji su najudaljeniji od granica modela i formiranjem novih modela dok se ne zadovolje pretpostavljeni kriterijumi (obično broj podataka koji se smatra izuzetkom).

Pored ovakve vrste modela, moguće je formirati i statističke modele.

2.2.4.2 Statističke metode

Kod statističkih metoda primenjuje se sledeća pretpostavka: regularni podaci se javljaju u regionima sa visokom verovatnoćom na koju ukazuju statistički modeli, dok se anomalije javljaju u regionima sa niskom verovatnoćom [19].

Statističke metode se mogu podeliti u dve grupe: metode za koje se pretpostavlja način na koji su podaci raspoređeni (parametarske metode) i metode kod kojih se ne pretpostavlja raspodela po kojoj se podaci javljaju (neparametarske metode).

Jedan od najjednostavnijih testova koji je moguće primeniti i na višedimenzionalne podatke opisan je u [70]. U njemu se formira tzv. *box* dijagram sa pet karakterističnih vrednosti razmatranog niza: najmanja vrednost, donja četvrtina, sredina, gornja četvrtina i najveća vrednost. Detekcija se obavlja na osnovu vizuelne inspekcije opisanih dijagrama. Za višedimenzionalne podatke kao meru rastojanja između podataka autori predlažu korišćenje Mahalanobis rastojanja.

Parametarske statističke metode

Kod parametarskih metoda se pretpostavlja statistička raspodela kao osnova formiranja statističkog modela. Na osnovu vrste raspodele, oni se dalje mogu podeliti na [19]: 1) modele *Gauss*-ove raspodele, 2) regresione modele i i 3) modele mešavine raspodela.

Jedan od prvih testova koje koristi model *Gauss*-ove raspodele predložen je još davne 1931. u [104]. U njemu se anomalijom smatra podatak koji ima verovatnoću van intervala $[\mu-3\sigma, \mu+3\sigma]$. U ovaj interval ulazi 99.7% podataka *Gauss*-ove raspodele, pa je analogija sa metodom detekcije jasna.

Još jedan od postupaka koji pretpostavlja model u obliku *Gauss*-ove raspodele (kod malog broja podataka, $N<30$, Studentove *t* raspodele) je tzv. *Grubbs*-ov metod [43] kod kog se računa *z* statistika (razlika razmatrane veličine u nizu i srednje vrednosti niza, podeljenih sa standardnom devijacijom), koja se upoređuje sa određenim pragom značajnosti Studentove *t* raspodele. Ovaj jednostavan test,

koji ne zahteva parametre, može se koristiti isključivo jednom na jednom skupu podataka. [96] je unapredio ovaj metod tako da se može koristiti više puta nad istom serijom podataka. Za primenu na višedimenzionalne podatke, u [70] se predlaže statistika u obliku Mahalanobis razlike između razmatranog podatka i srednje vrednosti čitavog skupa podataka, nad kojom se dalje primenjuje *Grubbs*-ov test.

Ukoliko se raspolaže malim brojem podataka, predlaže se upotreba Studentove *t* raspodele umesto *Gauss*-ove (širi repovi raspodele daju veću verovatnoću podacima udaljenijim od srednje vrednosti). Među popularnim testovima za detekciju anomalija su *t* test i, za višedimenzionalne podatke, *Hotteling*-ov t^2 test [74].

U mnogim radovima koristi se model linearne regresije. Problem kod ovog tipa modela je što se uticaj podataka sa anomalijama prenosi i na sam fitovan model. Odbacivanjem podataka koji najviše odstupaju od formiranog modela, do stabilizovanja parametara modela (potrebno je definisati graničnu vrednost), iterativno se mogu detektovati izuzeci. Modifikacija metode predstavljene u [113] je formiranje linearne regresione krive sa kriterijumom da se minimizuje medijana razlika vrednosti i rezultata modela [113], čime se smanjuje broj računskih operacija u procesu formiranja modela.

Modeli podataka mogu biti lokalnog karaktera, tj. mogu se odnositi samo na određenu grupu podataka umesto na ceo razmatrani skup. U [95] je metoda mešavine Gausovih raspodela *Gaussian mixture models*) primenjena na niz EEG podataka kako bi se otkrili potpisi nekih bolesti (npr. epilepsije). Nakon formiranja modela (broj Gausovih raspodela je određen iterativno) granična vrednost pripadanja nekom od modela u obliku verovatnoće određuje da li je novi podatak izuzetak ili ne.

Neparametarske statističke metode

Najpoznatija neparametarska statistička metoda je metoda histograma. Ova metoda se najviše koristi u procesu detekcije hakerskih upada u računarske mreže zbog velike količine podataka koji pristižu u realnom vremenu [32]. Za potrebe formiranja histograma potrebni su regularni podaci (ne može se očekivati da postoji dovoljna količina podataka sa anomalijama, mada je bilo i pokušaja zasnovanih na takvim očekivanjima). Izbor širine štapića histograma je ključni detalj ove metode. Ukoliko su oni uski, javiće se velika količina lažnih alarma, dok bi u suprotnom slučaju neka anomalija mogla proći neopaženo.

Povećanje broja dimenzija podataka kod statističkih metoda uvodi kako problem zahtevnog procesiranja podataka, tako i problem rastojanja između podataka u prostoru (manja gustina podataka), čime se smanjuje osetljivost metoda za detekciju izuzetaka. Postoji nekoliko načina kako je moguće boriti se sa podacima sa velikim brojem dimenzija. Neki od njih su pretprocesiranje podataka [2] i smanjenje dimenzionalnosti [33].

Povećana složenost modela sa povećanjem dimenzionalnosti podataka koji se razmatraju može se prevazići upotrebom metode *principal component analysis* (PCA) [33], [86]. Nažalost, osnovnom PCA metodom ne mogu da se modeliraju nelinearne zavisnosti između podataka. Ovaj metod se može iskoristiti u pretprocesiranju bilo koje grupe podataka kod kojih je izražena linearna veza da bi se smanjila dimenzionalnost. U [33] predlaže se da se kao granična vrednost za izbor broja glavnih komponenti (*principal components*) uzme granica od 85% zbira svih sopstvenih vrednosti.

2.2.4.3 Klasifikacija

Veštačke neuralne mreže (*artificial neural networks*, ANN) predstavljaju algoritme koji se mogu prilagoditi potrebama regresije (opisano u Poglavlju 2.2.3.3) ili klasifikacije podataka koji se koriste za učenje (treniranje) mreže. Učenjem mreže se određuju parametri ANN čime se obezbeđuje da ANN za nove ulazne podatke pruži informaciju o klasi kojoj pripadaju. ANN model se može shvatiti kao fitovani nelinearni model podataka. Ukoliko su podaci u skupu za učenje ANN obeleženi kao regularni ili neregularni, moguće je primeniti neki od algoritama za nadgledano učenje. Ukoliko to nije mogućen, potrebno je primeniti neki od algoritama za nenadgledano učenje.

Razni autori imali su različite ideje kako da iskoriste moćan alat kao što je ANN za potrebe detekcije izuzetaka. Tako je, na primer, u [82] na ispitivanju vremenskih serija dobijenih merenjem vibracija kod aviona ANN upotrebljena da predvidi merenu vrednost iz n prethodnih vrednosti. Ukoliko se merena vrednost mnogo razlikuje od simulirane, u pitanju je izuzetak. U [59] koristi se auto-asocijativna ANN koja ima usko grlo na mestu jednog od skrivenih slojeva. Na taj način je omogućeno da se ulazni podaci preslikavaju u same sebe sa redukovanim redudantnim osobinama. Ovaj tip ANN se često koristi kada je potrebno primeniti postupak sličan PCA na nelinearno zavisnim podacima. Autor je upoređivao ulazne vrednosti i vrednosti koje su rezultati ANN i na osnovu njihove razlike tumačio da li je u pitanju outlajer ili regularan podatak.

Ponekad je potrebno podatke klasifikovati iako oni ne pokazuju nikakvu vrstu zavisnosti. Jedan od načina za to je primenom SVM (*support vector machines*) algoritma. Ovaj algoritam omogućava razdvajanje klasa podataka višedimenzionalnom ravni. U mnogim slučajevima to nije moguće direktno, pa je potrebno kernel funkcijama transformisati podatke u podatke sa više dimenzija od originalnih. Minimalno povećanje računske kompleksnosti postupka omogućeno je upotrebom skalarnog proizvoda i specifičnih kernel funkcija (sigmoidna, *Gauss*-ova, polinomna, itd.). U [111] je opisana primena SVM pri detekciji izuzetaka kod medicinskih podataka.

2.2.4.4 Grupisanje

Nenadgledane ANN se primenjuju kada na podacima u skupu za učenje mreže ne postoje oznake da li je u pitanju izuzetak ili regularan podatak. Postavljen problem je moguće definisati i preko problema grupisanja podataka, kod koga se podaci na osnovu karakterističnih osobina grupišu u određeni broj grupa. *Self-organizing maps* (SOM) [68] je najpoznatiji predstavnik nenadgledanih ANN. Osnovni zadatak SOM je da višedimenzionalne podatke mapira preko specifičnih težinskih faktora u nisko dimenzionalnu mapu čvorova (obično dvodimenzionalnu). Određivanje težinskih faktora sprovodi se učenjem SOM (iterativnim menjanjem težinskih koeficijenata do konvergencije). U primerima datim u [99] i [126] SOM je formirana iz regularnih podataka, a ukoliko je novi podatak (podatak koji se ispituje) udaljen od najbližeg čvora (*best matching unit*, BMU) više nego što je limitirano graničnom vrednosti, u pitanju je izuzetak. Sa druge strane, u [51] postupak određivanja outlajera se svodi na ispitivanje greške mapiranja:

$$g(x, m_i) = \frac{1}{1 + \left(\frac{\|x - m_i\|}{a} \right)^2},$$

gde je a srednja vrednost svih rastojanja između podataka za učenje SOM i odgovarajućih BMU.

2.2.4.5 Mašinsko učenje (*machine learning*)

Većina algoritama koji se mogu iskoristiti kod problema detekcije izuzetaka odnose se na brojčane podatke, dok se samo nekolicina može primeniti kod opisnih podataka. Ponekad je izuzetao složen zadatak razviti proceduru za izračunavanje sličnosti ili različitosti između podataka, a to predstavlja tek prvi korak u primeni metoda predstavljenih u prethodnim odeljcima. Jedan od načina kako se ovaj problem može prevazići je upotreba stabala odluke (*decision trees*). U [60] i [105] prikazani su primeri upotrebe C4.5 algoritma za otkrivanje grešaka i neočekivanih unosa u bazu podataka. Pouzdanost stabala odluke u velikoj meri zavisi od toga da li stabla obuhvataju sve podatake. Ukoliko su isuviše kompleksna, stabla odluke mogu dovesti do neželjenih efekata pri upotrebi, kao što je *over fitting* (rezultati modela u sebi sadrže i modelirane slučajne greške). Zbog toga ih je potrebno očisti od suvišnih detalja, kao što je pokazano u [60] i [105]. Ukoliko se želi fleksibilan sistem, u koji bi se mogla prema potrebi dodavati nova pravila, može se primeniti drugi algoritam mašinskog učenja, nazvan *rule-based system*. Ovakvi sistemi se sastoje od niza pravila koja služe za izvođenje nekog zaključka ili kao pomoć za donošenje odluke.

2.2.4.6 Hibridne procedure

Kod hibridnih procedura koristi se više od jedne metode za detektovanje izuzetaka u podacima. Hibridne procedure se formiraju zbog dva cilja: upotrebom jedne procedure eliminišu se slabosti druge i upotrebom više procedura na istom skupu podataka povećava se pouzdanost procedure.

U [82] se, pored ANN modela, koristi i *k-means* algoritam koji služi da se detektuju globalne razlike u potpisu vremenskih serija vibracija aviona, što ANN modelom nije omogućeno. Autori članka [50] upotrebili su niz kernel funkcija (*Gauss*-ovih raspodela) da bi modelirali ponašanje sistema mernih uređaja u avionu, dok su prethodno primenom PCA postupka smanjili broj dimenzija podataka na dva.

Poslednjih godina ispituje se mogućnost da se na istom problemu primeni više metoda za detekciju outlajera. S obzirom na to da svaki od postupaka ima svoje prednosti i slabosti, pretpostavlja se da bi se kombinacijom više metoda dobio novi kvalitet u rezultatima. Tako je nastao JAM sistem (*Java agents for meta-learning*) [108] u kom se koristi pet metoda: tri stabla odluke (ID3, CART i C4.5), jedan *rule-based* system (Ripper) i Bajesov klasifikator (*naive Bayes classifier*). U eksperimentima je pokazano da svaka od metoda ima određene specifičnosti i da pokazuje dobre rezultate na nekim grupama podatka, dok na nekim dominiraju mane algoritama, i da je uz dodatno tumačenje rezultata moguće proceniti i podeliti podatke tako da se dobiju optimalne performanse sistema.

U [17] je opisana primena tri klasifikaciona algoritma na problem detekcije upotrebe zemljišta sa satelitskih snimaka. Autori su koristili stablo odluke, *k-NN* i linearni model. Definitivna odluka se dalje iz rezultata tri pomenute metode donosi konsenzusom. Konsenzus je rigorozniji metod od, na primer, metode većine, jer je potrebno da sve tri metode proglase podatak izuzetkom, pa da se on tako i označi. Kod donošenje odluke većinom potrebno je da većina donese odluku da je podatak izuzetak (u ovom slučaju dva od tri).

Detekcija izuzetaka predstavlja disciplinu u razvoju u mnogim oblastima, a naročito u onim kod kojih ne postoje poznate relacije između podataka, već se one otkrivaju i formiraju na osnovu podataka koji su često i sami opterećeni greškama. U hidrotehnici često postoje relacije između podataka koje se ogledaju u fizičkim ili statističkim zavisnostima, pa se metode za otkrivanje izuzetaka primenjuju jedino u slučajevima kada je relacije potrebno otkriti i modelirati.

2.2.5 Kvalitet hidro-meteoroloških veličina

Hidro-meteorološke veličine kao što su nivo vode u vodotocima, brzina vode, brzina vetra, visina/intenzitet padavina ili temperatura vazduha obično su merene na specifičnim lokacijama i organizovane u vremenskim serijama. Potrebe za pouzdanim podacima navele su naučnike da osmisle metode za proveru kvaliteta podataka koje se uglavnom odnose na pojedine veličine i njihove karakteristike bez uvida u relacije sa merenim podacima drugih veličina. Iako se često proces provere kvaliteta podataka ne navodi kao primaran (nije cilj naučnog doprinosa rada), u radovima se primenjeni postupak provere kvaliteta podataka navodi kao neophodan za dobijanje kvalitetnih rezultata analiza koje su predmet rada. U daljem tekstu se navodi nekoliko primera.

2.2.5.1 Kiše

U [115] opisano je nekoliko jednostavnih procedura za proveru podataka o kišama merenim kišomerima sa klackalicom (*tipping-bucket raingauges*). Procedure se odnose na analizu intervala izmeĎu klackanja klackalice. Na osnovu nekoliko jednostavnih testova i predloženih graničnih vrednosti, moguće je proceniti da li je mernje regularno ili ne. Testovi su podeljeni u dve grupe, i to za pojedinačne kišomere i za mreže kišomera, a odnose se na detektovanje brzog, sporog i ubrzavajućeg klackanja klackalice.

Još jedna procedura za kontrolu kvaliteta merenih kiša zemaljskim kišomerima predstavljena je u [61]. Danski meteorološki institut (DMI) razvio je procedure za automatsku i ručnu proveru izmerenih intenziteta kiša koje se dalje koriste za modeliranje hidroloških pojava softverskim paketom MOUSE. Jednostavna automatska procedura predviĎa da se kao neregularan označi svaki podatak kojim je registrovano više od 2mm/min. Na taj način je sprečena pojavapodataka o nerealno velikim intenzitetima kiša. Ručna procedura se sastoji iz nekoliko koraka: 1) provera izmerenih vrednosti sa bliskim kišomerom, 2) konsultovanje sa kartama intenziteta kiša i provera putanja iznenadnih oluja, 3) provera hijetograma za koji se smatra da mora da ima uzlaznu i silaznu granu bez naglih skokova, i, ukoliko i dalje podatak deluje sumnjivo pre njegovog odbacivanja, 4) provera istorijskih serija.

U radu [42] predstavljen je projekat VOLTAIRE (*Validation of Multisensors Precipitation Fields and Numerical Modeling in Mediterranean Test Sites*). Srž ovog projekta predstavlja razvoj validacionih procedura za radarski merene intenzitete padavina. Kako je rezultat radarskog snimka slika, većina procedura se zasniva na pretraživanju piksela i registrovanju sumnjivih detalja. Neke od procedura su uklanjanje ehoa od objekata na zemlji (zgrade, stubovi, itd.), uklanjanje radijalnih anomalija, ublažavanje efekata usled slabog odbijanja radarskih talasa zbog jakih kiša, itd. Za potrebe ovog projekta uglavnom su korišćene metode koje se koriste u filtriranju radarskih signala, dok su metode koje bi koristile zemaljske merne stanice samo najavljene.

Kvalitet radarskih snimaka kiša koji se koriste kao ulaz u hidrološki model je tema rada [30]. Da bi se proverio kvalitet podataka sprovedeno je nekoliko procedura: 1) provera pojedinačnih piksela prikazanih u formi tabele u cilju otkrivanja grubih grešaka, 2) statistička provera intenziteta kiša registrovanih u toku nekoliko nedelja, 3) uporeĎivanje sa podacima registrovanim na zemaljakim stanicama o tome da li je bilo kiše ili ne, 4) vizuelna inspekcija radarskih snimaka, 5) uporeĎivanje radarskih snimaka sa vremenskim serijama dobijenim sa zemaljskih kišomera, 6) kontrola simuliranih protoka (dobijenih hidrološkim modelom) sa merenim protocima na mernim stanicama.

2.2.5.2 Hidraulički parametri (nivoi, brzine i protoci)

Jedan predlog vrednovanja merenih podataka dao je Bertrand Krajevski u [79]. On predlaže sedam koraka u procesu vrednovanja, koji se mogu prilagoditi i za automatski rad: 1) status senzora

(ON/OFF/HOLD), 2) fizičke granice merne veličine, 3) statistički realne granice merne veličine, 4) vreme od poslednjeg održavanja ili provere merila, 5) gradijent signala, 6) redudantno merenje, ako postoji, i 7) analitička redudantnost, ako postoji. Ocene kvaliteta podataka koje su diskretne: A – dobar kvalitet, B – sumnjiv podatak i C – loš kvalitet. Do krajnje ocene se kroz sedam predloženih testova dolazi odabiranjem najlošije (favorizovanjem).

Pjatišek, Janis i Omon u [90] predlažu jednostavan način kako se mogu proveriti podaci koji su prikupljeni merenjem hidrauličkih veličina u kanalizacionom sistemu kada nema kiše i kada je izražen karakteristični šablon tečenja u fekalnoj kanalizaciji. Oni su odredili šablone (uključujući neizvesnost u obliku intervala) protoka za vreme kada nema kiše uzimajući u obzir i nedeljnu i sezonsku neravnomernost. Veliki nedostatak ovakvog načina vrednovanja podataka je to što se kao anomalije registruju i podaci dobijeni za vreme kišnih epizoda.

Modeliranje kanalizacionog sistema za potrebe detektovanja grešaka pri merenju opisano je u radu [10]. Da bi se izbegao problem modela zasnovanog na fizičkim karakteristikama sistema (*Hydroworks*) koji je preglomazan, kanalizacioni sistem je podeljen na delove i svaki deo je modeliran linearnim sistemom definisanim ulazima, stanjima i izlazima. Problem nelinearnosti modela je rešen upotrebom rasplinutih skupova tako što su pojedine situacije modelirane posebno, da bi se rezultati sabrali preko vrednosti pripadnosti rasplinutom skupu.

U radu [89] detekcija ekstremnih protoka i neregularnog rada senzora u kanalizacionom sistemu zasniva se na korišćenju hidrološkog (kiša-oticaj) i hidrauličkog, bilansnog modela. Hidrološki model se koristi da bi se izračunao dotok u svaki čvor mreže, dok se hidraulički model koristi da bi se izračunao protok u svakom čvoru i u svakom trenutku. Pošto su korišćene izuzetno jednostavne relacije (bazirane na linearnim rezervoarima), Kalmanov filter je upotrebljen da se model koriguje merenjima. Za potrebe detekcije neregularnosti u radu senzora upotrebljen je test opisan u [127]. Kao promenljiva na osnovu koje se odlučivalo o regularnosti merenja koristio se odnos verodostojnosti (*likelihood*) da se komponenta Kalmanovog filtera (inovacija) javlja u obliku Gausove raspodele sa srednjom vrednosti jednakom nuli.

U radu [47] koriste se modeli formirani regresionom analizom nad istorijskim vremenskim serijama merenih kiša i protoka u kanalizacionom sistemu u gradu Nantu, Francuska. Utvrđeno je da se bolja slaganja postižu ukoliko se interval semplovanja podataka poveća na nekoliko sati, pa i na ceo dan. Na taj način se umanjuje nelinearnost veza između podataka i postiže bolje uklapanje u formirani linearni model.

Kinosita u svom radu [66] daje neka uputstva o proveri i korekciji podataka. Uputstva se odnose na upotrebu statističkih alata (statističkih raspodela, linearne regresije, itd.), ali i hidrauličkih i hidroloških relacija (najviše vezanih za QH krivu i transformaciju poplavnog talasa, i zaprimenu Meningove formule).

2.2.5.3 Temperatura

U [27] opisan je metod otkrivanja anomalija u vremenskim serijama temperature vazduha. Razvijen je dinamički stohastički model na principima Markovog lanca, kojim se modeliraju temperature na osnovu temperature izmerene prethodnog dana, dana u godini i vremena u toku dana. Analogija sa hidrotehničkim veličinama se može uočiti kod tečenja u kanalizacionom sistemu u vreme kada nema padavina. Greške, koje su u [27] podeljene na grube, srednje i sofisticirane, moguće je odrediti jedino kada ne postoji značajno odstupanje od predefinisanog šablona podataka (što se može dogoditi, na primer, za vreme oluja i naglog pada temperatura).

2.3 Softverska rešenja

Da bi se vrednovanje podataka učinilo što efikasnijim, razvijana su softverska rešenja bazirana na različitim konceptima. Neki su orijentisani isključivo ka ručnoj validaciji, dok neki imaju i automatsku opciju, kojom je moguće delove vremenskih serija proveriti. Zajedničko za sva rešenja je to da su bazirana na ekspertskom znanju i da se aplikacije koriste samo kao alat.

Implementacija sistema za validaciju predstavlja poseban izazov jer se zahteva da automatska metoda postigne, ako ne i prestigne, efikasnost eksperta koji bi je obavio van realnog vremena [107]. Interesantan način implementacije u svom radu predstavili su Koneho i saradnici. [23]. U njihovom radu su razmatranu fizičku veličinu povezali sa raspoloživim znanjima i informacijama preko tzv. saznajnih blokova (*cognitive units*). Saznajni blokovi (merni senzori, podslivovi, simulacije, blokovi za povezivanje) povezani su u strukturu koja simulira putanju vode po slivu. U svakom bloku se izvršava neki od zadataka. Na primer, u bloku koji se odnosi na senzore proveravaju se osnovne relacije validnosti izmerenog podatka (vreme pristizanja podatka, minimalna/maksimalna vrednost, itd.). Iako je u radu spomenuta neodreĎenost podataka koji se vrednuju, u sistemu se pretpostavljaju egzaktne vrednosti.

U radu [34] prikazana je implementacija *parity space* metode na vremenskim serijama merenih hidrotehničkih veličina. Softversko rešenje obuhvata grafički interfejs, grafički prikaz vremenskih serija pomoću dijagrama, modul za formiranje modela, i modul za ručno prepravljanje rezultata automatske primene. Ova moćna metoda pretpostavlja linearni model sistema i transformaciju matrice transformacije linearnog modela tako da se veličine prevedu u prostor reziduala, po mogućnosti tako orijentisanih da greška pri merenju jedne veličine bude drugačije orijentisana u prostoru od grešaka u drugim.

Pored ideje i matematičko-logičke podrške, podjednako važan aspekt kod projektovanja sistema za vrednovanje merenih podataka je mogućnost njegove implementacije u sistem merenja, transfera i prikupljanja podataka. Bez mogućnosti za efikasnu implementaciju, procedura razvijena za potrebe vrednovanja podataka je osuĎena na ograničenu primenu sa jednokratno odreĎenim parametrima. Različite implementacije se mogu sresti u literaturi. Baza podataka informacionog centra Indije (*Environmental Information Center*, EIC) [134] predviĎa proveru kvaliteta podataka različitim validacionim testovima u nekoliko nivoa, i to:

- *Nivo 0*: Generalna provera formata, vremena semplovanja, jedinica, intervala u kom se podatak nalazi, itd;
- *Nivo 1*: Proverava se jedan podatak u pogledu nastanka i transformacije;
- *Nivo 2*: Provera podatka kroz vreme i prostor – analiza trenda, korelacije, itd;
- *Nivo 3*: *rule--based* analiza podataka koji su u relaciji sa podatkom koji se proverava; i
- *Nivo 4*: Provera kvaliteta podataka na osnovu statističkih zavisnosti, korelacija, itd.

Nakon testiranja podataka, u bazu podataka se upisuju kako ocene testova, tako i kvalitet podataka na osnovu rezultata testova.

Efikasan sistem za obradu velikog broja podataka dobijenih merenjem različitih mernih veličina u nuklearnoj fizici predstavljan je u [85]. Podaci se sakupljaju, nad podacima se obavljaju jednostavne

i brze analize, a rezultati se što brže vraćaju korisniku. Sistem je dizajniran prema Observer šemi programiranja (*Observer pattern*).

Dizajn ekspertskog sistema, predstavljenog u [80], pretpostavlja mogućnost da se ekspertsko znanje akumulira postepeno. Na taj način je omogućeno da se sistem uči uz pomoć novih informacija koje je prikupio u toku rada.

Baza podataka sa ugrađenim alatima za određivanje kvaliteta hidroloških podataka WISKI [100] predstavlja kompleksan sistem koji je, pre svega, orijenstisan ka organizaciji, čuvanju i dostupnosti podataka (uključujući i vizuelizaciju), dok se alati za određivanje kvaliteta mogu sažeti u četiri nivoa: 1) nivo samih podataka, 2) nivo transfera podataka, 3) nivo celog sistema, i 4) IT nivo. Provera ograničenja i provera gradijenta samo su neke od metoda koje su implementirane u softver. Neke metode se mogu pokrenuti da rade automatski, dok se neke metode moraju pokrenuti uz asistenciju eksperta.

Ukoliko se vrednovanje podataka sprovodi van realnog vremena, uz mogućnost upotrebe alata za vizuelizaciju podataka, radna platforma WISKI baze podataka predstavlja komercijalni produkt za održavanje, vrednovanje i popravku kvaliteta podataka hidroloških baza podataka. Sistem je opremljen brojnim alatima za proveru podataka preko hidroloških i hidrauličkih zavisnosti (statističke zavisnosti, krive protoka, itd.), kao i sistema za uklanjanje šuma i grešaka.

Tabela 2.1: Ocena kvaliteta podatka po [64] i [100]

Nivo kvaliteta	Ocena kvaliteta	Skraćenica	Opis
Primarni	Dobar	G	Oznake se dodeljuju podacima automatski ili ručno.
	Nedostajući	M	
	Neproveren	U	
	Izračunat	E	
	Sumnjiv	S	
	Itd.		Postoji još oznaka na raspolaganju.
Sekundarni	Kompletan	C	Oznake se dodeljuju izračunatim podacima automatski ili ručno.
	Nekompletan	I	
	Nedostajući	M	
	Izmenjen	Ed	Oznake se dodeljuju izmenjenim podacima automatski ili ručno.
	U intervalu	WR	Oznake se dodeljuju i izmerenim i izračunatim podacima automatski ili ručno.
	Iznad gornje granice	BL>	
	Ispod donje granice	BL<	
	Bez oznake	NR	
	Itd.		Postoji još oznaka na raspolaganju.
Tercijarni	Sneg	d	Oznake kao dodatne informacije za padavine.
	Poreklo	t	
	Podeljen	a	Automatski dodata oznaka.
	Itd.		Postoji još oznaka na raspolaganju.

Na WISKI bazu podataka se nadovezuje i softversko rešenje za kontrolu kvaliteta podataka NIKLAS [72]. U ovom softverskom rešenju se primenjuje hijerarhijska procedura testiranja podataka, od jednostavnih ka komplikovanim testovima, predložena od strane [119] i potvrĐena od strane [81]. Testovi koji se nad podacima primenjuju u ovom softverskom rešenju su:

1. detekcija nedostajućih podataka;
2. detekcija fizički nemogućih vrednosti;
3. detekcija konstantnih vrednosti (ravnih linija);
4. detekcija podataka van granica odreĐenih graničnom vrednosti (*threshold*);
5. detekcija nula;
6. detekcija podataka koji imaju neobično veliku vrednost (iako oni mogu biti i regularni);
7. detekcija podataka koji imaju neobično nisku vrednost (iako oni mogu biti i regularni).

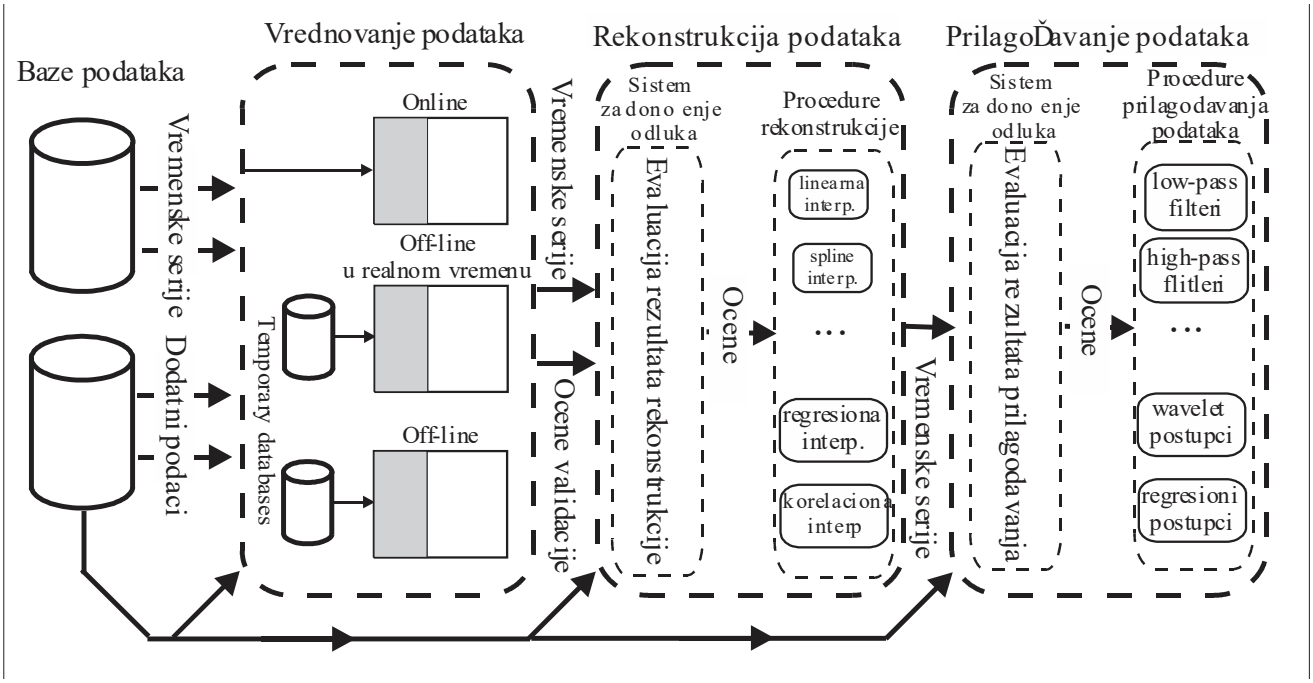
Vidi se da različite potrebe vode ka različitom dizajnu sistema za validaciju podataka. Sistem za validaciju hidrotehničkih podataka takoĐe je voĐen odreĐenim zahtevima, ali i ograničenjima, koja su postavljena da bi se implementacija učinila robusnijom.

———————— · ————————

Iako je problematika kvaliteta podataka prepoznata u nauci i praksi pre više decenija, ne postoji univerzalan pristup ovoj temi. Dosadašnji napori, čiji je kratak prikaz naveden u ovom poglavlju, bili su usmereni isključivo ka rešavanju specifičnih problema. Rešenja koja se odnose na specifične probleme vrednovanja podataka su fokusirana na ograničene informacije koje se mogu dobiti iz vrednovanih podataka, kao što su rastojanja izmeĐu podataka, statističke zavisnosti ili neki oblici relacija (npr. linearni modeli). Cilj ove doktorske teze je upravo da se prikaže razvijena metodologija kojom je moguće obuhvatiti sve informacije o vrednovanom podatku, a pogotovu informacije vezane za njegovu vezu sa ostalim merenim podacima u fizičkom sistemu.

3. Problematika vrednovanja podataka

Od svog nastanaka do upotrebe hidrotehnički podatak treba da prođe niz transformacija koje su pod pojmom „priprema podataka" ilustrovane na slici 3.1. Tri koraka, koja u ovom procesu slede jedan drugi, jasno su odvojena: vrednovanje podataka, rekonstrukcija podataka i prilagođavanje podataka potrebama korisnika. Operater koji upravlja podacima i komunicira sa bazama podataka prenosi ih iz modula za vrednovanje u modul za rekonstrukciju, a zatim u modul za prilagođavanje. On je važan deo sistema i na slici 3.1 predstavljen je strelicama.



Slika 3.1: Priprema podataka

Vrednovanje podataka je prvi korak u procesu poboljšanja kvaliteta podataka. Da bi se napravila efikasna rekonstrukcija podataka i omogućilo podešavanje podataka, odnosno prilagođavanje podataka korisniku, modul za validaciju (slika 3.1) mora da pruži izvesne informacije o kvalitetu i pouzdanosti svakog izmerenog podatka. Ocena vrednovanja, koja je binarna (nule i jedinice), kontinualna (0-100%) ili opisna (dobro, neizvesno ili loše), rezultat je rada modula za vrednovanje. Originalne serije podataka i rezultati (ocene) vrednovanja podataka, sačuvani kao meta-podaci, jesu ulazne veličine za modul za rekonstrukciju podataka.

U modulu za rekonstrukciju se podaci koji nedostaju i podaci sa niskim ocenama kvaliteta zamenjuju pouzdanijim podacima, ukoliko je to moguće. Različite tehnike, kao što su interpolacija podataka, upotreba redundantnih podataka, itd, mogu se koristiti u ručnom, polu-automatskom ili automatskom režimu rada.

Nakon rekonstrukcije podataka modul za podešavanje podataka mora da prilagodi i transformiše podatke kako bi bili pogodniji za određenu, konkretnu upotrebu. Ponovno uzorkovanje podataka, njihovo filtriranje i statistički proračuni samo su neke od tehnika koje se, ukoliko je potrebno, mogu upotrebiti u ovom modulu. Forma ovog trećeg radnog modula zavisi od potreba i specifičnosti svake aplikacije. U jednoj primeni se, na primer, iz baze podataka o protocima u kanalizacionom sistemu traže srednji dnevni protoci, dok je u drugoj primeni interesantan, na primer, petominutni protok tokom noćnih sati.

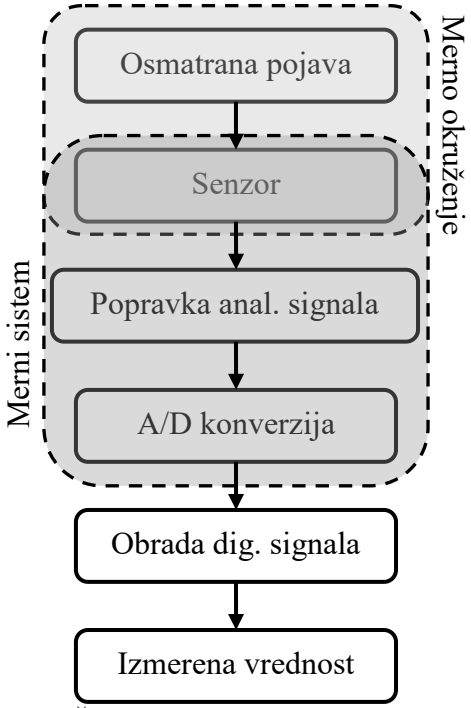
Kada se završi rad prva dva modula podaci se mogu preneti u skladište podataka zajedno sa meta-podacima (cela istorija) koji sadrže transformaciju koju su doživeli i stepen kvaliteta koji će budućem korisniku ukazati na pouzdanost podataka. Treći modul se pokreće za svaku aplikaciju ili korisnika posebno, po potrebi. U zavisnosti od potreba, ovaj modul se može dopunjavati novim funkcijama.

Tema ove doktorske disertacije je upravo razvoj i implementacija prvog modula sa slike 3.1, modula za vrednovanje podataka, čiji se rezultati dalje mogu upotrebiti pri rekonstrukciji i prilagoĐavanju podataka upotrebi. Vrednovanje hidrotehničkih podataka može se opisati kao: procedura kojom je moguće proceniti kvalitet podatka na osnovu greške merenja i načina upotrebe, tj. informacija koje iz podataka slede, uzimajući u obzir sva raspoloživa sredstva

Vrednovanje podataka predstavlja disciplinu koja je blisko vezana za mnoge delove hidrotehničke struke. Teorija merenja hidrotehničkih veličina i modeliranje hidrotehničkih procesa dve su glavne grane hidrotehnike na kojima se zasniva algoritam za vrednovanje podataka predložen u ovoj disertaciji. U ovom poglavlju se sumiraju neki od važnih detalja hidrotehničke teorije i prakse koji imaju za cilj da uvedu čitaoca u problematiku vrednovanja podataka i moguće načine da se neka uočena uska grla i problemi prevaziĐu.

3.1 Merenje hidrotehničkih veličina

Merenje predstavlja proces upoređivanja neke veličine sa jedinicom mere te iste veličine. Rezultat procesa merenja je odnos merene veličine i jedinice mere koji se obično izražava u kombinaciji sa korišćenom jedinicom mere. Proces merenja hidrotehničkih veličina mernim senzorima može se opisati jednostavnim algoritmom prikazanim na slici 3.2.



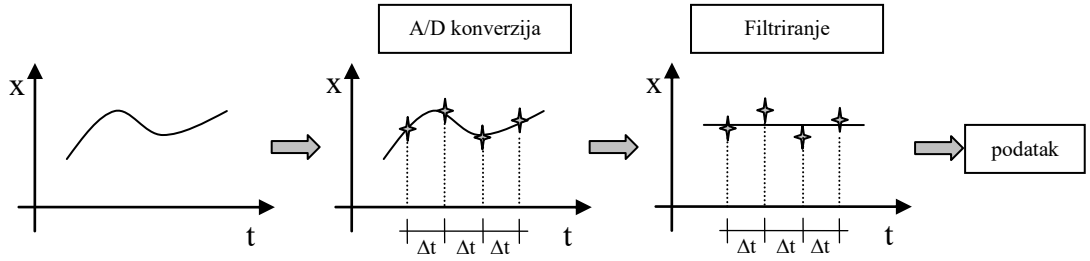
Slika 3.2: Šematski prikaz sistema za merenje

Iz okruženja se putem senzora registruju odreĐeni fenomeni za koje je senzor dizajniran i pretvaraju u izlaznu veličinu. Izlazna veličina senzora može biti pomeraj nekog mehaničkog sistema (kazaljka na manometru, pisač po mernoj traci, itd.), električni signal (napon, jačina struje ili frekvencija), itd. Da bi se obezbedio nesmetan rad mernog uređaja obično se merno mesto odabere i pripremi tako da

se spreči uticaj neželjenih efekata okoline na merenje. Ne sme se zanemariti činjenica da je uticaj senzora i okruženja obostran, tj. da senzor ponekad menja merno okruženje u kom se nalazi (npr. uronjen senzor brzine menja strujnu sliku). U procesu tumačenja rezultata merenja potrebno je efekte ovakvog delovanja uzeti u obzir. Ukoliko je izlaz mernog pretvarača neka električna veličina obično se od uticaja temperature štiti merenjem preko Vinstonovog mosta (polumosta), dok se šum eliminiše analognim *bandwidth* ili *lowpass* filterom [91]. Neke od mera obezbeđenja kvalitetnog merenja mogu značajno uticati na rezultat merenja i u amplitudnom i u frekventnom domenu. Na primer, merenje pritiska preko creva ili nivoa preko mernog bunara koji deluje kao filter i neutrališe visoke frekvencije merene veličine [13].

Digitalizacija signala je sledeći korak u procesu merenja i ona se ogleda u uzorkovanju kontinualnog signala u određenim vremenskim intervalima (analogno-digitalna konverzija – A/D konverzija). A/D konverzija je proces koji se obavlja uređajima koji se zovu A/D konvertori i bazirana je na određivanju vrednosti analognog signala u diskretnim vremenskim trenucima. Proces A/D konverzije praćen je brojnim ograničenjima, kao što su problem rezolucije konvertora, nelinearnosti ili vremena uzorkovanja [65].

Nakon A/D konverzije kontinualni signal transformisan je u digitalni (prilagođen upotrebi pomoću računara). Naknadno procesiranje signala uključuje transformaciju signala preko kalibracione krive, filtriranje, određivanje statističkih parametara, itd. Na kraju mernog procesa kao rezultat očekuje se registrovana vrednost merne veličine koja je reprezentativna za osmatranu pojavu (slika 3.3).



Slika 3.3: Proces transformacije analognog signala u digitalni i transformacija digitalnog signala

Može se zaključiti da je merni sistem izuzetno kompleksne prirode i da rezultat merenja u velikoj meri zavisi kako od adekvatnog izbora svake od komponenti, tako i od interakcije svih komponenti mernog sistema. Procedura merenja u hidrotehnici se, kao što je opisano u prethodnim paragrafima, sastoji iz više koraka i u svakom od koraka se mogu javiti problemi, pogotovo ako je merenje automatsko i bez stalne kontrole posade mernog mesta. Vremenom, ukoliko se ne otkriju i adekvatno ne dokumentuju, problemi nastali u toku merenja bivaju zaboravljeni, dok trag koji su ostavili u registrovanom izmerenom podatku ostaje zabeležen.

Izmerena vrednost je rezultat mernog procesa i pretpostavlja se da je reprezentativna slika merene veličine. U zavisnosti od adekvatnog izbora merne lokacije, mernog sistema, načina obrade i transporta signala, merena vrednost može u sebi imati dodatne neželjene komponente, kao što su greške ili neodređenost. Postojanje neodređenosti uslovljeno je tipom podataka i izborom merne metode i merne skale.

3.1.1 Tipovi podataka dobijeni merenjem

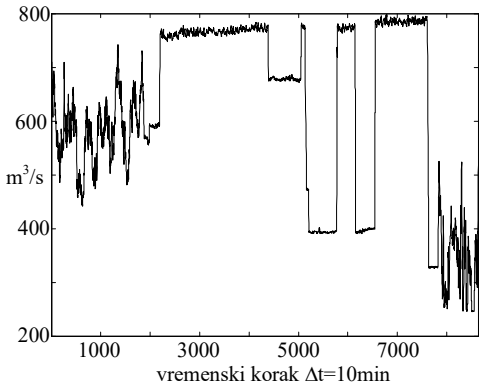
Podaci predstavljaju interpretaciju rezultata osmatranja neke pojave. Da bi se analizi te pojave na osnovu podataka pristupilo sveobuhvatno i analitički, potrebno je odrediti tip podataka kojim se raspolaže. Tipovi podataka mogu se podeliti u više kategorija, koje se međusobno prepliću. Neke od

kategorija su: 1) kvalitativni ili kvantitativni, 2) egzaktni ili neodređeni, 3) kategorički ili numerički, itd.
Iako ređe, pored kvantitativnog podatka, hidrotehničku veličinu moguće je opisati kvalitativnim tipom podatka. Kvalitativni tip podataka ne predstavlja veličinu brojevima, već opisom koji se može izraziti jezikom. Kao primer bi se mogao navestinivo reke opisan atributima nizak, normalan ili visok. Podatak ovog tipa može se dobiti transformacijom kvantitativnih podataka i njihovom klasifikacijom u određene kategorije. Veoma važno uočiti da kod nekih kvalitativnih opisa postoji poredak, dok kod drugih ne postoji. Kod navedenog primera nivoa reke očigledno je da poredak postoji ukoliko se definiše kriterijum na koji se poredak odnosi. Kriterijumom bi se dobila mogućnost da se odredi koja je vrednost najbolja, a koja najgora od mogućih alternativa. Važna karakteristika kvalitativnog opisa veličina je teškoća u određivanju veličine rastojanja (razlike) između podataka.

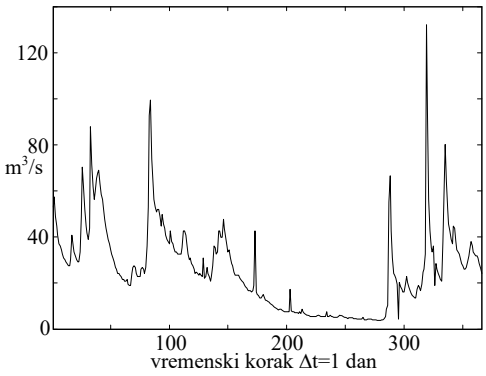
Kvantitativni tip podataka, za razliku od kvalitativnog, izražava se brojevima, a ne jezičkim odrednicama. Međutim, iako je neki podatak predstavljen kao broj, karakteristike podataka određuju način na koji se podatak može interpretirati. Na primer, jedinstveni matični broj građanina (JMBG) ili broj registarskih tablica vozila, iako se izražava numeričkim vrednostima, ne omogućava da se podaci upoređuju numeričkim relacijama ili da se dodatne karakteristike određuju primenjivanjem npr. računskih operacija (sabiranje, oduzimanje, itd.). Sa druge strane, kvantitativni tip podataka obuhvata i podatke dobijene merenjem kojima je moguće manipulisati matematičkim alatima. Osnovna dva tipa podataka koji spadaju u ovu grupu kvantitativnih podataka su diskretni i kontinualni tip podataka. Diskretni tip podataka uglavnom se odnosi na pobrojavanje. Primer bi bio diskretni podatak o broju razdvojenih enetiteta, npr. broj agregata hidroelektrane ili broj potrošača priključenih na jedan čvor vodovodne mreže. Iako je ovaj tip podataka uglavnom ograničen na predstavljanje prirodnim brojevima, njih je moguće upoređivati ili je moguće njima manipulisati matematičkim operacijama. Kontinualni podatak zavisi od načina merenja i instrumenata koji su u tu svrhu korišćeni. Naime, podatak o površini nekog područja može se razlikovati u zavisnosti od načina njegovog određivanja Uz kontinualni tip podataka neophodno je priložiti i podatak o preciznosti ili neizvesnosti koji podatak nosi.

Veliku ulogu u definisanju tipa podataka ima i skala po kojoj se podaci mogu poređati. Ponekad skala ne postoji ili nije bitna za dalja razmatranja i analize, ali ponekad su podaci prirodno poređani po određenoj skali, kao što je slučaj sa vremenskim serijama gde se kao skala koristi vremenska linija. Osnovna tri tipa skala su intervalska skala, skala sa razmerom i kružna skala. Intervalska skala predstavlja skalu kod koje se podaci nalaze u predefinisanom relativnom redu, dok apsolutna pozicija nije poznata. Kod skale sa razmerom, da bi razmera postojala mora postojati jedan fiksni podatak u odnosu na koji se razmere mogu definisati. Kada se beleže vremenski podaci, kao što su nedelje, dani u mesecu, sati, itd, može se koristiti kružna skala.

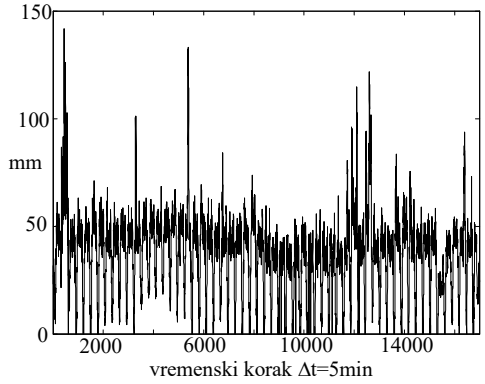
Uglavnom je tip podataka koji se koristi u skladu sa količinom informacija kojom se raspolaže o razmatranoj veličini, što znači da se jedna veličina, u zavisnosti od toga koliko informacija o njoj postoji, može opisati na više načina. Kao primer bi se mogao navesti podatak o padavinama na nekom području. Mogu se pretpostaviti tri tipa podataka o padavinama: kontinualni podaci o intenzitetu kiše, sumarni podaci o npr. dnevnim količinama padavina i binarni podaci o tome da li je bilo padavina ili ne. Svaki od navedena tri tipa podataka nosi određenu količinu informacija i ima specifičnu upotrebnu vrednost.



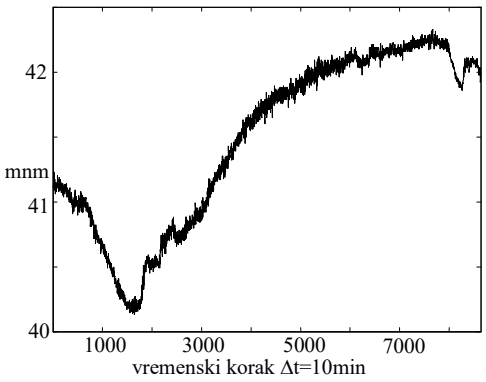
A: Protok kroz agregat hidroelektrane



B: Protok na meren na vodotoku



C: Brzina vode u kanalizacionom kolektoru



D: Nivo vode nizvodno od hidroelektrane

Slika 3.4: Vremenske serije merenih hidrotehničkih veličina

Iako su kvantitativni podaci više zastupljeni u tehničkoj praksi, ponekad je mnogo podesnije upotrebiti kvalitativne podatke. Da bi se kvalitativni podaci prilagodili radu sa računarima, oni se moraju predstaviti u numeričkoj formi (transformisati u kvantitativni tip podataka). Često je ta transformacija standardizovana, ali se ponekad način transformacije mora i posebno definisati. Primer standardizovane transformacije bi bila transformacija kvalitativne informacije u boju, a dalje boje u kod (npr. crvena je u RGB kodu [1, 0, 0]). Primer nestandardizovane transformacije (na primer, prilagoĐene samo nekom slivu) bila bi klasifikacija kiša na one koje bi izazvale oticaj i na one koje ne bi izazvale.

Tradicionalan pristup podacima je preko egzaktnih vrednosti (*crisp value*). Egzaktna vrednost podrazumeva da se podatak može predstaviti jednim brojem, oznakom klase kojoj pripada ili nekom kvalitativnom veličinom koja mu se pripisuje. Naime, ukoliko su tip podatka i merna skala diskretni, postoje merne metode kod kojih ne figuriše neodreĐenost, već se rezultat merenja može iskazati egzaktnom vrednosti. Na primer, metodom prebrojavanja se može doći do tačnog broja diskretnih vrednosti kao rezultata merenja (npr. broj potrošača vodovodnog sistema). U tom slučaju postoji samo greška koja se može napraviti u brojanju, dok je rezultat metode bez neodreĐenosti. Nasuprot diskretnim merenjima, u slučaju kontinualne merne skale, pored greške, neizostavna je i neodreĐenost (npr. potrošnja po potrošaču vodovodnog sistema).

Većina podataka kojima se opisuju hidrotehničke veličine proizvod su kvantitativnih merenja kontinualnih veličina pa su samim tim neizvesne (neodreĐene) prirode. Kao ilustracija, na slici 3.4 prikazane su neke od vremenskih serija sa izmerenim hidrotehničkim veličinama. Ne smeju se

zaboraviti ni podaci koji u sebi nemaju neodređenost (npr. podatak da li je pumpa upaljena ili ugašena). Neodređeni podaci mogu se opisati pre skupom vrednosti nego jedinstvenim brojem.

Nasuprot činjenici da je merna metoda nosilac neodređenosti, uzrok greške je uglavnom kompleksna pojava koja se može pripisati mnogobrojnim faktorima koji figurišu pri merenju. Takođe, za razliku od neodređenosti, svaki podatak, bez obzira na njegove karakteristike ili karakteristike merne metode kojom je dobijen, može sadržati grešku. Iako je često korisno znati uzrok greške, sistem za vrednovanje podataka predstavljen u ovoj disertaciji ne razmatra uzrok, već isključivo postojanje greške u podatku.

3.1.2 Podela grešaka u podacima

Da bi se došlo do dobro osmišljenog sistema za vrednovanje neophodno je obuhvatiti sve karakteristike konkretnih vremenskih serija. Zato je potrebno imati dobru bazu znanja o svim aspektima koji utiču na proces merenja, kao i dostupne metode za predikciju podataka i dodatne informacije koje se u metodama javljaju kao ulazne veličine. Tradicionalno se vrednovanju podataka pristupa upotrebom logike i iskustva, što često zahteva vizuelni prikaz merenih vrednosti i ostalih raspoloživih podataka. Stoga, postoje dva puta kojim se može ići u procesu vrednovanja podataka:

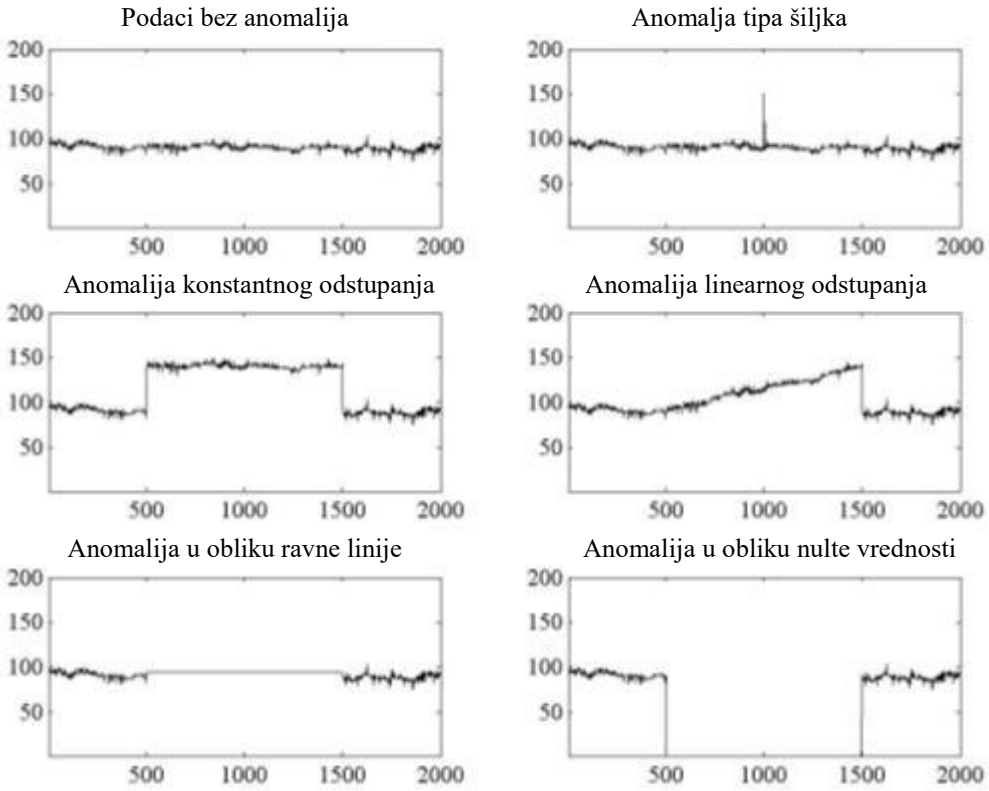
- iskustvenom formulacijom relacija između podataka i vizuelizacijom podataka;
- matematičkom formulacijom relacija između podataka.

Prva grupa metoda za vrednovanje predstavlja tradicionalne metode za vrednovanje podataka i otkrivanje grešaka. Grube greške, na primer, greške sa nultom vrednošću, vrednosti sa anomalijama u obliku ravne linije, vrednosti koje izlaze iz okvira apsolutnog minimuma ili maksimuma ili nestanak struje, samo su neke anomalije koje stručnjak lako može da prepozna jednostavnim vizuelnim pregledom grafičkog prikaza podataka. Za primenu prve grupe metoda korisno je iskustvo u prepoznavanju specifičnih „potpisa" grešaka u podacima.

Druga grupa metoda uslovljena je postojanjem dodatnih informacija i podataka i matematičkim prikazom relacija između merenih vrednosti. Formiranjem matematičkih formulacija odnosa između merenih vrednosti otvara se mogućnost da se formirani alati bazirani na relacijama koriste u automatskom ili polu-automatskom režimu, za razliku od ručnih postupaka koji su vezani za vizuelizaciju podataka.

Ponekad je moguće ispitati, i na kraju izolovati, neke razloge za pojavu podataka sa greškama. Ali za neke greške ili ne postoji logično objašnjenje, ili su uzroci toliko složeni da njihovo utvrđivanje nije vredno truda. Prema tome greška može predstavljati indikator nepoželjnog događaja, ali se često samo radi o nepoželjnoj vrednosti uzorka koja mora da se ispravi ili odbaci.

Tradicionalnim pristupom detekciji grešaka, koji podrazumeva vizuelnu inpekciju podataka, mogu se uočiti neki specifični „potpisi" određenih grešaka sistema za merenje. Na slici 3.5 dati su neki od njih [12].



Slika 3.5: Tipovi anomalija u izmerenim podacima

Greška tipa šiljka je vrlo čest tip anomalije. Njena osnovna karakteristika je kratko trajanje (obično samo jedna merena vrednost) i, nakon što se pojavi, sistem nastavlja da beleži ispravne uzorke. Na neke tehnike merenja, kao što je ultrazvučno merenje brzine Dopler-postupkom, utiče slučajni proces protoka čestica, pa je taj postupak merenja osetljiv na greške tog tipa. Ostali uzroci anomalije tipa šiljka mogu biti mehanički kvarovi mernog uređaja, povećana osetljivost na određene pojave, itd.

Greške sa konstantnim i linearnim pomerajem izaziva nešto što ima dugotrajniji uticaj na merni sistem. Mogući razlozi za ovo su povećan mehanički histerezis, greška u krivoj kalibracije, postepena ili brza promena mikro-lokacije merenja, nestanak struje, itd. Greške gde se tokom merenja javljaju konstantna vrednost ili nula lako se prepoznaju u serijama podataka i mogu da imaju veze, osim sa nekim neželjenim okolnostima, i sa podešavanjem merne opreme. Pored predstavljenih tipova grešaka, greškom se može smatrati i degradacija signala usled šuma.

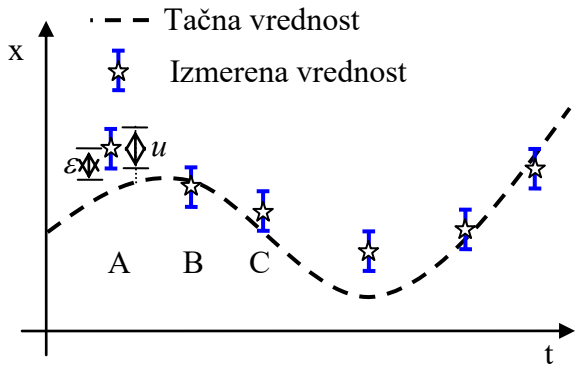
Pored podele prema obliku greške, za potrebe formiranja nekih specifičnih postupaka detekcije grešaka u podacima [57], greške se mogu podeliti na: aditivne i multiplikativne. Aditivni tip greške podrazumeva da zbir anomalije i tačne vrednosti čini izmerenu vrednost, dok je kod multiplikativnog tipa izmerena vrednost jednaka umnošku tačne vrednosti i nekog parametra.

Iako se navedeni tipovi grešaka mogu lako uočiti uz pomoć vizuelne inspekcije podataka, postoje i greške koje nemaju specifičan „potpis" kada se grafički prikažu. Takve greške se teško otkrivaju tradicionalnim vizuelnim tehnikama, već se moraju primeniti složenije metode bazirane na relacijama između merenih podataka.

Predloženom metodologijom uglavnom se tretiraju kvantitativni podaci, mada se može primeniti i na kvalitativne ukoliko se mogu povezati matematičkim relacijama sa drugim merenim podacima. Takođe, u ovoj doktorskoj disertaciji pretpostavlja se da greška u podatku ne mora imati specifičan „potpis", i njeno prisustvo u podatku detektuje se pomoću matematičkih relacija sa drugim merenim podacima. To znači da je moguće detektovati grešku iako ona ne pripada nijednoj navedenoj kategoriji grešaka. Sa druge strane, predloženi pristup ne sužava prostor odluke pri detekciji uzroka greške.

3.2 Greške i neodređenost

Osnovne neželjene komponente koje se mogu prepoznati u svakoj merenoj vrednosti su merna greška i merna neodređenost. Na slici 3.6 slikovito je prikazana razlika između greške i neodređenosti izmerene vrednosti. Neodređenost (u) je karakteristika merene veličine koja ne zavisi od toga koliko je merena vrednost bliska sa tačnom, već označava oblast (anvelopu) u kojoj se tačna vrednost očekuje (što ne mora uvek biti tačno).



Slika 3.6: Neodređenost (prikazana u formi intervala) i greška merenja

Veličina greške (ε) upravo označava koliko je izmerena vrednost bliska tačnoj. Na slici 1.6 prikazana je aproksimacija greške kao odstupanje tačne vrednosti od sredine intervala izmerene vrednosti. S obzirom na toda tačna vrednost nije dostupna, uglavnom je moguće proceniti samo granice u kojima se veruje da se greška nalazi. One bi u slučaju prikazanom na slici 3.6 iznosile:

$$\varepsilon = [\underline{\varepsilon}, \overline{\varepsilon}] = \left[\min\left(\left| x^T - \left[x^M_{\min}, x^M_{\max} \right] \right| \right), \max\left(\left| x^T - \left[x^M_{\min}, x^M_{\max} \right] \right| \right) \right],$$

gde su x^T tačna vrednost, a $\left[x^M_{\min}, x^M_{\max} \right]$ merena vrednost u obliku intervala. Na slici 3.6 su izdvojene tri izmerene vrednosti, prikazane u obliku intervala, obeležene sa A, B i C. Podatak A ima neodređenost u i vrednost greške ε u egzaktnom obliku ukoliko se ona računa od srednine intervala. Kod vrednosti B je u izmerenu vrednost uključena i tačna vrednost, pa se može postaviti relacija između greške i neodređenosti [93]:

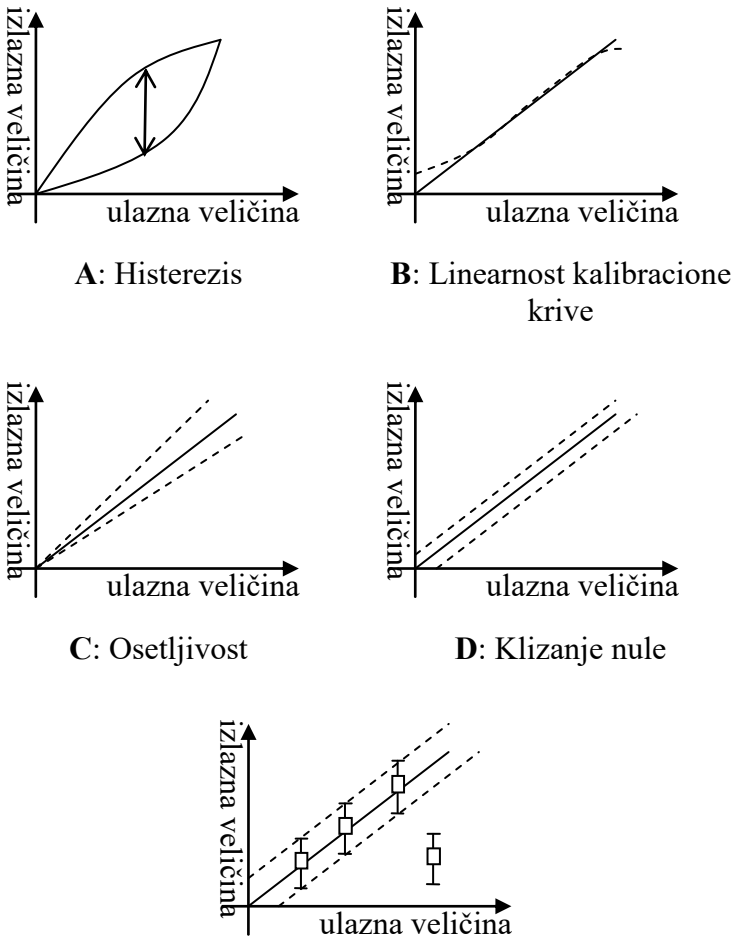
$$|\varepsilon| \le |u|$$

Ova relacija važi isključivo u slučaju kada je tačna vrednost uključena u interval izmerene vrednosti (granični slučaj je merena vrednost C), što se u praksi često i pretpostavlja (zanemaruje se mogućnost postojanja greške kao kod vrednosti A) kada se merenje obavi pažljivo, prema najboljoj mernoj praksi.

Za razliku od greške, za koju je potrebno znati tačnu vrednost da bi je odredili, veličina neodređenosti se procenjuje na osnovu brojnih faktora koji utiču na proces merenja (merna oprema, merno okruženje, itd.). Često nije moguće imati u vidu sve izvore neodređenosti (neizvesnosti) i grešaka kod merenja, ali je neke izvore moguće detektovati. Izvore neodređenosti je moguće podeliti u dve grupe [35]: neodređenost zbog samog mernog instrumenta, i neodređenost zbog niza drugih uticaja koji utiču na merenje. Neodređenost zbog mernog instrumenta se može podeliti na: neodređenost usled rezolucije samog instrumenta, u_0, i neodređenost usled konstrukcije mernog instrumenta, u_c. Neodređenost usled rezolucije mernog instrumenta se obično računa kao polovina najmanjeg podeoka na mernoj skali ($u_0 = \pm 0.5 \times res$). Sa druge strane, neodređenost usled konstrukcije mernog instrumenta se može izračunati preko neodređenosti koje potiču od komponenti mernog instrumenta. Ukupna neodređenost samog mernog instrumenta iznosi:

$$u_d = \sqrt{u_0^2 + u_c^2}.$$

Neodređenost usled ostalih uticaja koji utiču na merenje zavisi od brojnih faktora koji često postaju poznati tek nakon što se merni uređaj pusti u rad i kada se eksperimentalno utvrde njegove karakteristike pri testiranju, kalibraciji i upotrebi.



A: Histerezis

B: Linearnost kalibracione krive

C: Osetljivost

D: Klizanje nule

E: Ponovljivost

Slika 3.7: Izvori grešaka u fazi kalibracije mernog instrumenta [35]

Nastanak grešaka se, kao i poreklo neodređenosti, može povezati sa: kalibracijom mernog instrumenta (slika 3.7), samim merenjem i obradom sirovih merenih podataka. Kalibracija mernog instrumenta predstavlja usklađivanje izlaza mernog uređaja sa vrednošću merene veličine. Kao što je prikazano na slici 3.7, može se izdvojiti pet tipova grešaka koje potiču od kalibracije [35]. Kalibracijom se detektuje neodređenost, dok se pokušava da se eliminišu greške u podacima. Iako je kalibracija mernog uređaja sprovedena na najbolji mogući način, kalibracione karakteristike se, nažalost, u toku vremena menjaju, pa je merni uređaj potrebno rekalibrisati. Period rekalibracije proizvođač mernog instrumenta obično istakne, ali pošto on zavisi od uslova upotrebe merila, često se rekalibracija sprovodi u intervalima koji su određeni na osnovu empirijskih pokazatelja rada uređaja.

Izvori grešaka i neodređenosti pri samom merenju prema [35] navedeni su u tabeli 3.1.

Tabela 3.1: Izvori grešaka pri merenju [35]

	Izvor greške
1	Uslovi u sistemu za osmatranje
2	Kontakt mernog uređaja sa mernim
3	medijumom
4	Konverzija signala
5	Izlaz i transfer signala
6	Uslovi u sistemu koji se osmatra
7	Instalacija sistema za osmatranje
8	Spoljni uticaji
9	Prostorna varijabilnost osmatrane veličine
10	Vremenska varijabilnost osmatrane veličine
11	Promena kalibracionih karakteristika uređaja
itd.	

Greške koje potiču iz navedenih izvora teško su predvidljive. Uglavnom se njihovo prisustvo uoči tek nakon što je merni uređaj pušten u rad i nakon analize prikupljenih podataka. Zbog toga je potrebno predvideti da će se u toku rada mernog uređaja javiti greške koje nisu uračunate u neizvesnost prikupljenog podatka. Uzrok tako nastalih grešaka je možda moguće detektovati i otkloniti, ali to je moguće jedino ako i kada se takve greške uoče.

3.2.1 Klasifikacija i prikazivanje grešaka i neodređenosti

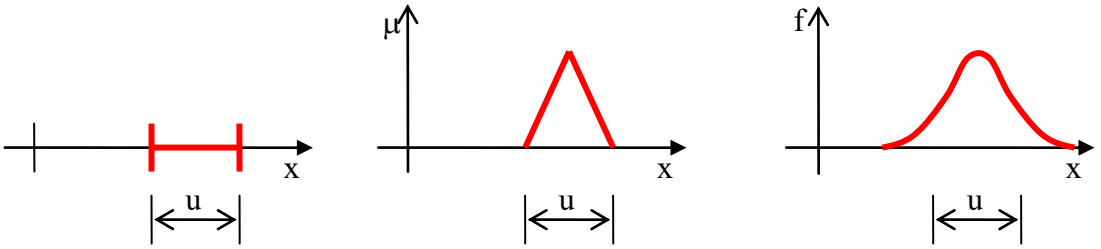
Ne ulazeći u način kako merna greška nastaje, greške se mogu podeliti na: grube greške, sistematske greške i slučajne greške. Greška kao kvalitativna osobina može se kvantitativno izraziti na dva načina, kao apsolutna greška i kao relativna greška. Apsolutna greška predstavlja apsolutnu vrednost razlike između izmerene i stvarne vrednosti:

$$\varepsilon_{aps} = \left| x_{izm} - x_{stv} \right| .$$

Nasuprot apsolutnoj, relativna greška predstavlja relativni odnos apsolutne greške i izmerene vrednosti:

$$\varepsilon_{rel} = \frac{\left| x_{izm} - x_{stv} \right|}{x_{izm}} .$$

Neodređenost se može matematički prikazati na više načina od kojih su svakako najčešće korišćeni intervali (slika 3.8A), rasplinuti skupovi (slika 3.8B) i statističke raspodele (slika 3.8C).



A: Interval **B**: Rasplinuti skup **C**: Statistička raspodela

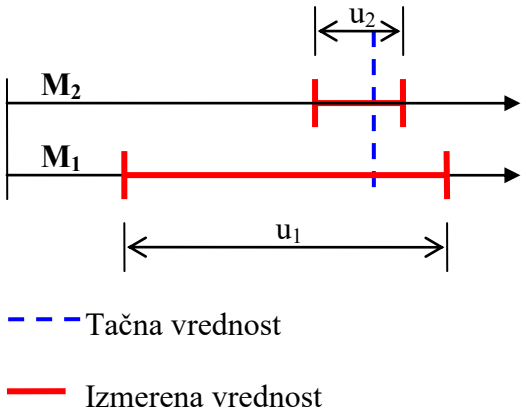
Slika 3.8: Matematički prikaz neodređenosti merenog podatka

Utisak o neodređenosti kod intervala je nedvosmisleno širina intervala, kod rasplinutog skupa je na slici 3.8 neodređenost predstavljena preko širine baze, dok se kod statističke raspodele neodređenost vezuje za neku karakterističnu vrednost funkcije raspodele, npr. varijansu ili interval kvantila od 95%. Intervalska matematika, teorija mogućnosti (*possibility theory*) i statistička teorija predstavljaju matematičke alate za manipulaciju i transformaciju navedenih oblika u kojima je moguće prikazati neodređene veličine.

3.2.2 Interpretacija merenih podataka

Merenje predstavlja niz procedura kojim se dolazi do podataka. Rezultat merenja, stoga, predstavlja podatak o vrednosti merene veličine. Do podatka o vrednosti merene veličine uglavnom se može doći na više načina, odnosno pomoću više različitih mernih metoda. Ukoliko se merne metode posmatraju samo sa stanovišta neodređenosti, čime se pretpostavlja da merna metoda nema u sebi grešku, svrha merenja se svodi na smanjenje neizvesnosti vrednosti merene veličine.

Na slici 3.9 prikazani su rezultati merenja pomoću dve merne metode, M_1 i M_2. Na slici su veličine prikazane preko intervala, a neizvesnosti širinom intervala, u_1 i u_2. Metoda M_2 je nesumnjivo doprinela manjoj neizvesnosti u pogledu vrednosti merene veličine.

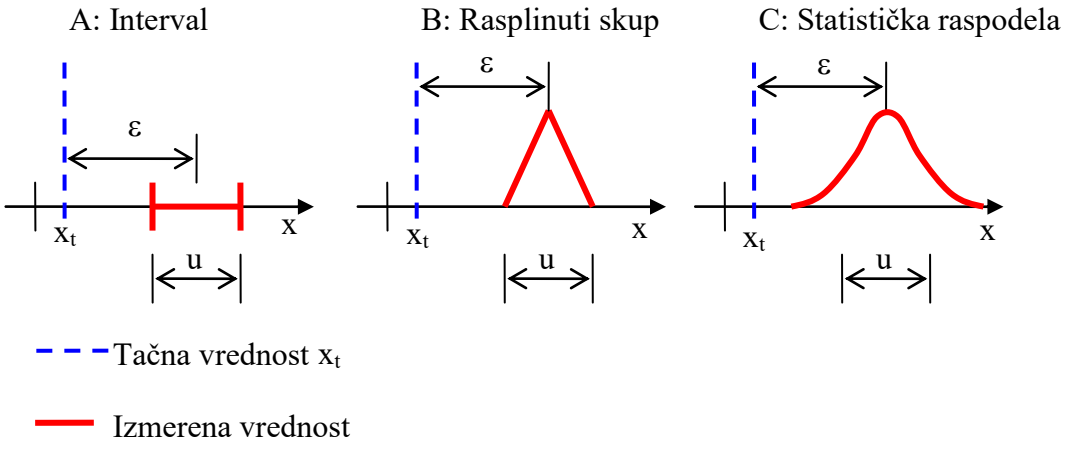


Slika 3.9: Primer rezultata merenja pomoću dve merne metode M_1 i M_2

Važno je podvući da se uglavnom ne raspolaže tačnom vrednosti već je ona dodata uz pretpostavku da je merene vrednosti uključuju.

Pored intervala, neizvesnost se može matematički opisati na još dva načina, kako je prikazano na slici 3.10: rasplinutim skupom i statističkom raspodelom.

Greška merenja može biti tolika da kompromituje izmerenu vrednost i da uprkos smanjenoj neizvesnosti pruži pogrešnu informaciju o razmatranoj pojavi.

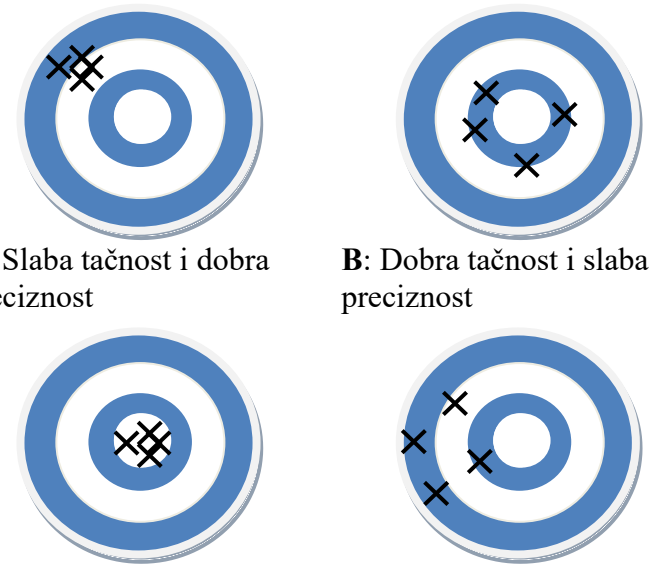


Slika 3.10: Greška i neodreĎenost merenog podatka

Greška merenja je na slici 3.10 označena kao rastojanje tačne vrednosti od sredine intervala, od vrednosti rasplinutog skupa sa najvišom vrednosti pripadnosti ($\mu(x)=1$) ili od srednje vrednosti statističke raspodele. Pošto tačna vrednost nije poznata, u procesu provere vrednosti podatka potrebno je proceniti veličinu greške koju on sadrži.

3.2.2 Tačnost i preciznost merenja

Razmatrane karakteristike koje prate merni podatak u pogledu grešaka i neodreĎenosti mogu se iskazati i kao njegova tačnost i preciznost. Iako su možda na prvi pogled ta dva termina sinonimi, ukazuju na dve potpuno različite karakteristike podatka koji je dobijen merenjem. Čest primer u literaturi [35] je vezan za pogaĎanje mete strelicama u pikadu. Na slici 3.11 prikazana su četiri karakteristična slučaja.



A: Slaba tačnost i dobra preciznost

B: Dobra tačnost i slaba preciznost

C: Dobra tačnost i dobra preciznost

D: Slaba tačnost i slaba preciznost

Slika 3.11: Tačnost i preciznost na primeru pikada

Kada se logika iz pikada prenese na podatke dobijene merenjima, može se zaključiti da se preciznost odnosi na neizvesnost merene veličine, dok se tačnost odnosi na postojanje greške u merenim podacima. Veličine graničnih vrednosti grešaka, koje se u procesu eksploatacije mogu pripisati neodređenosti, mogu se odrediti u toku konstrukcije mernog uređaja, testiranja, kalibracije i na osnovu uslova u kojima merni uređaj radi. U toku upotrebe, greške u podacima mogu se samo pretpostaviti i posredno odrediti, na osnovu rezultata merenja merilima više tačnosti.

———————— · ————————

U ovoj doktorskoj disertaciji polazi se od činjenice da podatak može imati bilo koji od tri navedena oblika, tj. da može, ali i ne mora imati grešku, a cilj predloženog algoritma za vrednovanje je upravo da se postojanje greške otkrije. Za razliku od greške, pretpostavlja se da podatak mora imati neodređenost koja se može predstaviti u jednom od tri oblika: intervalom, rasplinutim skupom ili statističkom raspodelom.

3.3 Primeri grešaka u podacima

U nastavku se navode tri primera koji se odnose na hidrotehničke probleme iz različitih oblasti. Iako sistem za vrednovanje nije zamišljen za potrebe laboratorijskih merenja, kontrolisani uslovi i ponovljivost koji karakterišu ovakav način merenja mogu se upotrebiti za proveru ideje i rada sistema. S obzirom na to da se sistem za vrednovanje podataka oslanja na matematičke relacije koje se mogu primeniti u procesu provere karakteristika podataka, navedeni primeri oslikavaju tu mogućnost u svetlu neodređenosti koju matematički modeli osmatranih procesa nose. Veličina neodređenosti modela direktno utiče na čvrstinu relacija koje se mogu primeniti i ponekad, iako je moguće u laboratorijskim uslovima ostvariti kontrolisane uslove merenja, nije moguće obezbediti adekvatan matematički opis merenog procesa („All models are wrong, some are useful―[George Box, 1979]). Sa druge strane, kod merenja sprovedenih na terenu, gde se meri više veličina istovremeno, nudi se mogućnost da se različite merene veličine upoređuju ina taj način proveravaju (vrednuju).

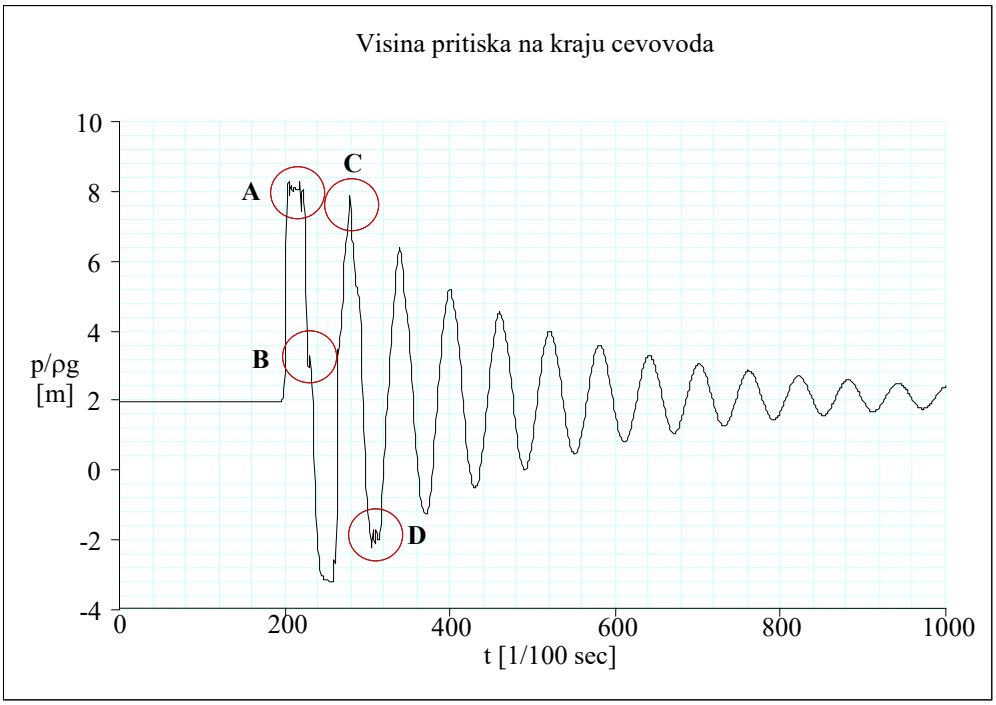
Prvi primer predstavlja laboratorijsku simulaciju hidrauličkog udara koja se osmatra merenjem protoka, pritisaka i otvorenosti zatvarača. Drugi primer odnosi se na hidrološki ciklus na nekom slivu koji se osmatra uz pomoć niza merenih veličina kao što su nivoi, protoci, temperature, vlažnost vazduha, itd. Treći primer je primer merenja parametara rada kanalizacionog sistema (nivoa, brzina i parametara kvaliteta).

3.3.1 Primer 1 – Laboratorijska instalacija za izazivanje hidrauličkog udara

Hidraulički udar je pojava koja se javlja pri nagloj promeni graničnog uslova u sistemima pod pritiskom, kada nastaje talas visokog ili niskog pritiska koji putuje duž cevovoda. Nakon refleksije od nekog diskontinuiteta talas se odbija i prelazi u talas niskog, odnosno talas visokog pritiska. Talasi visokog i niskog pritiska se smenjuju sve dok sile otpora ne ublaže pojavu hidrauličkog udara.

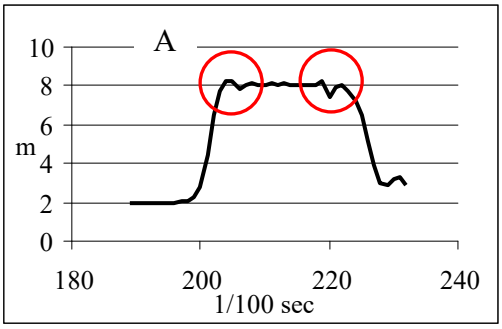
Merenje hidrauličkog udara predstavlja složen zadatak (čak iako se merenje obavlja u laboratoriji) koji mora biti u skladu sa nekoliko kriterijuma: 1) merni uređaji moraju biti dovoljno precizni, kako u opsegu redovnog operativnog rada cevovoda, tako i za ekstremne vrednosti pritiska i protoka, 2) vreme odziva mernih uređaja mora biti u skladu sa promenama merenih vrednosti, 3) treba imati u vidu da mehaničke karakteristike mernih uređaja mogu značajno da utiču na rezultat merenja (na primer, vibracije membrane senzora pritiska usled impulsa pritiska), 4) brze promene hidrauličkih parametara (pritiska i protoka), kao i promene graničnih uslova, zahtevaju brzo uzorkovanje koje ponekad mora biti reda veličine i nekoliko stotina Hz, i 5) potrebno je obezbediti dobro dihtovanje svih delova cevovoda jer usled pojave negativnog talasa pritiska gasovita faza, koja biva uvučena u cevovod, može da zapuši priključke mernih uređaja i priguši pritiske koje je potrebno registrovati (gustina vazduha je oko 1000 puta manja od gustine vode), itd.

Merenje je obavljeno na kraju cevovoda (neposredno ispred zatvarača). koji počinje rezervoarom, a završava se zatvaračem. Pritisak se meri DRUCK sondom, koja je prethodno kalibrisana na instalaciji za kalibraciju sondi za pritisak, brzinom uzorkovanja od 100 Hz. Nakon akvizicije podataka (vremenska serija je prikazana na slici 3.12) uočene su anomalije u podacima u prve dve periode oscilovanja pritiska. Nakon prve dve periode, vizuelnim postupkom je potvrĎeno da su podaci regularni (odgovaraju karakteristikama pojave hidrauličkog udara).

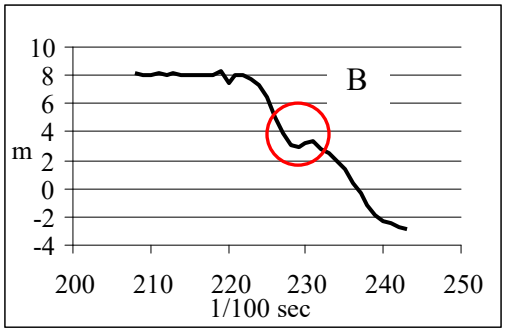


Slika 3.12: Visina pritiska na kraju cevovoda pri izazvanom hidrauličkom udaru

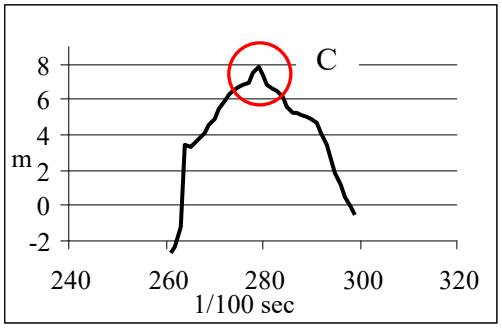
Detektovane anomalije su u krupnijoj razmeri prikazane na slici 3.13. Anomalija **A** dogodila se pri prvom maksimumu i sumnja se da je njen uzrok vibriranje membrane DRUCK sonde za merenje pristiska. Anomalija **B** dogodila se u toku prelaska sa pozitivnog u negativan pritisak i najverovatnije je uzrokovana zarobljenim vazduhom na mestu priključka sonde za merenje pritiska. Za anomaliju **C** se sumnja da je nastala usled lošeg odziva mernog uređaja, dok se za anomaliju **D** pretpostavlja da je nastala usled vibracija membrane sonde (isto kao i anomalija **A**).



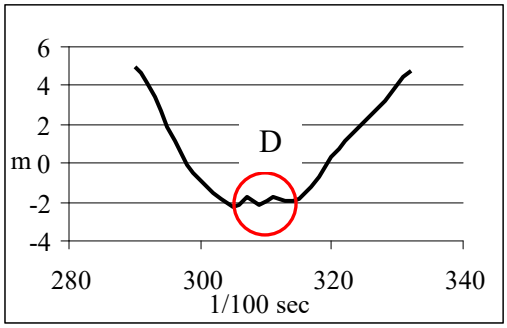
A: Anomalija usled vibracije membrane mernog uređaja



B: Anomalija zbog zarobljenog vazdušnog mehura



C: Anomalija usled lošeg odziva mernog uređaja



D: Anomalija usled vibracije membrane mernog uređaja

Slika 3.13: Pojedine anomalije u zapisu merenog pritiska pri izazvanom hidrauličkom udaru (slika 1.10)

Vrednovanje merenih podataka dobijenih merenjem hidrauličkog udara obavljeno je vizuelnom proverom grafičke predstave podataka. Obeležene su potencijalne greške u podacima koje ne moraju biti greške merenja, već mogu biti i registrovane pojave koje nisu dovoljno istražene.

Procedura vrednovanja podataka, osmišljena za navedenu vremensku seriju, uključuje sve aspekte koji utiču na merenje i koji određuju karakteristike podataka koji su prikupljeni. Najznačajniji su tehnički aspekti, koji predviđaju vibracije membrane i neadekvatan odziv mernog uređaja, ekspertsko mišljenje koje pretpostavlja postojanje vazdušnih mehurova u sistemu i mogućnost da se neki od njih zaglavi u mernom uređaju, kao i relacione karakteristike koje opisuju karakteristični oblik vremenske serije pritiska kod pojave hidrauličkog udara. Ovakav način vrednovanja podataka (vizuelnom inspekcijom grafičke predstave podataka) tradicionalno je najčešći. Osnovni problemi vezani za njega su što je za njegovo sprovođenje potrebno angažovati iskusnog eksperta, i što se može obaviti na relativno malom obimu podataka.

Kod laboratorijskih merenja, gde je eksperiment ponovljiv, uočene anomalije u podacima najčešće imaju za cilj da upute na moguće probleme, da bi se u nekom ponovljenom merenju one izbegle. Kod merenja na terenu merenja se retko mogu ponoviti pod identičnim uslovima, jer se uslovi neprestano menjaju.

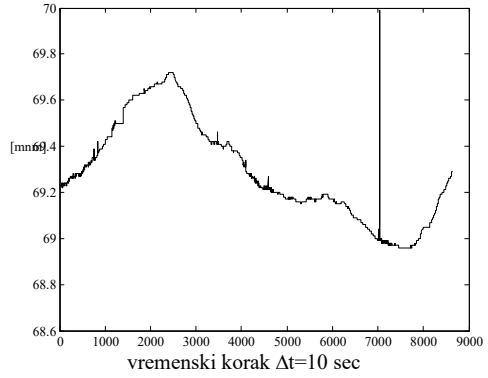
3.3.2 Primer 2 – Osmatranje hidrološkog ciklusa i rada hidroelektrana

Osmatranje hidrološkog ciklusa je proces merenja svih veličina koje utiču na kretanje vode u prirodi. Pored hidroloških veličina (protoci i nivoi u vodotocima i podzemnim akviferima), neretko je za potrebe predviđanjai analizu kretanja vode u prirodi potrebno meriti i meteorološke veličine, kao što su padavine, vlažnost vazduha, brzina i pravac vetra, osunčanost, temperatura vazduha, atmosferski

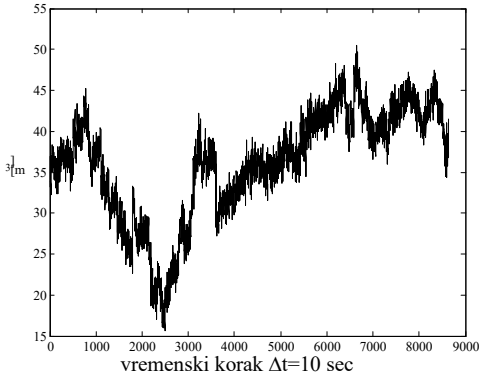
pritisak, oblačnost, debljina snežnog pokrivača, evaporacija, itd. Složenost problema matematičkog modeliranja hidrološkog ciklusa i potreba za predviĎanjima ukazuju na potrebu za istorijskim podacima. Osnovni kriterijum za upotrebu i analizu istorijskih podataka, pored činjenice da moraju biti regularni (bez grešaka i anomalija), jeste njihova uporedivost. Za razliku od laboratorijskih merenja koja se obavljaju u kontrolisanim (što znači ponovljivim) uslovima, okruženje u kome se obavlja terensko merenje (a hidrološka merenja spadaju u terenska) neprestano se menja. Zbog toga vrednovanje podataka mora biti usmereno ne samo na trenutno stanje mernog okruženja i osmatranog fenomena, već se mora obratiti pažnja i na izmenjene uslove merenja na koje se ponekad ne može uticati. Rezultat vrednovanja bi, zbog toga, morao da ukaže i na promene uslova merenja.

Lako se zaključuje da, iako je podatak o izmerenoj veličini regularan, on se ne može uporeĎivati sa istorijskim podacima koji su mereni na istoj lokaciji ako je ona izmenjena. Zbog toga je neophodno da se uz svaki podatak vodi evidencija o načinu kako je on nastao i koje su okolnosti vladale u toku njegovog nastanka.

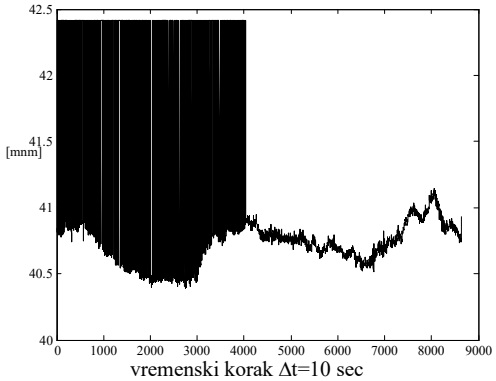
Hidrološke veličine se mogu meriti različitim metodama i na mernim lokacijama izabranim po različitim kriterijumima. Problemi kod merenja mogu biti mnogobrojni, od neadekvatnog mernog mesta (npr. merenje brzine vetra u predelu sa bujnom vegetacijom koja u odreĎeno doba godine menja pravac i brzinu strujanja vetra), do kvarova na instrumentima i sistemima za transfer podataka. S obzirom da su moderne hidrološke stanice pozicionirane uglavnom na nepristupačnim mestima, bez ljudske posade, a transfer podataka se obavlja automatski, postoji opasnost da se, ukoliko nije predviĎena validacija podataka i postojanje meta-podataka, zauvek izgubi informacija, ne samo o tome da li je podatak regularan, već i informacija o okolnostima u kojima je on nastao. Zbog toga se može desiti da se umanji pouzdanost baze istorijskih podataka ili čak da se u procesu analiziranja i predviĎanjakoriste podaci koji nisu regularni ili uporedivi.



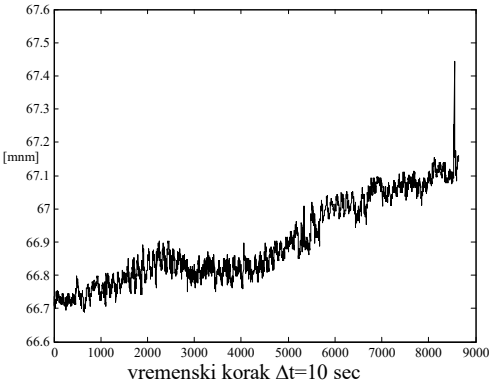
A: Kota nivoa na Dunavu



B: Protok kroz turbinu br. 2



C: Kota donje vode



D: Kota gornje vode

Slika 3.14: Merene hidrološke veličine na hidroenergetskom sistemu Djerdap

Hidrološke veličine se mogu vrednovati i na osnovu drugih veličina sa kojima su u relaciji. To mogu biti veličine koje se prikupljaju za potrebe upravljanja radom hidroelektrana. Na slici 3.14 prikazan je niz veličina koje se prikupljaju za potrebe upravljanja hidroenergetskim sistemom Djerdap. Vremenske serije veličina od interesa prikupljane su u intervalu od 10 sekundi i skladištene u istorijskoj bazi podataka. Anomalije, uočljive čak i golim okom na prikazanim dijagramima, predstavljaju grube greške u merenjima. Mereni podaci sa grubim i ostalim greškama, bez vrednovanja, ostaju pohranjeni u bazi podataka i predstavljaju neupotrebljive nizove, osim uz prethodnu (mukotrpnu) obradu od strane korisnika. Proces vrednovanja podataka pružio bi pomoć u vidu ocene kvaliteta podataka i značajno pojednostavio njihovo korišćenje.

Vrednovanje podataka za osmatranje jednog složenog tehničkog sistema, kao što je hidroenergetski sistem, mora obuhvatiti brojne fenomene koji ponekad i nemaju adekvatnu matematičku formulaciju (na primer, stvaranje depresionog levka na ulaznoj građevini pri startu turbine). Uzrok tome je ili nedostatak ulaznih podataka za matematičke modele (geometrija rečnog korita, koeficijenti korisnih dejstava agregata, itd.), ili nedovoljno znanje o samom fenomenu. Zbog toga je u procesu vrednovanja potrebno primeniti i neke dodatne alate kao što su *machine learning* tehnike ili veštačka inteligencija.

3.3.3 Primer 3 – Merenja količina vode i parametara kvaliteta u kanalizacionom sistemu

Kanalizacioni sistem predstavlja složen infrastrukturni skup objektata koji imaju za cilj da prikupe i sprovedu otpadne i kišne vode do postrojenja za prečišćavanje otpadnih voda, odnosno do recipijenta. Složene hidrauličke i hemijske procese koji se odvijaju u toku transporta vode kroz sistem cevi, kanala, sabirnih rezervoara i crpnih stanica moguće je osmatrati merenjem hidrauličkih i hemijskih parametara. Uslovi koji vladaju u kanalizacionom sistemu predstavljaju loše okruženje kako za mernu opremu, tako i za ostale komponente mernog sistema, pa se mogu očekivati anomalije u prikupljenim podacima.

Pokazatelji stanja u kanalizacionom sistemu mogu se podeliti na hidrauličke (nivo vode, brzina vode, protok) i hemijske parametre (temperatura, pH, elektroprovodnost, redoks potencijal, itd.). Izbor metode merenja ovih parametara zavisi od mnogih faktora, među kojima su uslovi koji vladaju na mernom mestu, mogućnost redovnog održavanja merne opreme, zahtevana tačnost merenja i neodređenost metode, itd. Na neke od metoda je moguće osloniti se i prilagoditi ih radu bez ljudske posade. Tada merni uređaj bez stalnog ekspertskog nadzora šalje podatke o izmerenim parametrima. Pošto se merno okruženje u lošim uslovima, kakav je kanalizacioni sistem, neprestano menja i loše utiče na mernu opremu, pouzdanost merenja vremenom opada. Zbog toga se popravka pouzdanosti merenja mora sprovoditi redovnim održavanjem i kalibracijom mernih uređaja. Službe za održavanje mernih sistema u kanalizaciji ponekad broje i više desetina tehničkih lica, a održavanje se uglavnom sprovodi prema preporuci proizvođača opreme. Pošto se oprema ugrađuje na merna mesta od kojih svako ima određene specifičnosti, na osnovu rezultata naknadnog vrednovanja podataka moguće je sprovesti analizu i proceniti adekvatan period održavanja i kalibaracije mernih uređaja, što bi čitav sistem učinilo pouzdanijim, a proces održavanja zasigurno jeftinijim.

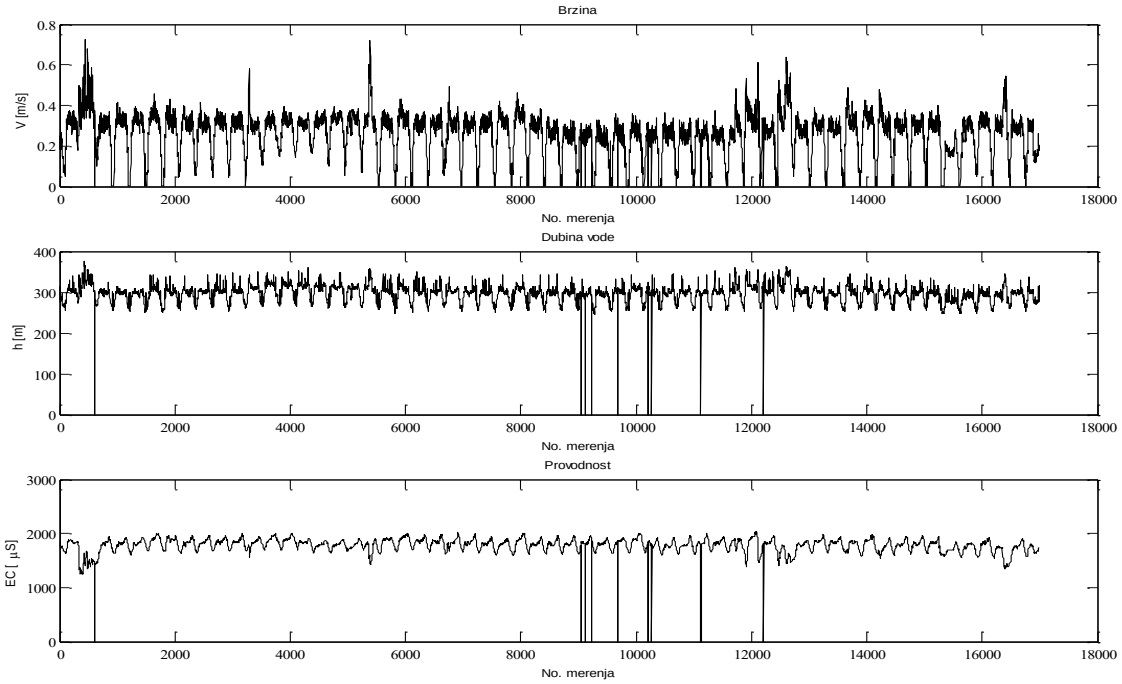
S obzirom da je za sveobuhvatno osmatranje kanalizacionog sistema potrebno merenje velikog broja hidrauličkih i hemijskih parametara (ponekad i više stotina njih), njihova provera i vrednovanje mogu se, između ostalog, sprovesti upoređivanjem raspoloživih izmerenih vrednosti. Dodatne mogućnosti u procesu validacije daje upotreba hidrauličkih modela i modela kvaliteta baziranih na fizičkim principima (*first principle based models*). Fizički bazirani hidraulički modeli, poput modela dinamičkog talasa (EpaSWMM [97]), kinematičkog talasa (EpaSWMM [97] ili BEMUS [29]), pomoću jednačina održanja mase i količine kretanja povezuju hidrauličke parametre, koji se na taj

način mogu dovesti u vezu. TakoĎe, i modeli kvaliteta (QUAL2E [118]) u sebi sadrže matematičke relacije koje oponašaju hemijske procese koji se odvijaju u kanalizacionoj vodi. Imajući u vidu da su matematički modeli samo konceptualni prikaz modeliranih procesa i fenomena i u velikoj meri zavise od količine raspoloživih informacija na osnovu kojih su formirani, u obzir se mora uzeti i neodreĎenost kako samog modela, tako i ulaznih podataka i kalibracionih parametara [16].

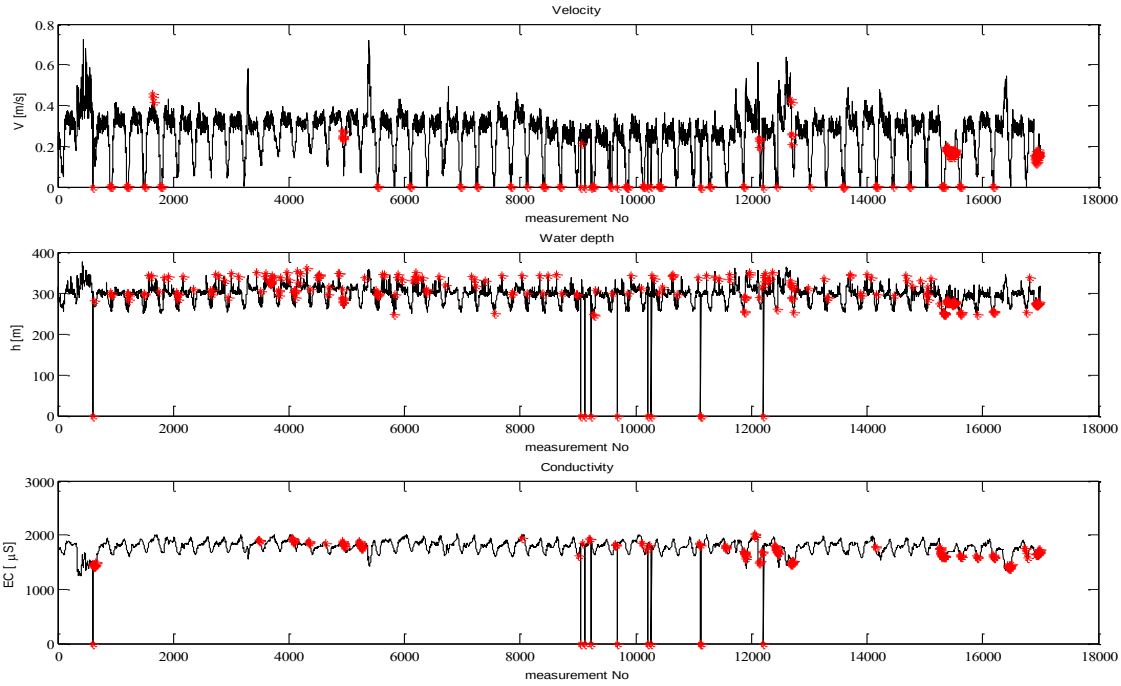
Na slici 3.15 prikazani su rezultati merenja nivoa, brzine i provodnosti pomoću automatske merne stanice postavljene na jednom od izliva beogradske kanlizacije (Višnjica). Merenje nivoa i brzine je sprovodeno ultrazvučnom metodom, dok se merenje provodnosti obavlja industrijskom mernom sondom koja je postavljena direktno u tok kanalizacione vode. Merenja se obavljaju i šalju u sabirnu bazu podataka u intervalima od po 5 minuta.

Neke od anomalija se mogu uočiti i vizuelnom inspekcijom podataka, dok je za detekciju ostalih, manje uočljivih anomalija potrebno upotrebiti specifične matematičke alate (slika 3.16). Povremeno registrovanje nula vrednosti, anomalije u vidu pikova, povećan šum, itd, stalni su pratilac rezultata merenja. Uzrok navedenih anomalija može biti različit, od povremenih grešaka u radu sistema, do ispada sistema zbog prestanka napajanja ili kvara mernog uređaja.

Kod velikog broja podataka (ukoliko je interval uzorkovanja 5 minuta, u bazu podataka pristiže 12x24=288 podataka u jednom danu sa jednog mernog instrumenta) vrednovanje podataka je potrebno sprovesti automatski ili poluautomatski jer se, zbog količine potaka koji pristižu, ne može očekivati da se validacija obavi ručno. Sistem za automatsko vrednovanje podataka se mora osmisliti i dizajnirati tako da daje maksimalno dobre rezultate bez čestog uključivanja eksperta u proces podešavaja parametara metoda vrednovanja.



Slika 3.15: Rezultati merenja hidrauličkih parametara i parametara kvaliteta vode u kanalizacionom sistemu sa vremenskim korakom Δt=5min



Slika 3.16: Rezultati vizuelnog vrednovanja podataka izmerenih na beogradskom kanalizacionom sistemu sa vremenskim korakom Δt=5min (markerima su obeleženi podaci sa greškom)

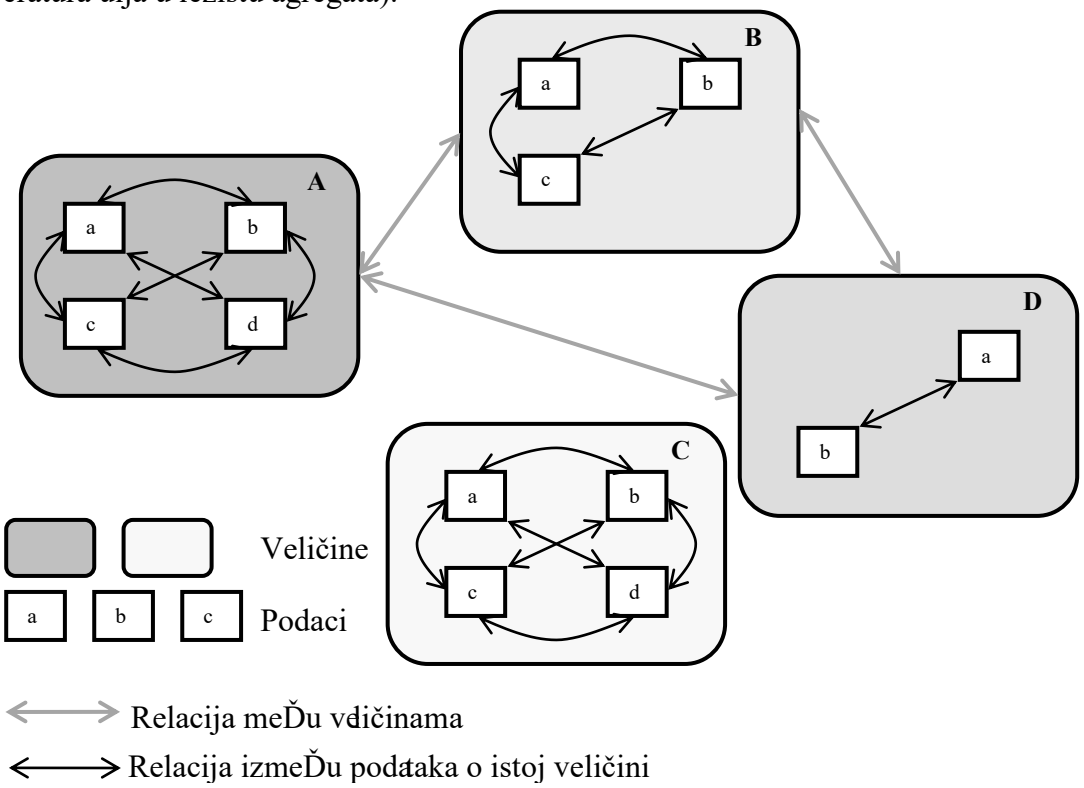
U ovom poglavlju prikazana su tri primera koja ilustruju prikupljanje podataka u tri različita sistema. Sistem za vrednovanje podataka prikazan u ovoj disertaciji podrazumeva postojanje relacija između podataka, pa se istraživačka merenja, čije je rezultate tek potrebno protumačiti i preinačiti u matematičke relacije, ne mogu ovim sistemom vrednovati. Ipak, pošto se istraživačka merenja uglavnom sprovode u laboratorijama pri kontrolisanim uslovima, eksperiment je često moguće ponoviti pri sličnim uslovima, što omogućava uspostavljanje relacija između ponovljenih rezultata merenja i vrednovanje podataka predloženim sistemom.

3.4 Relacije između podataka

Da bi se greške u toku merenja detektovale i kvantifikovale, potrebno je problemu pristupiti sa dve strane: sa strane neodređenosti i procenjenih grešaka kod izmerenih podataka analizom merne procedure, i sa strane relacija koje se mogu uspostaviti između izmerenih podataka i raspoloživih informacija. U opštem slučaju podatak nije usamljen, već je deo grupe informacija koja postoji o veličini koju reprezentuje. Na slici 3.17 prikazana je opšta slika sistema za osmatranje. Osmatra se određeni broj veličina (A, B, C, D, itd.) različitim metodama (a, b, c, d, itd.). Postoji maksimalno $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ $(r=2)$ relacija koje je moguće uspostaviti između podataka (ukoliko se pretpostavi da između svaka dva podatka postoji jedna veza). Postojanje i broj relacija između veličina, sa druge strane, zavisi od mnogih faktora: postojanja fizičke ili logičke zavisnosti, lokacije merenja, itd. Relacije između podataka omogućavaju formulisanje metoda (alata), čiji će rezultat (koji se naziva predikcija) da pruži informaciju o oblasti gde se očekuje vrednost nekog podatka.

Primer jednog hidrotehničkog sistema, prema kome bi veličina A bila snaga na agregatima merena pomoću četiri merne metode, veličina B nivo meren pomoću tri merne metode, a veličina D protok meren pomoću dve merne metode. Veličina označena sa C predstavlja veličinu koja nije u vezi sa veličinama A, B i D, bilo zato što je srodna veličina koja je po prostoru merena na lokaciji takvoj da

se pomenuta veza gubi, ili predstavlja veličinu koja nije u fizičkoj vezi sa prve tri veličine (na primer, C je temperatura ulja u ležištu agregata).



Slika 3.17: Šematski prikaz sistema za osmatranje

Pored relacija između merenih podataka, mora se naglasiti i postojanje relacija između merenih podataka i podataka o fizičkom sistemu u kom su oni izmereni. Te relacije se, pre svega, odnose na fizička i istorijska ograničenja koja diktiraju sistem i osobine veličina koje se mere. Prema datom primeru, ograničenja za veličinu A bi bila vezana za instalisanu snagu turbina i principe upravljanja, za veličinu B bi bila vezana za kotu normalnog uspora brane i minimalnu kotu u akumulaciji, dok bi za veličinu D to bila ograničenja vezana za propusnu moć sprovodnih objekata.

Za razliku od činjenice da se na greške i neodređenost u podacima može uticati izborom merne opreme, adekvatnom obradom sirovih podataka, itd, na osnovu relacija između podataka moguće je jedino proceniti njihov kvalitet (ne može se uticati na povećanje kvaliteta podatka), što je često i dovoljno.

Vrednovanje podataka upravo predstavlja proces kvantifikovanja kvaliteta podatka na osnovu relacija razmatranog podatka sa podacima iste vremenske serije, podacima drugih vremenskih serija i ostalim korisnim informacijama. Može se zaključiti da se bez dodatnih podataka i relacija između podataka ne može sprovesti vrednovanje, pa je samim tim usamljeni podatak lišen mogućnosti da bude proveren. Metodologija vrednovanja podataka prikazana u ovoj doktorskoj disertaciji je stoga razvijena u skladu sa sledećim principom:

> Metodologija za vrednovanje podataka mora da omogući kombinaciju svih raspoloživih relacija, kako onih koje se mogu uspostaviti između samih merenih podataka, tako i onih koje se mogu uspostaviti između merenih podataka i informacija o fizičkim i ostalim karakteristikama sistema.

3.4.1 Relacije između podataka – matematički prikaz

Relacije između podataka, u hidrotehnici poznate i kao matematički modeli, predstavljaju veze koje postoje između podataka, takve da se pomoću neke karakeristike (*feature*) jednog podatka može opisati neka karakteristika drugog podatka sa kojim je u relaciji. Relacije između podataka se mogu opisati pomoću opšte formule:

$$R: \quad F(Z,\theta),$$

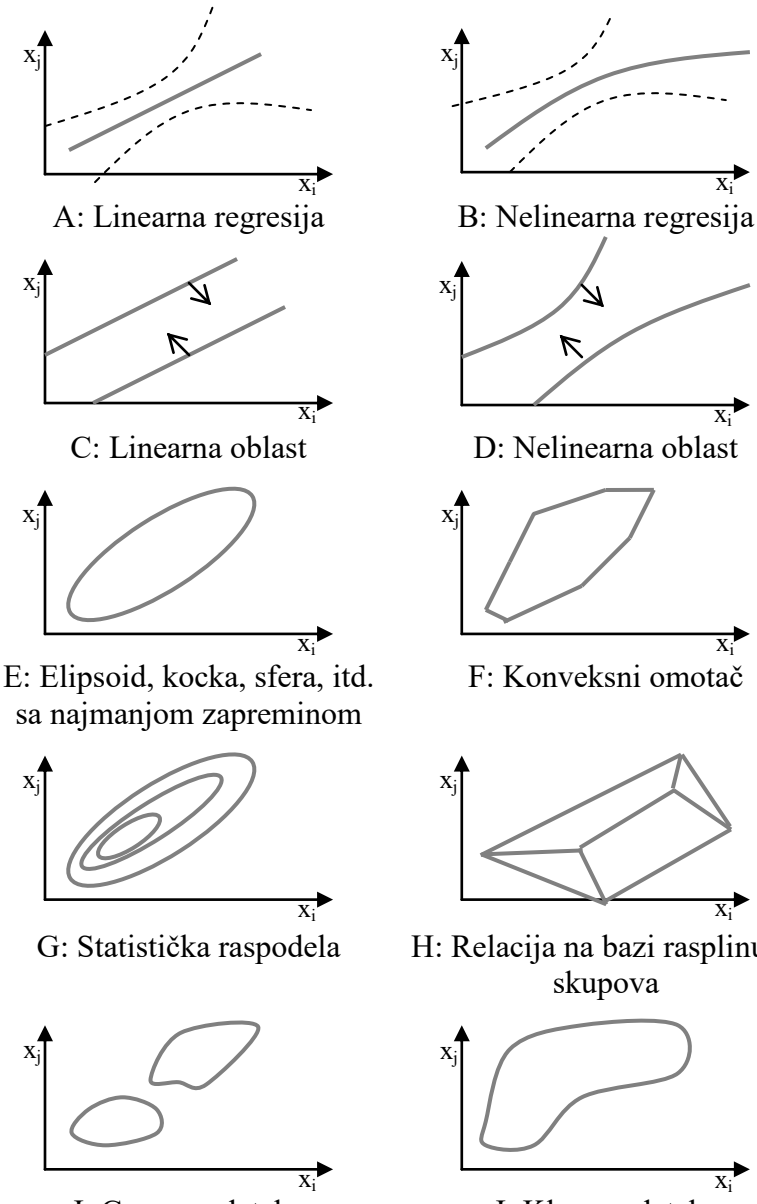
gde je F skup različitih funkcija u relaciji f_i, Z niz veličina koje učestvuju u relaciji, a θ niz parametara relacije čije se vrednosti dobijaju kalibracijom. U procesu upotrebe relacije za potrebe predikcije, veličine koje učestvuju u relaciji (Z) dele se na ulazne veličine (X) i veličine koje predstavljaju rezultat (Y):

$$M_{R,Y}: \quad Y^{M_R} = F(X,\theta),$$

gde $M_{R,Y}$ predstavlja jedan oblik relacije R (koji se može nazvati metoda) kod koje je X niz ulaznih vrednosti relacija, a Y^{M_R} niz rezultata metode $M_{R,Y}$. U hidrotehničkoj praksi se metode iz relacija izvode prema praktičnim potrebama i mogućnostima, pa se, na primer, iz relacije koja vezuje kišu i oticaj, kiša uglavnom smatra ulaznom veličinom, dok se oticaj izračunava. Takva relacija se koristi u situacijama kod kojih je oticaj teško ili nemoguće izmeriti. Ukoliko postoje podaci i o kiši i o oticaju, što je slučaj kada se sprovodi vrednovanje podataka, mora se razmatrati i druga formulacija relacije, tj. kolika bi bila kiša koja bi izazvala izmereni oticaj.

Na slici 3.18 prikazano je deset načina na koje je moguće matematički povezati podatke pomoću relacija. Relacije su prikazane u dve dimenzije zbog toga što se one najjednostavnije mogu izraziti na dvodimenzionalnom medijumu kao što je papir ili monitor (vizuelno je moguće prikazati tri dimenzije, dok je upotrebom boja, oblika, veličina ili simbola grafički prikaz moguće obogatiti i većim brojem dimenzija).

Svaki od navedenih načina matematičkog iskazivanja relacija između podataka može se proširiti na proizvoljan broj dimenzija (ili smanjiti na jednu dimenziju). Razlog zbog koga su linearna regresija i linearna oblast izdvojene, umesto da se tumače kao specijalni slučajevi nelinearne regresije i nelinearne oblasti, jeste niz razvijenih matematičkih postupaka koji su zasnovani na linearnoj vezi između podataka (intervali poverenja, PCA, itd.).



A: Linearna regresija B: Nelinearna regresija

C: Linearna oblast D: Nelinearna oblast

E: Elipsoid, kocka, sfera, itd. sa najmanjom zapreminom F: Konveksni omotač

G: Statistička raspodela H: Relacija na bazi rasplinutih skupova

I: Grupe podataka J: Klase podataka

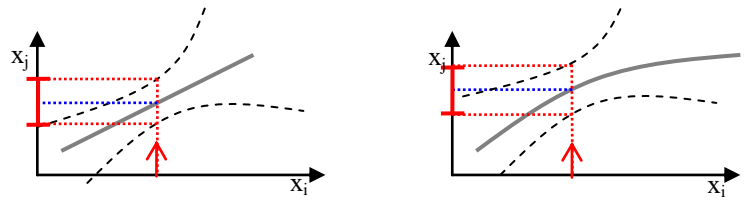
Slika 3.18: Matematička formulacija relacija između podataka

Relacija između podataka se može predstaviti kao put kojim se neka karakteristika jedne veličine transformiše u neku karakteristiku iste ili neke druge veličine. Važno je napomenuti da se u relacijama ne kriju isključivo numeričke vrednosti podataka koji se razmatraju, već se relacijama mogu tretirati podaci iskazani u drugim oblicima, kao što je klasa kojoj podatak pripada (slika 3.18J). Do konačnog oblika navedenih relacija (algoritma) može se doći na različite načine. Neki od njih su prikazani u tabeli 3.2.

Tabela 3.2: Neki načini odreĎivanjaparametara modela (relacija)

Tip relacije	Način formiranja relacije
Linearna regresija	• Metoda najmanjih kvadrata • *Maximum likelihood estimation* • *Principal component regression* • itd.
Nelinearna regresija	• Metoda najmanjih kvadrata • ANN • SVR • itd.
Linearna oblast	• Formiranje oblasti na osnovu linearne regresije • SVR • itd.
Nelinearna oblast	• Formiranje oblasti na osnovu linearne regresije • Intervali poverenja • SVR • itd.
Elipsoid, kocka, sfera, itd. najmanje zapremine	• *Rotating calipers method* • *An $O(n^2)$-time algorithm* • *Applet's algorithm* • *Megiddo's linear-time algorithm* • itd.
Konveksni omotač	• *Jarvis march algorithm* • *Graham scan algorithm* • *Divide and conquer algorithm* • itd.
Statistička raspodela	• *Goodness of fit methods*
Rasplinuti skupovi	• *Fuzzy logic methods*
Grupe podataka	• *k-means* • *V-means* • QT • *Fuzzy c-means* • itd.
Klase podataka	• ANN • SVM • *Naive Bayes* • *k Nearest Neighbor* • itd.

Ukoliko se podaci posmatraju numerički (i kvalitativni podaci se mogu transformisati u numeričku formu), može se zaključiti da je jedinica informacije koja se prenosi u svim prikazanim relacijama interval. Egzaktna vrednost je samo specijalan slučaj koji se može ostvariti samo kod linearne ili nelinearne regresije (slika 3.19A i 3.19B), uz zanemarivanje neodreĎenosti samog modela. Pored intervala, na primer, kod statističkih raspodela, jedinica informacije obogaćena je i podatkom o verovatnoći (slika 3.19G).



A: Linearna regresija

B: Nelinearna regresija

C: Linearna oblast

D: Nelinearna oblast

E: Elipsoid, kocka, sfera, itd. sa najmanjom zapreminom

F: Konveksni omotač

G: Statistička raspodela

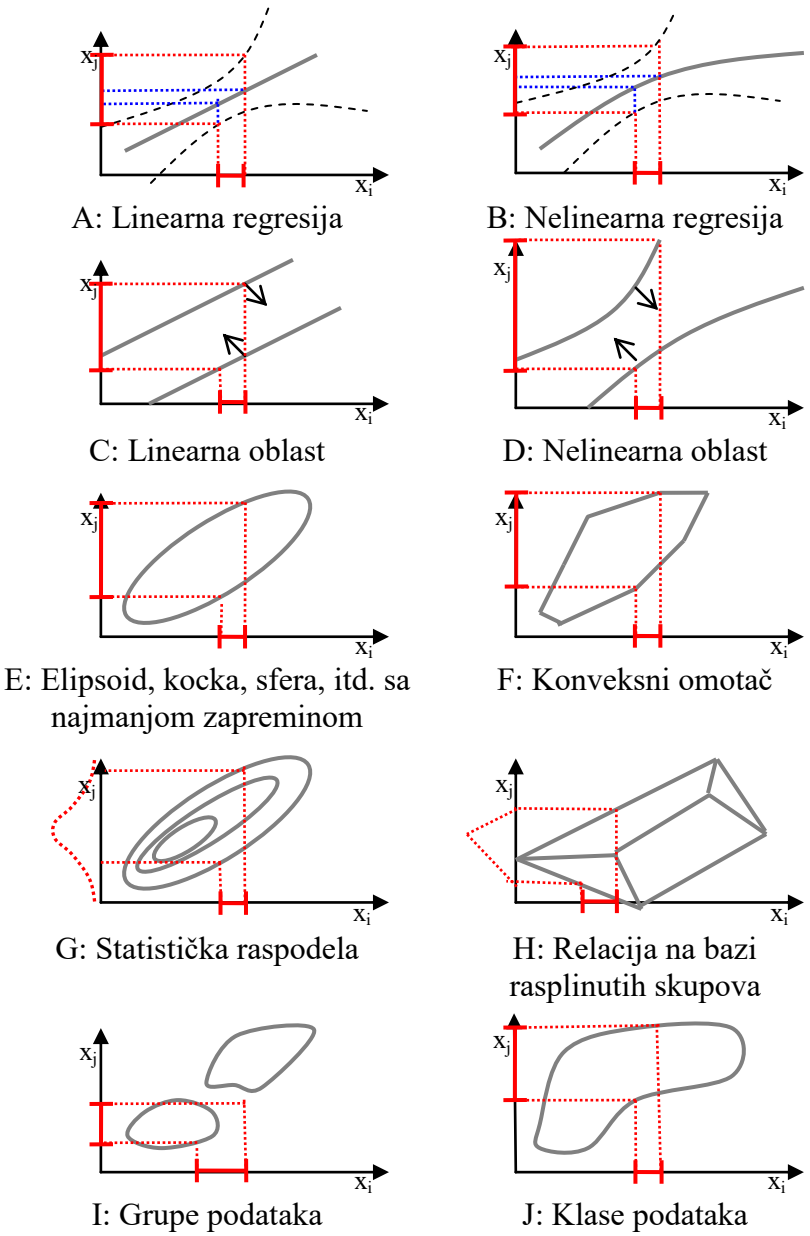
H: Relacija na bazi rasplinutih skupova

I: Grupe podataka

J: Klase podataka

Slika 3.19: Interval x_j kao rezultati primene kada je ulazna veličina x_i egzaktna

U praksi, ulazne vrednosti su neodreĎene. Na slici 3.9 prikazana je propagacija neodreĎenih ulaznih vrednosti u obliku intervala. Na slici je naglašena i greška koja se javlja ukoliko se ne uzme u obzir neodreĎenost samog modela (slika 3.20A i 3.20B) kod linearne i nelinearne regresije.



Slika 3.20: Rezultati primene relacija ukoliko je ulazna veličina x_i u obliku intervala

Najmanju jedinicu informacije, interval, relacije mogu obogatiti dodatnim informacijama kojima se mogu bliže opisati neke karakteristike neodređene veličine. Dodavanjem funkcije pripadnosti u relaciji na bazi rasplinutih skupova (*membership function*) moguće je formirati rasplinuti skup (slika 3.20H), a dodavanjem funkcije gustine verovatnoće, raspodelu verovatnoće (slika 3.20G). Pored podele prema obliku u kom se mogu matematički prikazati, relacije se mogu podeliti i na:

1. relacije bazirane na fizičkim zakonima;
2. statističke relacije;
3. *data mining* relacije.

Relacije bazirane na fizičkim zakonima svode se na zakone održanja i obično se matematički formulišu sistemima diferencijalnih jednačina. Diferencijalne jednačine se dalje rešavaju, a rezultat je u vidu regresije. Statističke relacije podrazumevaju uvođenje pojma verovatnoće. Rezultat

statističkih modela su verovatnoće vrednosti razmatranih veličina, regresione krive ili oblasti (slika 3.18A) dobijene uz određene statističke pretpostavke (npr. greška raspoređena po normalnoj raspodeli). *Data mining* predstavlja proces otkrivanja određenih pravila ili šablona u podacima. Metodama klasifikacije ili grupisanja podaci se svrstavaju u određene klase ili grupe na osnovu svojih karakteristika ili se formira regresiona kriva koja aproksimira vezu između podataka.

Oblik koji se odabere za potrebe formiranja relacije između podataka zavisi, pre svega, od postojanja veze između podataka – fizičke, statističke ili neke koju je tek potrebno otkriti *data mining* metodama.

3.4.2 Neizvesnost i greške ulaznih veličina, parametara modela i relacija između podataka

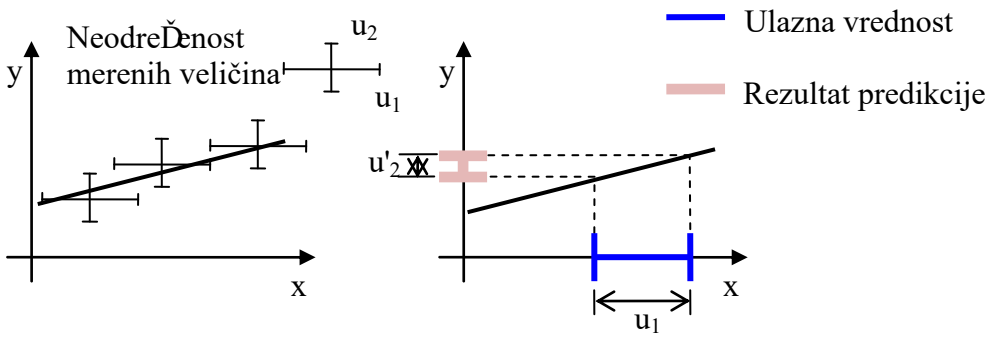
Ključni faktor kod vrednovanja podataka predstavljaju relacije između podataka (u hidrotehnici se još zovu i matematički modeli). Relacija između podataka predstavlja zavisnost kojom se karakterisika od interesa nekog podatka (najčešće njegova vrednost) može izračunati na osnovu poznavanja odgovarajućih informacija od kojih ova zavisi (karakteristika drugih podataka).

Osnovno svojstvo svih veličina koje figurišu u matematičkom modelu koji je u sastavu metode za vrednovanje podataka je postojanje neodređenosti i potencijalne greške. Postojanje potencijalne greške i neodređenosti može se pretpostaviti kod svih komponenti neke relacije: ulaznih vrednosti, parametara i matematičkih jednačina. Da bi se procesom vrednovanja detektovale greške u podacima, potrebno je smanjiti mogućnost da greške jednačina na neki način doprinesu intenzitetu ili umanje intenzitet grešaka u rezultatima, već da se na rezultate reflektuju isključivo greške u ulaznim podacima. Zbog toga je neophodno smanjiti na minimum postojanje grešaka u jednačinama relacija koje se koriste.

Sa druge strane, formiranje relacija uglavnom podrazumeva konceptualnu prirodu razmatranog problema. Konceptualizacijom problema modeliraju se samo najznačajniji uticaji, dok se posredni, manje bitni uticaji zanemaruju i uglavnom se smatraju šumom. S obzirom na to da na merene podatke utiču i ti posredni uticaji koji se ne uključuju u konceptualnu sliku problema, adekvatna reprezentacija matematičkog modela mora da uvrsti i te posredne nemodelovane uticaje. Jedan od načina je kroz modeliranje neodređenosti samog modela, pored neodređenosti ulaznih veličina.

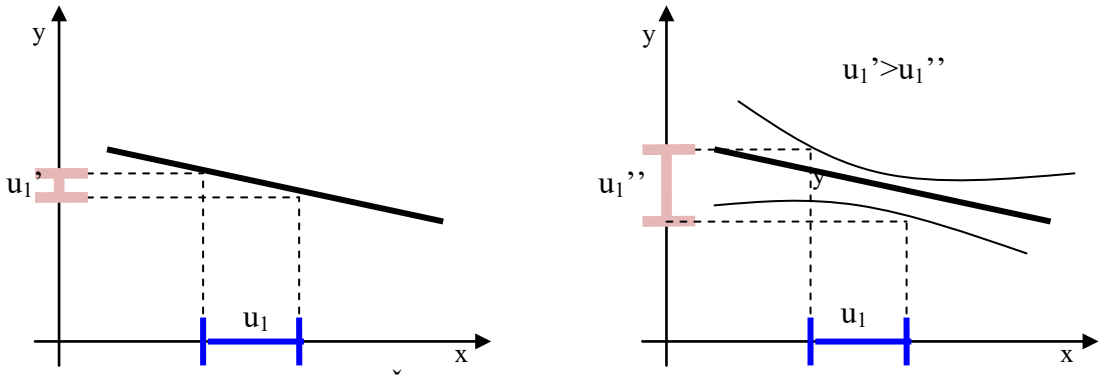
Uvrštavanjem neizvesnosti u komponete matematičkih modela, modeli prestaju da budu deterministički, tj. za iste ulazne vrednosti ne dobijaju se pri svakoj simulaciji isti rezultati. Ukoliko su matematički modeli formirani u determinističkom obliku, a parametri i ulazne veličine u neodređenom obliku, metodama za propagaciju neodređenosti može se obaviti transfer neizvesnih ulaznih veličina i parametara u neizvestan rezultat determinističkog modela [16].

Često se zanemaruje postojanje grešaka i neizvesnosti kod samog modela, tj. matematičkih jednačina i nejednačina koje ga sačinjavaju. Zbog toga se, na primer, u deterministčkim regresionim modelima sa izvodom manjim od jedinice može dogoditi paradoks da izračunate vrednosti (rezultati predikcije) imaju manju neodređenost od ulaznih veličina, ako se u obzir uzme samo neizvesnost ulaznih veličina, ali ne i neizvesnost samog matematičkog modela. Ovo je prikazano na slici 3.20, na kojoj se vidi kako je model nastao i kako se pri upotrebi modela za ulazne vrednosti x neodređenosti u_1 kao rezultat y dobijaju vrednosti manje neodređenosti nego što je neodređenost veličina y, $u_2' < u_2$. Ovaj paradoks nastaje zato što je pri formiranju modela informacija o neodređenosti veličine y osrednjena i na taj način izgubljena.



Slika 3.20: Propagacija neodređene veličine prikazane u obliku intervala preko linearnog modela

Na slici 3.21 je prikazan regresioni model sa rezultatima u slučaju kad u model nije uključena neodređenost (slika 3.21A), i u slučaju kad jeste (slika 3.21B). Može se primetiti da se neodređenost rezultata modela povećava.



A: Linearni model bez neodređenosti samog modela

B: Linearni model sa prikazanom neodređenosti samog modela

Slika 3.21: Transformacija neodređenog podatka u neodređeni rezultat

Grešku modela je teško razdvojiti od neodređenosti modela, tj. teško je razdvojiti da li je uzrok neadekvatnog rešenja modela njegova neodređenost ili greška. Uzroci postojanja grešaka i neodređenosti mogu se svrstati u više grupa. Neke od njih su:

1. upotreba uzorka umesto cele populacije podataka na osnovu kojih se formira ili kalibriše model (npr. intervali poverenja statističkih modela);
2. konceptualizacija modeliranog fenomena i konceptualni prikaz matematičkim aparatom;
3. kalibracija modela, tj. podešavanje parametara modela.

Problem koji nastaje usled toga što se kod linearne i nelinearne regresije regresiona linija formira pod pretpostavkom da se njom predstavljaju relacije razmatranih veličina dobijenih iz uzorka, može se rešiti formiranjem intervala poverenja. Intervali poverenja pružaju mogućnost da se, na osnovu činjenice da se u procesu formiranja modela koristio samo raspoloživi uzorak iz populacije podataka, definiše oblast u kojoj je moguće postojanje rešenja. Neizvesnost i greška samog matematičkog modela leže i u tome što je matematički model samo konceptualni prikaz modeliranog procesa, o kome često nisu ni dostupne sve informacije. Adekvatnim izborom tipa modela moguće je smanjiti neodređenost modela, ali podatak o neizvesnosti rezultata često nije dostupan. U susret ovom problemu izašla je grupa autora [53] koji su primenili Bajesovu statističku teoriju i osrednjavanjem većeg broja modela (*Bayesian model averaging*, BMA) pokušali da izračunaju matematičko očekivanje i varijansu rezultata modela. S obzirom na to da jednačine matematičkog modela predstavljaju samo njegov okvir, kalibracijom se određuju vrednosti parametara, tj. smanjuje se

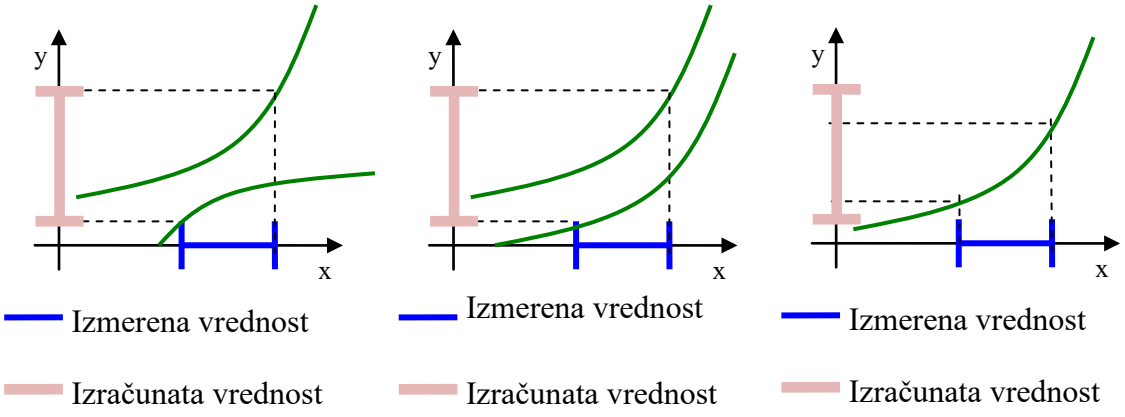
greška modela. Sa druge strane, neodređenost podataka koji se koriste za kalibraciju mora se uračunati u sam model.

3.4.3 Modeliranje neodređenosti modela

U procesu vrednovanja podataka relacije između podataka (matematički modeli) omogućavaju da se izračuna mereni podatak koji je potrebno vrednovati. Mapiranjem neodređenih ulaznih vrednosti preko matematičkih jednačina matematičkog modela dobijaju se i neodređene vrednosti rezultata. Neizvesnost (neodređenost) i greške modela mogu ponekad značajno uticati na rezultate modela.

Neodređenost samih jednačina predstavlja poseban izazov za modeliranje i može se, između ostalog, sprovesti na tri načina (slika 3.22):

1. formiranjem modela u neodređenom obliku;
2. izračunavanjem parametara modela u neodređenom obliku, tako da se u njima sadrži i neodređenost jednačina modela; i
3. kompenzacijom, tj. dodavanjem neodređenosti na rezultat modela.



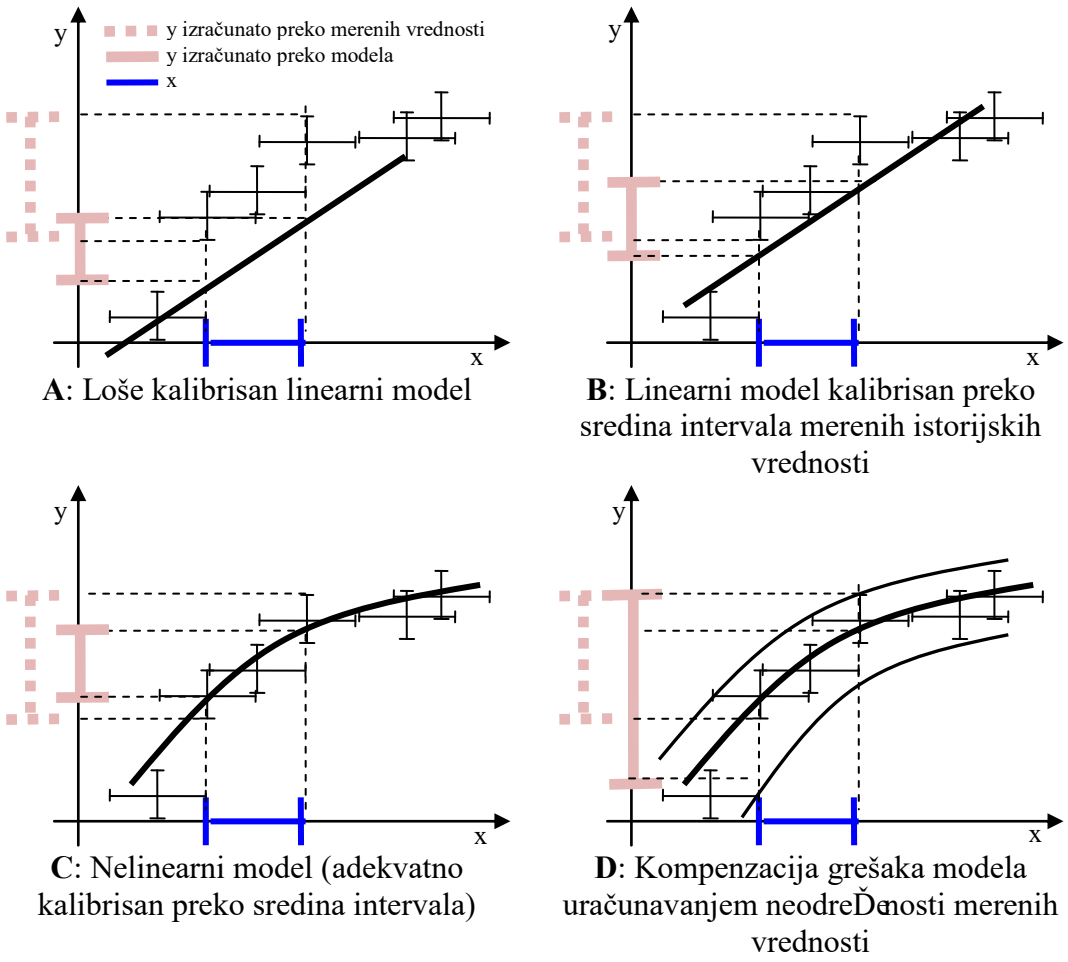
Slika 3.22: Modeliranje neodređenosti matematičkih modela

3.4.4 Greške modela

Relacije koje se mogu uspostaviti između podataka sadrže izvesnu neodređenost, ali mogu sadržati i greške. Neodređenost u relacijama potiče od nemogućnosti da se adekvatno opišu svi procesi koji učestvuju u vezi između dve veličine, a izražava se konceptualnom prirodom modela. Greške samih relacija potrebno je razdvojiti od grešaka rezultata koje mogu biti uzrokovane izborom približne numeričke metode ili greškama u ulaznim vrednostima ili parametrima modela (uticaj grešaka u podacima na parametre linearnih i nelinearnih regresionih modela detaljno je opisan u [39]).

Greške u rezultatima zbog modela, između ostalog, potiču i od neadekvatne kalibracije modela i gubitka informacija konceptualizacijom strukture modela. Gubitak informacija konceptualizacijom se može povezati sa izborom samog koncepta modela i neuključivanjem neizvesne prirode podataka koji se koriste za formiranje modela ili njegovu kalibraciju. Prilikom izbora načina konceptualizacije obično se žrtvuje neki aspekt procesa koji povezuje merene veličine, čime se delom unosi i neizvesnost u sam model. Sa druge strane, neuključivanjem neodređene prirode podataka koji se koriste za formiranje ili kalibraciju modela gubi se informacija o neodređenosti koju podaci sa sobom nose. Na slici 3.23 prikazani su neki regresioni modeli koje je moguće formirati na osnovu

merenih istorijskih vrednosti. Potrebno je primetiti da je ulazna veličina tako odabrana da postoje merene vrednosti koje odgovaraju njenim granicama.



A: Loše kalibrisan linearni model

B: Linearni model kalibrisan preko sredina intervala merenih istorijskih vrednosti

C: Nelinearni model (adekvatno kalibrisan preko sredina intervala)

D: Kompenzacija grešaka modela uračunavanjem neodređenosti merenih vrednosti

Slika 3.23: Greške modela i kompenzacija grešaka

Na slici 3.23A prikazan je linearni model koji nije adekvatno kalibrisan. Ulazna vrednost u obliku intervala se transformiše u rezultat, takođe u obliku intervala. Rezultat sadrži grešku koja proističe iz činjenice da se deo informacija koje se odnose na postojanje neodređenosti kod podataka na osnovu kojih se formira model gube ovakvim načinom modeliranja. Na slici 3.23B prikazan je linearni model, adekvatno kalibrisan na osnovu sredina intervala neodređenosti merenih veličina kojim se greška smanjuje, ali ne i otklanja. Kalibracijom matematičkog modela kojim se izračunava vrednovani podatak otklanja se sistematska greška modela (*bias*), ali i dalje postoji neizvesnost samog modela i druge greške. Na slici 3.23C prikazan je nelinearni model kojim se greška dodatno smanjuje, ali ona ne može biti manja od $0.5u_y$, gde u_y predstavlja širinu intervala u y pravcu.

Greška modela se može opisati na različite načine, kao što su statistički (srednja vrednost i varijansa), širina intervala maksimalne greške, itd. Postojanje greške u modelu uglavnom je neprihvatljivo, pa se one mogu kompenzovati uračunavanjem granica neodređenosti. S obzirom da i kalibracione vrednosti imaju neodređenost, potrebno ju je uračunati u neodređenost samog modela. Na slici 3.23D prikazan je model kod koga je greška kompenzovana uračunavanjem neodređenosti merenih vrednosti.

Prirodno se nameće pitanje izbora odgovarajuće relacije između podataka. Jednostavan odgovor bi bio da je za potrebe vrednovanja potrebno odabrati onu relaciju koja ima najmanju neodređenost i

najmanju grešku. Ovakav princip nije najbolji u opštem slučaju izbora relacija između podataka. Na primer, ponekad se koncept same relacije koristi za isticanje nekih karakterisitka podataka ili za uklanjenje nekih neželjenih efekata. Tako se kod formiranja kalibracione krive regresionom metodom kojom se pretpostavlja raspored odstupanja podatka od rezultata regresije u vidu tzv. normalne raspodele $N(0,\sigma)$ posredno uklanjaju pretpostavljene slučajne greške u podacima.

Relacije između podataka i njihova matematička formulacija predstavljaju ključni element sistema za vrednovanje podataka predstavljenog u ovoj doktorskoj disertaciji. Formiranje relacija između podataka, kalibracija matematičkih modela proizašlih iz relacija i upotreba kalibrisanih modela podrazumevaju kompenzaciju neodređenosti samog modela, kao i ulaznih podataka i parametara modela. Ovakav pristup modeliranju predstavlja novinu nasuprot tradicionalnom pristupu kod koga se kalibracija matematičkog modela sprovodi uz pretpostavku da u kalibracionim podacima postoji isključivo slučajna greška.

3.5 Kvalitet merenih podataka

Pre nego što se počne sa vrednovanjem podataka potrebno je odrediti metrički sistem po kome će se taj proces sprovesti. Naime, vrednost podatka (*value of data*) nije jednoznačna i jedan podatak može imati više vrednosti. Vrednost koja je reprezentativna za pojedinog korisnika i vrstu odluke koju je potrebno doneti određuje se na osnovu poznavanja čitavog procesa koji je podatak prošao i upotrebe za koju je namenjen. Vrednost podatka se, dalje, može povezati sa opštim terminom – kvalitetom podatka.

Prema dokumentima [116] i [117] Agencije za zaštitu životne sredine EPA, kvalitet podataka je definisan kao "skup osobina i karakteristika podataka koje se zasnivaju na mogućnostima da zadovolje očekivanja i potrebe korisnika, tj. kupca". Korisniku su podaci potrebni da bi doneo određene odluke i upravo pogodnost podatka da pomogne u donošenju ispravne odluke direktno je vezana za njegov kvalitet [24]. Ovakva postavka problema kvaliteta podataka vodi ka potrebi da bude definisan kvalitet odluke. Opšta karakteristika odluke koja je direktno vezana za kvalitet jeste ispravnost odluke. Nažalost, u velikom broju slučajeva ispravnost odluke nije poznata u momentu donošenja odluke, već se ona može oceniti tek kasnije prema posledicama koje je odluka proizvela. Nešto slabija karakteristika odluke, koja se može proceniti i u trenutku njenog donošenja, jeste neranjivost odluke [24]. Neranjivost odluke ogleda se, pre svega, u nepostojanju objektivnih tehničkih, naučnih, ekonomskih ili političkih prepreka da se odluka sprovede. Ukoliko je potrebno, mogu se primeniti i strožiji kriterijumi, na primer, ekonomski, socijalni, itd, i u skladu sa njima sprovesti procena neranjivosti odluke. Provera kvaliteta podataka (validacija podataka) je, prema tome, proces kojim se ocenjuje pogodnost podataka za potrebe donošenja neke odluke.

U literaturi se sreće nekoliko definicija kvaliteta podataka:

1. GIS Rečnik[3]: kvalitet podataka se odnosi na stepen izuzetnosti, demonstriran od strane podataka u odnosu na prikazivanje konkretnog stvarnog fenomena.
2. Vlada Britanske Kolumbije[4]: stanje potpunosti, validnosti, konzistentnosti i tačnosti koje čini podatke pogodnim za konkretnu upotrebu.

[3] *GIS Rečnik* - http://www.fw.umn.edu/FW5620/glossary.htm
[4] Vlada Britanske Kolumbije- http://www.cio.gov.bc.ca/other/daf/IRM_Glossary.htm

3. Rečnik izraza kontrole kvaliteta[5]: ukupnost osobina i karakteristika podataka koja se odnosi na njihovu sposobnost da zadovolje određenu namenu; zbir stepena izuzetnosti za faktore koji su povezani sa podacima.
4. Rečnik izraza o kvalitetu podataka[6] objavljen od strane Međunarodne asocijacije za kvalitet informacija i podataka (IAIDQ)[7]: pogodnost podataka za korišćenje; informacije koje ispunjavaju zahteve svojih autora, korisnika i administratora.

U samim definicijama koje potiču iz različitih izvora vidi se da je kvalitet podataka usko vezan za korišćenje podatka, tj. odluku koja se na osnovu podatka donosi. To znači da jedan podatak ima različit kvalitet za dobijanje jedne informacije u odnosu na dobijanje neke druge informacije.

Dakle, podatke možemo vrednovati na osnovu tri kriterijuma:

1. koliko je podatak redak;
2. koliki je rizik za upotrebu (informaciju koja se dobija) za koju je namenjen ukoliko podatak nije odgovarajući; i
3. tačnost i preciznost kojom reprezentuje pojavu koju predstavlja.
4.

Prema prvom kriterijumu, vrednost podatka raste ukoliko je podatak redak, tj. ukoliko je pojava koju reprezentuje neponovljiva ili retko ponovljiva. Prema tom kriterijumu najmanju vrednost imaju podaci koji su dobijeni u laboratorijama. Naime, uslovi koji se u laboratorijama mogu ostvariti omogućavaju da se svaka pojava može ponoviti. Izuzetak, naravno, predstavljaju eksperimenti koji se mogu obaviti samo ograničen broj puta (npr. jednom) i slučajna otkrića. Veću vrednost od laboratorijskih, prema prvom kriterijumu, imaju podaci dobijeni u realnim sistemima pri praćenju i osmatranju procesa. Svaka pojava se u tom slučaju može posmatrati kao neponovljiva i svaka istorijska vrednost dobijena osmatranjem je praktično mogla biti izmerena samo jednom. Najveću vrednost imaju registrovani retki fenomeni u realnim sistemima ili prirodi. Vrednost tih pojava je uglavnom neprocenjiva (npr. katastrofalne poplave, havarije, itd.).

Prema drugom kriterijumu, vrednost i važnost odluke koja se donosi na osnovu razmatranog podatka ima ključnu ulogu. Potrebno je posmatrati to koliko je podatak ključan za donošenje informacije i koliko je informacija osetljiva na razmatrani podatak. Retka hidrološka pojava može potpuno izmeniti sliku o hidrologiji nekog područja i dovesti do informacija koje vode odlukama koje mogu inicirati velike investicije u infrastrukturu. Jedan od načina kako se može vrednovati podatak prema drugom kriterijumu je tzv. Bajesov faktor rizika [78]:

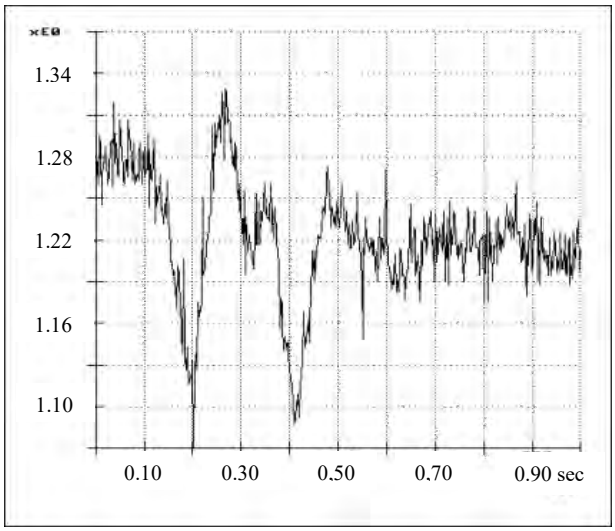
$$\tau = \frac{(C_{10} - C_{00})P_{H_0}}{(C_{01} - C_{11})P_{H_1}},$$

gde su C_{00} i C_{11} dobit ako se se postupi ispravno (podatak koji je od značaja se prihvati kao da je od značaja i obrnuto), C_{10} i C_{01} cena ukoliko je odlučeno pogrešno, a P_{H_0} i P_{H_1} su uslovne verovatnoće da se na osnovu raspoloživih podataka d došlo do zaključka da je razmatrani podatak od značaja (H_0) ili ne (H_1):

[5] Rečnik izraza kontrole kvaliteta - http://www.hanford.gov/dqo/glossaries/Glossary_of_Quality_Assurance_Terms1.pdf
[6] Rečnik izraza o kvalitetu podataka - http://iaidq.org/main/glossary.shtml
[7] Međunarodne asocijacije za kvalitet informacija i podataka (IAIDQ) - http://iaidq.org/

$$P_{H_0} = p(d|H_0) \text{ i } P_{H_1} = p(d|H_1) .$$

Treći kriterijum razmatra tačnost i preciznost podatka kojom se opisuje osmatrana pojava. Prema ovom kriterijumu, podatak se vrednuje ne samo u odnosu na to koliko je tačna vrednost veličine koja oslikava neku pojavu, već i da li je ona za nju reprezentativna, tj. da li pokazuje ono što se od nje i očekuje. Ponekad se neki detalj u podacima može smatrati šumom ili greškom ukoliko se merenje razmatrane pojave obavlja neadekvatno.



Slika 3.24: "Uhvaćen" relativno mali hidraulički udar pri merenju pritiska u distributivnoj vodovodnoj mreži

Na slici 3.24 prikazan je rezultat merenja pritiska u distributivnoj vodovodskoj mreži sa detaljem hidrauličkog udara, verovatno izazvanim manipulacijom zatvarača. Ukoliko cilj nije upotreba podataka za analizu hidrauličkog udara (ispitivanje gubitaka u mreži, opasnost od zagaĐenja vode u mreži, itd.) karakterističan zapis bi se smatrao šumom.

———————— · ————————

U ovoj tezi razmatra se vrednovanje podataka prema drugom i trećem kriterijumu, koje vodi ka dobijanju relevantnih podataka sa visokim kvalitetom, koji bi se dalje mogli upotrebiti u procesu vrednovanja podataka prema prvom kriterijumu. Treći kriterijum koristi se kao primarni jer je bolje nemati neki podatak nego imati nepouzdan ili netačan podatak.

4. Metodologija vrednovanja podataka

U poglavlju *Pregled literature* prikazano je trenutno stanje kako u hidrotehničkoj praksi, tako i u domenu naučnih istraživanja u oblasti vrednovanja podataka. Jedan od utisaka koji se sam po sebi nameće jeste da su primeri prikazani u naučnim radovima uglavnom orijentisani isključivo ka jednoj metodi koja se smatra najprikladnijom za raspoloživi set podataka, dok se zanemaruje mogućnost postojanja drugih metoda. Rezultat koji primenjene metode vrednovanja daju se eventualno poboljšava pre-procesiranjem podataka metodama filtriranja, smanjenja dimenzionalnosti, itd. Ukoliko se pri vrednovanju podataka i koristi više metoda procedure za njihovo kombinovanje uglavnom su vezane za „ekspertsko" tumačenje njihovih rezultata. Takođe se može primetiti da u literaturi ne postoji zajednička osnova svih metoda koja bi predstavljala radni okvir iz kog bi mogle da budu izvedene sve metode navedene u literaturi.

U ovom poglavlju se opisuje procedura za vrednovanje podataka koja predstavlja platformu pomoću koje je moguće upotrebiti i međusobno kombinovati bilo koju metodu za vrednovanje iz literature, i po potrebi razviti složenije metode sa više mogućnosti. Predloženi sistem predstavlja drugi korak u proceduri pripreme podataka za upotrebu, prikazanoj na slici 3.1. Osnovne prednosti metodologije koja se predstavlja su:

- predstavljanje podataka u neodređenom obliku (egzaktne vrednosti su samo specijalni slučaj);
- pristup vrednovanju podataka baziran na različitim (najboljim) relacijama između podataka i raspoloživim informacijama;
- mogućnost kombinovanja informacija koje se mogu upotrebiti za vrednovanje;
- lako uvođenje novih informacija u proceduru vrednovanja i novih veličina u sistem za vrednovanje;
- mogućnost da se proces vrednovanja dekomponuje;
- mogućnost praćenja istorije vrednovanja i svih međurezultata.

Neizvesnost (neodređenost) merenih vrednosti predstavlja prateću osobinu svakog podatka. Način merenja, izbor mernog uređaja, karakteristike merne lokacije, itd. utiču na preciznost vrednosti podatka merene veličine. Radni okvir predstavljen u ovoj tezi zasnovan je na neodređenoj predstavi merenih podataka, dok se egzaktne vrednosti razmatraju isključivo kao specijalni slučajevi. Akcenat je stavljen na intervalski zapis neodređenosti merenih veličina, dok se cela procedura može proširiti i na rasplinute skupove, kao i na statističke raspodele.

Tradicionalna praksa vrednovanja zasniva se na upoređivanju sa definisanim etalonom. Na primer, vrednovanje voća na pijaci se može sprovesti u odnosu na količinu (masu ili zapreminu) koja se izražava preko definisanog sistema mera (gram, kilogram, itd.) ili u odnosu na kvalitet, gde se ono može, na primer, svrstati u određenu klasu kvaliteta. Vrednovanje merenih hidrotehničkih podataka zasnovano je na upoređivanju merenog podatka sa etalonom formiranim na osnovu raspoloživih informacija o načinu merenja, karakteristikama merne veličine, uslovima koji vladaju u procesu merenja, drugim merenim vrednostima, itd. Kao što se može definisati više načina vrednovanja kod voća na pijaci, tako se može definisati i više načina vrednovanja merenog hidrotehničkog podatka. U predstavljenom radnom okviru koristi se procedura formiranja etalona za vrednovanje podataka u pogledu njegove vrednosti, koja iskazuje koliko je mereni podatak blizak tačnoj vrednosti. Radni okvir je zasnovan na matematičkim relacijama između vrednovanog podatka i raspoloživih informacija na osnovu kojih se formira etalon za upoređivanje sa vrednovanim podatkom.

Često se za potrebe vrednovanja izdvajaju informacije iz različitih izvora, informacije koje se nalaze u različitim oblicima ili informacije koje imaju različit nivo neizvesnosti i tačnosti (manje ili više izraženu grešku). U predloženom radnom okviru sve vrste informacija se pomoću matematičkih metoda i procedura transformišu u oblik u kom se mogu upoređivati sa merenim podatkom. Uz pomoć statističke teorije transformisane informacije kombinuju se u cilju dobijanja reprezentativne i verodostojne vrednosti etalona sa kojom je moguće uporediti mereni podatak.

Ponekad se naknadno dolazi do nekih informacija koje mogu pomoći u vrednovanju merenog podatka. Kod mnogih procedura koje se mogu naći u literaturi potrebno je u tom slučaju ponoviti čitav postupak vrednovanja sa uvedenim novim informacijama. U predloženom radnom okviru uvođenje novih informacija u proceduru vrednovanja moguće je obaviti sprovođenjem samo onih matematičkih operacija koje se odnose na novu informaciju.
Kompleksni sistemi, sa više merenih vremenskih serija koje je potrebno vrednovati, zahtevaju veliku angažovanost ukoliko je potrebno da jedan čovek ima kontrolu nad celim sistemom. Predloženi sistem ima mogućnost da bude prostorno i organizaciono dekomponovan. Prostorna dekompozicija podrazumeva podelu sistema na delove prema mestu na kom se nalaze (pretpostavlja se da se relacije mogu uspostaviti između veličina merenih bliskim mernim stanicama). Sa druge strane, organizaciona dekompozicija daje mogućnost da više aktera rade zasebne poslove (modeliranje, predprocesiranje, itd.).

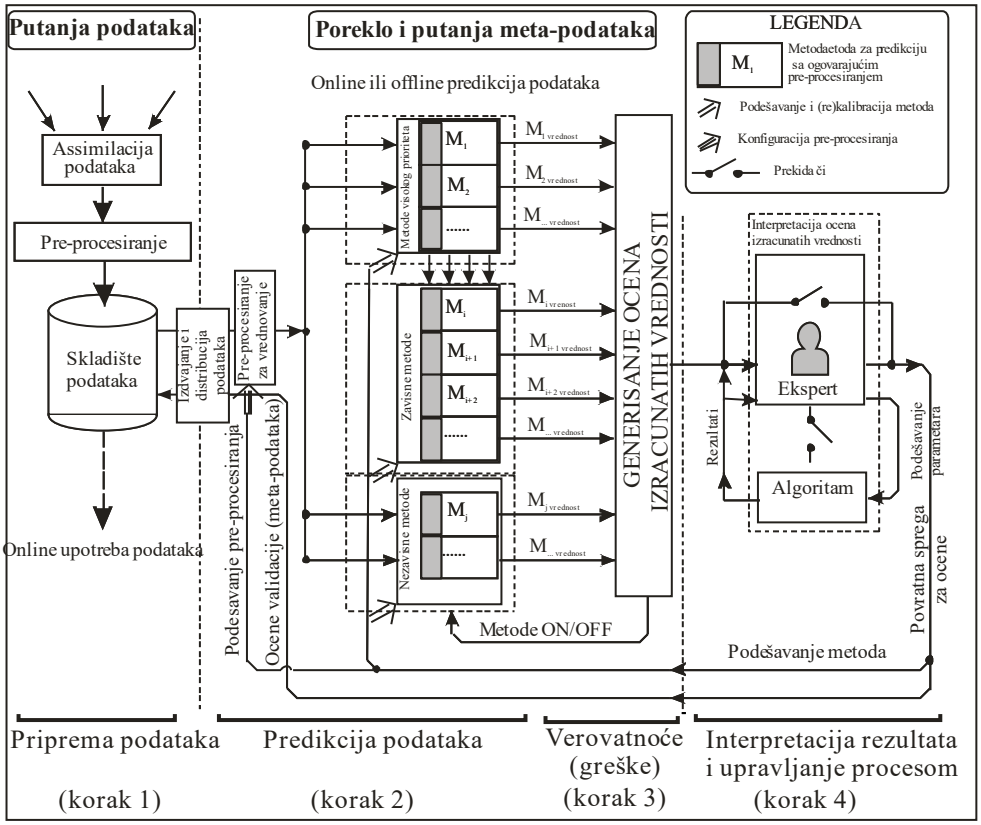
4.1 Radni okvir istema za vrednovanje podataka

Ključni faktor kod vrednovanja podataka je postojanje dodatnih informacija i, uz to, relacija kojima se vrednovani podatak može povezati sa dodatnim informacijama. Svaku relaciju je potrebno tako formulisati da se kao njen rezultat dobije predikcija vrednovanog podatka, tj. izračunata vrednost. Postojanje više relacija između podataka vodi ka više izračunatih vrednosti.

Gledano iz ugla matematičke formulacije relacija između podataka, vrednovanje može da obavi stručnjak (ručno), ili se mogu iskoristiti statistički alati i veštačka inteligencija (automatski), ili to može da obavi stručnjak uz pomoć statističkih alata i veštačke inteligencije (polu-automatski). Automatska validacija je jedini način da se proceni kvalitet podataka u realnom vremenu, dok se za pripremu istorijskih podataka ili tumačenje delikatnih pojava može koristiti polu-automatska ili ručna validacija, uz pomoć tehnika vizuelizacije podataka.

Platforma za vrednovanje podataka (prvi korak procedure prikazane na slici 3.1) razvijena u ovoj disertaciji implementirana je i testirana u MatLab okruženju i prikazana je na slici 4.1. Procedura za vrednovanje podataka može se podeliti na četiri koraka:

1. korak 1: priprema i distribucija podataka;
2. korak 2: generisanje izračunatih vrednosti vrednovanog podatka koristeći više metoda;
3. korak 3: izračunavanje verovatnoća grešaka;
4. korak 4: interpretacija (tumačenje) izračunatih vrednosti i donošenje odluke o kvalitetu podatka.



Slika 4.1: Procedura za vrednovanje podataka

Korak 1: Priprema i distribucija podataka
Procedura počinje prikupljanjem i grupisanjem podataka iz različitih izvora, kao i osnovnom prethodnom obradom koja obuhvata normalizaciju podataka, jednostavnu proveru podataka (npr. da li je vrednost podatka numerička vrednost) i skladištenje podataka u relacionoj bazi podataka. Iz relacione baze podataka operater „po zahtevu― distribuira podatke i njihove meta-podatke modulima za vrednovanje i prenosi rezultate vrednovanja nazad do baze podataka u vidu novih meta-podataka. U ovom koraku se formiraju podaci u neodreĎenom obliku (interval, rasplinuti skup, statistička raspodela) iz karakteristika merne metode. Na primer, ukoliko je rezultat merne metode egzaktna vrednost (što je čest slučaj u praksi), potreban je podatak o neodreĎenosti merne metode da bi se formirao podatak u neodreĎenom obliku.

Iako izuzetno važan, ovaj korak se ne razmatra u disertaciji. Pretpostavlja se da su podaci potrebni za rad sistema raspoloživi i adekvatno složeni u neki format (CSV, relacionu bazu podataka, itd.).

Korak 2: Predikcija vrednosti vrednovanog podatka pomoću formiranih relacija
U ovom delu sistema za vrednovanje podataka se metodama formiranim na osnovu relacija izmeĎu podataka izračunavaju predikcije vrednovanog podatka. Metoda uključuje kako relacije izmeĎu podataka, tako i procedure pretprocesiranja koje poboljšavaju karakteristike razvijenih metoda. Za predikciju vrednovanog podatka koristi se više metoda koje kao rezultat daju intervale u kojima se vrednovani podatak očekuje. U narednom koraku se formira sud o tome da li predikcije vrednovanih podataka sadrže greške ili ne.

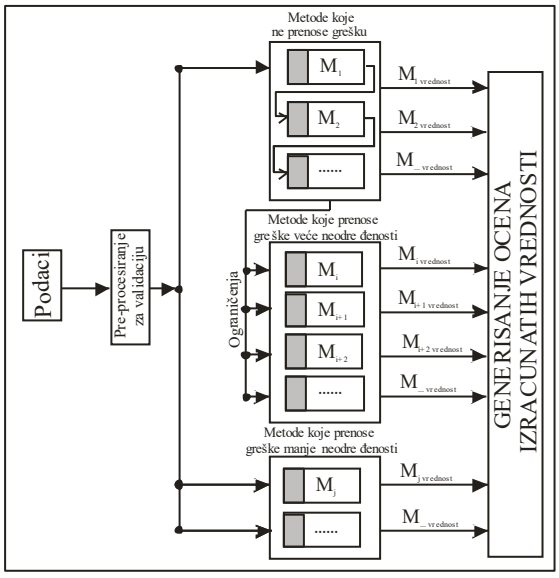
Korak 3: Ocene izmerenih vrednosti i predikcija vrednosti vrednovanog podatka

U ovom koraku izračunavaju se težinski koeficijenti koji odgovaraju verovatnoćama predikcija koje su izračunate metodama iz prethodnog koraka. Ukoliko ulazne vrednosti metode kojom se izračunava predikcija nekog podatka sadrže greške, težinski koeficijenti imaju niske vrednosti, dok, ukoliko ulazne vrednosti ne sadrže greške, težinski koeficijenti imaju visoke vrednosti. Takođe, težinski koeficijenti koji odgovaraju metodama sa manjom neodređenosti imaju veće težinske koeficijente od metoda sa većom neodređenosti.

Korak 4: Interpretacija upoređenih izmerenih podataka i njihovih predikcija
Na kraju procesa potrebno je uporediti izmerene i izračunate vrednosti, protumačiti rezultate upoređivanja i izračunati parametre po kojima se sprovodi vrednovanje podataka kao i reprezentativnu izračunatu vrednost.

4.2 Predikcija merenih veličina

Na slici 4.2 prikazan je samo korak 2 sa slike 4.1, sa grupom metoda za izračunavanje vrednovanog podatka koje se mogu primeniti u ručnoj, polu-automatskoj ili automatskoj proceduri za vrednovanje. Obično je učinak bolji ukoliko se metode koriste ručno ili polu-automatski, mada se automatskim radom ponekad mogu postići zadovoljavajući rezultati. Svi podaci moraju da prođu opštu prethodnu obradu pre nego što uđu u lanac metoda za predikciju. Proveravanje konzistenosti vremenske skale, generisanje dodatnih privremenih vremenskih serija sa različitim vremenskim korakom korišćenjem različitih tehnika interpolacije ili agregacije ili obeležavanje podataka u skladu sa dodatnim informacijama koje su prikupljene (da li je dan ili noć, da li je pumpa uključena ili isključena, da li je vreme kišovito ili ne, itd.) samo su od neke procedura koje se mogu sprovesti radi povećanja efikasnosti prenosa i obrade podataka u okviru glavnog dela procedure vrednovanja – izračunavanja vrednovanog podatka različitim metodama na osnovu dodatnih informacija.



Slika 4.2: Predikcije vrednovanog podatka različitim metodama i generisanje ocena poređenja merenih i vrednosti predikcija (korak 2 i korak 3)

Na slici 3.17 u Poglavlju 3 šematski je prikazana struktura mernog okruženja sa akcentom na relacijama između merenih podataka. Ukoliko postoje relacije između podataka, pored izmerene vrednosti moguće je izračunati i predikciju merenog podatka. Koliko relacija postoji, toliko je predikcija moguće dobiti. Kao ulazne vrednosti metoda za izračunavanje vrednovanog podatka mogu se koristiti kako informacije koje se ne vrednuju, tako i podaci koji se vrednuju. Ukoliko se koriste podaci koji se ne vrednuju (npr. granice u kojima se očekuje merni podatak koje zavise od

geometrije sistema), one u relacije ulaze kao fiksni parametri (npr. $h \geq h_{min}$, $h \leq h_{max}$, gde su h_{min} i h_{max} parametri geometrije sistema u kojoj se obavlja merenje veličine h). Takođe, te relacije daju ograničenja ostalim relacijama koje kao ulazne vrednosti koriste podatke sa potencijalnom greškom. Na slici 4.2 metode izvedene iz takvih relacija označene su kao metode koje ne prenose grešku. One generišu ograničenja koja služe kao jedan od ulaza metodama sa visokom neodređenosti. Metode sa niskom neodređenosti kao rezultat daju vrednosti u okviru ograničenja, pa nema potrebe za uvođenjem ograničenja kao ulaznih veličina.

Relacije za predikciju podataka se moraju pravilno pripremiti i primeniti. One predstavljaju ključni element i određujukvalitet samog sistema za vrednovanje. Primena zavisi kako od strukture metode, tako i od njenog položaja u sistemu. Mogu se izdvojiti dve kategorije relacija: relacije čiji rezultat ne sadrži grešku i relacije čiji rezultat sadrži grešku. Ukoliko su ulazne veličine relacije podaci koji se ne vrednuju već se smatraju tačnim, rezultat tih relacija nema grešku. Dve kategorije relacija validacije imaju poseban položaj u sistemu (slika 4.1). Detaljan primer procesa formiranja relacija za jedan sistem za vrednovanje prikazan je u trećem primeru u petom poglavlju.

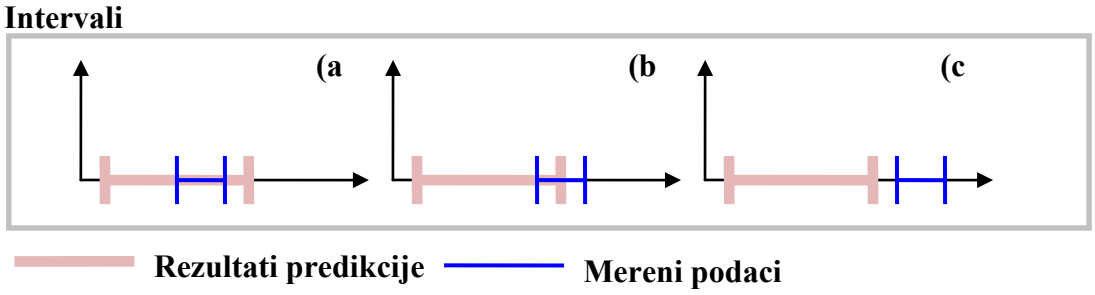
Izbor najboljih relacija za potrebe vrednovanja svodi se na uočavanje najbolje hidrotehničke prakse u odnosu na zahteve za eliminaciju grešaka u modelima i modeliranju neodređenosti modela. To znači da izbor i način formiranja metoda za predikciju iz relacija između podataka ne zavisi od ostatka sistema za vrednovanje, odnosno da su rezultati predikcija ulaz u naredni modul validacije, tj. generisanje ocena vrednosti predikcija.

4.2 Generisanje ocena izmerene vrednosti i rezultata predikcije

Nakon što se sprovede predikcija merenih podataka pomoću metoda izvedenih iz relacija koje ih vezuju, potrebno je uporediti izmerene i vrednosti dobijene predikcijom. S obzirom na to da su kalibracijom metoda za predikciju uklonjene greške metoda i modelirane neodređenosti relacija, greška u podatku koji predstavlja ulaznu vrednost metode direktno se preslikava na rezultat metode. Zbog toga nije moguće jednostavno uporediti izmerene i izračunate vrednosti već je u isto vreme potrebno proceniti i mogućnost da se u predikcijama nalaze greške. U ovom odeljku se predlaže metoda kojom se izračunava verovatnoća izračunatog podatka na osnovu drugih merenih podataka sa kojim je u vezi i na osnovu karakteristika relacija koje se u procesu vrednovanja koriste.

4.2.1 Upoređivanje izmerene vrednosti i njene predikcije

Izračunavanje veličine koja se vrednuje matematičkim modelom naziva se predikcija i njen cilj je da se dobije vrednost predikcije koja se može uporediti sa izmerenom. Na slici 4.3 prikazani su odnosi neodređenih veličina u obliku intervala. Pretpostavlja se da merene vrednosti imaju manju neodređenost nego oblasti u kojima se mogu nalaziti, mada to nije uvek slučaj.



Slika 4.3: Odnosi veličina kod intervala

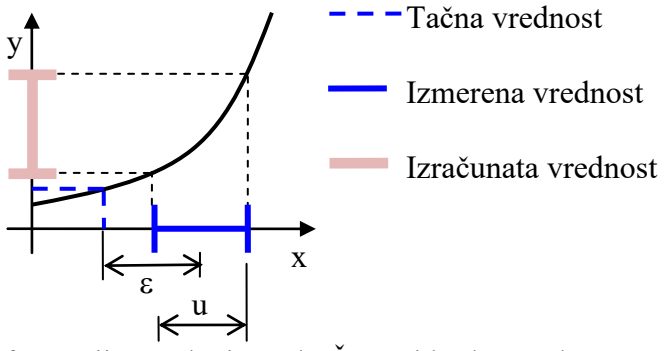
U obrnutom slučaju, tj. ukoliko je neodređenost merenih vrednosti veća od oblasti u kojima je predviđeno da se merene vrednosti nalaze, došlo bi se u situaciju da je bolje izračunati neki podatak nego ga izmeriti. Gledano isključivo preko neizvesnosti merenih i izračunatih vrednosti podataka, merenja i služe da bi se smanjila neodređenost vrednosti koje je moguće izračunati.

Na slici 4.3 prepoznaju se po tri tipa odnosa za neodređene veličine izražene u obliku intervala (označeni malim slovima a, b i c). Merena vrednost se može sadržati u intervalu dobijenom metodama za vrednovanje (slika 4.3a), može se delimično sadržati u njemu (slika 4.3b) ili biti van njega (slika 4.3c).

4.2.2 Preslikavanje greške i neodređenosti ulaznih vrednosti u rezultat modela

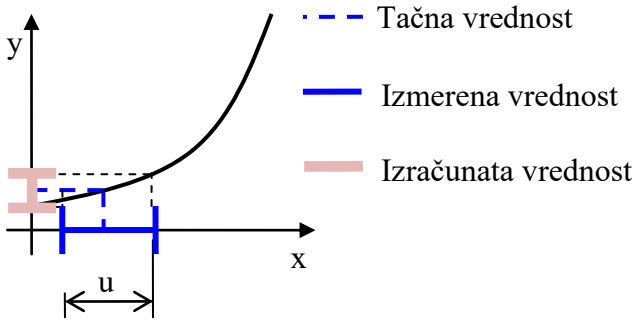
Često su neke od ulaznih vrednosti matematičkih modela takođe vrednosti koje je potrebno vrednovati. Transformacijom ulaznih vrednosti, pomoću jednačina matematičkog modela i kalibrisanim parametrima, u izračunatu veličinu koja se vrednuje istovremeno se preslikavaju i neodređenost i greška tih ulaznih vrednosti.

Greška i neodređenost koji postoje u ulaznim vrednostima matematičkog modela se prenose i na rezultat modela (slika 4.3). Ukoliko ona ne bi postojala, tj. ukoliko bi se pretpostavilo da greška u ulaznim podacima ne postoji, kao rezultat metode za vrednovanje dobila bi se oblast u kojoj se očekuje tačna vrednost merene veličine (slika 4.5). U odsustvu greške u ulaznim podacima, neodređena vrednost bi se preslikala takođe u neodređenu vrednost, ali bez prisustva greške, tj. tačna vrednost bi pripadala rezultatu u odliku intervala, rasplinutog skupa ili statističke raspodele. U tom slučaju bi za vrednovanje podataka bilo potrebno odabrati jednu metodu, onu koja ima najmanju neodređenost, i pomoću nje izračunati vrednost.



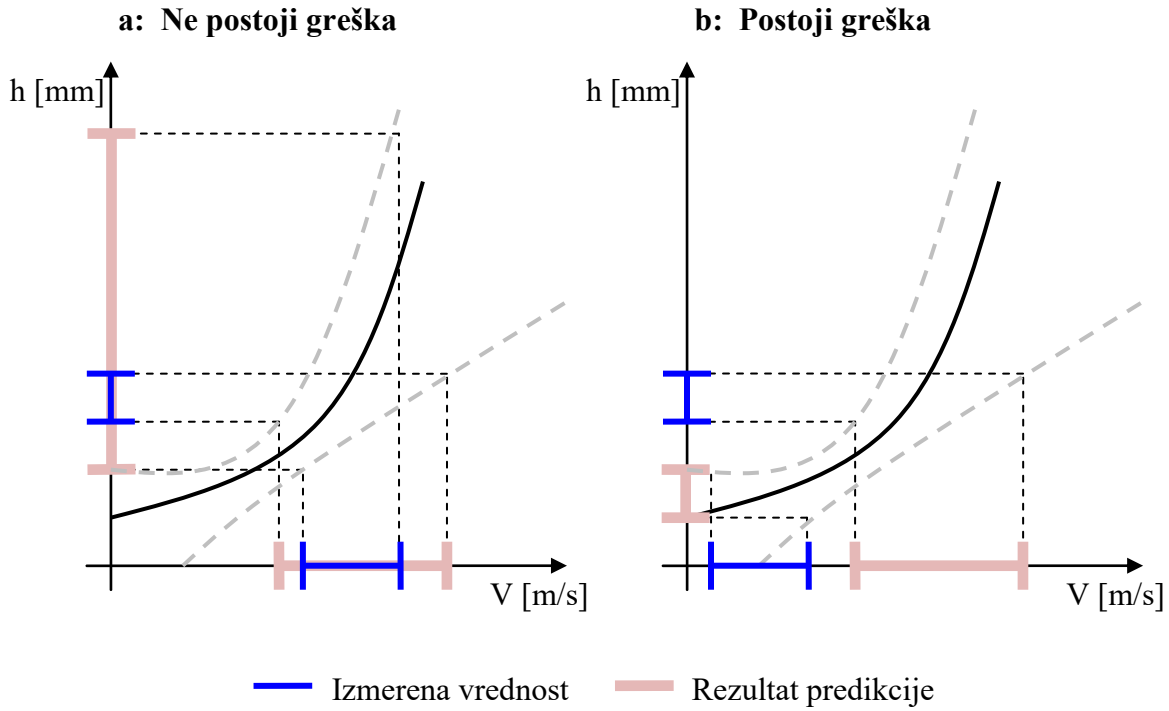
Slika 4.4: Transformacije greške i neodređenosti kod metoda za predikciju podataka

Na slici 4.4 prikazana je transformacija neodređenosti i greške u procesu predikcije veličine koja se vrednuje. Matematički model prikazan je preko nelinearne krive zbog jednostavnosti, dok je tačna vrednost, koja je poznata, prikazana zbog ilustracije celog procesa. U slučaju da je model predstavljen preko oblasti, kao na slici 3.18C ili 3.18D, egzaktna (tačna) vrednost bi takođe bila preslikana u interval. Matematičkim principima propagacije neodređenosti može se relativno lako neodređena vrednost preslikati u neodređeni rezultat matematičkog modela koji je uglavnom istog oblika kao i ulazni podaci.



Slika 4.5: Transformacija neodređenosti u odsustvu greške kod metoda za predikciju podataka

S obzirom na to da je cilj vrednovanja da se odredi rizik od upotrebe nekog izmerenog podatka u pogledu neizvesnosti i greške koju sa sobom nosi, a u skladu sa načinom upotrebe, ne sme se zanemariti greška koja se transformiše u procesu modeliranja neke veličine. Takođe, pošto je otkrivanje i kvantifikovanje prisustva greške u merenom podatku jedan od zadataka procesa vrednovanja podataka, dodatne informacije o njenom prisustvu se mogu dobiti i posredno, preko rezultata matematičkih modela metoda za vrednovanje.

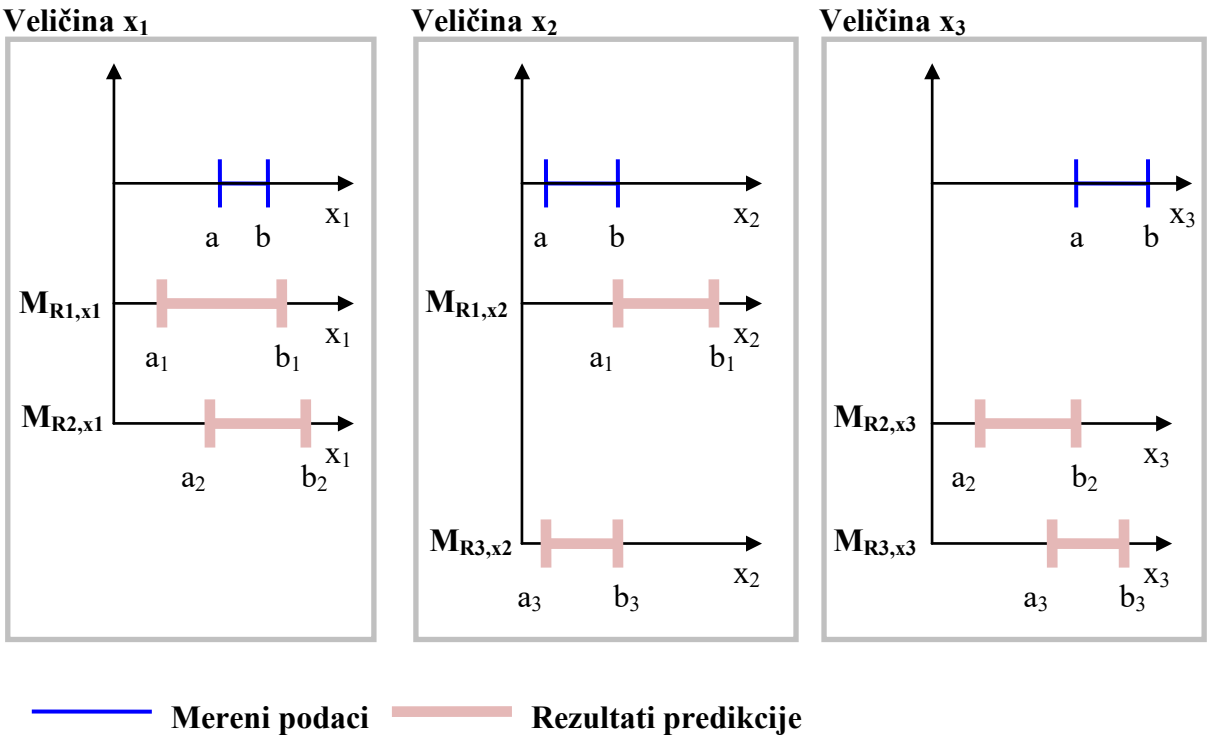


Slika 4.6: Dva slučaja kod vrednovanja dve veličine jednom metodom, relacija između podataka prema 3.18D

Na slici 4.6 prikazana su dva slučaja, jedan, kada ne postoji greška u podacima, i drugi, kada postoji greška u jednom merenom podatku. Ukoliko su obe merene vrednosti bez grešaka koje nisu pokrivene neodređenošću, merene vrednosti nalaze se unutar intervala izračunatih vrednosti. Time je pokazano da su mereni podaci smanjili neodređenost vrednosti koja je izračunata. Sa druge strane, ukoliko neka od veličina (u ovom primeru, brzina) sadrži grešku, i rezultat metode za predviđanje kod koje je veličina sa greškom ulazna vrednost sadrži grešku. S obzirom da tačna vrednost veličina koje se vrednuju nije poznata, jedini dokaz o tome da postoji greška u podacima jeste odstupanje izračunatih od merenih vrednosti. Nažalost, kada se vrednuju samo dve veličine sa jednom relacijom

nije moguće odrediti koja veličina (veličine) sadrži grešku. Pretpostavlja se da bi se sa više relacija između vrednovanih podataka i dodatnih informacija mogla bliže odrediti oblast u kojoj se očekuje vrednost merenog podatka, pomoću čega bi se sprovelo njegovo vrednovanje.

Kada se vrednovani podatak izračuna pomoću više metoda za vrednovanje, može se očekivati da se rezultati neće podudarati. Razlog tome su različite osobine korišćenih metoda (osetljivosti, neodređenosti matematičkog modela, itd.), i neodređenosti i greške u ulaznim vrednostima matematičkog modela metoda za vrednovanje. Na slici 4.7 grafički su prikazani vrednovani podaci x_1, x_2 i x_3 i rezultati tri relacije između podataka R_1, R_2 i R_3. Pored navedenih veličina, na raspolaganju su još i osobine metoda za predikciju kao što su osetljivost, monotonost, itd.



Slika 4.7: Primer rezultata metoda za predikciju podataka u obliku intervala

Nakon što se pomoću metoda za predikciju izračunaju očekivane vrednosti podataka (slika 4.7), proces interpretacije rezultata metoda za predikciju može se podeliti u dva koraka:

1. određivanje reprezentativne oblasti (intervala, rasplinutog skupa ili statističke raspodele) u kojoj se očekuje merena vrednost;
2. određivanje rizika od upotrebe merenog podatka.

Reprezentativnu oblast u kojoj se očekuje neki podatak potrebno je odrediti na osnovu karakteristika merenih podataka, rezultata metoda za predikciju podataka i njihovog međusobnog odnosa. Takođe su na raspolaganju i karakteristike metoda za predikciju podataka.

4.2.3 Verovatnoća kvaliteta podatka i reprezentativna vrednost

Vrednovanje izmerenog podatka moguće je izvršiti na osnovu izračunate očekivane vrednosti tog istog podatka (treba naglasiti da tačna vrednost nije poznata). Na nesreću, u izračunatoj vrednosti krije se greška koja potiče od ulazne vrednosti relacije na osnovu koje je vrednost vrednovanog

podatka i izračunata. Prema tome, raspoloživi podaci na osnovu kojih je moguće proceniti vrednost podatka su:

1. druge merene veličine (neke od njih imaju grešku i neodređenost);
2. izračunate vrednosti vrednovane veličine pomoću relacija sa drugim merenim veličinama; i
3. karakteristike relacija, tj. matematičkih modela (npr. osetljivost).

Pošto veličina grešaka kod ulaznih vrednosti relacija nije poznata (ako je poznata onda nije potrebno sprovoditi vrednovanje), greška koja je relacijom preneta na njen rezultat takođe nije poznata. Uz pretpostavku da je sistemom za vrednovanje podržano više relacija, pa da se samim tim dobija i više rezultata relacija (izračunate vrednovane veličine), javlja se potreba da se greške u izračunatim vrednovanim veličinama kvantifikuju i/ili eliminišu.

U predloženoj metodologiji za vrednovanje podataka uvodi se dodatna promenljiva koja označava koliko je izmereni podatak u skladu sa izračunatim, uz pretpostavku da je neodređenost izmerene vrednosti uvek manja od neodređenosti izračunate. Da se radi o egzaktnim vrednostima, ispitivala bi se jednakost izmerenih i izračunatih vrednosti. Uvođenje dodatne promenljive se može primeniti na bilo koje svrsishodno merenje, tj. merenje koje smanjuje neizvesnost neke veličine. Ukoliko izračunata vrednost ima manju neodređenost od izmerene, može se pretpostaviti da je merenje nesvrsishodno i da je bolje upotrebiti izračunatu vrednost.

Izračunavanje (predikcija) veličina koje se vrednuju podrazumeva izračunavanje vrednovanog podatka metodama za predikciju. Metode za predikciju često predstavljaju kombinaciju matematičkog modela vrednovane veličine i pretprocesiranja specifičnog za razmatranu metodu. Poželjno je da matematički modeli u okviru metoda za predikciju budu dizajnirani u skladu sa najboljom modelarskom praksom, sa obaveznim uvidom u trajanje izvršavanja modela (jer ono može da ugrozi implementaciju) i neodređenosti samog modela.

Neka je X skup od n podataka koji se vrednuju $X=\left[x_1,\ldots,x_i,\ldots,x_n\right]$, pri čemu podaci x_i mogu da budu od iste ili različitih merenih veličina ili čak podaci mereni istim mernim instrumentom, samo u drugim vremenskim presecima. Neka je R skup relacija (slika 3.17) za predikciju $R=\left[R_1,\ldots,R_i,\ldots,R_m\right]$, iz kojih su izvedene metode za predikciju M_{R_j,x_i}, koje se sastoje od procedure pripreme podataka i matematičkog modela za računanje nekog od vrednovanih podataka x_i. Maksimalan broj metoda za predikciju koji se može formirati za n podataka koji se vrednuju (Poglavlje 3.4) iznosi $\binom{n}{r}=\frac{n!}{r!(n-r)!}$, gde je $r=2$, u slučaju da je uspostavljena relacija između svaka dva vrednovana podatka.

Rezultati metoda za predikciju $x_i^{M_{R_j,x_i}}$ koji predstavljaju izračunati podatak x_i mogu se predstaviti matricom rezultata metoda za predikciju:

$$X^R:\quad \begin{array}{c|ccc} & x_1 & \cdots & x_n \\ \hline R_1 & x_1^{M_{R_1,x_1}} & \cdots & x_n^{M_{R_1,x_n}} \\ \cdots & \cdots & \cdots & \cdots \\ R_m & x_1^{M_{R_m,x_1}} & \cdots & x_n^{M_{R_m,x_n}} \end{array}.$$

U slučaju da su veličine koje se vrednuju predstavljene intervalima, matrica rezultata predikcije izgleda ovako:

$$X_{\text{int}}^{R}: \quad \begin{array}{c|cccc} & \left[a_{x_1}, b_{x_1}\right] & \cdots & \left[a_{x_n}, b_{x_n}\right] \\ \hline R_1 & \left\lfloor a_{x_1^{M_{R_1,x_1}}}, b_{x_1^{M_{R_1,x_1}}} \right\rfloor & \cdots & \left\lfloor a_{x_n^{M_{R_1,x_n}}}, b_{x_n^{M_{R_1,x_n}}} \right\rfloor \\ \cdots & \cdots & \cdots & \cdots \\ R_m & \left\lfloor a_{x_1^{M_{R_m,x_1}}}, b_{x_1^{M_{R_m,x_1}}} \right\rfloor & \cdots & \left\lfloor a_{x_n^{M_{R_m,x_n}}}, b_{x_n^{M_{R_m,x_n}}} \right\rfloor \end{array},$$

gde su a i b granice intervala. Posebna vrsta metoda za predikciju koja ne sadrži vrednovane podatke kao ulazne veličine (metoda čiji rezultat nema grešku) može, ali i ne mora da postoji za svaki vrednovani podatak $M^{'} = \left[M_1^{'}, \ldots, M_i^{'}, \ldots, M_k^{'}\right], k \le n$. Ove metode nisu izvedene iz relacija između podataka, već se odnose na relacije vrednovanog podatka sa podacima koji se ne vrednuju. Rezultati ovih metoda nemaju grešku, ali uglavnom imaju značajnu neizvesnost. Njihovi rezultati služe da ograniče rezultate metoda koji mogu imati grešku. Ukoliko se vrednovane veličine prikazuju u obliku intervala, rezultati ovih metoda će izgledati ovako:

$$M^{'}: \quad \left\lfloor a_{x_1^{M_1^{'}}}, b_{x_1^{M_1^{'}}} \right\rfloor, \quad ..., \quad \left\lfloor a_{x_1^{M_i^{'}}}, b_{x_1^{M_i^{'}}} \right\rfloor, \quad ..., \quad \left\lfloor a_{x_1^{M_k^{'}}}, b_{x_1^{M_k^{'}}} \right\rfloor \quad .$$

Metoda M_{R_j,x_i} kojom se izračunava x_i i dobija se njegova predikcija $x_i^{M_{R_j,x_i}}$ sastoji se od matematičke formulacije relacija (jednačina), parametara metode θ i ulaznih vrednosti $X_{x_i}^{M_{R_j,x_i}}$, koji predstavljaju skup svih vrednosti koje učestvuju u relaciji R_j bez veličine x_i:

$$M_{R_j,x_i}: x_i^{M_{R_j,x_i}} = F\left(X_{x_i}^{M_{R_j,x_i}}, \theta\right),$$

gde θ predstavlja niz parametara $\theta = \left[\theta_1, \ldots, \theta_2, \ldots, \theta_r\right]$ čije se vrednosti (egzaktne ili neodređene) procenjuju u procesu kalibracije. Rezultati predikcije $x_i^{M_{R_j,x_i}}$ sadrže neizvesnost i grešku koja potiče od podataka koji se vrednuju $X_{x_i}^{M_{R_j,x_i}}$, a učestvuju u metodi M_{R_j,x_i} kao ulazne vrednosti. Greške metode M_{R_j,x_i} se ne prenose na rezultat, što je omogućeno pripremom metoda opisanom u poglavlju 3.

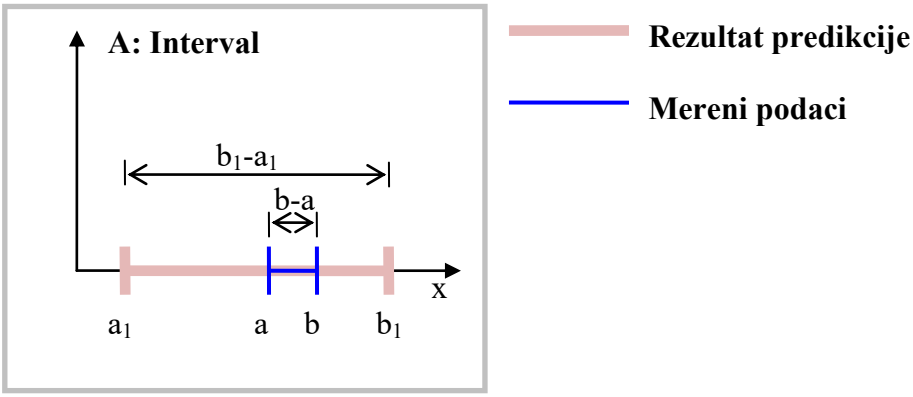
Verovatnoća izmerene vrednosti x_i koja je izmerena mernom metodom MM_i i ostalih raspoloživih informacija, od kojih su najznačajniji drugi izmereni podaci koji učestvuju u metodama za predikciju veličine x_i (X_{x_i}), mogu se izraziti višedimenzionalnom raspodelom iz koje sledi verovatnoća $p\left(x_i, X_{x_i}, MM_i\right)$. Verovatnoća $p\left(x_i, X_{x_i}, MM_i\right)$ može se razložiti na verovatnoće pojedinih promenljivih:

$$
\begin{aligned}
&p\left(x_i, X_{x_i}, MM_i\right) = \\
&p\left(x_i, X_{x_i} \mid MM_i\right) \times p\left(MM_i\right) = \\
&p\left(x_i \mid X_{x_i}, MM_i\right) \times p\left(X_{x_i} \mid MM_i\right) \times p\left(MM_i\right) = \\
&p\left(x_i \mid X_{x_i}, MM_i\right) \times p\left(X_{x_i}\right) \times p\left(MM_i\right)
\end{aligned},
$$

jer su dodatne informacije X_{x_i} i merna metoda x_i (MM_i) nezavisne veličine, pa je $p\left(X_{x_i} \mid MM_i\right) = p\left(X_{x_i}\right)$. Dalje sledi:

$$
\begin{aligned}
&p\left(x_i \mid X_{x_i}, MM_i\right) \times p\left(X_{x_i}\right) \times p\left(MM_i\right) = \\
&p\left(x_i \mid X_{x_i}\right) \times p\left(x_i \mid MM_i\right) \times p\left(X_{x_i}\right) \times p\left(MM_i\right) = \\
&\int p\left(x_i \mid x_i^{M_{R,x_i}}, X_{x_i}^{M_{R,x_i}}\right) p\left(x_i^{M_{R,x_i}} \mid X_{x_i}^{M_{R,x_i}}\right) dx_i^M \times p\left(x_i \mid MM_i\right) \times p\left(X_{x_i}\right) \times p\left(MM_i\right)
\end{aligned},
$$

gde je $p\left(x_i \mid X_{x_i}\right) = \int p\left(x_i \mid x_i^{M_{R,x_i}}, X_{x_i}^{M_{R,x_i}}\right) p\left(x_i^{M_{R,x_i}} \mid X_{x_i}^{M_{R,x_i}}\right) dx_i^M$. Integral označava zbir za sve metode M_{R_j,x_i}, $j = 1,\ldots,m$ koje se izvode iz relacija u kojima učestvuje vrednovana veličina x_i. Priorne verovatnoće merne metode ($p\left(MM_i\right)$) i dodatnih informacija drugih merenih veličina ($p\left(X_{x_i}\right)$)potrebno je pretpostaviti prema unapred poznatim uslovima merenja i uslovima pod kojima su prikupljene dodatne informacije. Ukoliko nema poznatih informacija moguće je pretpostaviti istu, nepromenljivu, verovatnoću za sve merne metode i izmerene vrednosti. Komponente integrala, verovatnoće $p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)$, mogu se izračunati direktno iz intervala velilčina koje su izmerene i koje predstavljaju predikciju vrednovanog podatka:



$$
P\left([a,b] \mid [a_1,b_1]\right) = \frac{P\left([a,b] \cap [a_1,b_1]\right)}{P\left([a_1,b_1]\right)}, \text{ za } P\left([a_1,b_1]\right) > 0
$$

$$
P\left([a_1,b_1]\right) = \frac{P\left([a_1,b_1] \cap [a',b']\right)}{P\left([a',b']\right)},
$$

gde je $\left[a^{'},b^{'}\right]$ interval moguićh vrednosti

$$P\left(\left[a,b\right]|\left[a_1,b_1\right]\right)=\begin{cases}1-\dfrac{\max\left(0,a-a_1\right)+\max\left(0,b_1-b\right)}{\left(b_1-a_1\right)}, & a_1<b \\ & b_1>a \\ 0 & \end{cases} \tag{4.1}$$

Sa druge strane, veličine $p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)$ nije moguće direktno izračunati, već ih je potrebno proceniti. Jedan od načina bi bio da se veličine $p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)$ izračunaju maksimiziranjem verodostojnosti njihove sume po svim veličinama i svim metodama:

$$J=\max\sum_i\sum_j p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right).$$

UvoĐenjem pomoćne varijable (varijabla koja nije osmotriva) P_{x_i} koja predstavlja verovatnoću x_i na osnovu svih izračunatih vrednosti $x_i^{M_{R_j,x_i}}$, i karakteristike relacije izmeĐu podataka koja predstavlja izvod po ulaznim vrednostima, $p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)$ se može se izračunati kao:

$$p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)=\sum_{X_{x_i}^{MR_j,x_i}}\left(\frac{\partial M\left(X_{x_i}^{M_{R_j,x_i}},\theta\right)}{\partial X_{x_i}^{M_{R_j,x_i}}}\right)^{-1} P_{x_i},$$

gde se računa zbir po svim ulaznim podacima $X_{x_i}^{M_{R_j,x_i}}$ koji učestvuju u metodi M_{R_j,x_i} koja kao rezultat daje izračunatu vrednost vrednovane veličine $x_i^{M_{R_j,x_i}}$. Ukoliko je matematički model linearan, veličina izvoda je jednaka koeficijentu pravca, a ukoliko nije linearan može se linearizovati:

$$M_{R_j,x_i}:x_i^{M_{R_j,x_i}}\approx M_{R_j,x_i}\left(X^0\right)+\sum\frac{\partial M_{R_j,x_i}\left(X_{x_i}^{M_{R_j,x_i}}\right)}{\partial X_{x_i}^{M_{R_j,x_i}}}X_{x_i}^{M_{R_j,x_i}}.$$

Veličina P_{x_i} verovatnoća x_i na osnovu svih izračunatih vrednosti $x_i^{M_{R_j,x_i}}$ može se izračunati preko:

$$P_{x_i}=\sum_j p\left(x_i \mid x_i^{M_{R_j,x_i}},X_{x_i}^{M_{R_j,x_i}}\right)\times\frac{p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)}{\sum_{M_j} p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)},$$

gde je zbir po svim izračunatim vrednostima vrednovane veličine x_i ($x_i^{M_{R_j,x_i}}$), metodama za predikciju M_{R_j,x_i} u kojima učestvuje veličina x_i.

Problem izračunavanja $J = \max \sum_i \sum_j p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)$ može se rešiti iterativno, *expectation maximisation* (EM) metodom [26]. Nakon pretpostavljanja početnih vrednosti $p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)$ ili P_{x_i} potrebno je iterativno sprovesti dva koraka:

E korak:
$$p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right) = \sum_{X_{x_i}^{M_{R_j,x_i}}} \left(\frac{\partial M_{R_j,x_i}\left(X_{x_i}^{M_{R_j,x_i}}, \theta\right)}{\partial X_{x_i}^{M_{R_j,x_i}}} \right)^{-1} P_{x_i} \text{ , i} \tag{4.2}$$

M korak:
$$P_{x_i} = \sum_j p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right) \times \frac{p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)}{\sum_j p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)},$$

gde je $p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)$ izračunato prema (4.1). Nakon što su izračunate verovatnoće $p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)$ moguće je odrediti verovatnoću da je mereni podatak dobro izmeren i reprezentativnu oblast rešenja u kojoj se očekuje mereni podatak.

Verovatnoća regularnosti izmerenog podatka može se izračunati pomoću formule:

$$x_i^{grade} = \sum_j w_i^{M_{R_j,x_i}} \times p\left(x_i \mid x_i^{M_{R,x_i}}, X_{x_i}^{M_{R,x_i}}\right), \tag{4.3}$$

gde je $p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)$ predstavljeno težinskim koeficijentom $w_i^{M_{R_j,x_i}}$.

Ukoliko podatak ne zadovoljava kriterijume formirane na osnovu načina njegove upotrebe, moguće je umesto izmerenog podatka za upotrebu koristiti izračunatu reprezentativnu oblast.

4.2.4 Reprezentativna vrednost izračunatog podatka

Verovatnoća reprezentativne izračunate vrednosti može se izračunati preko izraza:

$$p(x_i^M \mid X_{x_i}^M) = \sum_{j=1}^{k} p(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}, x_i^{M_{R_j,x_i}}) p(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}),$$

gde x_i^M predstavlja reprezentativnu vrednost vrednovanog podatka x_i dobijenu predikcijom pomoću metoda M_{R_j,x_i} , $j = 1 \ldots k$. Matematičko očekivanje reprezentativne izračunate vrednosti se dalje može izraziti kao:

$$E[x_i^M \mid X_{x_i}^M] = \sum_{M_j} x_i^{M_{R_j,x_i}} p(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}) \text{ ,} \tag{4.4}$$

a varijansa kao:

$$Var[x_i^M \mid X_{x_i}^M] = \sum_j (Var[x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}, x_i^{M_{R_j,x_i}}] + x_i^{M_{R_j,x_i}})^2 p(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}) - E[x_i^M \mid X_{x_i}^{M_{R_j,x_i}}]^2 \ , \qquad (4.5)$$

gde izraz $x_i^{M_{R_j,x_i}} = E[x_i^{M_{R_j,x_i}} \mid X_{x_i}^M, x_i^{M_{R_j,x_i}}]$ predstavlja reprezentativnu vrednost izračunatu samo pomoću metode M_{R_j,x_i}, tj. vrednost $x_i^{M_{R_j,x_i}}$, a ($\sum_{j=1}^{k} p(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}) = 1$) predstavlja težinski koeficijent koji treba pripisati *j*-tom metodu. To znači da se reprezentativna vrednost može izračunati kao težinski osrednjene izračunate vrednosti, pri čemu su težinski koeficijenti podaci o verovatnoćama izračunatih vrednosti $p(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}})$.

Ukoliko su izračunate vrednosti $x_i^{M_{R_j,x_i}} = \left[a_j, b_j\right]$, $j = 1 \ldots k$ intervali, reprezentativna vrednost se može izračunati pomoću izraza:

$$E[x_i^M \mid X_{x_i}^M] = \left[\sum_{j=1}^{k} p(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}) \times a_j, \sum_{j=1}^{k} p(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}) \times b_j \right] \ .$$

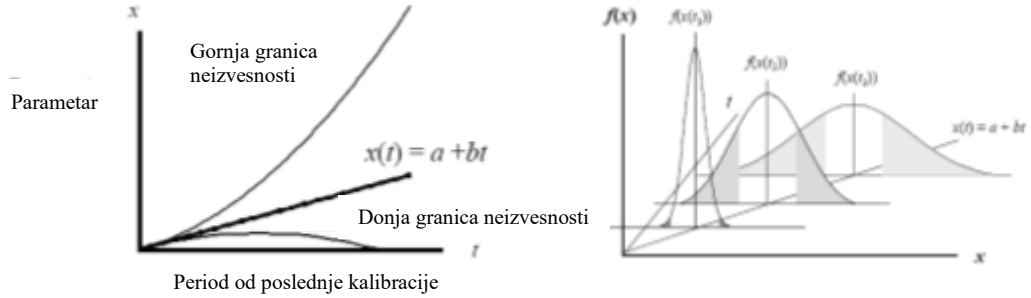
4.2.5 Pouzdanost merne metode

Verovatnoća $P(x_i \mid MM_i)$ označava mogućnost da se veličina x_i dobija mernom metodom MM_i. Ovim članom (MM_i) se može modelirati npr. verovatnoća da će se dobiti dobra merena vrednost u pogledu perioda od poslednjeg održavanja i kalibracije. Kao primer bi se mogla navesti promena pouzdanosti merne metode kroz vreme:

$$P(MM_i, t \mid x_i) = \frac{P(x_i \mid MM_i, t) \times P(MM_i)}{P(x_i)} \ .$$

S obzirom da se merna metoda bira unapred, verovatnoća se u toku merenja može modelirati kroz vreme. Verovatnoću izmerenog podatka na koju utiče merna metoda moguće je izračunati na bazi promene kalibracionih karakteristika [135], oblaganja mernih senzora nečistoćama [22] ili kvara komponenti mernog uređaja u funkciji vremena.

Kalibracijom se ostvaruje veza između izlazne veličine mernog uređaja i veličine koja se želi odrediti. Kalibracija mernog uređaja se ostvaruje upoređivanjem izlaza mernog uređaja sa poznatim vrednostima veličine koja se meri i formiranjem tzv. kalibracione krive koja predstavlja matematički zapis veze dve veličine.



Slika 4.8: Statistički opis rasta neizvsnosti veličine $f(x)$ kroz vreme [135]

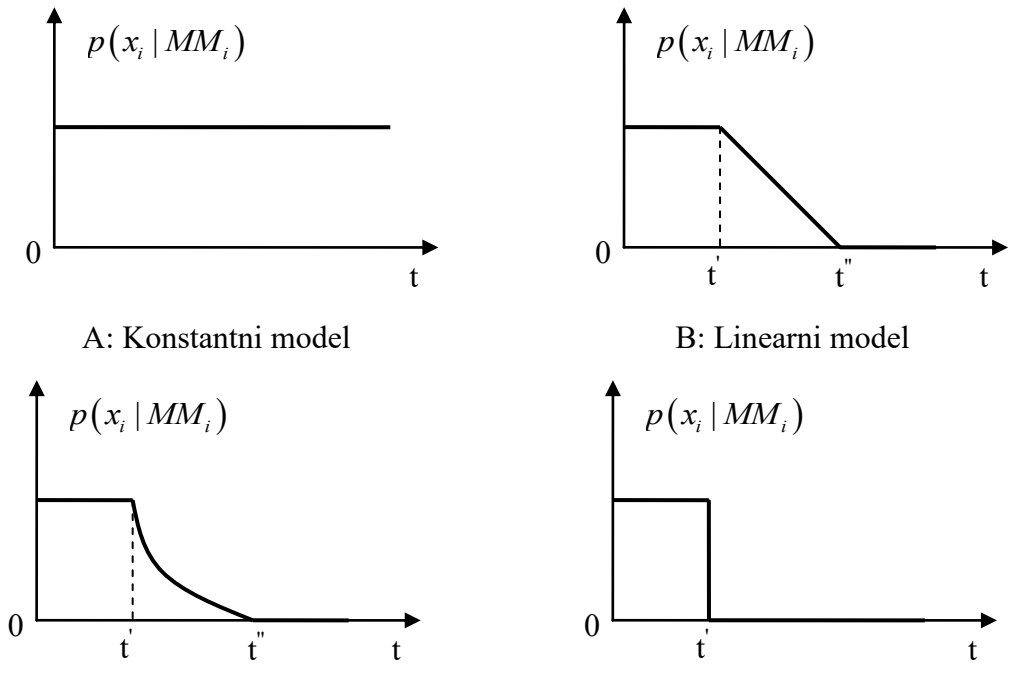
Veza između dve veličine često nije jednoznačna, već je za formiranje kalibracione krive potrebno upotrebiti neki statistički alat, kao što je, na primer, regresija. Formiranjem regresione krive izlazna vrednost mernog uređaja se transformiše, o čemu će biti reči kasnije u ovom poglavlju. Sa druge strane, kroz vreme se veza između dve veličine menja, pa se povećanje neizvesnosti može prikazati kao na slici 4.8. U isto vreme predviđa se i povećanje greške, i to u obliku linearne funkcije $x(t) = a + b \times t$.

Kao mera neodređenosti u zavisnosti od proteklog vremena od poslednje kalibracije se u [135] predlaže tzv. merna pouzdanost (*measurement reliability*), koja se može izraziti pomoću relacije:

$$R(t) = \int_{L_1}^{L_2} f\left(x(t)\right)dx ,$$

gde su L_1 i L_2 granice tolerancije odstupanja kalibracione krive. Ovaj podatak se može odrediti povremenom proverom kalibracionih karakteristika mernog uređaja, i preko njega predvideti optimalan period između dve kalibracije.

U nedostatku istorijskih podataka, promena pouzdanosti merne metode kroz vreme može se opisati i predefinisanim modelima. Na slici 4.9 prikazani su primeri modela konstantne vrednosti, linearne promene, polinomne ili eksponencijalne promene i *step* promene verovatnoće $p(x_i \mid MM_i)$.



A: Konstantni model B: Linearni model

C: Polinomni (eksponencijalni) model D: *Step* model

Slika 4.9: Modeli uslovne verovatnoće $p(x_i \mid MM_i)$

4.3 Ocena kvaliteta podatka – donošenje odluke

Na osnovu rezultata predložene metode i načina upotrebe podatka potrebno je doneti odluku da li je izmereni podatak korektan ili ne. Uz pretpostavku da je merenje obavljeno metodom koja omogućava podatak sa adekvatnom neodređenosti za razmatranu upotrebu, potrebno je proceniti

kako prisustvo greške u podatku, tako i adekvatnost metode za vrednovanje. Rezultati predložene metode kao rezultat pružaju brojne parametre na osnovu kojih bi se mogao izvući zaključak o regularnosti merenog podatka x_i:

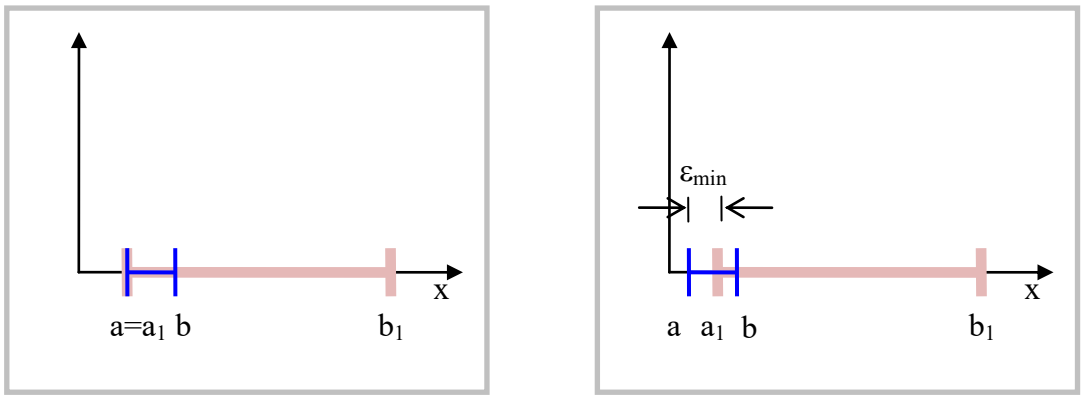
1. izračunate vrednosti podatka koji se vrednuje različitim metodama $x_i^{M_j}$;
2. podatkei o uslovnoj verovatnoći izmerene vrednosti $p_{x_i}^{x_i^{M_{Rj\text{-}xi}}} = p\left(x_i \mid x_i^{M_{Rj\text{-}xi}}, X_{x_i}^{M_{Rj\text{-}xi}}\right)$;
3. podatke o uslovnoj verovatnoći izračunatih vrednosti $w_i^{M_{Rj\text{-}xi}} = p\left(x_i^{M_{Rj\text{-}xi}} \mid X_{x_i}^{M_{Rj\text{-}xi}}\right)$;
4. podatak o ukupnoj verovatnoći kvaliteta izmerenog podatka $p\left(x_i, X_{x_i}, MM_i\right)$;
5. podatak o reprezentativnoj vrednosti $E[x_i^M \mid X_{x_i}^M]$.

Ocene kvaliteta merene vrednosti mogu se dobiti iz podatka o ukupnoj verovatnoći izmerenog podatka $p\left(x_i, X_{x_i}, MM_i\right)$, dok se oblast u obliku intervala u kojoj se mereni podatak očekuje može proceniti na osnovu njegove reprezentativne vrednosti $p\left(x_i, X_{x_i}, MM_i\right)$.

Veličina $p\left(x_i, X_{x_i}, MM_i\right)$ u velikoj meri zavisi od neizvesnosti metoda koje se koriste pri predikciji merenog podatka, pa se u slučaju značajne neizvesnosti metoda koje se koriste ne može povući jednoznačna granica koja bi odvojila regularne od neregularnih podataka. Ova situacija bi se mogla prevazići izračunavanjem dva parametra. Prvi bi razdvojio sigurnu grešku koju podatak sadrži i neizvesnost metode u pogledu mogućnosti da detektuje grešku određenog intenziteta, dok bi se drugim parametrom izdvojila informacija o grešci, a izgubila informacija o neodređenosti predikcija vrednovanih podataka.

4.3.1 Parametar sigurne greške u merenom podatku

Parametar sigurne greške predstavlja minimalnu grešku koju mereni podatak poseduje. Na slikama 4.10A i 4.10B prikazana su dva slučaja: A) kada je minimalna greška jednaka 0 i B) kada je minimalna greška jednaka $\varepsilon_{x_i}^{\min}$.



A: Minimalna greška merenog podatka jednaka je nuli

B: Minimalna greška merenog podatka jednaka je $\varepsilon_{x_i}^{\min}$

Slika 4.10: Minimalna greška podatka

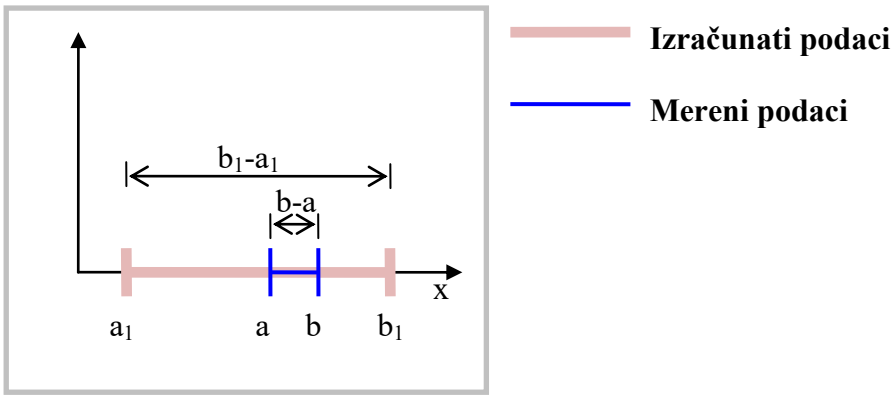
Veličina $\varepsilon_{x_i}^{\min}$ može se izračunati preko veličine neodreĎenosti merenog podatka i rezultata predikcije:

$$\varepsilon_{x_i}^{\min} = \left(1 - p_{x_i}^{norm}\right) \times (b - a), \tag{4.6}$$

gde je $p_{x_i}^{norm}$ veličina normirane verovatnoće. Normirana verovatnoća $p_{x_i}^{norm}$ predstavlja normiranu ukupnu verovatnoću. Normirana verovatnoća podatka se može izračunati pomoću formule

$$p_{x_i}^{norm} = \left(\sum_{x_i} w_i^{M_j} \frac{p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)}{\max\left(p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)\right)} \right) \times p\left(x_i \mid MM_i\right) \times p\left(X_{x_i}\right) \times p\left(MM_i\right), \tag{4.7}$$

pa je $p_{x_i}^{norm}$ u granicama izmeĎu $p_{x_i}^{norm} = \left\lfloor 0, p\left(x_i \mid MM_i\right) \times p\left(X_{x_i}\right) \times p\left(MM_i\right) \right\rfloor$ ili $p_{x_i}^{norm} = [0,1]$ kada važi $p\left(x_i \mid MM_i\right) = p\left(X_{x_i}\right) = p\left(MM_i\right) = 1$. Maksimalna uslovna verovatnoća može se izračunati preko izraza prikazanog na slici 4.11.



$$\max\left(p\left(x_i \mid x_i^{M_j}, X_{x_i}^{M_j}\right)\right) = \max\left(P\left([a,b] \mid [a_1,b_1]\right)\right) = \frac{b-a}{b_1-a_1}$$

Slika 4.11: Maksimalne uslovne verovatnoće izmerenog podatka

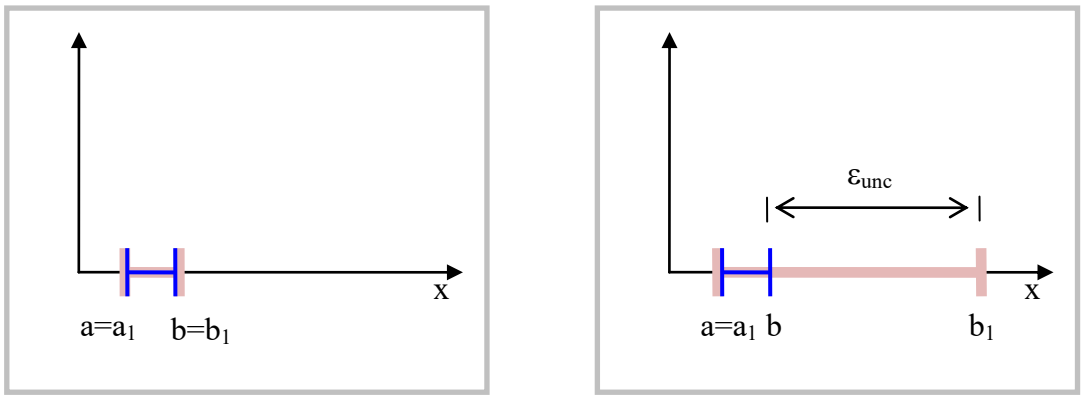
Normiranjem verovatnoća izgubila se informacija o neizvesnosti relacija koje su učestvovale u procesu vrednovanja, pa je opseg mogućih vrednosti postao jedinstven. Na taj način postalo je moguće povući jedinstvenu granicu za određeni mereni podatak kojim bi se ograničila minimalna prihvatljiva greška. Minimalna greška koja postoji u podatku može biti neotkrivena ukoliko je neizvesnost metoda za predikciju podataka visoka.

4.3.2 Parametar neizvesnosti metode za vrednovanje izmerenog podatka

Uticaj neizvesnosti sistema za vrednovanje može se izraziti preko maksimalnih uslovnih verovatnoća $\max\left(p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)\right)$ koje je moguće izračunati preko izraza prikazanog na slici 4.11. Neizvesnost sistema za vrednovanje $x_i^{M_j,unc} = [0,1]$ velilčine x_i može se izračunati pomoću formule

$$x_i^{M_{R_j,x_i},unc} = \sum_M w_i^{M_{R_j,x_i}} \times \max\left(p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)\right),$$

gde je $w_i^{M_{R_j,x_i}} = p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)$ težinski koeficijent uz izračunatu vrednost izmerenog podatka x_i uz pomoć relacije M_j. Granice u kojima se kreće $x_i^{M_{R_j,x_i},unc}$ su između 0 za veliku neodređenost i 1 kada nema neodređenosti. Informacija o verovatnoći pruža mogućnost da se izračuna neizvesnost greške kod podatka. Na slici 4.12 prikazana su dva slučaja: A) neizvesnost greške je jednaka nuli i B) neizvesnost greške je jednaka ε_{unc}.



A: Neizvesnost greške merenog podatka jednaka je nuli

B: Neizvesnost greške merenog podatka jednaka je ε_{unc}

Slika 4.12: Neizvesnost greške podatka

Veličina $\varepsilon_{x_i}^{unc}$ se može izračunati preko veličine $x_i^{M_j,unc}$ i veličine neodređenosti merenog podatka:

$$\varepsilon_{x_i}^{unc} = \left(\frac{1}{x_i^{M_{Rj},x_i,unc}} - 1 \right) \times (b-a) \qquad (4.8)$$

.

Potrebno je napomenuti da se neizvesnost računa kao ~~„~~najveća" neizvesnost, jer se ne uzima u obzir koja je pozicija izmerene veličine u odnosu na interval rezultata predikcije. To praktično znači da se raspoloživim metodama za predikciju može detektovati greška koja je $\varepsilon_{x_i} > \varepsilon_{x_i}^{unc}$. Ukoliko poznajemo minimalnu vrednost greške i njenu neizvesnost, moguće je grešku prikazati u obliku intervala kao

$$\varepsilon = [\varepsilon_{\min}, \varepsilon_{\min} + \varepsilon_{unc}], \text{ za } \varepsilon_{\min} < 0$$

ili

$$\varepsilon = [\varepsilon_{\min} - \varepsilon_{unc}, \varepsilon_{\min}], \text{ za } \varepsilon_{\min} > 0 .$$

Da bi se poboljšalo vrednovanje podataka, tj. smanjila njihova neizvesnost, potrebno je formirati relacije između podataka koje su manje neizvesne. To se, između ostalog, može postići uključivanjem dodatnih informacija ili novih merenja u sistem.

Odluka o tome da li je izmereni podatak primeren za određenu upotrebu svodi se na proveru da li je vrednost podatka u granicama sa dovoljno visokom verovatnoćom i da li je prostor koji pokriva ta verovatnoća preterano širok usled velike neodređenosti drugih podataka na osnovu kojih se razmatrani podatak vrednuje i/ili velike neodređenosti samih relacija koje ih povezuju.

Ukoliko se izmereni podatak nalazi van prostora vrednosti sa visokom verovatnoćom u kojoj se očekuje na osnovu izračunatih vrednosti, on se mora proglasiti neregularnim. Međutim, ako se nalazi u predviđenom prostoru postavlja se pitanje da li je podatak adekvatno vrednovan. Podaci i relacije na osnovu kojih je neki podatak vrednovan mogu biti sa visokom neodređenosti pa samo vrednovanje može biti nedovoljno za neku vrstu upotrebe. Na primer, ukoliko je podatak o dubini

vrednovan na osnovu podatka o brzini koji je dobijen mernom metodom sa visokom neodređenosti, iako se mereni podatak, na primer, u obliku intervala, nalazi u intervalu izračunatog, širina izračunatog intervala može biti velika, tako da vrednovanje nije adekvatno.

4.4 Ocena rezultata metodologije za vrednovanje

Faktori na osnovu kojih se formira sistem za vrednovanje podataka mogu se podeliti u dve grupe: 1) raspoloživi podaci, informacije i znanja, i 2) relacije između podataka. Da bi se napravila razlika između neka dva sistema za vrednovanje podataka, potrebno je definisati načine upoređivanja i, ukoliko je to moguće, kriterijume rangiranja različitih sistema vrednovanja.

Tradicionalno se neka metoda može proveriti: 1) upoređivanjem rezultata metode sa merenim vrednostima, 2) upoređivanjem rezultata metode sa rezultatima složenije metode koja pruža veću tačnost ili 3) namernim uvođenjem anomalija u podatke za testiranje i proverom odgovora sistema. S obzirom na to da se pouzdanost podataka uglavnom ne meri, cilj procesa vrednovanja podataka je upravo generisanje jedne takve informacije, pa se ne može očekivati da se kvalitet metode proveri sa nečim što, u stvari, ne postoji. Kod nekih metoda merenja (npr. ultrazvučnih merila brzine i nivoa) može se generisati informacija o kvalitetu odbijenog signala kojim je dobijen podatak. Međutim, kvalitet signala samo delimično opisuje pouzdanost podatka.

U radovima [14], [15] i [11] navode se metode za upoređivanjesistema za vrednovanje po nekoliko osnova. Ukoliko se ocene pouzdanosti svedu na binarni oblik, i podaci se klasifikuju u dve grupe – pouzdani i nepouzdani, ocena kvaliteta se može preformulisati i svesti na prebrojavanje detektovanih anomalija, nedetektovanih podataka sa anomalijama i pogrešno detektovanih podataka bez anomalija:

$$p = \frac{N_{registered}}{N_{anomalies} + N_{missed} + N_{registered\ nonanomalies}}, \tag{4.9}$$

gde su $N_{registered}$ broj registrovanih anomalija, $N_{anomalies}$ ukupan broj podataka sa anomalijama, N_{missed} broj neregistrovanih podataka sa anomalijama, ($N_{anomalies}$ - $N_{registered}$), $N_{registered\ nonanomalies}$ registrovanih podataka bez anomalija. Može se primetiti da je ocena p jednaka jedinici jedino kada su registrovani svi i samo podaci sa anomalijama, tj. kada su $N_{registered\ nonanomalies} = 0$ i $N_{missed} = 0$.

Alternativno, ako podataka ima puno i ukoliko se zarad registrovanja svih podataka sa anomalijama mogu žrtvovati neki podaci bez anomalija, postupak određivanja ocene pouzdanosti se može preformulisati:

$$p_{false\ tolerant} = \frac{N_{registered}}{N_{anomalies} + N_{missed}}, \tag{4.10}$$

Sa druge strane, ukoliko su podaci retki i ukoliko se oni sa malim anomalijama mogu zadržati i proglasiti ispravnim, postupak određivanjaocene može se formulisati na drugi način:

$$p_{false\ sensitive} = \frac{N_{registered}}{N_{anomalies} + N_{registered\ nonanomalies}}, \tag{4.11}$$

Svi navedeni parametri zahtevaju neku referentnu metodu sa kojom bi rezultati testirane metode bili upoređeni. S obzirom da se ni za jednu metodu ne može apsolutno reći da je superiorna u odnosu na ostale i da se može proglastiti za referentnu, za potrebe testiranja razvijene metodologije u ovoj doktorskoj disertaciji korišćeni su rezultati ekspertskog vrednovanja na osnovu vizuelnog utiska i iskustva. Zbog toga se u narednom odeljku prikazuje rezultat istraživanja tradicionalnog pristupa vrednovanja – vrednovanja vizuelizacijom podataka.

4.5 Vrednovanje podataka vizuelizacijom

Tradicionalni pristup vrednovanju podataka vezan je za njihovu vizuelizaciju (grafički prikaz). Vizuelnom predstavom podataka se stvara slika ne samo o samom podatku i njegovoj veličini, već i o njegovom odnosu sa ostalim podacima iste i drugih merenih veličina. Na detekciju anomalija u podacima utiču:

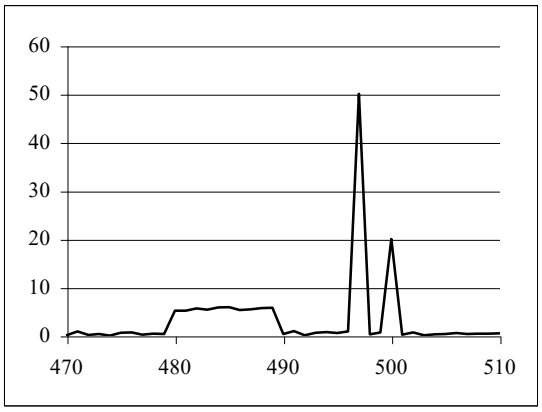
1. grafička prezentacija podataka i odabranih osobina podataka (*features*);
2. vizuelni doživljaj (*perception*) podataka;
3. način zaključivanja o regularnosti podataka.

Grafička prezentacija podataka predstavlja prvi korak u procesu vrednovanja podataka vizuelizacijom. Osnovni cilj grafičke prezentacije je izdvajanje osobina podatka koje su od interesa za proces vrednovanja. Ukoliko je grafičkom prezentacijom neka osobina sakrivena ili nedovoljno izražena, smanjuje se mogućnost da se u toku vizuelnog doživljaja i zaključivanja u pogledu regularnosti podatka donesu ispravni zaključci. Vizuelni doživljaj je vezan za psihološki fenomen obrade slike i način kako naš vizuelni sistem organizuje sirove podatke u konkretne objekte ili delove objekata. Zaključak o regularnosti nekog podatka se dalje donosi na osnovu znanja, iskustva i vizuelnog doživljaja grafičke predstave podataka.

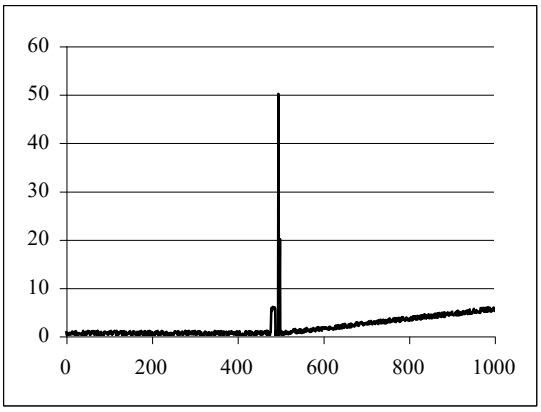
4.5.1 Grafička prezentacija podataka

Grafička prezentacija podataka (vizuelizacija podataka) predstavlja prvi korak u procesu detekcije anomalija u podacima. Ona predstavlja šematizovanu predstavu podataka kao kombinaciju definisanih atributa koji predstavljaju jedinice informacije [38]. Glavni cilj vizuelizacije podataka je prenos informacija vizuelnom percepcijom. Vizuelna reprezentacija podataka može se povezati sa mnogim vrstama podataka (npr. podacima dobijenim kliničkim ispitivanjem, podacima o terenu, itd.). Vizuelizacija merenih podataka prikupljenih za potrebe naučnih i praktičnih disciplina vezanih za hidrotehniku pripada grupi postupaka vizuelizacije naučnih podataka. Podaci vezani za hidrotehnku uglavnom su u formi vremenskih serija vezanih za prostor (*time-space series*).

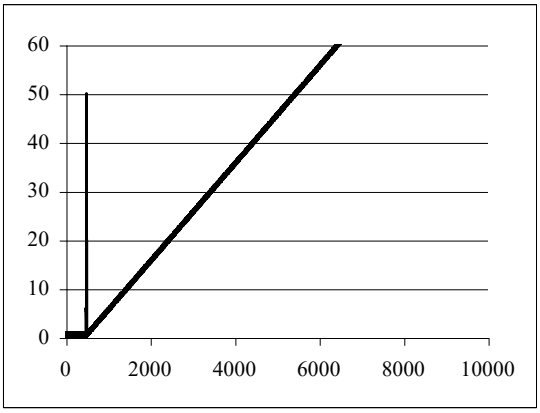
Veliki uticaj na donošenje zaključaka o postojanju anomalija u podacima ima način na koji su podaci i njihove osobine predstavljeni. Na slici 4.13 prikazana su tri dijagrama sa vremenskom serijom (konstantna vrednost + šum) kojoj su dodate tri anomalije: 1) konstantno odstupanje, 2) dve anomalije u obliku šiljka i 3) linearno odstupanje. Može se zaključiti da širina vremenske ose u kombinaciji sa rezolucijom monitora ima uticaj na donošenje zaključaka o postojanju i vrsti anomalija u podacima. Na slici 4.13A prikazan je odgovarajući deo vremenske skale tako da se lako uočavaju svi detalji vezani za prve dve anomalije, dok se treća anomalija ne može vizuelno izolovati. Na slici 4.13B prve dve anomalije uočavaju se sa manje detalja, ali se treća anomalija jasnije može detektovati. Na slici 4.13C prve tri anomalje vide se kao jedna, dok je treća anomalija izuzetno uočljiva.



A: x=[470,510], y=[0,60]



B: x=[0,1000], y=[0,60]



C: x=[0,10000], y=[0,60]

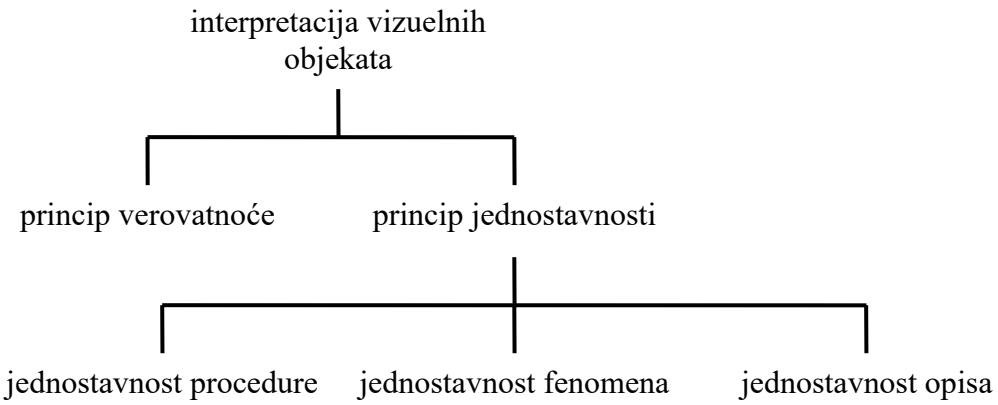
Slika 4.13: Testiranje algoritma za prikaz vremenske serije u zavisnosti od veličine vremenske skale u MsExcel softverskom paketu

Postoje različiti načini da se podaci prikažu grafički, a način kako se oni mogu interpretirati proučava posebna grana psihologije koja se bavi vizuelnim doživljajem.

4.5.2 Vizuelni doživljaj

Proučavanje reakcije čoveka na vizuelnu stimulaciju počelo je tzv. *Gestalt* školom psihologije [76].

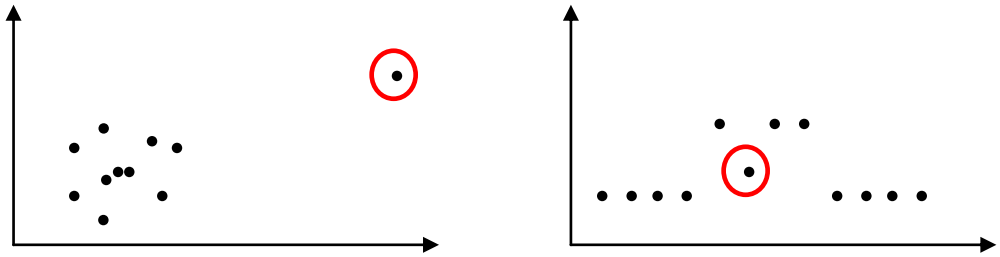
Dva osnovna pristupa kojima se objašnjava reakcija čoveka na vizuelni stumulans su princip jednostavnosti i princip najverovatnije interpretacije (*likelihood*), slika 4.14 [120]. Princip najverovatnije interpretacije zahteva poznavanje dodatnih informacija, pa i zaključak o tome šta je video čovek donosi na osnovu iskustva. Princip jednostavnosti je formalizovan tzv. teorijom strukturne informacije (*structural information theory*, SIT). SIT predstavlja teoriju o ljudskoj percepciji i načinu na koji vizuelni sistem čoveka organizuje i prihvata vizuelne stimulacije različitog tipa [71]. Prema principu jednostavnosti, informacija se prima kao najjednostavnija moguća reprezentacija. Najjednostavnija moguća reprezentacija se odabira tako što se prvo uočavaju sve pravilnosti i hijerarhijska organizacija vizuelnih objekata i njihovih delova.



Slika 4.14: Interpretacija vizuelnog doživljaja

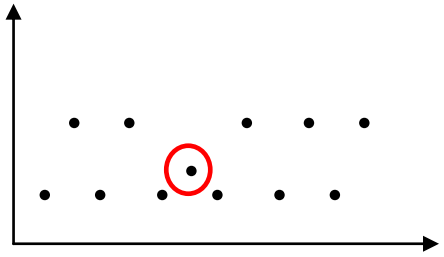
Teorijom strukturne informacije može se objasniti činjenica da je neke anomalije u podacima lakše vizuelno uočiti od drugih i uputiti na to šta se može uraditi da se uočljivost anomalija poboljša.

Na slici 4.15 prikazano je šest grupa podataka iz kojih je potrebno izdvojiti jedan ili više podataka koji nisu regularni. Na slici 4.15A grafički je prikazano više podataka koji se mogu svrstati u jednu grupu i jedan podatak koji očigledno ne pripada grupi. Ukoliko se podaci na ovoj slici ispitaju u pogledu postojanja neregularnih podataka, očigledan kandidat je podatak koji ne pripada grupi. Na slici 4.15B prikazana je grupa podataka koji su poređani u simetričnom redu. Podatak koji odudara iz simetrije je, u ovoj grupi podataka, očigledan kandidat za anomaliju. Slična situacija je i kod grupe podataka prikazane na slici 4.15C gde su prikazani podaci koji se ponavljaju (periodični podaci). Na slici 4.15D prikazana je serija podataka koji se javljaju van nekog jednostavnog šablona. Kod ovakvih serija nije očigledno izdvajanje anomalija iz podataka. Dvosmislene podatke (slike 4.15E i 4.15F) predstavljaju podaci kod kojih se javlja više podataka ili grupa podataka koji su različiti pa nije očigledno koji podaci pripadaju grupi regularnih, a koji grupi neregularnih podataka.

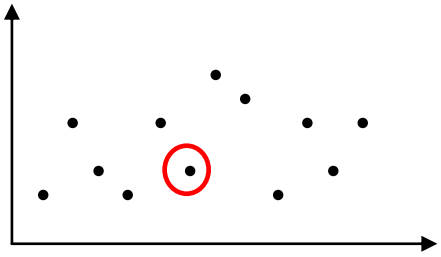


A: Grupisani podaci

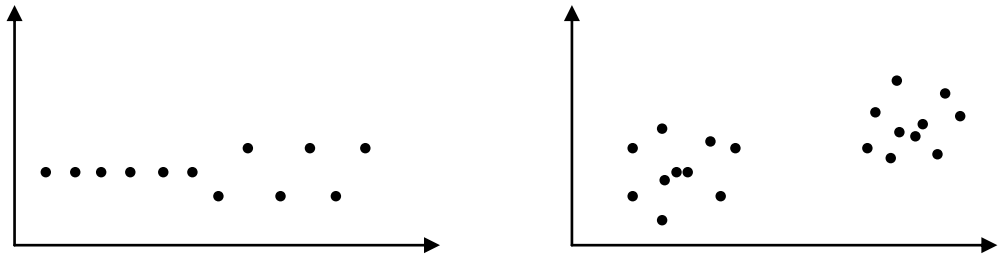
B: Simetrični podaci



C: Podaci koji se ponavljaju (periodični)



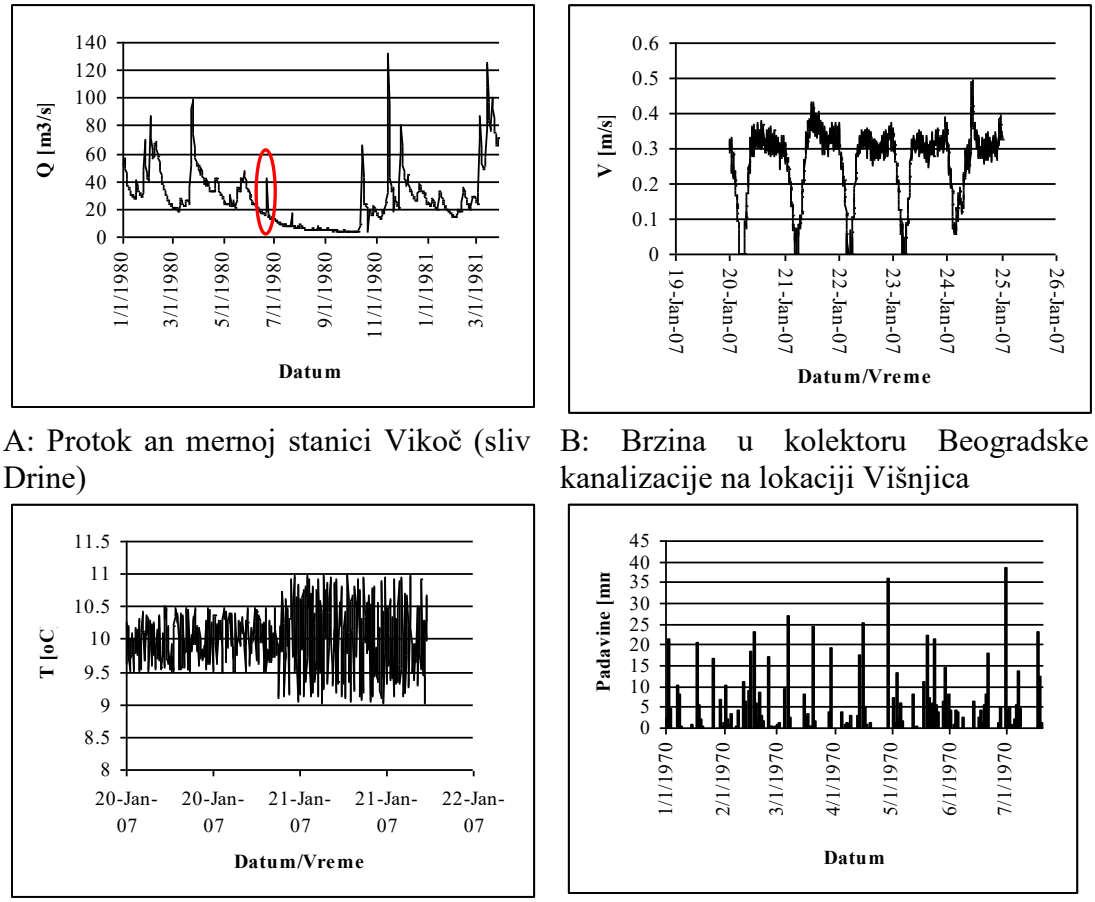
D: Podaci van nekog očiglednog šablona



E: Dvosmisleni podaci F: Dvosmisleni podaci

Slika 4.15: Registrovanje anomalija u podacima vizuelnom percepcijom

Može se zaključiti da se anomalije u podacima lakše uočavaju kod podataka kod kojih su izraženi bliskost, ponavljanje ili simetrija (antisimetrija). Hidrotehnički podaci mogu se naći u skoro svim navedenim oblicima. Na slici 4.16 prikazani su primeri serija hidrotehničkih veličina kod kojih se mogu uočiti pravilnosti (slike 4.16A i 4.16B), dvosmislenih podataka (slika 4.16C) i podataka koji se javljaju van nekog očiglednog šablona (slika 4.16D).



A: Protok an mernoj stanici Vikoč (sliv Drine)

B: Brzina u kolektoru Beogradske kanalizacije na lokaciji Višnjica

C: Temperatura vode u kanalizacionom kolektoru

D: Padavine na mernoj stanici Nova Varoš (sliv Drine)

Slika 4.16: Primeri serija podataka koji se koriste u hidrotehnici

Mora se zapaziti da su vremenske serije merenih podataka uglavnom opterećene šumom, kao i da se postojanje šablona (simetričnosti, periodičnosti i ponavljanja) uočava i tamo gde su jedinice šablona skalirane (slika 4.16A).

Transformacijom podataka modelima i izdvajanjem i kombinovanjem osobina podataka (*feature engineering*) može se poboljšati utisak o postojanju pravilnosti između podataka. Postojanje pravilnosti u grafičkoj predstavi podataka može se matematički oceniti pomoću algorithmic information theory – AIT [106], koja proučava odnos između načina prikaza i količine informacija koju prikaz nosi [102].

Ukoliko se u serijama podataka ne mogu uočiti pravilnosti, izdvajanje podataka sa anomalijama predstavlja izuzetno komplikovan zadatak u procesu vizuelnog vrednovanja podataka. Zbog toga se često zaključivanje o regularnosti nekog podatka oslanja, pored vizuelne percepcije, i na dodatna znanja i iskustva o merenoj veličini, uslovima merenja, itd.

4.5.3 Zaključivanje o regularnosti podatka

Zaključivanje o regularnosti podatka se u velikoj meri oslanja na već postojeća znanja, iskustva i dostupne korisne informacije. Zbog toga je vizuelni doživljaj samo jedan od činilaca u procesu vrednovanja podataka. Iako se principom najverovatnije interpretacije vizuelnog doživljaja unosi i deo iskustvenog znanja, zaključak o regularnosti podataka donosi se i na osnovu dodatnih informacija koje često i nisu grafički prikazane. Na taj način se unosi dodatni kvalitet u odluku da li je podatak anomalija ili ne. Dodatne informacije upućuju na pojedine podatke ili delove vremenskih serija na koje je potrebno obratiti posebnu pažnju pri vizuelnom vrednovanju podataka. Dodatne informacije se mogu grupisati na sledeći način:

- alarmi (nestanak struje, pad napona baterija, registrovane greške, itd.);
- informacije o okruženju (kiša/sneg/grad);
- druge merene veličine (trendovi, zajednički pikovi više veličina, itd.);
- ograničenja (fizičke granice, itd.);
- itd.

Često se brojne dodatne informacije teško prate, a ponekad i različito tumače od strane različitih eksperata. Zbog toga se rezultati tradicionalnog pristupa vrednovanju podataka razlikuju od eksperta do eksperta. Uzrok tome su različiti nivoi znanja, različito iskustvo, itd. Takođe, često je teško dokumentovati, a kasnije i protumačiti proces vrednovanja koji je sproveden na osnovu iskustva. Upotreba nekog definisanog algoritma na bazi relacija između podataka, i vrednovanje podataka na deterministički način, pomoću računara, uneli bi uniformnost u proces vrednovanja, jednostavno dokumentovanje procesa vrednovanja i mogućnost da se različita vrednovanja podataka uporede.

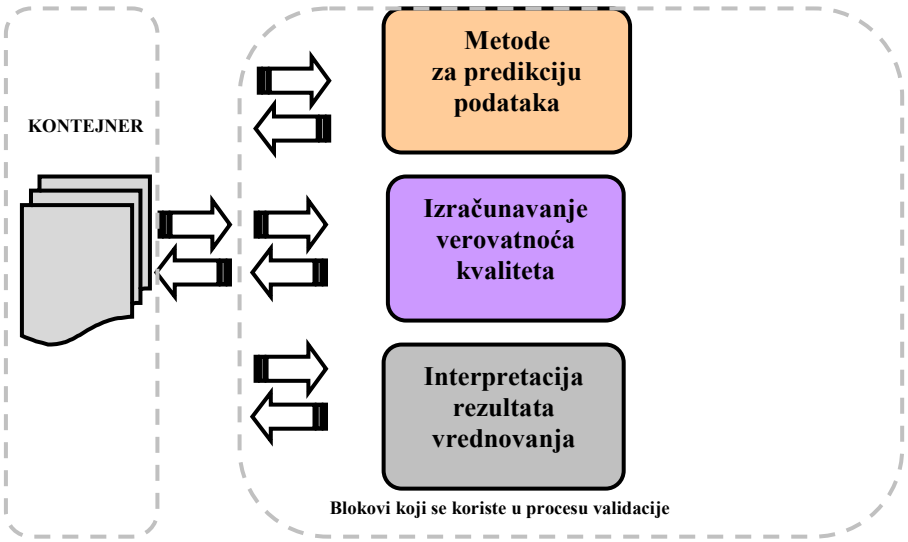
Primer zaključivanja o regularnosti podatka na osnovu vizuelne percepcije bi se mogao vezati za podatak o protoku zaokružen na slici 4.16A. Naime, u vremenskoj seriji se vizuelno može uočiti šablon koji se može dalje povezati sa podacima o kišama. Kada kiša padne javlja se, sa određenim zakašnjenjem, skok u protoku (poplavni talas). Protok dalje tokom suvog vremena eksponencionalno opada. Ukoliko se uoči skok u vremenskoj seriji protoka kada nije bilo kiše, ili ako skok izostane kada kiše ima, može se zaključiti da je u pitanju neregularan podatak. Sistem za vrednovanje predložen u ovoj disertaciji koristi logiku iza vizuelne percepcije u obliku matematički formulisanih relacija između podataka. Tako se u matematičkom modelu kiša-oticaj nalaze matematičke formule koje opisuju upravo skok protoka kada kiša padne i eksponencijalni pad protoka kada kiše nema. Ukoliko to nije postignuto matematičkim modelom, onda model nije adekvatan i treba ga odbaciti.

4.6 Implementacija metodologije u MatLab-u

Metodologija je implementirana u programskom paketu MatLab. Pošto procedura validacije podataka zavisi, pre svega, od karakteristika osmatranog procesa, a dalje i od raspoloživih dodatnih informacija i znanja, jasno je da se za svaku vremensku seriju koja predstavlja osmatrani proces mora obezbediti specifičan proces sastavljen od odgovarajućih relacija i metoda. Da bi se sistem za validaciju i obradu podataka učinio što fleksibilnijim i prilagodljivijim, osmišljen je tako da se za svaku vremensku seriju može sastaviti iz delova (blokova) koji se mogu uklapati u celinu koja bi bila prilagođena vremenskoj seriji čiji se podaci vrednuju. To znači da je obezbeđeno da se arhitektura postupka validacije za svaku vremensku seriju može osmisliti i implementirati pomoću tipskih blokova koji predstavljaju objekte za obradu i vrednovanje podataka, pri čemu podaci koji se vrednuju, podaci koji se odnose na proces vrednovanja i rezultati međukoraka obrade i vrednovanja treba da su u svakom trenutku dostupni.

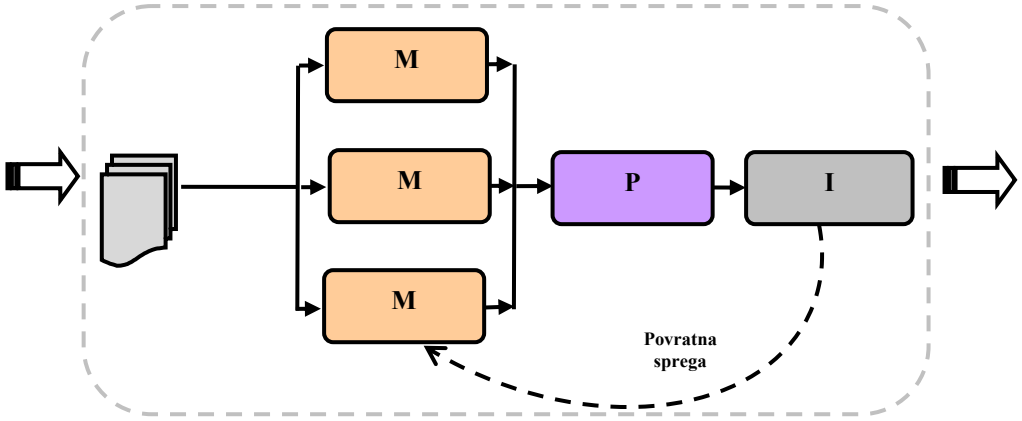
Sisitem za vrednovanje sastoji se iz tri dela:

1. *kontejnera*, tj. objekta koji čuva podatke u procesu vrednovanja (u njemu se nalaze podatak koji se vrednuje, parametri metoda za izračunavanje predikcija, ukoliko je potrebno izabrani istorijski podaci, rezultati međukoraka, dodatne informacije koje se koriste u procesu vrednovanja, itd.);
2. *blokova* u kojima se obavlja neki segment vrednovanja (objekti za primenu metoda vrednovanja, objekti za računanje verovatnoća regularnosti podataka i objekti za interpretaciju rezultata validacije i automatsko upravljanje procesom);
3. *veza* koje označavaju redosled izvršavanja blokova procesa vrednovanja i koje zajedno sa blokovima čine strukturu sistema za vrednovanje.



Slika 4.17: Mesto kontejnera i blokova u sistemu za validaciju i obradu podataka

Na slici 4.17 prikazana je šema pozicija objekata u sistemu za vrednovanje, njihove međusobne relacije i njihova komunikacija. Za svaku vremensku seriju potrebno je dizajnirati raspored blokova u sistemu za vrednovanje, u koji se uključuju i blokovi za interpretaciju rezultata i upravljanje procesima vrednovanja (povratne sprege).



Slika 4.18: Primer rasporeda (strukture) blokova i veza u sistemu za vrednovanje podataka

Podaci koji su potrebni da se izvrši neki korak u procesu vrednovanja (predikcija metodama, izračunavanje verovatnoća ili interpretacija rezultata) nalaze se u kontejneru (vrednovani podatak, ostali podaci koji se kao ulazne veličine koriste u metodama za predikciju, parametri metoda, itd.).

Struktura sistema za vrednovanje je linearnog tipa, tj. ne postoje paralelena izvršavanja niti ciklične veze za pojedine podatke. Povratna sprega koja je naznačena na slici 4.18 označava funkcionalnost koja se ogleda u upravljanju sistemom u narednim vremenskim koracima, tako da se u zavisnosti od rezultata vrednovanja nekog podatka on, na primer, može isključiti iz procedure vrednovanja nekog drugog podatka.

5. Primeri primene razvijene metodologije

Metodologija predložena u ovoj tezi testirana je na tri primera:

1. hipotetičkom opštem primeru;
2. hipotetičkom hidrotehničkom primeru; i
3. na realnim podacima merenim na sistemu Beogradskog vodovoda i kanalizacije.

Prvi primer (hipotetički opšti primer) odabran je da bi se ilustrovao koncept predloženog algoritma i da se na slikovit način pokažu situacije kada algoritam daje dobre rezultate, a kada ti rezultati nisu zadovoljavajući. Drugi primer predstavlja hipotetički primer iz oblasti hidrotehnike. U ovom primeru sistem je testiran u odnosu na neke česte oblike grešaka u podacima. Treći primer je odabran tako da se na realnim merenim podacima prikažu rezultati predloženog algoritma i uporede sa rezultatima iz literature. Kod trećeg primera je sprovedena procedura pripreme sistema za vrednovanje koja se može sažeti u sledeća četiri koraka:

1. pretprocesiranje;
2. određivanje granica podataka;
3. definisanje relacija između podataka;
4. primena algoritma na odabrane podatke vremenske serije.

Pretprocesiranje podrazumeva neophodnu pripremu podataka koja se ogleda u konverziji podataka u odgovarajuće jedinice, usklađivanje formata podataka, detekciju konteksta u kome se podaci nalaze, itd. Takođe se u procesu pretprocesiranja određuju neki parametri vremenskih serija koji mogu pomoći u definisanju granica podataka ili relacija između podataka (nivo šuma, nivo autokorelacije, nivo korelacije, itd.). Nakon pretprocesiranja pristupa se određivanju granica koje podatak ne bi trebalo da prekorači. Granice mogu biti definisane na osnovu fizičkih ograničenja ili statistički. Definisanje relacija između podataka ne radi se ukoliko relacije ne postoje. Ukoliko se relacije mogu odrediti, u trećem koraku se razmatraju prvo relacije koje su bazirane na fizičkim zakonitostima, zatim statističke relacije, i na kraju *data mining* relacije. Relacije između veličina kalibrisane su na osnovu izabranog skupa podataka kod koga su ili prethodno ručno odstranjeni sumnjivi podaci na osnovu vizuelne inspekcije i iskustva, ili se pristupilo metodama koje je moguće koristiti u prisustvu anomalija u podacima kao što je robusna regresija. Relacije takođe treba validovati i odrediti im inverzni oblik. Nakon pripreme podataka, definisanja granica i relacija između podataka, podaci se trestiraju predloženim algoritmom.

5.1 Hipotetički primer – postavka problema

Vrednovanje podataka predloženim algoritmom prikazano je na hipotetičkom primeru[8] u kome učestvuju tri merene veličine, x_1, x_2 i x_3, predstavljene pomoću intervala (tabela 5.1).

Tabela 5.1: Izmerene vrednosti u obliku egzaktnih vrednosti i intervala

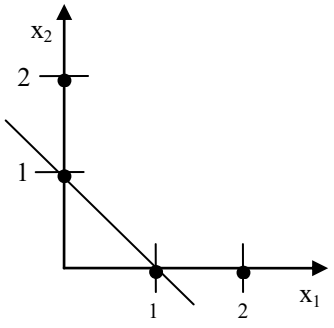
	x_1	x_2	x_3
Egzaktne vrednosti	0.25	0.75	1.25
Intervali	[0.24, 0.26]	[0.74, 0.76]	[1.24, 1.26]

[8] *Napomena*: Primer je tako odabran da se omogući egzaktna veza između merenih veličina. Treba naglasiti da se pretpostavlja da je relacija egzaktna, tj. da nema neodređenost, što u praksi nikada nije slučaj.

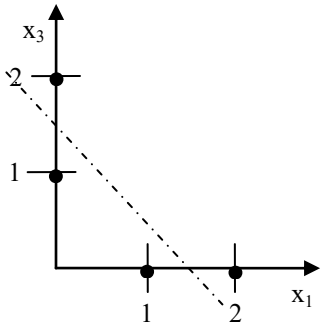
Merene veličine povezane su relacijama R_1, R_2 i R_3 prikazanim u tabeli 4.2 i grafički prikazanim na slici 5.1.

Tabela 5.2: Relacije R_1, R_2 i R_3 između veličina x_1, x_2 i x_3

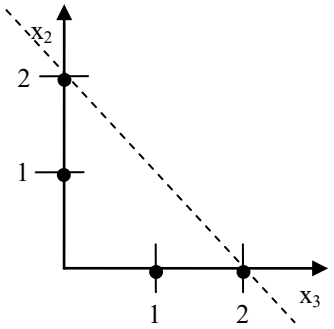
R_1	$x_1 + x_2 = 1$
R_2	$x_1 + x_3 = 1.5$
R_3	$x_2 + x_3 = 2$



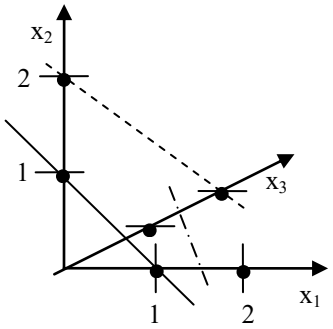
$R_1: x_1 + x_2 = 1$



$R_2: x_1 + x_3 = 1.5$



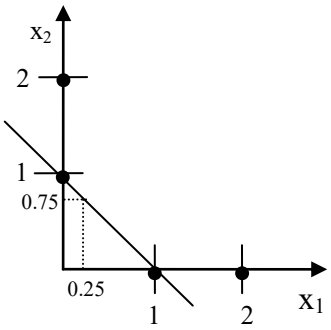
$R_3: x_2 + x_3 = 2$

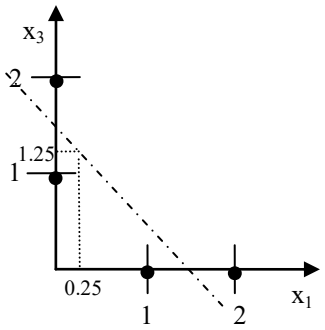


Prostorni prikaz ravni koje se formiraju relacijama R_1, R_2 i R_3

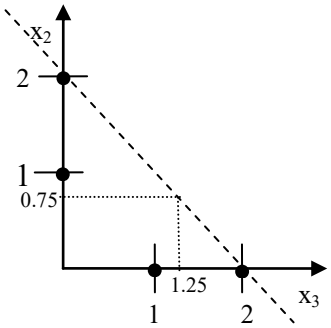
Slika 5.1: Grafička prezentacija relacija R_1, R_2 i R_3

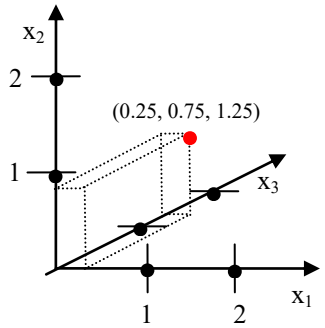
Svaka od merenih veličina učestvuje u po dve metode. U primeru se pretpostavlja da su metode egzaktne i da nemaju neodređenosti. Predikcije merenih veličina x_1 ($x_1^{M_{R_1,x_1}}$ i $x_1^{M_{R_2,x_1}}$) dobijaju se pomoću metoda M_{R_1,x_1} i M_{R_2,x_1}, x_2 ($x_2^{M_{R_1,x_2}}$ i $x_2^{M_{R_3,x_2}}$) pomoću metoda M_{R_1,x_2} i M_{R_3,x_2}, i x_3 ($x_3^{M_{R_2,x_3}}$ i $x_3^{M_{R_3,x_3}}$) pomoću metoda M_{R_2,x_3} i M_{R_3,x_3}. Izračunate vrednosti u egzaktnom obliku grafički su prikazane na slici 5.2, a izračunate u tabeli 5.3.





$R_1: \; x_1 + x_2 = 0.25 + 0.75 = 1$ $R_2: \; x_1 + x_3 = 0.25 + 1.25 = 1.5$





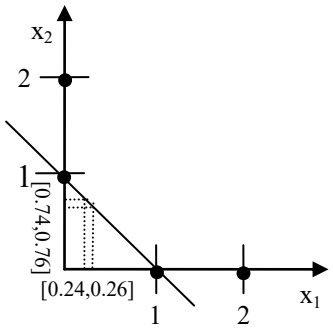
$R_3: \; x_2 + x_3 = 0.75 + 1.25 = 2$ Prostorni prikaz rešenja

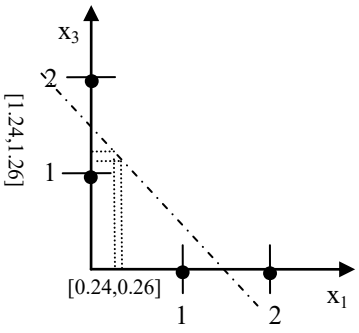
Slika 5.2: Primeri predikcije vrednosti merenih veličina x_1 ($x_1^{M_{R_1,x_1}}$ i $x_1^{M_{R_2,x_1}}$), x_2 ($x_2^{M_{R_1,x_2}}$ i $x_2^{M_{R_3,x_2}}$) i x_3 ($x_3^{M_{R_2,x_3}}$ i $x_3^{M_{R_3,x_3}}$) u egzaktnom obliku

Tabela 5.3: Izračunate vrednosti merenih veličina x_1 ($x_1^{M_{R_1,x_1}}$ i $x_1^{M_{R_2,x_1}}$), x_2 ($x_2^{M_{R_1,x_2}}$ i $x_2^{M_{R_3,x_2}}$) i x_3 ($x_3^{M_{R_2,x_3}}$ i $x_3^{M_{R_3,x_3}}$) u egzaktnom obliku

	x_1	x_2	x_3
R_1	$x_1^{M_{R_1,x_1}} = 1 - x_2 = 1 - 0.75 = 0.25$	$x_2^{M_{R_1,x_2}} = 1 - x_1 = 1 - 0.25 = 0.75$	
R_2	$x_1^{M_{R_2,x_1}} = 1.5 - x_3 = 1.5 - 1.25 = 0.25$		$x_3^{M_{R_2,x_3}} = 1.5 - x_1 = 1.5 - 0.25 = 1.25$
R_3		$x_2^{M_{R_3,x_2}} = 2 - x_3 = 2 - 1.25 = 0.75$	$x_3^{M_{R_3,x_3}} = 2 - x_2 = 2 - 0.75 = 1.25$

Ukoliko se merene veličine predstave intervalima, i izračunate vrednosti se dobijaju u obliku intervala. Na slici 5.3 prikazane su relacije i vrednosti u obliku intervala, u tabeli 5.4 su te vrednosti izračunate. Na slici 5.5 dobijeni rezultati grafički su uporeĐeni u formi matrice.





M_{R_1,x_2}:

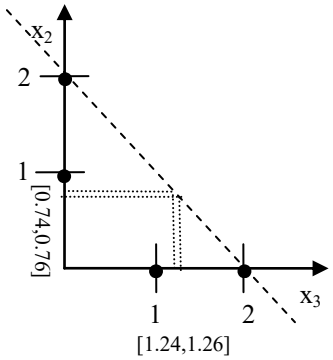
$x_2^{M_{R_1,x_2}} = 1 - x_1$

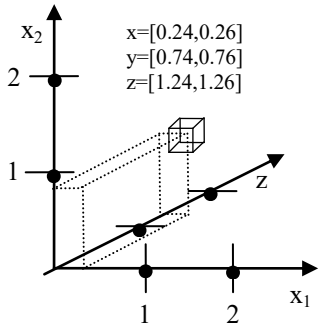
$= 1 - [0.24, 0.26] = [0.74, 0.76]$

M_{R_2,x_3}:

$x_3^{M_{R_2,x_3}} = 1 - x_1$

$= 2 - [0.24, 0.26] = [1.24, 1.26]$





M_{R_3,x_2}:

$x_2^{M_{R_3,x_2}} = 2 - x_3$

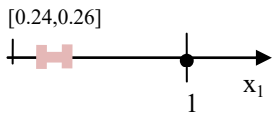
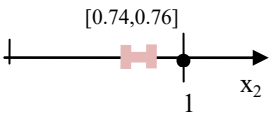
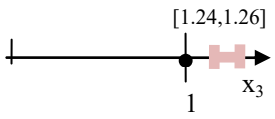
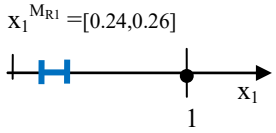
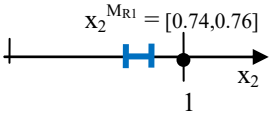
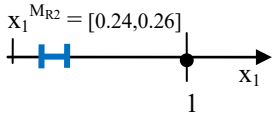
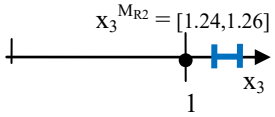
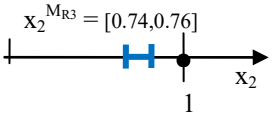
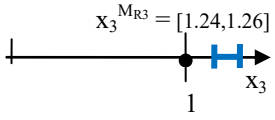
$= 2 - [1.24, 1.26] = [0.74, 0.76]$

Prikaz rešenja u prostoru

Slika 5.3: Primeri predikcije vrednosti merenih veličina x_1 ($x_1^{M_{R_1,x_1}}$ i $x_1^{M_{R_2,x_1}}$), x_2 ($x_2^{M_{R_1,x_2}}$ i $x_2^{M_{R_3,x_2}}$) i x_3 ($x_3^{M_{R_2,x_3}}$ i $x_3^{M_{R_3,x_3}}$) u obliku intervala

Tabela 5.4: Predikcije vrednosti merenih veličina x_1 ($x_1^{M_{R_1,x_1}}$ i $x_1^{M_{R_2,x_1}}$), x_2 ($x_2^{M_{R_1,x_2}}$ i $x_2^{M_{R_3,x_2}}$) i x_3 ($x_3^{M_{R_2,x_3}}$ i $x_3^{M_{R_3,x_3}}$) u obliku intervala

	x_1	x_2	x_3
R_1	M_{R_1,x_1} : $x_1^{M_{R_1,x_1}} = 1 - x_2 = 1 - [0.74, 0.76] = [0,24, 0.26]$	M_{R_1,x_2} : $x_2^{M_{R_1,x_2}} = 1 - x_1 = 1 - [0.24, 0.26] = [0.74, 0.76]$	
R_2	M_{R_2,x_1} : $x_1^{M_{R_2,x_1}} = 1.5 - x_3 = 1.5 - [1.24, 1.26] = [0.24, 0.26]$		M_{R_2,x_3} : $x_3^{M_{R_2,x_3}} = 1.5 - x_1 = 1.5 - [0.24, 0.26] = [1.24, 1.26]$
R_3		M_{R_3,x_2} : $x_2^{M_{R_3,x_2}} = 2 - x_3 = 2 - [1.24, 1.26] = [0.74, 0.76]$	M_{R_3,x_3} : $x_3^{M_{R_3,x_3}} = 2 - x_2 = 2 - [0.74, 0.76] = [1.24, 1.26]$

	x_1	x_2	x_3
Mereno			
R_1			
R_2			
R_3			

Slika 5.5: Predikcije vrednosti merenih veličina x_1 ($x_1^{M_{R_1,x_1}}$ i $x_1^{M_{R_2,x_1}}$), x_2 ($x_2^{M_{R_1,x_2}}$ i $x_2^{M_{R_3,x_2}}$) i x_3 ($x_3^{M_{R_2,x_3}}$ i $x_3^{M_{R_3,x_3}}$) u obliku intervala prikazane u formi tabele

Na osnovu izmerenih i izračunatih vrednosti koje se poklapaju testirana je metodologija, tj. lako se pokazuje:

1. da su podaci o slaganju izmerenih vrednosti i predikcija jednaki jedinici ($p(x_i \mid x_i^{M_{R_j,x_i}}) = 1$);
2. da su podaci o verovatnoćama su $p(x_i^{M_{R_j,x_i}} \mid X_i^{M_{R_j,x_i}}) = 1/k$, gde je $k = 2$ broj predikcija vrednosti $x_i^{M_{R_j,x_i}}$ merene veličine x_i ;
3. da je reprezentativna verovatnoća $p\left(x_i, X_{x_i}, MM_i\right) = 1$ izmerene vrednosti jednaka je jedinici;
4. da je reprezentativna izračunata vrednost $E[x_i^M \mid X_{x_i}^M]$ jednaka intervalu jednakom predikciji.

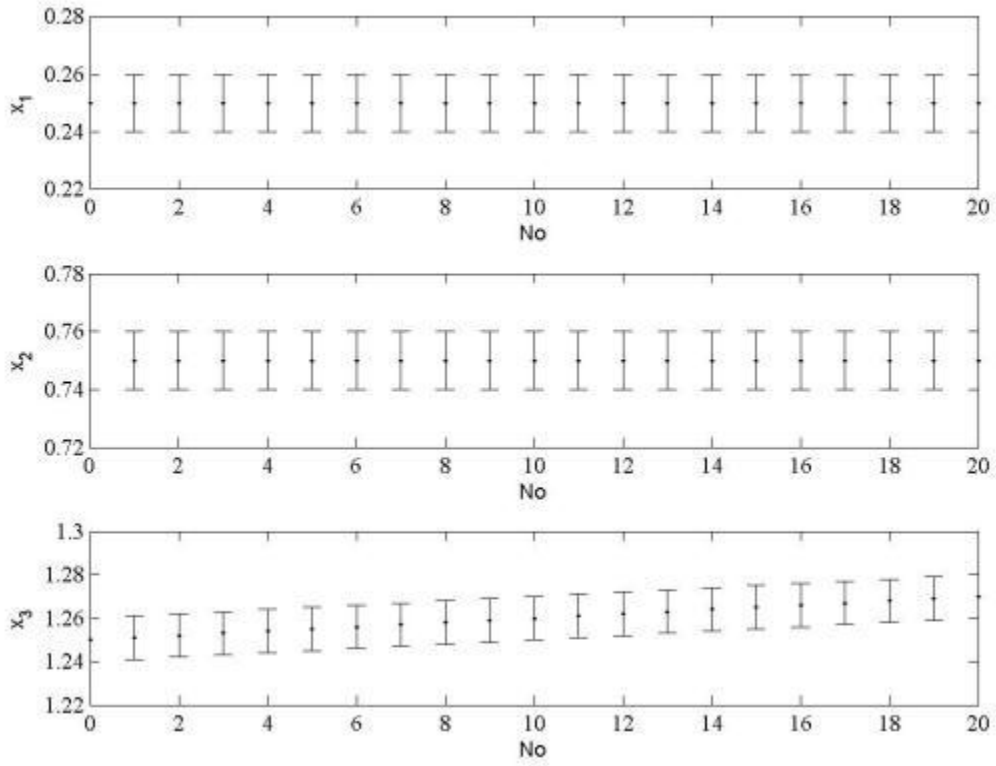
Ovaj hipotetički primer može se upotrebiti i da se potvrdi algoritam, tj. da se ispitaju njegovi rezultati kod pojave grešaka u merenim podacima i kod povećane neodreĎenosti modela koji se koristi u metodi za predikciju (neodreĎenost u relaciji izmeĎu podataka). U daljem tekstu su sprovedena tri numerička eksperimenta uz sledeće pretpostavke:

1. povećanjem greške ε_{x_k} veličine x_k smanjuju se $p(x_k \mid x_k^{M_{R_j,x_k}})$ i $p(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}})$ kod izračunatih vrednosti veličina kod kojih je x_k ulazna vrednost;
2. povećanjem neizvesnosti u_{x_k} metode M_{R_k,x_i} smanjuju se $p(x_i \mid x_i^{M_{R_k,x_i}})$ i $p(x_i^{M_{R_k,x_i}} \mid X_i^{M_{R_k,x_i}})$ kod predikcija vrednosti veličina koje se računaju modelom M_{R_k,x_i} ;
3. povećanjem greške dve veličine (npr. x_1 i x_2) sa različitim znakom za sličnu vrednost može se doći u situaciju da rezultat algoritma bude da je greška u trećoj (x_3).

5.1.1 Hipotetički primer – rezultati i diskusija

Numerički eksperiment 1

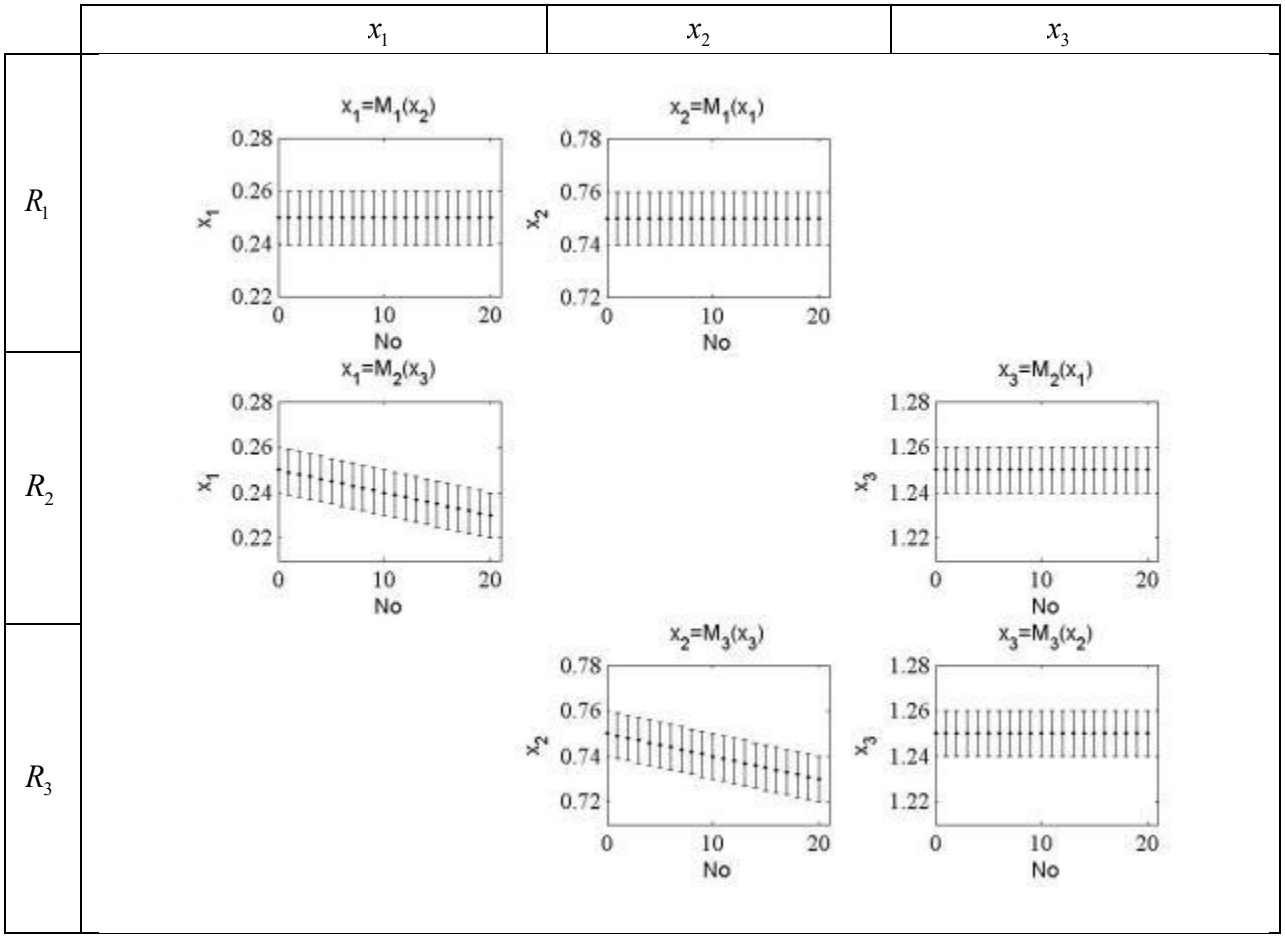
Ovaj numerički eksperiment je sproveden da bi se pokazalo kako se menjaju težinski koeficijenti rezultata modela, ocene merenih vrednosti i procenjene vrednosti sa povećanjem greške u jednoj od veličina. Zbog toga se namerno uvodi greška u veličinu x_3 , $x_3^{'} = x_3 + \varepsilon_{x_3}$, gde je $\varepsilon_{x_3} = [0, 0.001, \ldots, 0.02]$.



Slika 5.6: Dvadeset slučajeva merenih vrednosti veličina x_1, x_2 i x_3

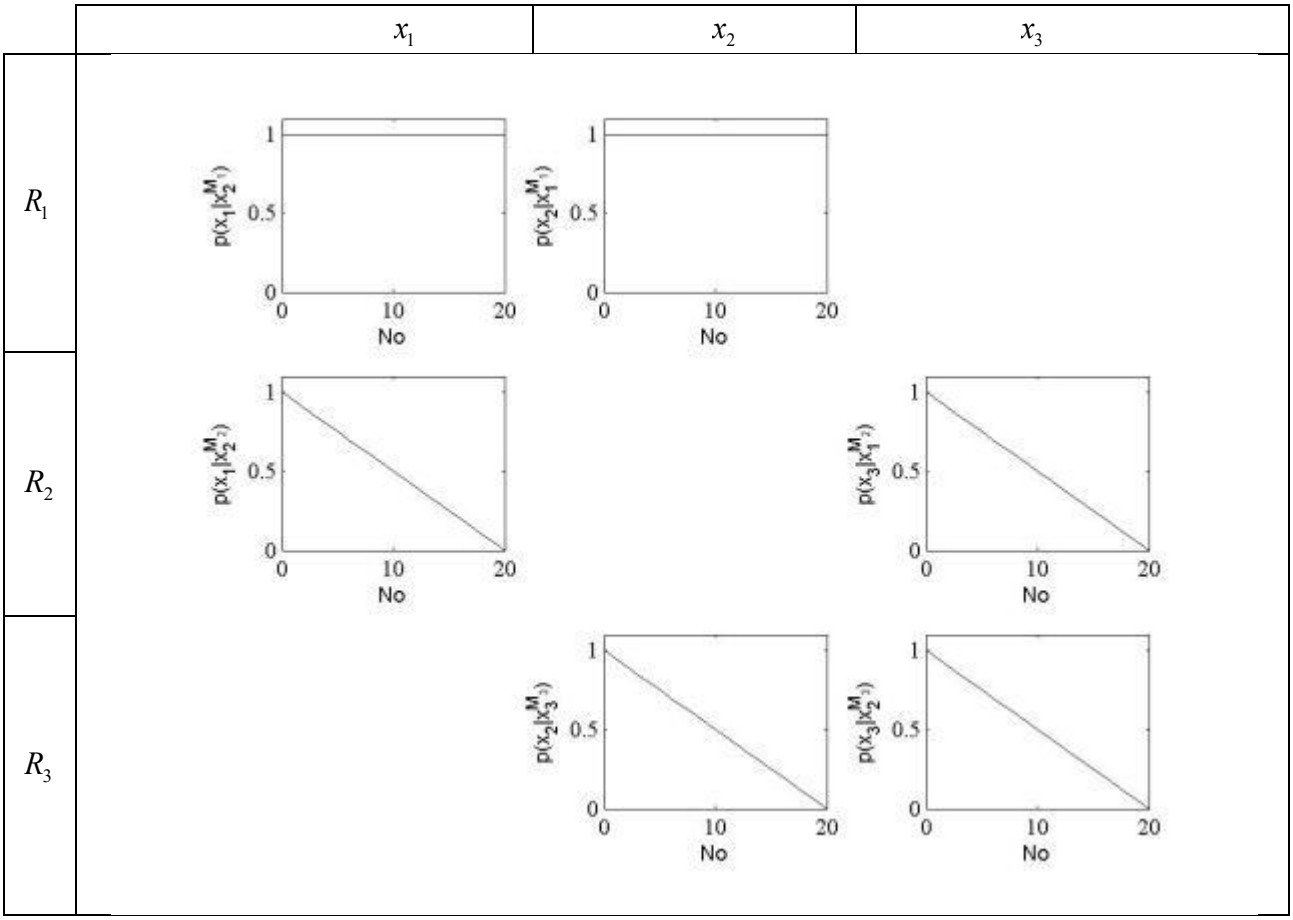
Na slici 5.6 prikazano je dvadeset slučajeva (ne računajući prvi kod koga je greška jednaka nuli) kod kojih veličine x_1 i x_2 ostaju konstantne, a veličina x_3 sadrži grešku. Može se videti da se u poslednjem slučaju x_3 čak ni ne seče sa intervalom koji predstavlja tačnu vrednost.

Kada se za svih dvadeset slučajeva pomoću metoda M_{R_1,x_1}, M_{R_2,x_1}, M_{R_1,x_2}, M_{R_3,x_2}, M_{R_2,x_3} i M_{R_3,x_3} izračunaju predikcije veličina koje se mere, dobijaju se rezultati prikazani na dijagramima slike 5.7. Kod prve relacije (R_1 - metoda M_{R_1,x_1} i M_{R_1,x_2}) nema promena jer su i veličine x_1 i x_2 ostale nepromenjene. Kod relacije R_2, tj. metoda M_{R_2,x_1} i M_{R_2,x_3}, uočava se da predikcija veličine x_1 sadrži grešku, zbog greške u x_3 kao ulaznom podatku. Isti slučaj je i sa relacijom R_3, tj. metodama M_{R_3,x_2} i M_{R_3,x_3} i veličinom x_2. Kod obe relacije pri predikciji veličine x_3 ne registruje se greška jer ulazne veličine (x_1 i x_2) ne sadrže grešku.



Slika 5.7: Rezultati izračunatih veličina x_1, x_2 i x_3 pomoću metoda izvedenih iz relacija R_1, R_2 i R_3

Verovatnoće slaganja izračunatih i izmerenih veličina ($p(x_i \mid x_i^{M_{R_j,x_i}})$ izračunate po izrazu 4.1) mogu se videti na slici 5.8. Kod metoda izvedenih iz relacije R_1 verovatnoće $p(x_1 \mid x_1^{M_{R_1,x_1}})$ i $p(x_2 \mid x_2^{M_{R_1,x_2}})$ su jednake jedinici jer se odgovarajući izračunati i izmereni intervali poklapaju. Što se tiče predikcija metoda izvedenih iz relacija R_2 i R_3, verovatnoće opadaju do nule. Razlog tome je ili izmerena vrednost sa greškom ($p(x_3 \mid x_3^{M_{R_2,x_3}})$ i $p(x_3 \mid x_3^{M_{R_3,x_3}})$) ili izračunata vrednost sa greškom ($p(x_1 \mid x_1^{M_{R_2,x_1}})$ i $p(x_2 \mid x_2^{M_{R_3,x_1}})$).



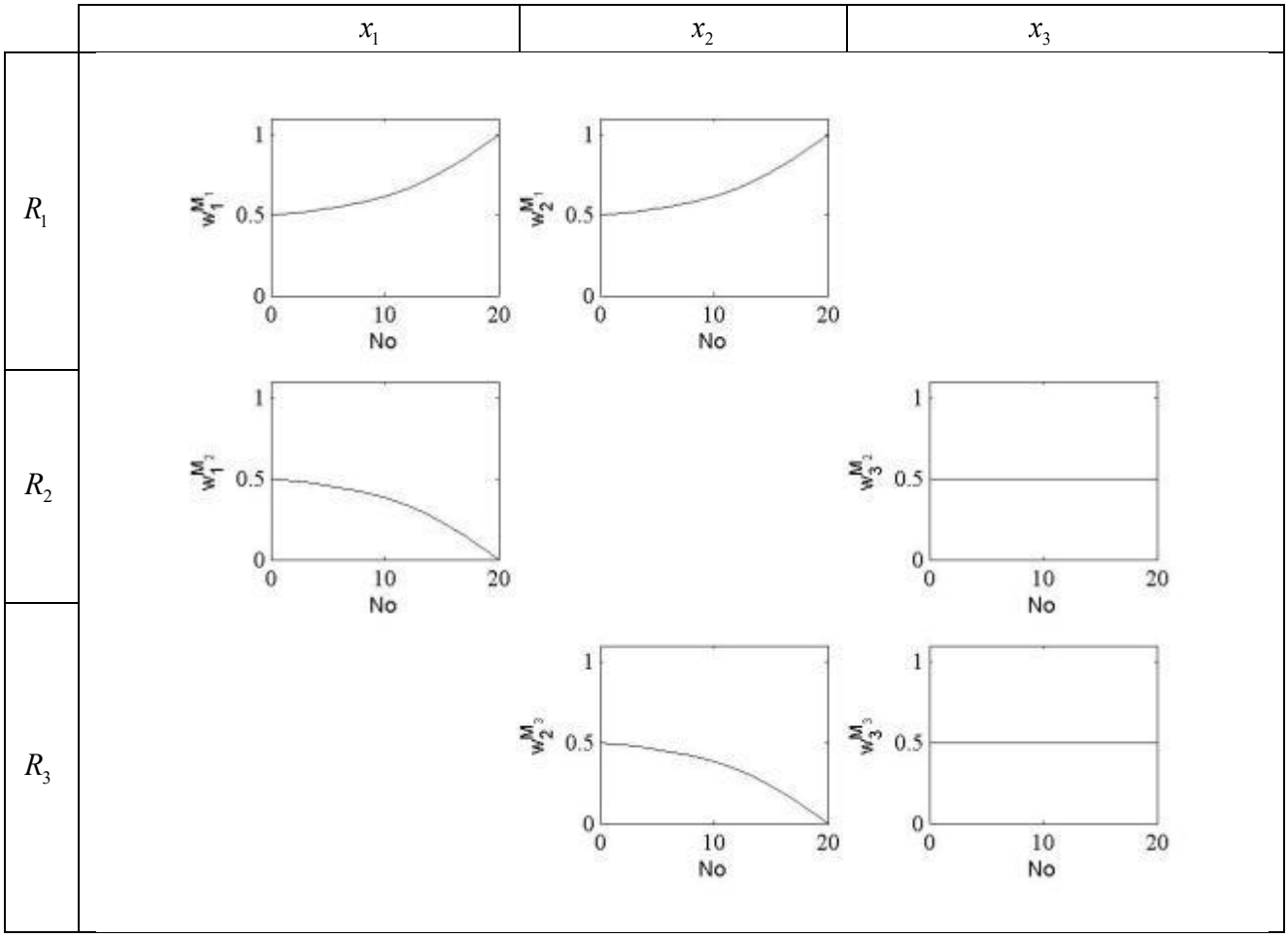
Slika 5.8: Uslovne verovatnoće slaganja izmerenih i izračunatih vrednosti

Izračunavanjem maksimalne verodostojnosti zbira slaganja izmerenih i izračunatih vrednosti:

$$J = \max \sum_{i=1}^{3}\sum_{j=1}^{3} p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right) = \max \sum_{i=1}^{3}\sum_{j=1}^{3} p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right) p\left(x_i^{M_{R_j,x_i}} \mid X_{x_i}^{M_{R_j,x_i}}\right)$$
$$= \max \sum_{i=1}^{3}\sum_{j=1}^{3} w_i^{M_{R_j,x_i}} \times p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right),$$

dobijaju se težinski koeficijenti kojima su slaganja pomnožena $w_i^{M_{R_j,x_i}}$, uz $\sum_j w_i^{M_{R_j,x_i}} = 1$.

Na slici 5.9 prikazani su težinski koeficijenti za svaku predikciju vrednovanih veličina izračunatih po izrazu 4.2. Vidi se da se kod predikcija vrednosti x_1 i x_2 pomoću metoda izvedenih iz relacija R_2 i R_3 sa porastom greške kod veličine x_3 smanjuju i težinski koeficijenti $w_1^{M_{R_2,x_1}}$ i $w_2^{M_{R_3,x_2}}$. Sa druge strane, težinski koeficijenti uz predikcije veličine x_3 pomoću metoda izvedenih iz istih relacija ostaju konstantni. Smanjenje težinskih koeficijenata uz predikcije veličina x_1 i x_2 pomoću metoda izvedenih iz relacija R_2 i R_3 prati povećanje odgovarajućih koeficijenata uz predikcije metodama izvedenim iz relacije R_1 zbog uslova $w_1^{M_{R_1,x_1}} + w_1^{M_{R_2,x_1}} = 1$ i $w_2^{M_{R_1,x_2}} + w_2^{M_{R_3,x_2}} = 1$.

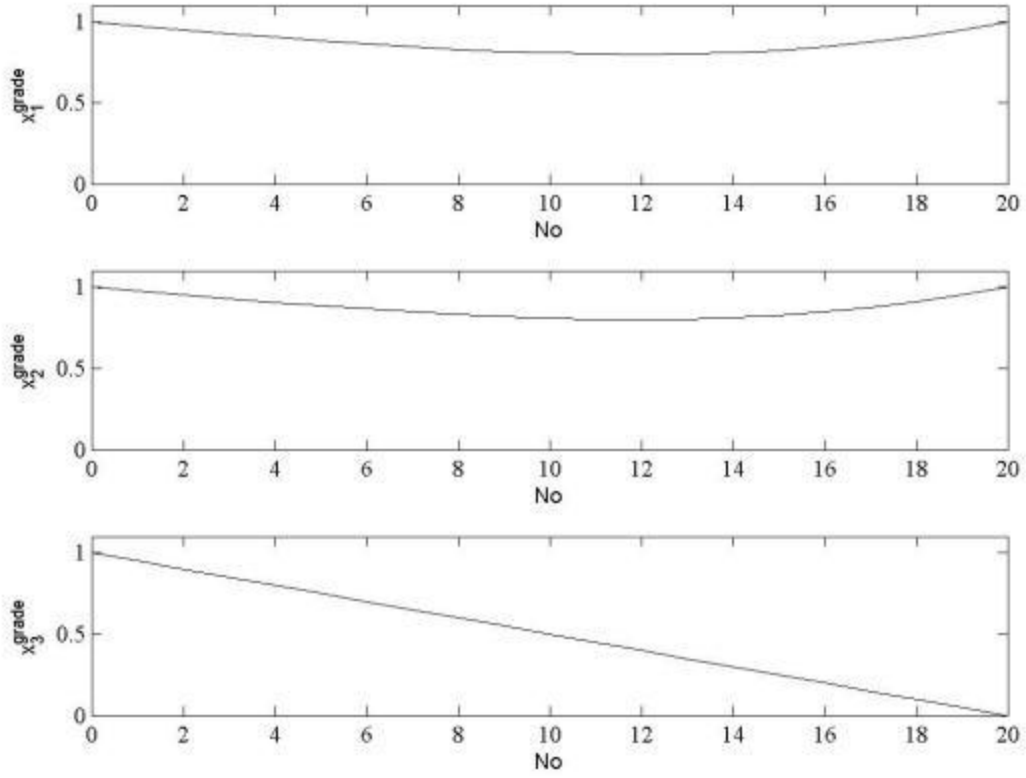


Slika 5.9: Težinski koeficijenti uz izračunate predikcije vrednovanih veličina $w_i^{M_{R_j,x_i}}$

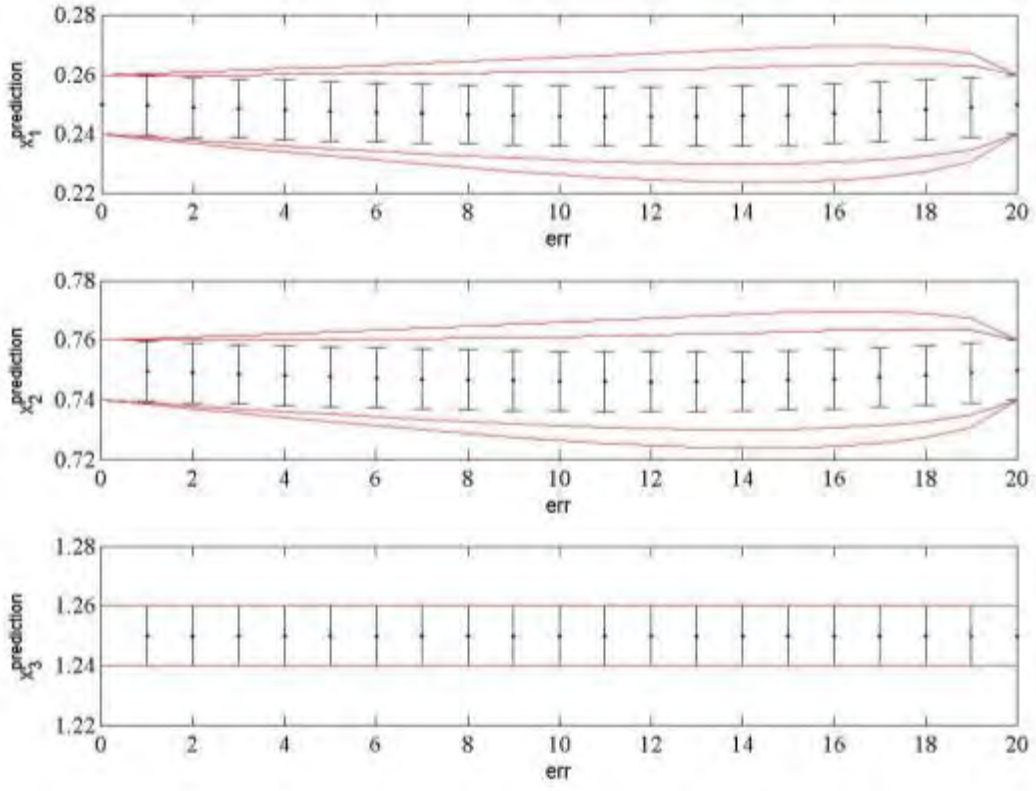
Nakon izračunavanja težinskih koeficijenata, izračunate su i verovatnoće regularnosti merenih podataka, izrazom 4.3, koje se mogu shvatiti i kao relativne ocene kvaliteta podataka:

$$x_1^{grade} = w_1^{M_{R_1,x_1}} \times p\left(x_1 \mid x_1^{M_{R_1,x_1}}, x_2\right) + w_1^{M_{R_2,x_1}} \times p\left(x_1 \mid x_1^{M_{R_2,x_1}}, x_3\right);$$
$$x_2^{grade} = w_2^{M_{R_1,x_2}} \times p\left(x_2 \mid x_2^{M_{R_1,x_2}}, x_1\right) + w_2^{M_{R_3,x_2}} \times p\left(x_2 \mid x_2^{M_{R_3,x_2}}, x_3\right);$$
$$x_3^{grade} = w_3^{M_{R_2,x_3}} \times p\left(x_3 \mid x_3^{M_{R_2,x_3}}, x_1\right) + w_3^{M_{R_3,x_3}} \times p\left(x_3 \mid x_3^{M_{R_3,x_3}}, x_2\right).$$

Na slici 5.10 vidi se da verovatnoće svih merenih veličina u početku opadaju, ali da se verovatnoće veličina x_1 i x_2 vrate na maksimalnu vrednost kada veličina x_3 potpuno izaĎe iz opsega tačne vrednosti. TakoĎe se može videti da su u apsolutnim vrednostima verovatnoće veličina x_1 i x_2 uvek više od verovatnoća veličine x_3, što navodi na zaključak da veličina x_3 sadrži grešku. Pretpostavlja se da bi se sa porastom broja podataka u odnosu na koje se neka veličina vrednuje, smanjilo i opadanje verovatnoća veličina koje su regularne.



Slika 5.10: Verovatnoće regularnosti merenih podataka



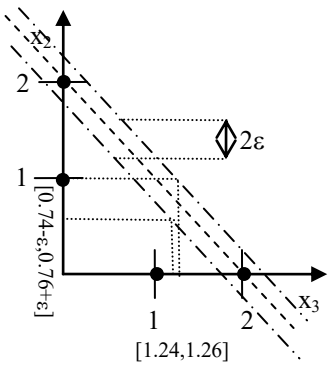
Slika 5.11: Predikcija vrednovanih veličina na osnovu izračunatih vrednosti

Pored verovatnoća regularnosti merenh podataka, moguće je izračunati i reprezentativnu vrednost i varijansu merenih podataka na osnovu izračunatih podataka i težinskih koeficijenata, izrazom 4.4 i 4.5. Na slici 5.11 može se primetiti da su očekivane vrednosti sve tri veličine tačne vrednosti, ali da kod veličina x_1 i x_2 varijansa predikcije raste, a zatim i opada, da bi se, na kraju, vratila u nulu. Razlog tome je razlika između vrednosti izračunatih pomoću dva modela koji se koriste za svaku od tih veličina. Kod veličine x_3 varijansa je jednaka nuli jer se obe predikcije poklapaju sa istim težinskim koeficijentima.

Ovim numeričkim eksperimentom pokazana je osnovna funkcija razvijenog algoritma koja se ogleda u izračunavanju težinskih koeficijenata uz uslovne verovatnoće slaganja predikcija i izmerenih vrednosti. Izračunati težinski koeficijenti uz rezultate relacija sa neregularnim parametrima kao ulaznim veličinama, umanjeni su u odnosu na rezultate relacija sa regularnim parametrima kao ulaznim vrednostima.

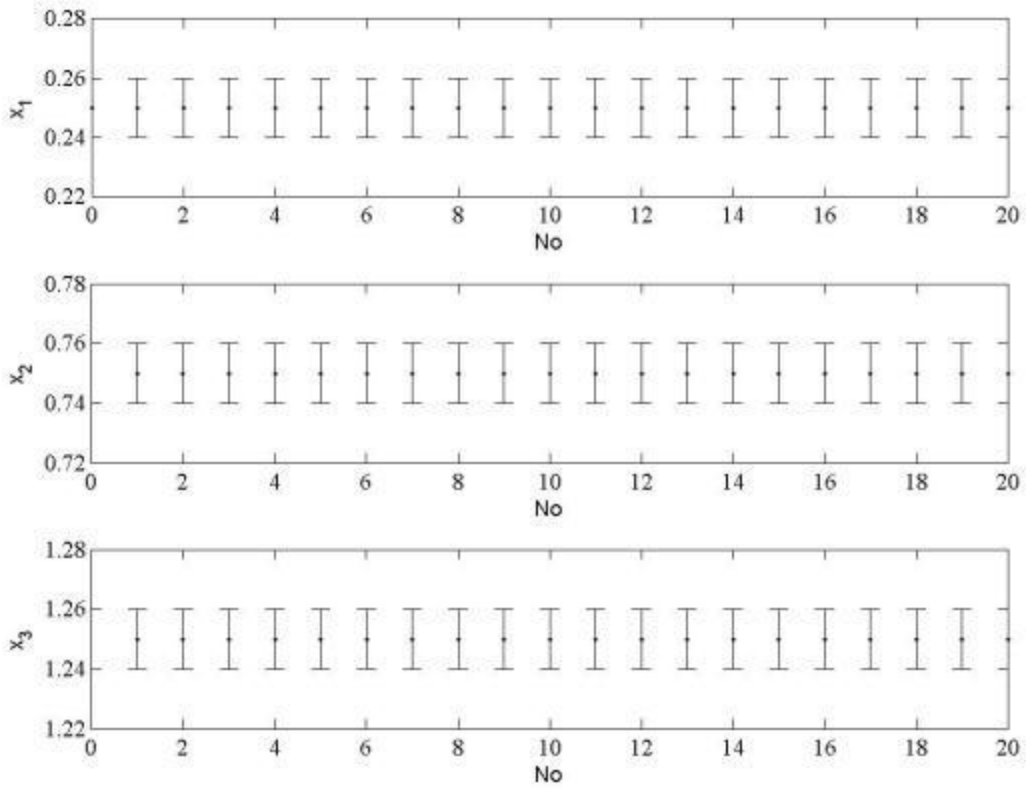
Numerički eksperiment 2

U ovom numeričkom primeru ispituje se ponašanje sistema za različite neodređenosti relacije R_3 iz koje su izvedene metode M_{R_3,x_2} i M_{R_3,x_3}. Pretpostavlja se povećanje neizvesnosti metoda M_{R_3,x_2} i M_{R_3,x_3}: $x_2 + x_3 = [2-\varepsilon, 2+\varepsilon]$, gde je $\varepsilon = [0, 0.001, \ldots, 0.02]$, kao što je prikazano na slici 5.12.



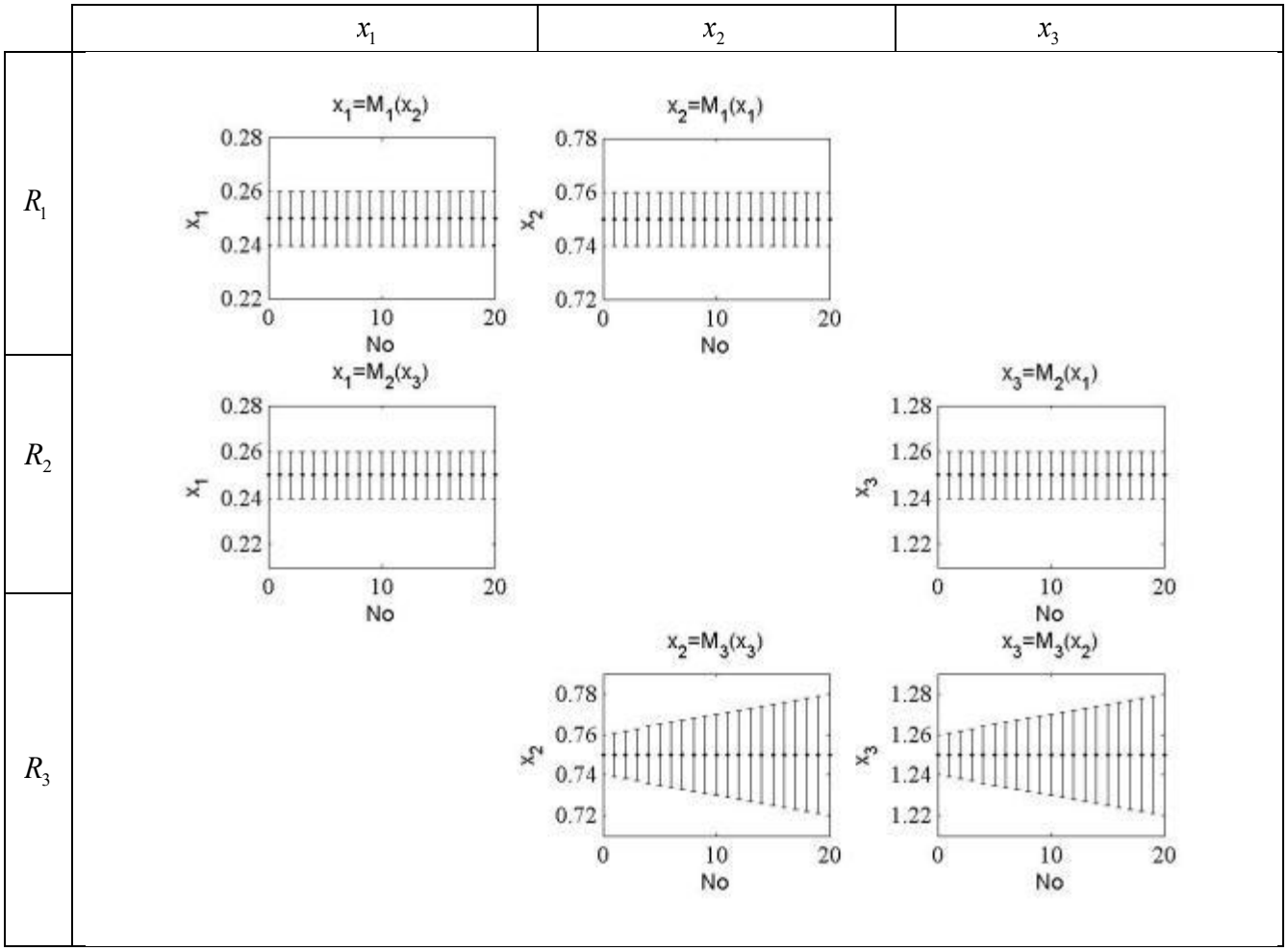
Slika 5.12: Relacija R_3 sa povećanom neodređenosti

Za tačne vrednosti veličina koje se vrednuju (slika 5.13), izračunate predikcije prikazane su na slici 5.14.



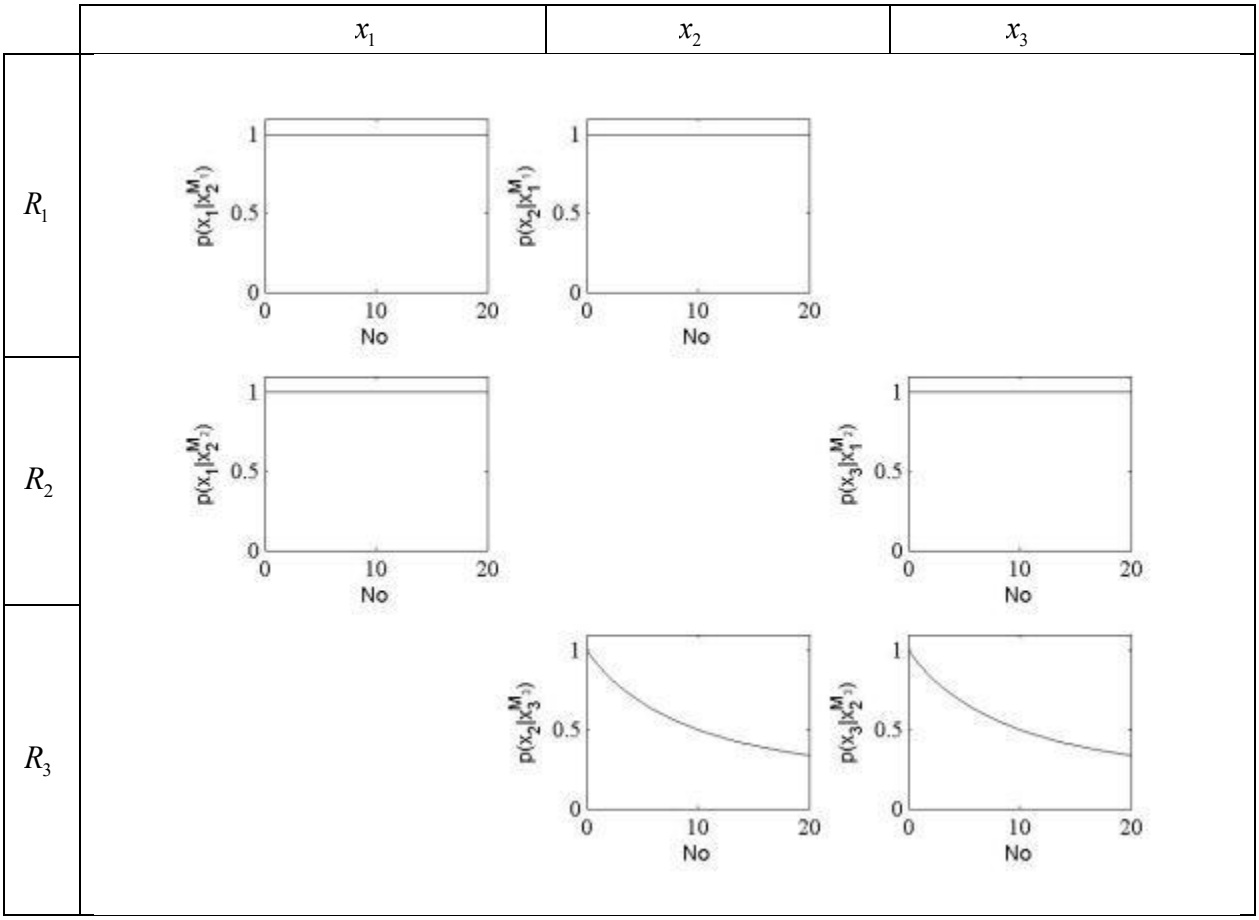
Slika 5.13: Veličine koje se vrednuju: $x_1 = [0.24, 0.26]$, $x_2 = [0.74, 0.76]$ i $x_3 = [1.24, 1.26]$

Na slici 5.14 se vidi da neizvesnost predikcija veličina x_2 i x_3 raste sa porastom neizvesnosti metoda izvedenih iz relacije R_3. S obzirom na to da se neizvesnost metoda izvedenih iz relacije R_3 povećava simetrično u odnosu na srednju liniju, simetrično je i povećanje neizvesnosti i predikcija.



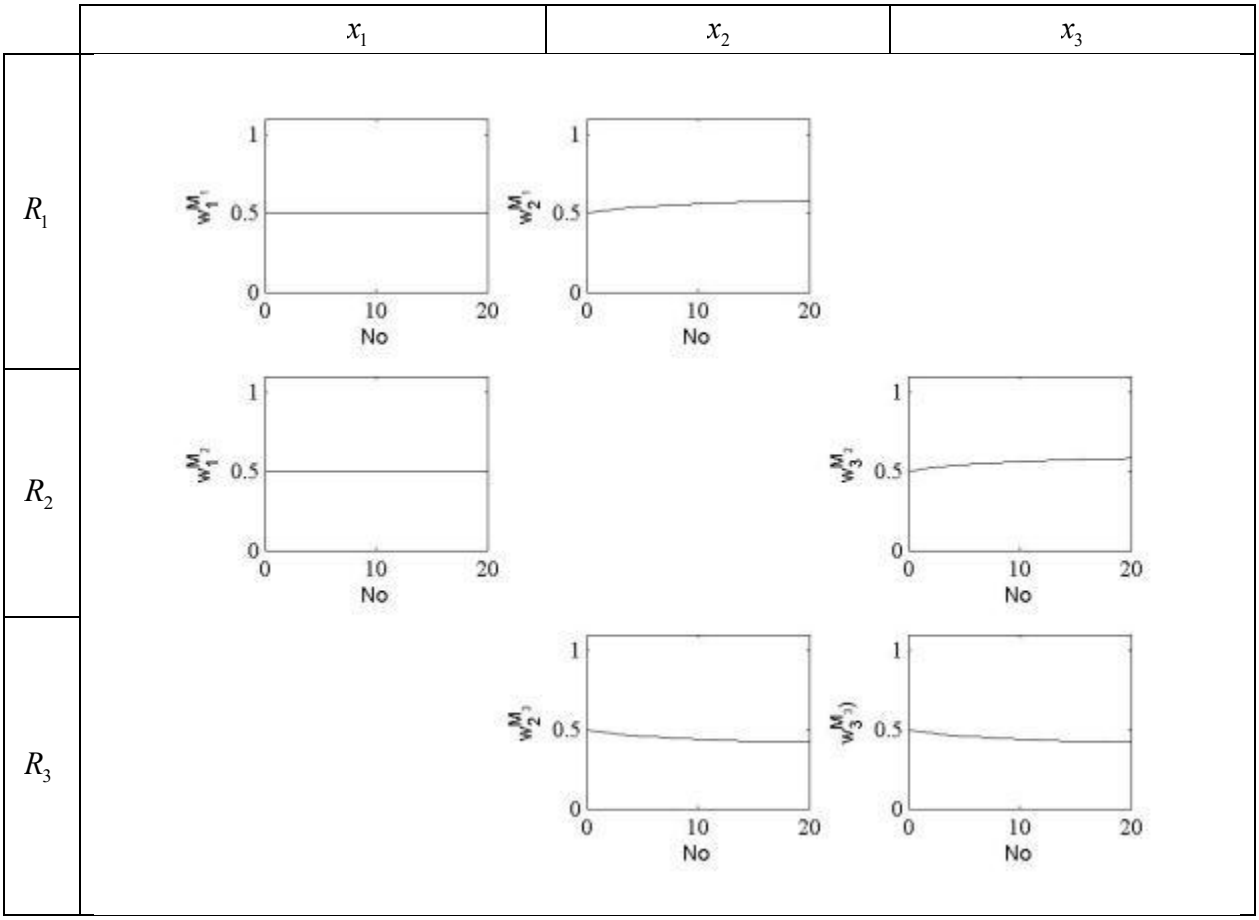
Slika 5.14: Izračunate vrednovane veličine

Povećana neizvesnost odražava se i na uslovne verovatnoće slaganja izmerenih veličina i njihovih predikcija, što se vidi na slici 5.15. Kod predikcija vrednosti metodama izvedenih iz relacije R_3 navedene uslovne verovatnoće opadaju asimptotski prema nuli. Nulu nikada neće dostići (što se može zaključiti iz definicije uslovne verovatnoće), a opadanje zavisi direktno od neodreĎenosti izračunate vrednosti.



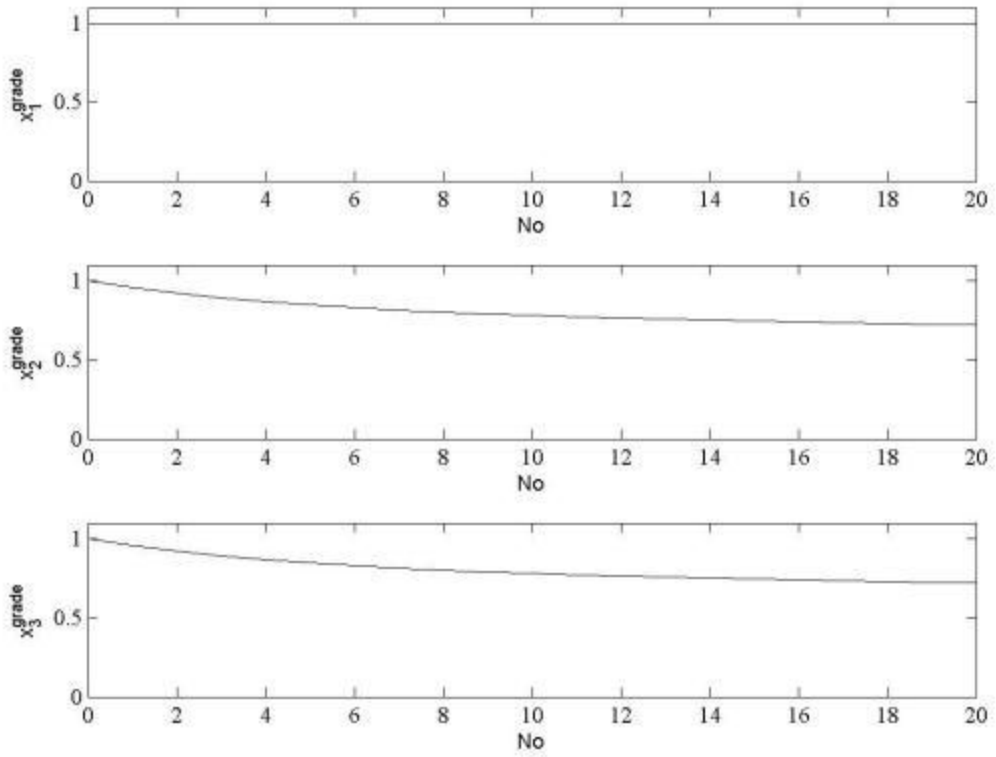
Slika 5.15: Uslovne verovatnoće slaganja izmerenih i izračunatih vrednosti

Težinski koeficijenti računati predloženim algoritmom prikazani su na slici 5.16. Vidi se da kod predikcija vrednosti metodama izvedenim iz relacije R_3, težinski koeficijenti opadaju sa porastom neodreĎenosti. Time se kažnjavaju manje pouzdane relacije, a nagraĎuju one sa manjom neodreĎenosti zbog uslova $\sum_j w_i^{M_{R_j \cdot x_i}} = 1$.



Slika 5.16: Težinski koeficijenti uz izračunate vrednosti

Ukupne verovatnoće izmerenih vrednosti mogu se dalje izračunati i prikazane su na slici 5.17. Vidi se da su ocene veličina x_2 i x_3 koje su vrednovane metoma izvedenim iz manje pouzdane relacije R_3 niže od odgovarajućih ocena veličine x_1. TakoĎe se može uočiti da je pad ocena u korelaciji sa neodreĎenosti relacije, čime se pokazuje da ocena vrednovanja zavisi od preciznosti (neodreĎenosti) relacije koja se koristi, pa se vrednovanje podataka može uporeĎivati iu tom pogledu.

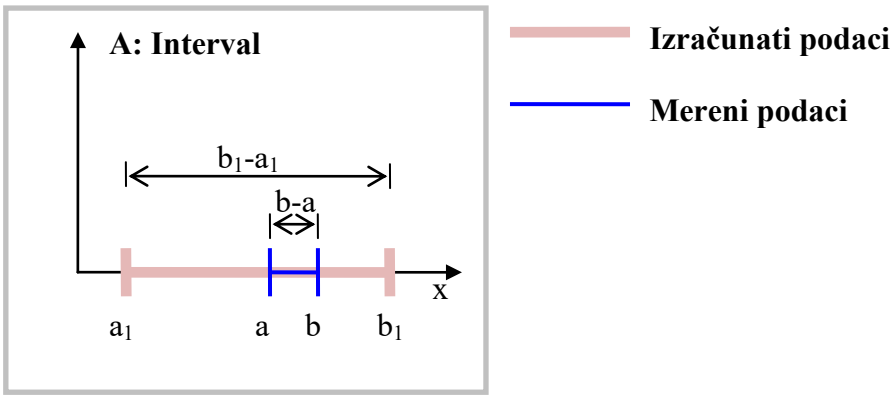


Slika 5.17: Ukupne verovatnoće izmerenih vrednosti

Povećanjem neodređenosti relacije R_3 smanjena je ukupna verovatnoća podataka koji u njoj učestvuju (x_2 i x_3), ali to ne znači da te vrednosti sadrže grešku. Da bi se proverilo da li vrednosti x_2 i x_3 zaista sadrže grešku izračunate su normirane verovatnoće, izrazom 4.7 $p^{norm}\left(x_2 \mid x_2^{M_{R_3,x_2}}, x_3\right)$ i $p^{norm}\left(x_3 \mid x_3^{M_{R_3,x_3}}, x_2\right)$, koje se odnose samo na odstupanje predikcije od izračunate vrednosti, bez uticaja neizvesnosti predikcije. Normirane verovatnoće $p^{norm}\left(x_2 \mid x_2^{M_{R_3,x_2}}, x_3\right)$ i $p^{norm}\left(x_3 \mid x_3^{M_{R_3,x_3}}, x_2\right)$ izračunate su deljenjem sa maksimalnim verovatnoćama za neodređenosti izmerenog podatka i njegove predikcije:

$$p^{norm}\left(x_2 \mid x_2^{M_{R_3,x_2}}, x_3\right) = \frac{p\left(x_2 \mid x_2^{M_{R_3,x_2}}, x_3\right)}{\max\left(p\left(x_2 \mid x_2^{M_{R_3,x_2}}, x_3\right)\right)} \text{ i } p^{norm}\left(x_3 \mid x_3^{M_{R_3,x_3}}, x_2\right) = \frac{p\left(x_3 \mid x_3^{M_{R_3,x_3}}, x_2\right)}{\max\left(p\left(x_3 \mid x_3^{M_{R_3,x_3}}, x_2\right)\right)}.$$

Maksimalne verovatnoće $\max\left(p\left(x_2 \mid x_2^{M_{R_3,x_2}}, x_3\right)\right)$ i $\max\left(p\left(x_3 \mid x_3^{M_{R_3,x_3}}, x_2\right)\right)$ računaju se pomoću izraza 5.1.



Slika 5.18: Uz izraz 4.8

$$\max\left(p\left(x_i \mid x_i^{M_{R_j,x_i}}, X_{x_i}^{M_{R_j,x_i}}\right)\right) = \frac{\overline{x}_i}{\overline{x}_i^{M_{R_j,x_i}}} = \frac{b-a}{b_1-a_1}, \tag{5.1}$$

gde je $x_i = [a,b]$ izmerena vrednost, a $x_i^{M_{R_j,x_i}} = [a_1,b_1]$ predikcija. Dalje se ukupne normirane verovatnoće mogu dobiti pomoću izraza:

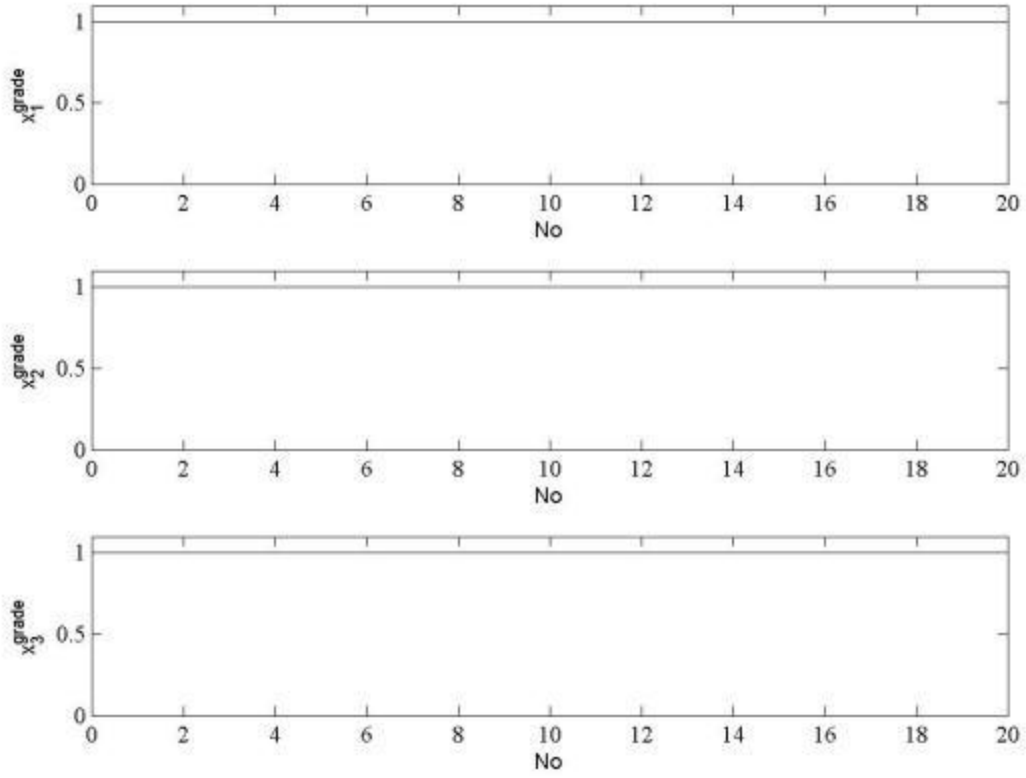
$$x_1^{grade} = w_1^{M_{R_1,x_1}} \times \frac{p\left(x_1 \mid x_1^{M_{R_1,x_1}}, x_2\right)}{\overline{x}_1 / \overline{x}_1^{M_{R_1,x_1}}} + w_1^{M_{R_2,x_1}} \times \frac{p\left(x_1 \mid x_1^{M_{R_2,x_1}}, x_3\right)}{\overline{x}_1 / \overline{x}_1^{M_{R_2,x_1}}};$$

$$x_2^{grade} = w_2^{M_{R_1,x_2}} \times \frac{p\left(x_2 \mid x_2^{M_{R_1,x_2}}, x_1\right)}{\overline{x}_2 / \overline{x}_2^{M_{R_1,x_2}}} + w_2^{M_{R_3,x_2}} \times \frac{p\left(x_2 \mid x_2^{M_{R_3,x_2}}, x_3\right)}{\overline{x}_2 / \overline{x}_2^{M_{R_3,x_2}}};$$

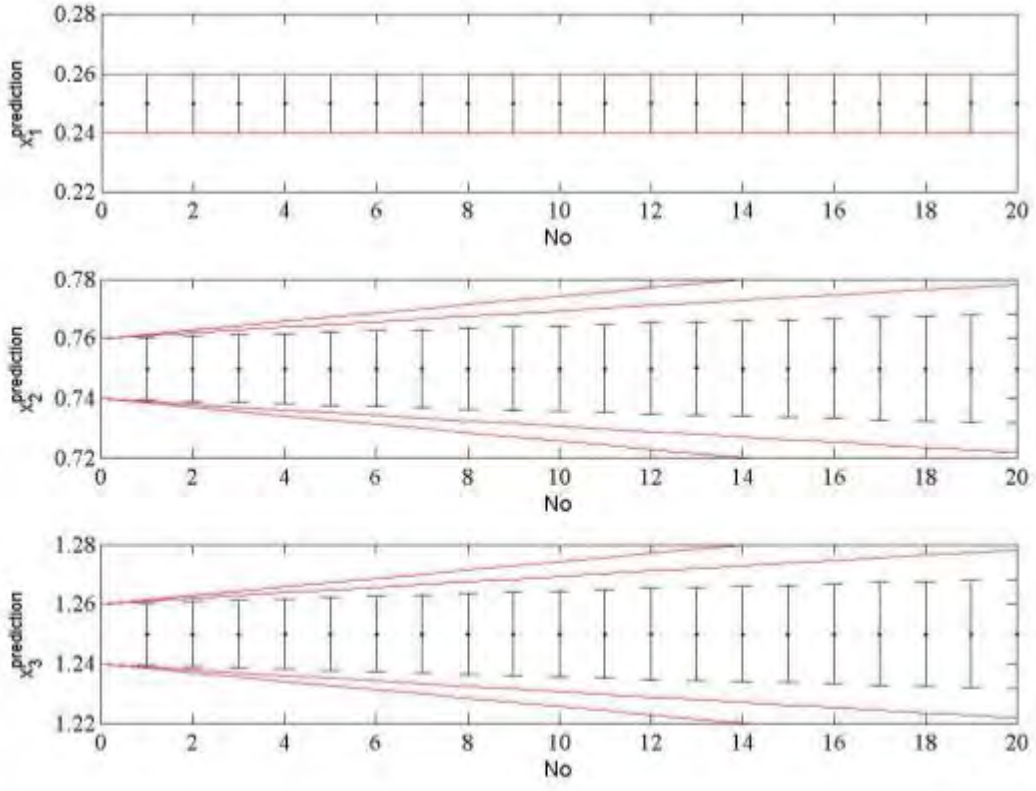
$$x_3^{grade} = w_3^{M_{R_2,x_3}} \times \frac{p\left(x_3 \mid x_3^{M_{R_2,x_3}}, x_1\right)}{\overline{x}_3 / \overline{x}_3^{M_{R_2,x_3}}} + w_3^{M_{R_2,x_3}} \times \frac{p\left(x_3 \mid x_3^{M_{R_2,x_3}}, x_2\right)}{\overline{x}_3 / \overline{x}_3^{M_{R_2,x_3}}},$$

gde $\overline{x}_1$, $\overline{x}_2$ i $\overline{x}_3$ predstavljaju širine intervala izmerenih vrednosti, a $\overline{x}_1^M$, $\overline{x}_2^M$ i $\overline{x}_3^M$ širine intervala predikcija odgovarajućim metodama.

Sa normiranim uslovnim verovatnoćama izmerenih podataka u odnosu na predikcije, verovatnoće podataka ne menjanju se bez obzira što se promenila neizvesnost relacije koja se koristi za vrednovanje. Na slici 5.19 vide se verovatnoće podataka koje su ostale jednake jedinici.



Slika 5.19: Normirane verovatnoće

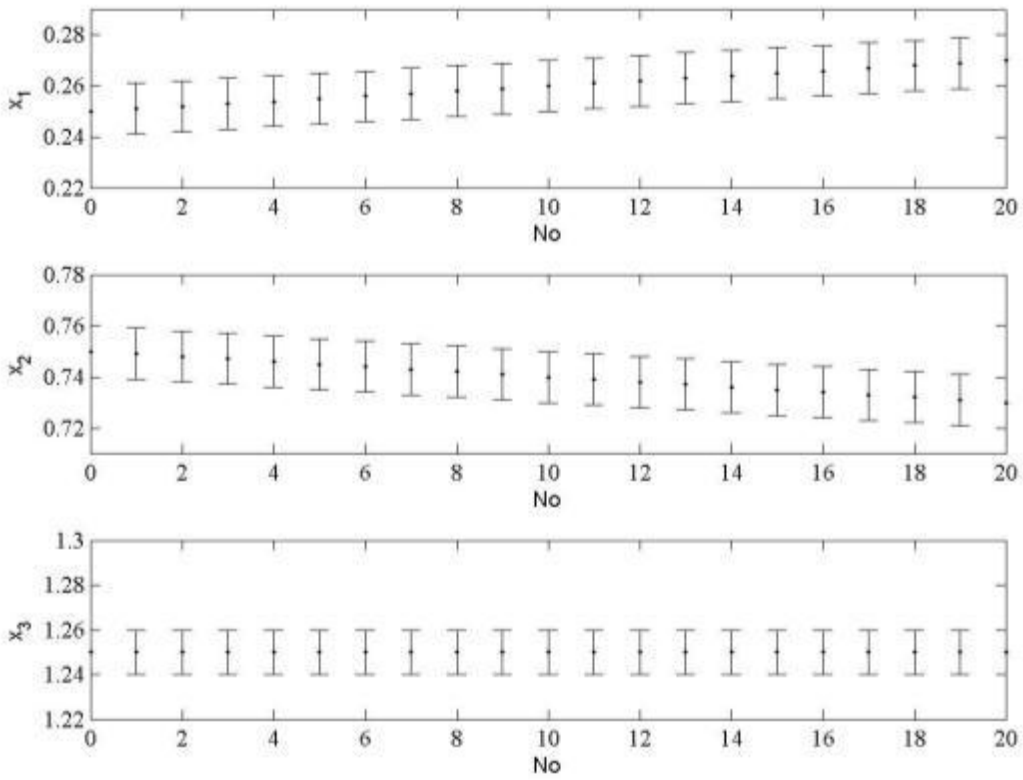


Slika 5.20: Izračunati intervali vrednovanih veličina sa standardnim devijacijama

Izračunati intervali, prikazani na slici 5.20, oslikavaju učešće neodređenosti relacije u predikciji. Što je relacija neodređenija, to su i očekivana vrednost i varijansa veće.

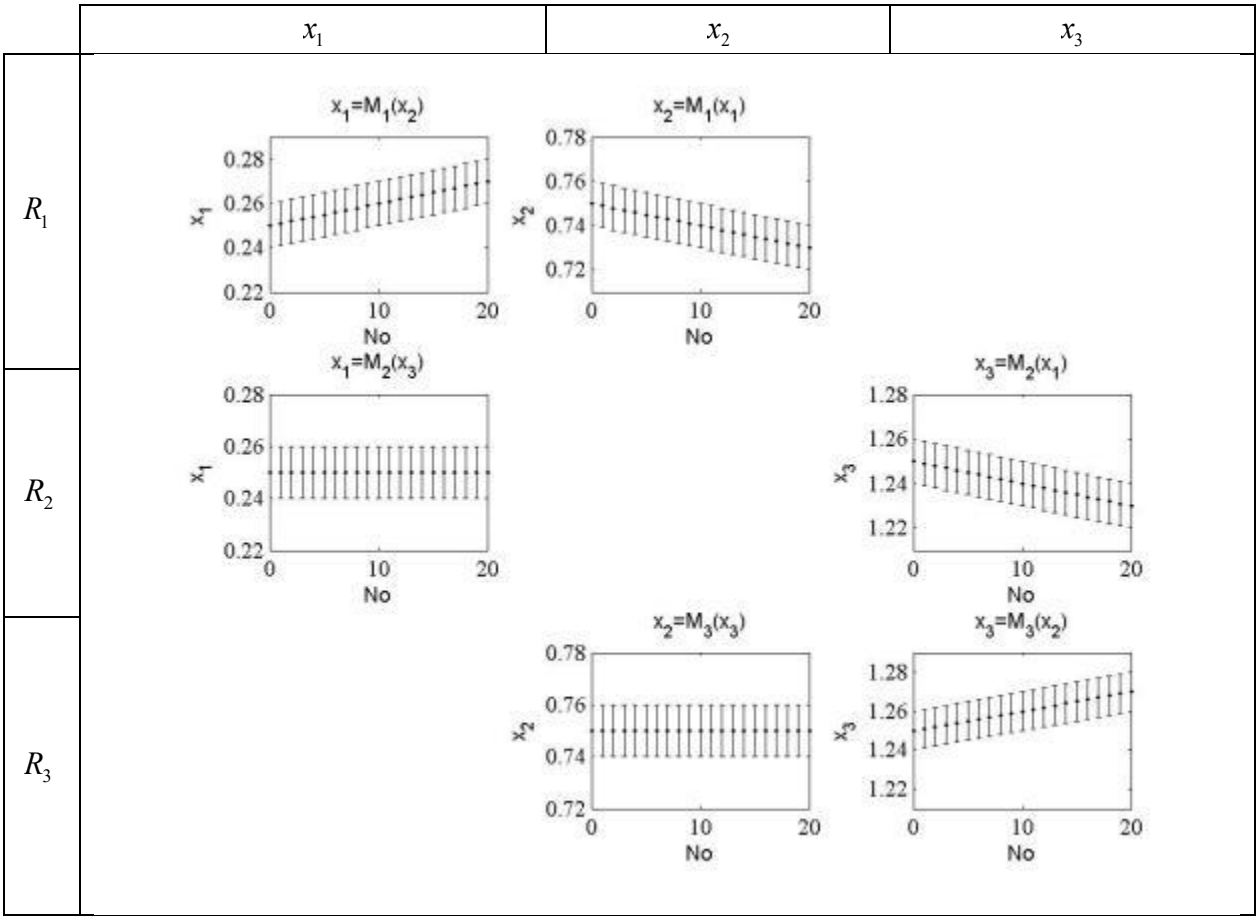
Numerički eksperiment 3

U ovom numeričkom eksperimentu simulirana je moguća greška sistema za vrednovanje podataka. Naime, ukoliko su greške u određenom odnosu, sistem može da pogreši i da grešku pripiše drugoj veličini. Pretpostavlja se greška u veličinama x_1 i x_2, $x_1^{'} = x_1 + \varepsilon$ i $x_2^{'} = x_2 + \varepsilon$, gde je $\varepsilon = [0, 0.001, \ldots, 0.02]$, dok je veličina x_3 tačna (slika 4.21).



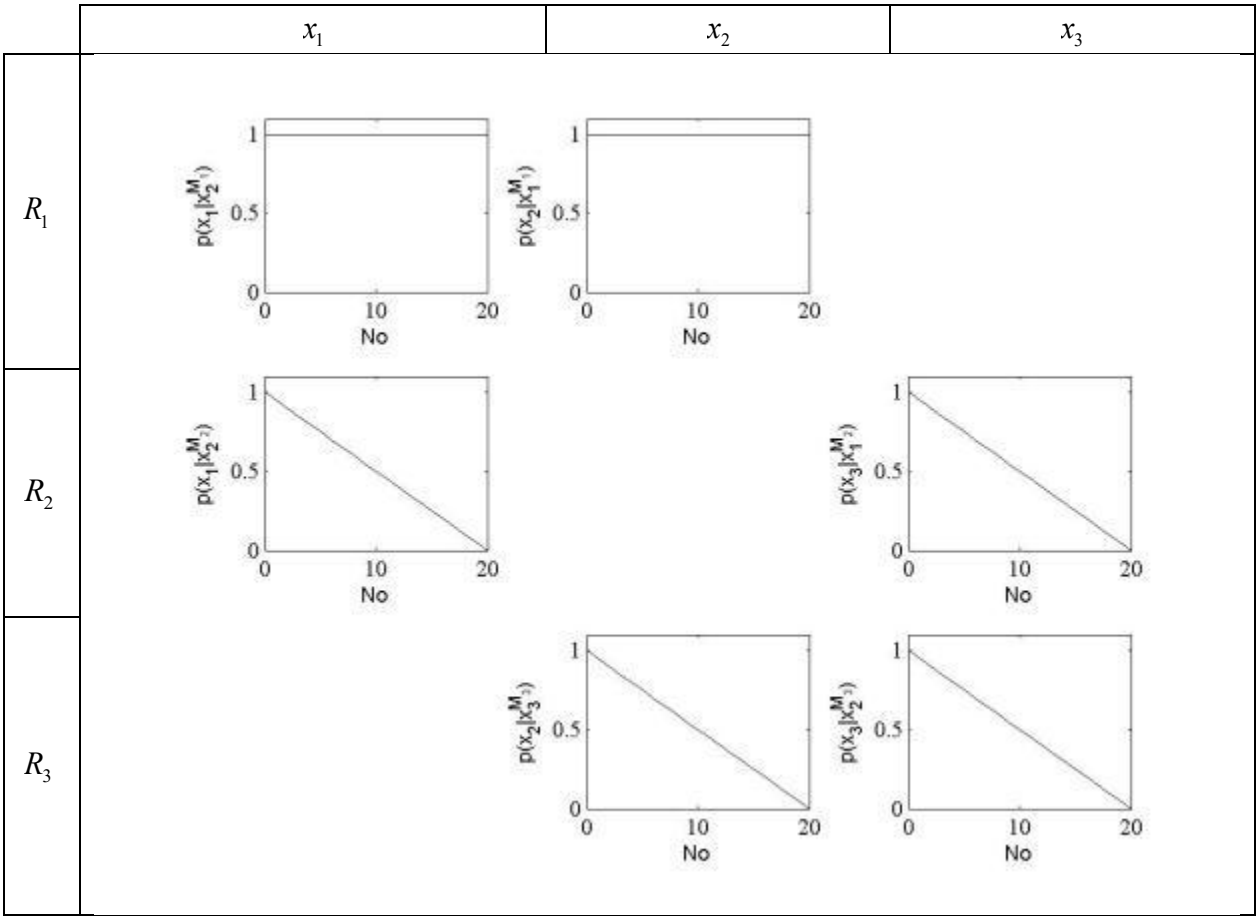
Slika 5.21:Niz vrednosti generisan za numerički primer 3

Rezultati metoda, tj. predikcije vrednovanih veličina, prikazani su na slici 5.22. Vidi se da rezultati metoda M_{R_1,x_1} i M_{R_1,x_2} relacije R_1 (veličine $x_1^{M_{R_1,x_1}}$ i $x_2^{M_{R_1,x_2}}$) odgovaraju vrednostima sa greškom, pa se razlikuju od tačnih vrednosti. To se događajer su unete greške u podacima veličina x_1 i x_2 podešene tako da ukažu na moguće nelogičnosti. Predikcija vrednosti x_1 metodom M_{R_2,x_1} ($x_1^{M_{R_2,x_1}}$) je tačna vrednost, s obzirom na to da je ulazna vrednost ove metode (x_3) tačna. Isto važi i za predikciju vrednosti veličine x_2 metodom M_{R_3,x_2}. Što se tiče izračunatih vrednosti veličine x_3 relacijama M_{R_2,x_3} i M_{R_3,x_3}, one odstupaju od tačnih vrednosti zbog toga što ulazne veličine ovih relacija sadrže grešku.



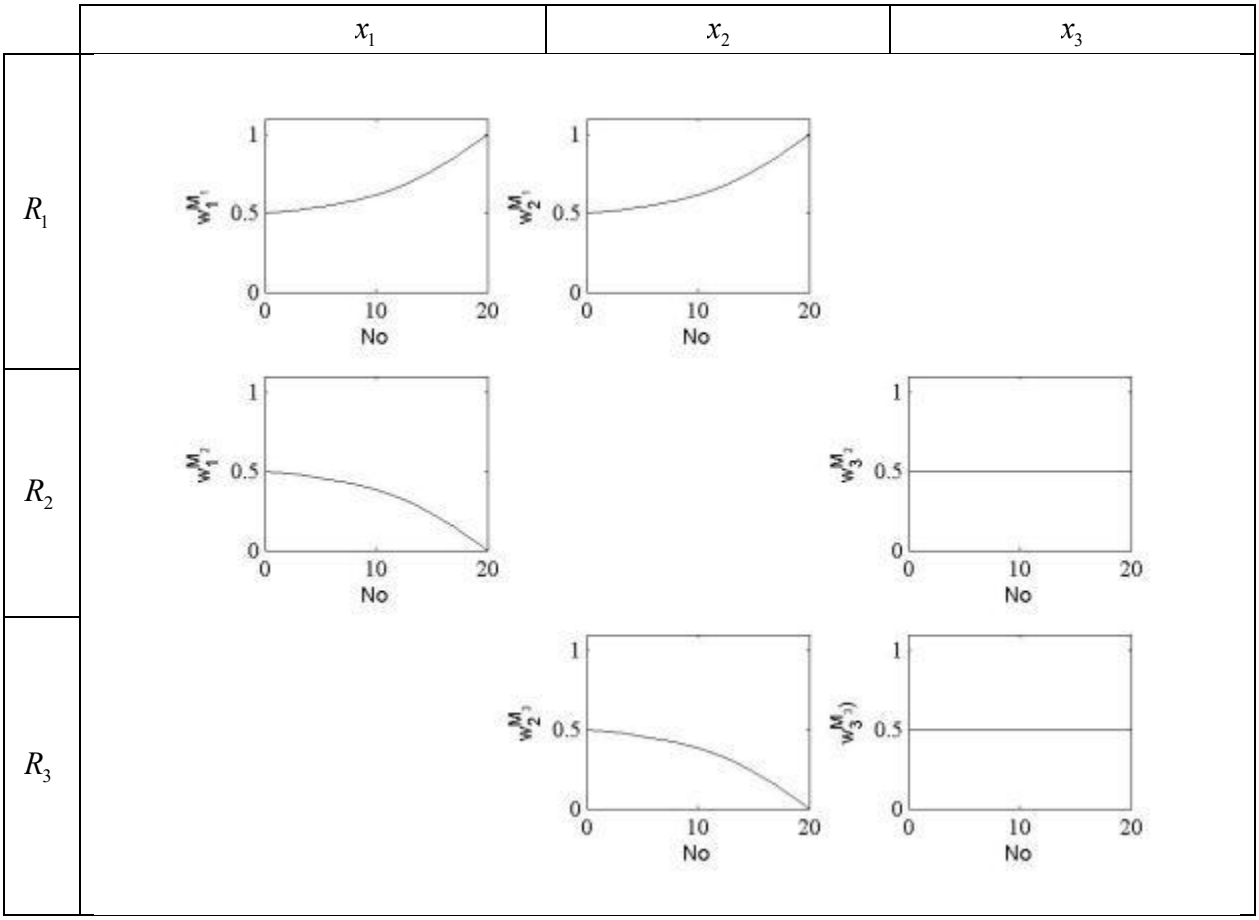
Slika 5.22: Izračunate vrednosti veličina x_1, x_2 i x_3

Usled nelogičnosti na koje je ukazano kod predikcija, verovatnoće slaganja izmerenih vrednosti i predikcija pokazuju dobro slaganje (visoka verovatnoća slaganja) za veličine x_1 i x_2, iako one imaju grešku. Loše slaganje se uočava kod relacija kojima se izračunava veličina x_3 (M_{R_2,x_3} i M_{R_3,x_3}). Sve verovatnoće slaganja prikazane su na slici 5.23.



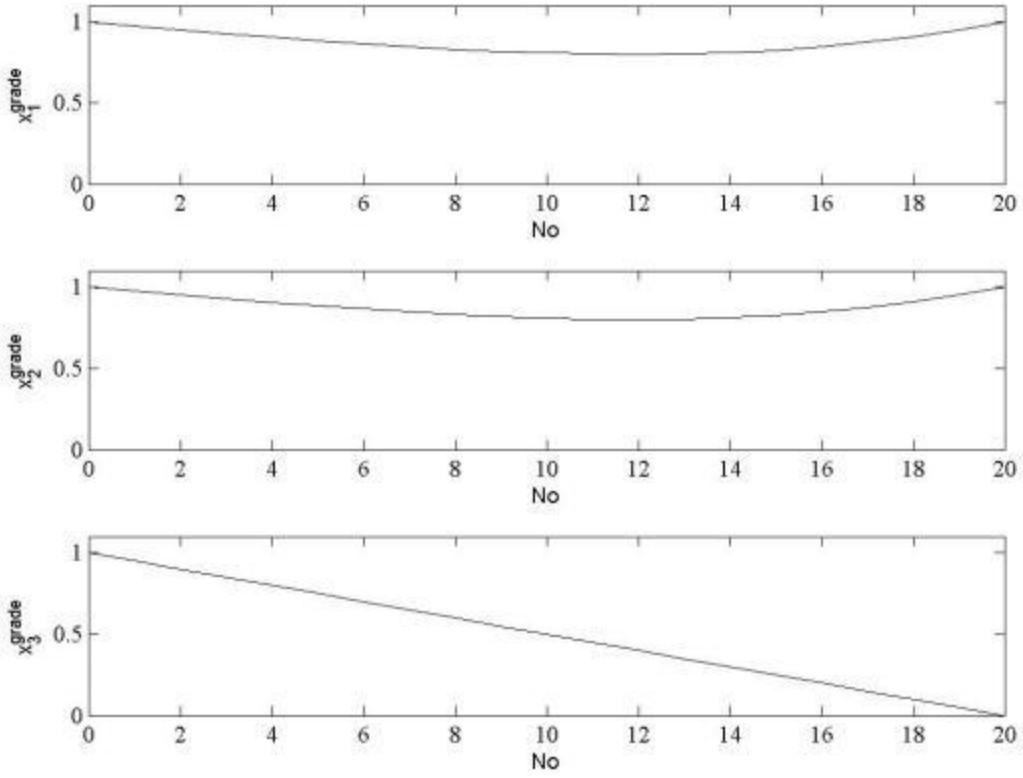
Slika 5.23: Verovatnoće slaganja izračunatih i izmerenih vrednosti

Preko verovatnoća slaganja izračunati su težinski koeficijenti koji maksimizuju verodostojnost zbira verovatnoća, prikazani na slici 5.24. Nelogičnost u izračunatim podacima prenosi se i na težinske koeficijente. Uočava se da sa rastom grešaka rastu težinski koeficijenti uz veličine $x_1^{M_{R_1,x_1}}$ i $x_2^{M_{R_1,x_2}}$, izračunatih metodama M_{R_1,x_1} i M_{R_1,x_2}, gde su ulazne vrednosti vrednosti sa greškama. Isti koeficijenti opadaju kod veličina $x_1^{M_{R_2,x_1}}$ i $x_2^{M_{R_3,x_2}}$, izračunatih metodama M_{R_2,x_1} i M_{R_3,x_2}, iako je ulazna veličina kod ovih relacija veličina x_3 koja ne sadrži grešku. Sa druge strane, težinski koeficijenti uz predikcije veličine x_3 izračunate metodama M_{R_2,x_3} i M_{R_3,x_3} stagniraju. Razlog tome je što su greške u predikcijama veličine x_3 simetrične.

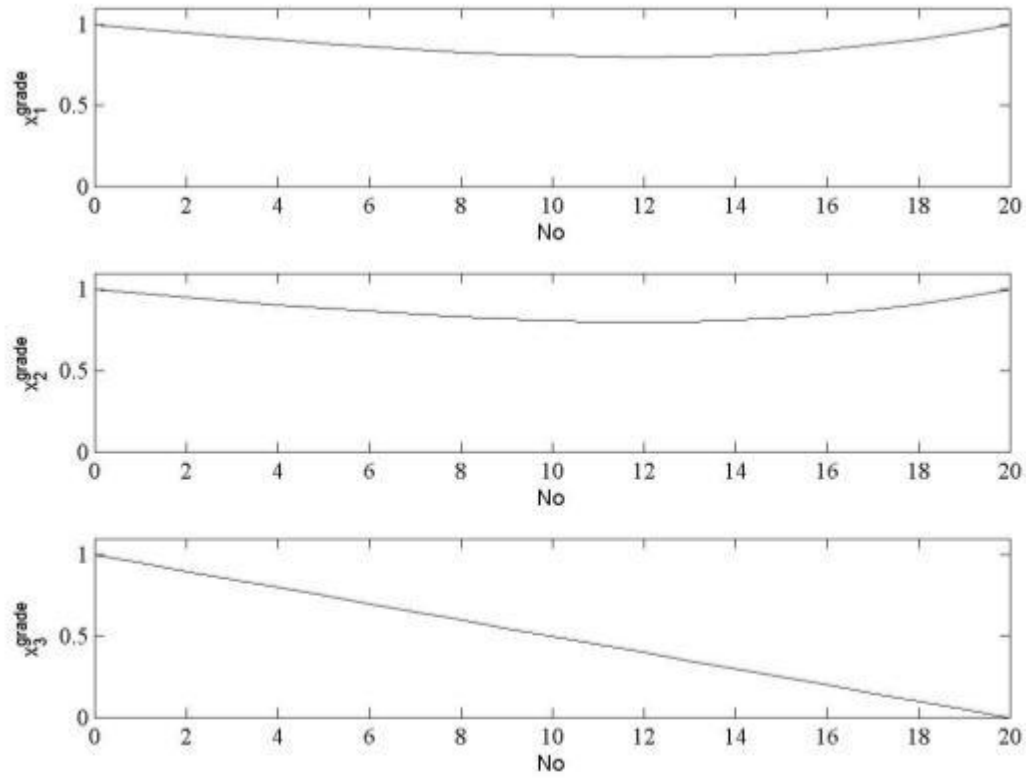


Slika 5.24: Težinski koeficijenti uz verovatnoće slaganja izmerenih i izračunatih vrednosti

Izračunate ukupne verovatnoće (ocene kvaliteta) prikazane na slici 5.25 nastavljaju da oslikavaju nelogične rezultate. Na slici 5.25 prikazane su ocene koje pružaju utisak da je veličina x_3 veličina koja sadrži grešku, za razliku od veličina x_1 i x_2 koje imaju visoke ocene.



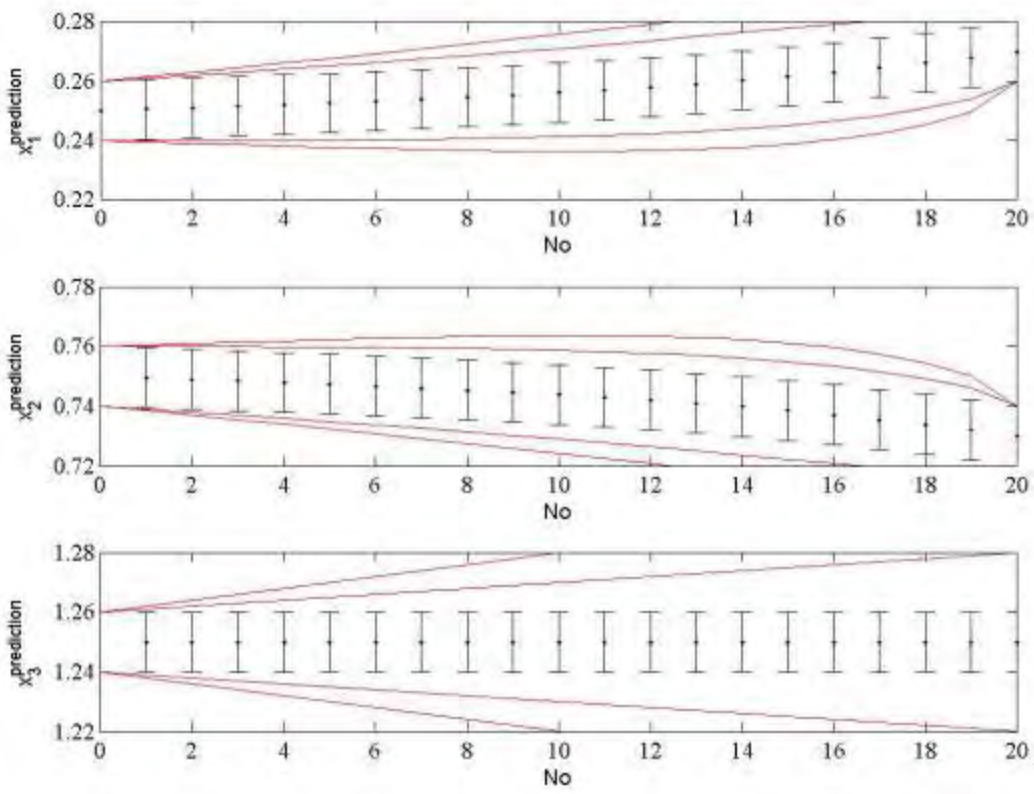
Slika 5.25: Ukupne verovatnoće izmerenih vrednovanih veličina (ocene)



Slika 5.26: Normirane ukupne verovatnoće

Nakon normiranja uslovnih verovatnoća (izraz 5.1) kojim se otklanja uticaj neodreĎenosti relacije, dobija se slična situacija, kao što se vidi na slici 5.26.

Kada se pogledaju izračunati intervali, situacija je takoĎe nelogična (slika 5.26). Izračunati intervali veličina x_1 i x_2 izgledaju kao dobro interpretirani, sa varijansama koje se povećavaju, a zatim i smanjuju. Što se tiče veličine x_3, očekivana vrednost koja odgovara tačnoj je proizvod simetričnih grešaka u izračunatim vrednostima metodama M_{R_2,x_3} i M_{R_3,x_3}, a povećavanje varijanse upravo ukazuje na postojanje grešaka u rezultatima ovih relacija.



Slika 5.26: Predikcija merenih vrednosti

5.1.2 Zaključak hipotetičkog primera

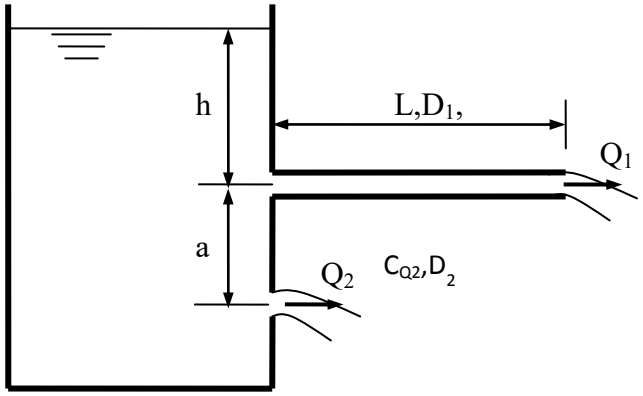
Prikazanim hipotetičkim primerom demonstriran je korak 3 algoritma za vrednovanje podataka, prikazanog na slici 4.1. Primećuju se sledeće karakteristike rezultata tog algoritma:

1. kažnjava se loše slaganje izmerenih i izračunatih vrednosti usled postojanja grešaka u podacima, smanjenjem težinskih koeficijenata uz rezultate relacija u kojima učestvuju te vrednosti;
2. kažnjava se povećana neodreĎenost relacija izmeĎu podatka, smanjenjem težinskih koeficijenata uz rezultate tih relacija;
3. iako je poslednjim numeričkim primerom demonstrirana situacija kada je algoritam dao pogrešne rezultate, potrebno je naglasiti da su greške u podacima tendenciozno podešene da bi se dobio taj efekat. U stvarnim situacijama, gde greške nisu namerno unete u podatke, to bi bio izuzetno redak slučaj. Pretpostavlja se da se mogućnot slične situacije smanjuje sa povećanjem broja veličina koje su uključene u vrednovanje.

5.2 Hipotetički hidraulički primer

Hipotetički primer iz oblasti hidrotehnike ima za cilj da demonstrira predloženu metodologiju na realnom sistemu, ali pri kontrolisanim uslovima. U ovom primeru demonstriraju se koraci 2 i 3 predloženog algoritma za vrednovanje prikazanog na slici 4.1, tj. formiranje metoda iz relacija i izračunavanja verovatnoća merenih podataka. Formiran sistem za vrednovanje testiran je na nekoliko čestih grešaka koje se mogu očekivati u procesu merenja: pojavu pikova, klizanje nule merača, naglo odstupanje i povećan šum.

Sistem na kome se mere hidrotehničke veličine sastoji se od rezervoara konačne visine bez dotoka iz kog ističe voda kroz cev određene dužine i kroz oštroivični otvor (slika 5.27) protocima Q_1 i Q_2.



Slika 5.27: Rezervoar iz kog ističe voda

Prečnici otvora i cevi iznose $D_1 = D_2 = 100$ mm, dužina cevi $L = 10\,\text{m}$, razmak između otvora i cevi iznosi $a = 1\,\text{m}$, a površina rezervoara $A = 10\,\text{m}^2$. Osmatranje sistema sprovodi se merenjem visine nadsloja vode h i protoka Q_1 i Q_2 mernim metodama sa neodređenostima $h^{unc} = 0.01$ m i $Q^{unc} = 0.0001\,\text{m}^3/\text{s}$. Između merenih veličina mogu se uspostaviti relacije prikazane u tabeli 5.5.

Tabela 5.5: Relacije koje se mogu uspostaviti između merenih veličina h, Q_1 i Q_2

Oznaka relacije	Matematički oblik relacije
R_1	$h = \frac{Q_1^2}{2gA_1^2}\left(\lambda\frac{L}{D_1}+1\right)$
R_1	$Q_2 = C_{Q_2}A_2\sqrt{2g\left(h+a\right)}$
R_1	$\left(\frac{Q_2}{C_{Q_2}A_2}\right)^2\frac{1}{2g}-a=\left(\lambda\frac{L}{D_1}+1\right)\frac{Q_1^2}{2gA_1^2}$

Koeficijent linijskog otpora λ računa se po formuli

$$\lambda = \begin{cases} \dfrac{64}{\text{Re}} & ,\text{Re} < 3000 \\ 0.115\left(\dfrac{k}{D_1}+\dfrac{60}{\text{Re}}\right)^{0.25} & ,\text{Re} \geq 3000 \end{cases}, \quad \text{Re} = \frac{\rho D_1 V_1}{\mu}.$$

Razmatra se deo operativnog rada rezervoara u procesu pražnjenja. Pražnjenje rezervoara može se opisati diferencijalnom jednačinom

$$A\frac{dh}{dt} = -Q_1 - Q_2$$

ili u numeričkoj formi

$$h^{t+1} = h^t - \frac{\Delta t}{A}\left(Q_1 + Q_2\right).$$

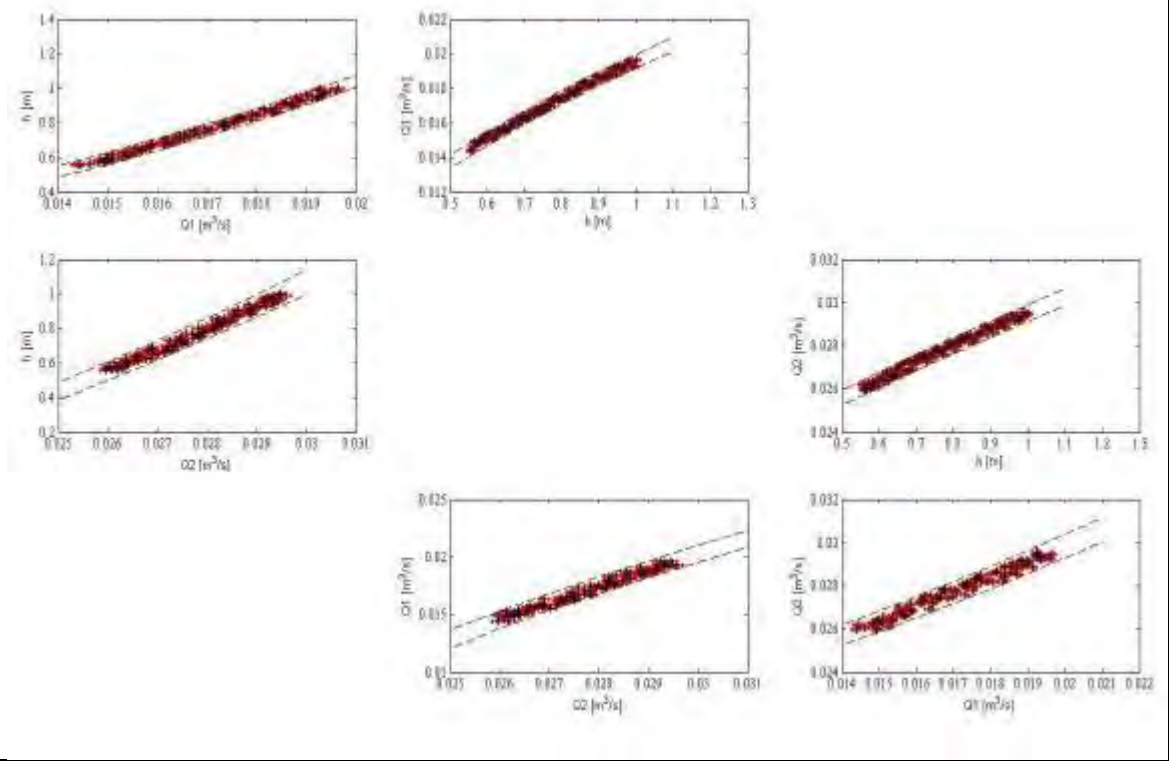
Za početnu vrednost $h^0 = 1$ m, $\Delta t = 1$ s i šum po uniformnoj raspodeli koji iznosi $h^{noise} = U\left(0, 0.0144\right)$ i $Q^{noise} = U\left(0, 1.443\times10^{-4}\right)$, generisani su podaci za kalibraciju metoda za predikciju vrednovanih podataka.

Na osnovu uspostavljenih relacija razvijene su metode koje se koriste za predikciju vrednovanih podataka. Metode za predikciju prikazane su u tabeli 5.6.

Tabela 5.6: Metode za predikciju merenih podataka

	h	Q_1	Q_2
R_1	$h = \frac{Q_1^2}{2gA_1^2}\left(\lambda\frac{L}{D_1}+1\right)$	$Q_1 = \sqrt{\frac{2ghA_1^2}{\left(\lambda L/D_1+1\right)}}$	
R_2	$h = \frac{1}{2g}\left(\frac{Q_2}{C_{Q_2}A_2}\right)^2 - a$		$Q_2 = C_{Q_2}A_2\sqrt{2g\left(h+a\right)}$
R_3		$Q_1 = \sqrt{\left(\frac{2gA_1^2}{\lambda L/D_1+1}\right)\left(\left(\frac{Q_2}{C_{Q_2}A_2}\right)^2\frac{1}{2g} - a\right)}$	$Q_2 = C_{Q_2}A_2\sqrt{2g\left(\left(\lambda\frac{L}{D_1}+1\right)\frac{Q_1^2}{2gA_1^2}+a\right)}$

Metode su kalibrisane, rezultati kalibracije prikazani su na slici 5.28, a kalibracioni parametri i metode u tabelama 5.7 i 5.8.

	h	Q_1	Q_2
R_1			
R_2			
R_3			

Slika 5.28: Kalibracija metoda za predikciju vrednovanih veličina

Tabela 5.7: Kalibracioni parametri metoda za predikciju

	h	Q_1	Q_2
R_1	h_unc=[-0.035,0.035]; (dodata neizvesnost)	Q_1_unc=[-0.0004,0.0004]; (dodata neizvesnost)	
R_2	C_Q=[0.59,0.61]; (parametar C_Q)		C_Q =[0.592,0.608]; (parametar C_Q)
R_3		C_Q =[0.588,0.61]; (parametar C_Q)	C_Q =[0.589,0.611]; (parametar C_Q)

Tabela 5.8:Kalibrisane metode za predikciju merenih veličina

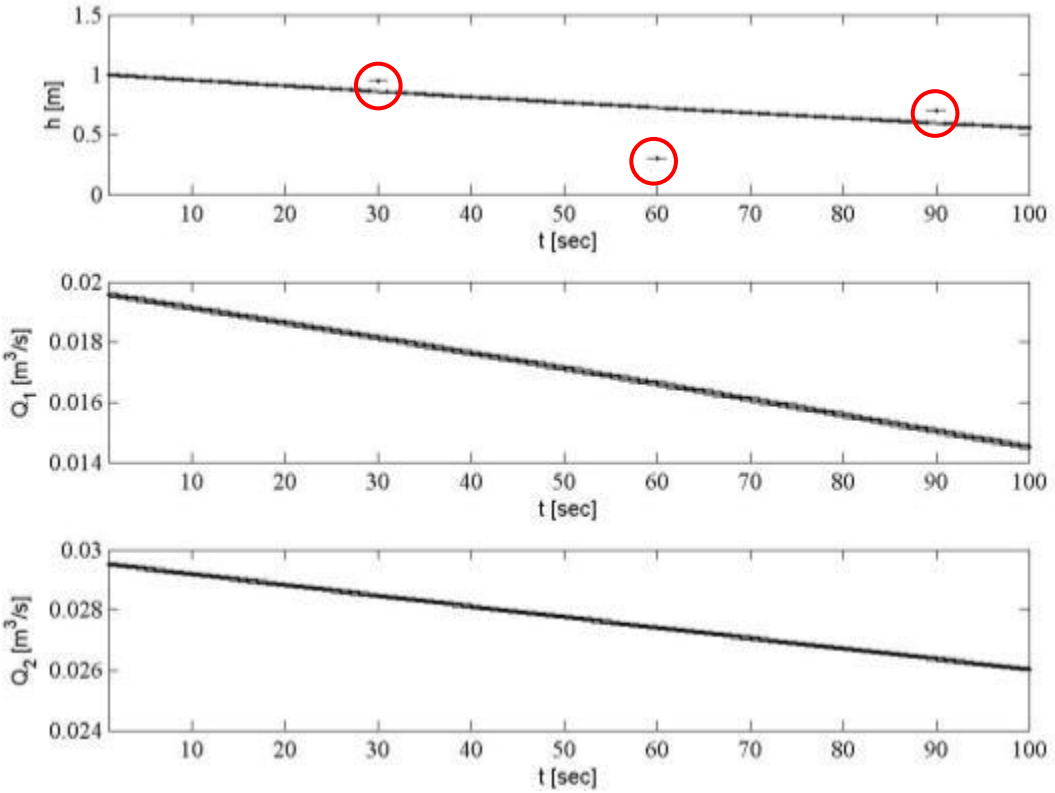
	h	Q_1	Q_2
R_1	$h=\frac{Q_1^2}{2gA_1^2}\left(\lambda\frac{L}{D_1}+1\right)+[\text{-}0.035,0.035]$	$Q_1=\sqrt{\frac{2ghA_1^2}{(\lambda L/D_1+1)}}+[\text{-}0.0004,0.0004]$	
R_2	$h=\frac{1}{2g}\left(\frac{Q_2}{[0.59,0.61]\times A_2}\right)^2-a$		$Q_2=[0.592,0.608]\times A_2\sqrt{2g(h+a)}$
R_3		$Q_1=\sqrt{\left(\frac{2gA_1^2}{\lambda L/D_1+1}\right)\left(\left(\frac{Q_2}{[0.588,0.61]\times A_2}\right)^2\frac{1}{2g}-a\right)}$	$Q_2=[0.589,0.611]\times A_2\sqrt{2g\left(\left(\lambda\frac{L}{D_1}+1\right)\frac{Q_1^2}{2gA_1^2}+a\right)}$

Nakon kalibracije metoda za predikciju sprovedeni su numerički eksperimenti. Razmatra se samo situacija kada se rezervoar prazni (za taj deo operativnog rada su i kalibrisane metode).

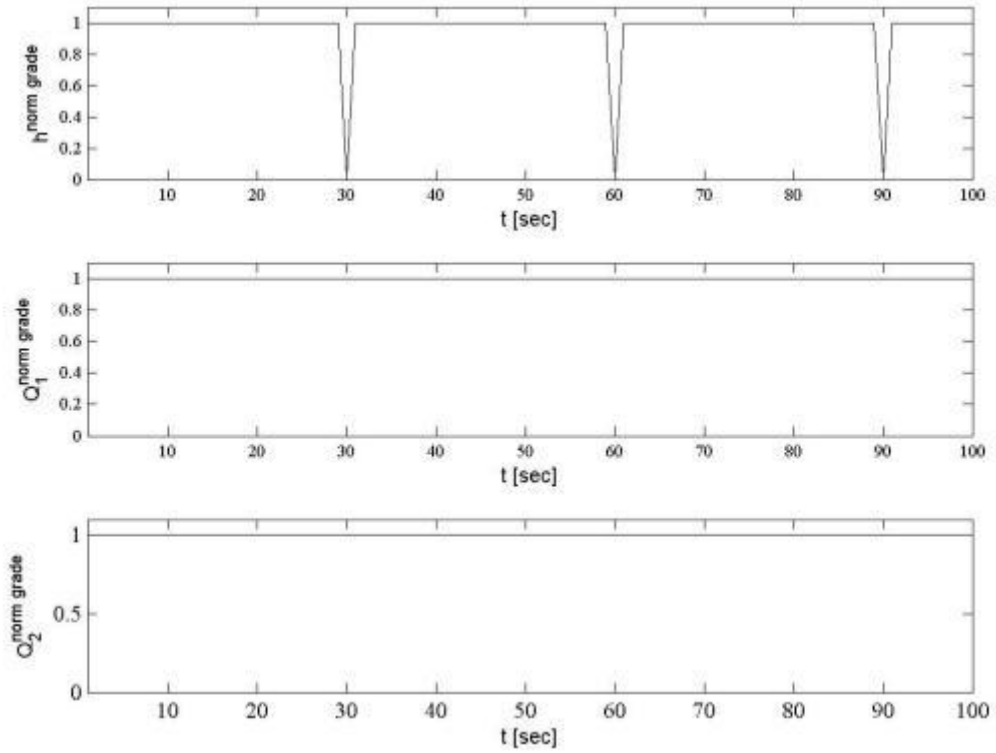
5.2.1 Hidrotehnički hipotetički primer – rezultati i diskusija

Pojava pikova

Često se pri merenju hidrotehničkih veličina u rezultatima merenja mogu uočiti pikovi koji predstavljaju greške. Podacima za testiranje sistema za vrednovanje dodati su pikovi u vremenskim trenucima $t_1 = 30\,\text{s}$, $t_2 = 60\,\text{s}$ i $t_3 = 90\,\text{s}$. Na slici 5.29 prikazani su generisani mereni podaci, dok su na slici 5.30 prikazane normirane verovatnoće regularnosti podataka. Može se primetiti da su za protoke Q_1 i Q_2 normirane verovatnoće jednake jedinici, što ukazuje na regularne podatke, dok se normirane verovatnoće nivoa h na mestima pikova spuštaju na nulu.



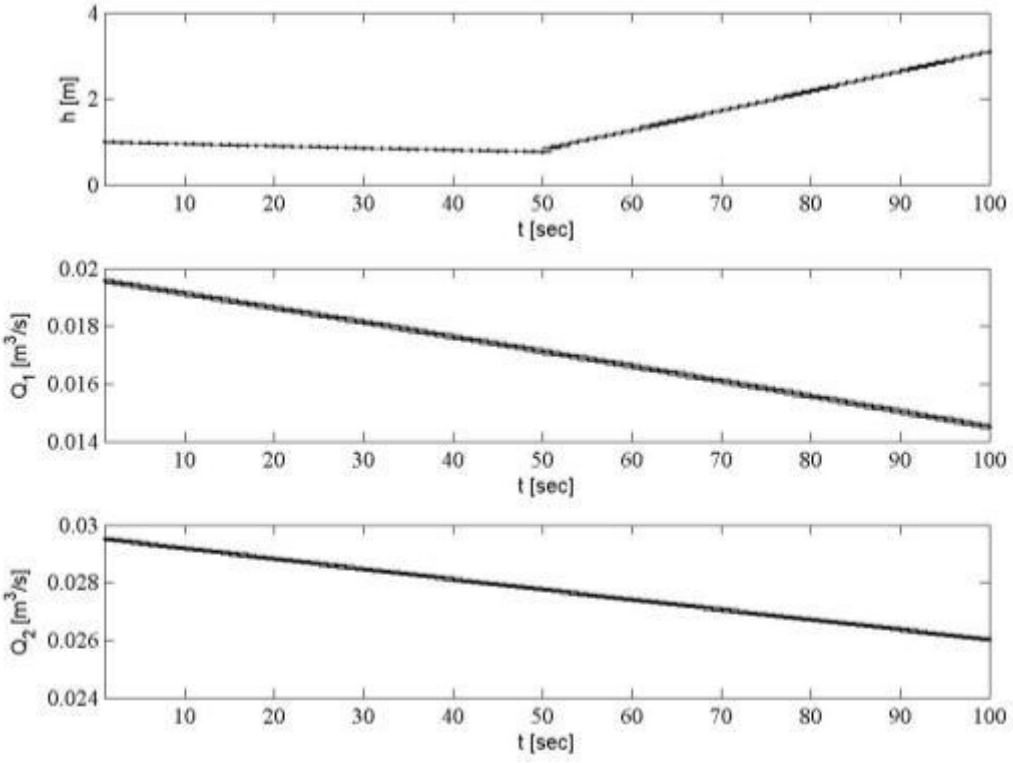
Slika 5.29: Vremenske serije za testiranje sa dodatim pikovima



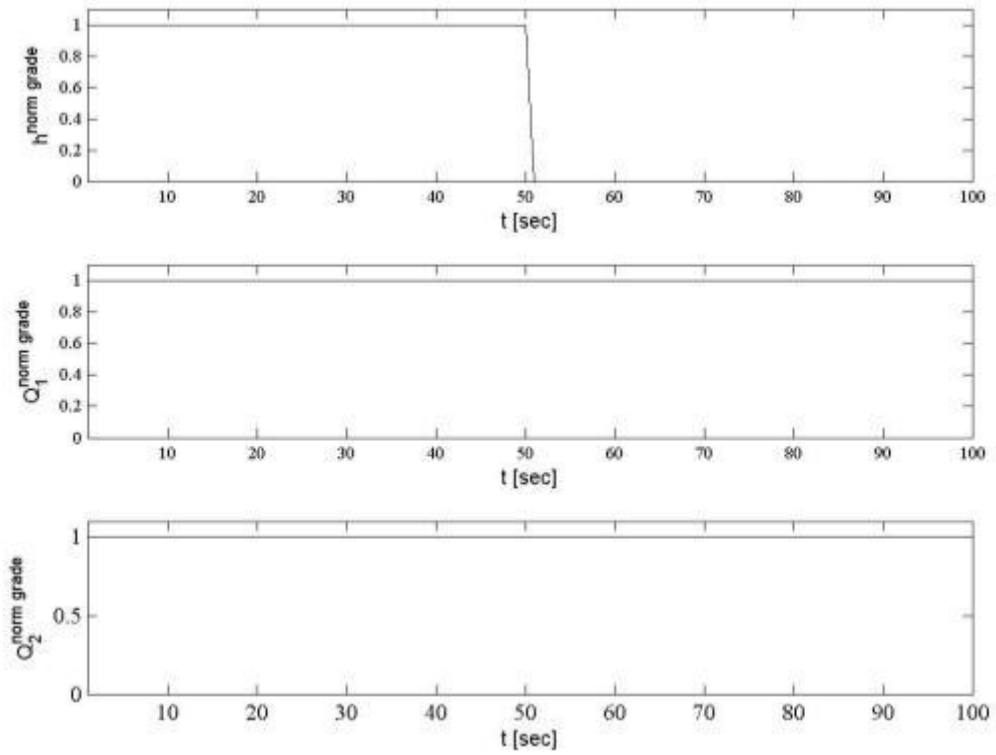
Slika 5.30: Normirane verovatnoće regularnosti podataka

Klizanje nule merača

Klizanje nule merača nivoa takoĎe je čest uzrok grešaka. Na slici 5.31 prikazane su ulazne vremenske serije sa simuliranim klizanjem nule merača nivoa sa početkom u $t_1 = 50\,\text{s}$. Na slici 5.32 normirane su verovatnoće regularnosti podataka. Primećuje se nagli pad verovatnoća regularnosti na nulu kod podataka o nivou, dok podaci o protocima ostaju regularni.



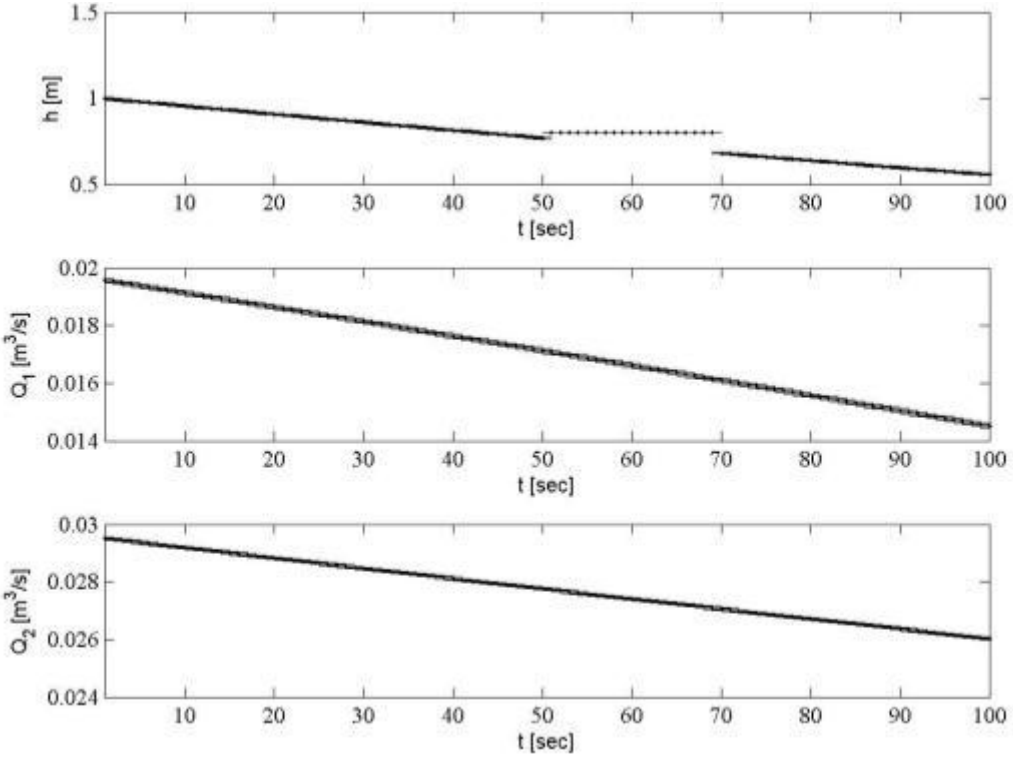
Slika 5.31: Vremenske serije za testiranje sa dodatim efektom klizanja nule



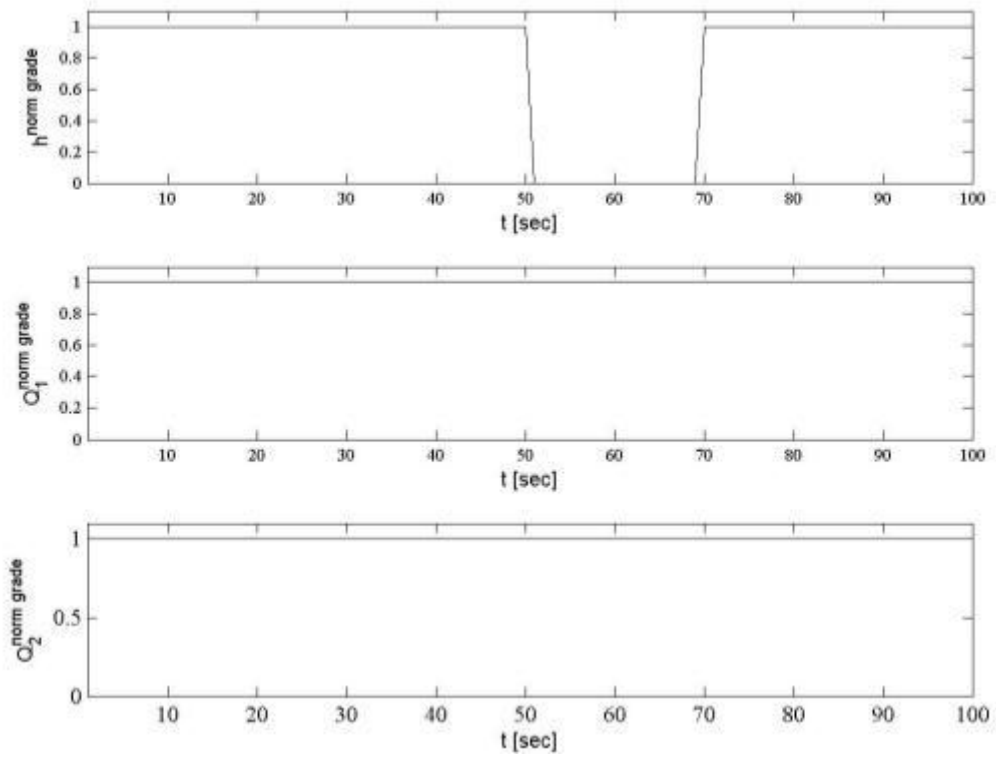
Slika 5.32: Normirane verovatnoće regularnosti podataka

Naglo odstupanje

Sledeća česta pojava kod merenja nivoa jeste nagla promena merenih vrednosti i zadržavanje odreĎene konstantne vrednosti tokom dužeg vremenskog perioda, nakon čega se vraća u normalan rad. Na slici 5.33 prikazana je simulacija naglog konstantnog odstupanja u vremenskom intervalu od $t_1 = 50\,\text{s}$ do $t_2 = 70\,\text{s}$.



Slika 5.33: Vremenske serije za testiranje sa dodatim efektom naglog odstupanja

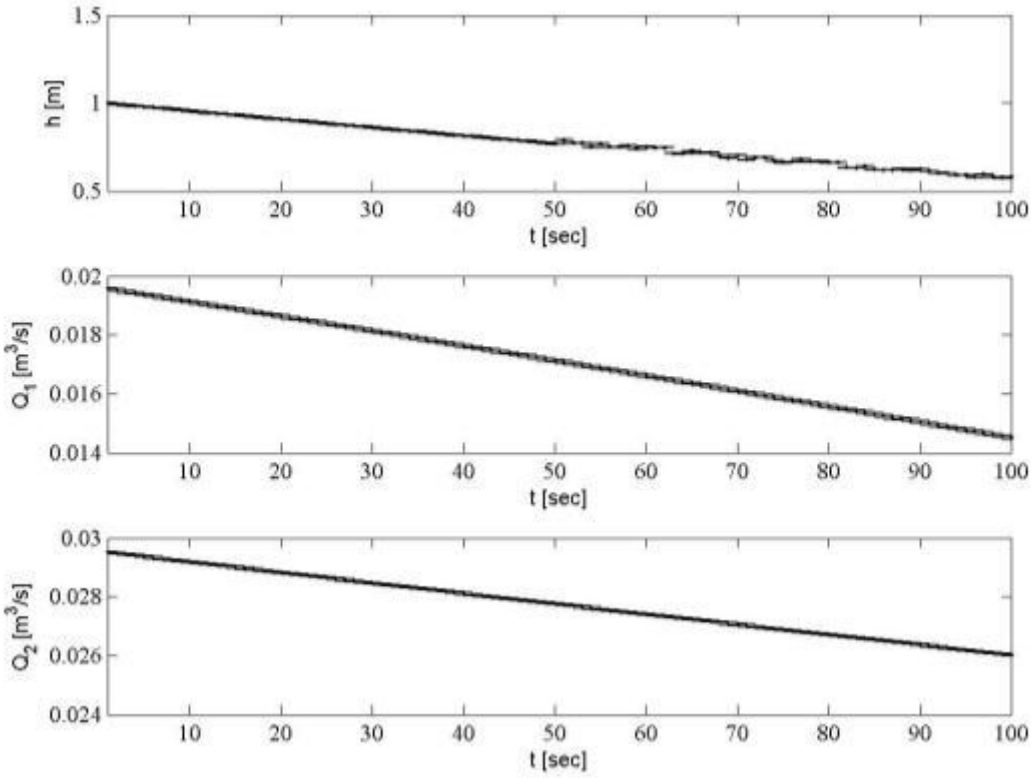


Slika 5.34: Normirane verovatnoće regularnosti podataka

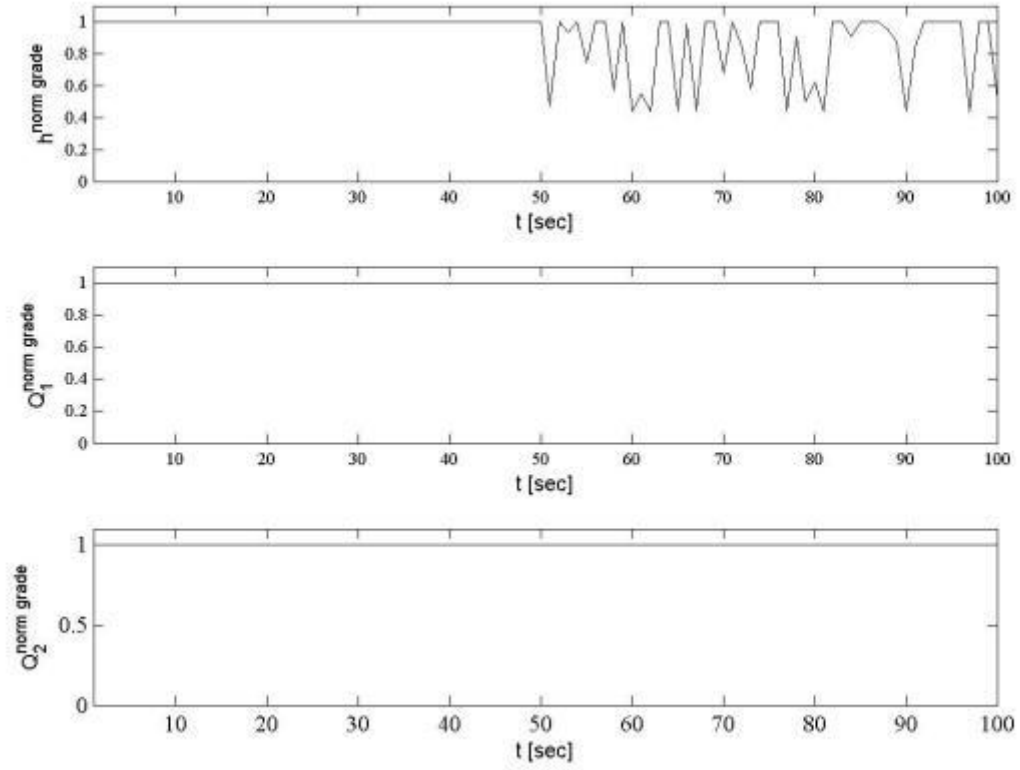
Na slici 5.34 prikazane su normirane verovatnoće regularnosti podataka i, kao i u prethodnim slučajevima, može se uočiti njihov pad za podatke o nivou vode na nulu, dok su verovatnoće koje se odnose na protoke ostale jednake jedinici.

Povećan šum

Povećan šum takoĎe predstavlja čestu pojavu koja se oslikava u merenim podacima nivoa vode rezervoara, bilo da su u pitanju efekti mernog okruženja (npr. talasi zbog vetra), bilo da je u pitanju šum koji potiče iz nekog drugog izvora. Na slici 5.35 prikazane su vremenske serije za testiranje sa dodatim efektom povećanog šuma sa početkom u vremenskom trenutku $t_1 = 50\,\text{s}$. Na slici 5.36 vidi se specifičan potpis unetog efekta na normiranim verovatnoćama regularnosti podataka nivoa vode, dok su verovatnoće koje odgovaraju protocima ostale visoke i jednake jedinici.



Slika 5.35: Vremenske serije za testiranje sa dodatim efektom povećanog šuma



Slika 5.36: Normirane verovatnoće regularnosti podataka

5.2.2 Zaključak hidrotehničkog hipotetičkog primera

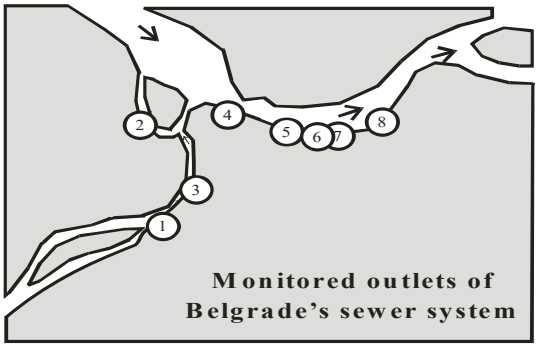
Ovim hidrotehničkim hipotetičkim primerom demonstrirani su korak 2 i korak 3 algoritma za vrednovanje podataka i pokazani su rezultati algoritma za neke specifične greške koje se mogu očekivati u procesu merenja. Može se zaključiti da algoritam daje smislene rezultate i da se iz izračunatih normiranih verovatnoća regularnosti podataka mogu povlačenjem granice odvojiti regularni od neregularnih podataka.

5.3 Realan primer merenja u kanalizacionom sistemu

U ovom primeru razmatraju se mereni podaci o parametrima količina i kvaliteta vode u kanalizacionom sistemu koji su prikupljeni na realnom sistemu u terenskim uslovima. Demonstriraju sesva četiri koraka algoritma prikazanog na slici 4.1: priprema podataka, predikcija podataka, izračunavanje verovatnoća grešaka u podacima i interpretacija rezultata i donošenje odluke.

5.3.1 Opis sistema

Sistem za osmatranje Beogradskog kanalizacionog sistema oformljen je 2006. godine sa ciljem da se pokrije 80% fekalnih i atmosferskih voda koje gravitiraju ka recipijentima, rekama Savi i Dunavu.

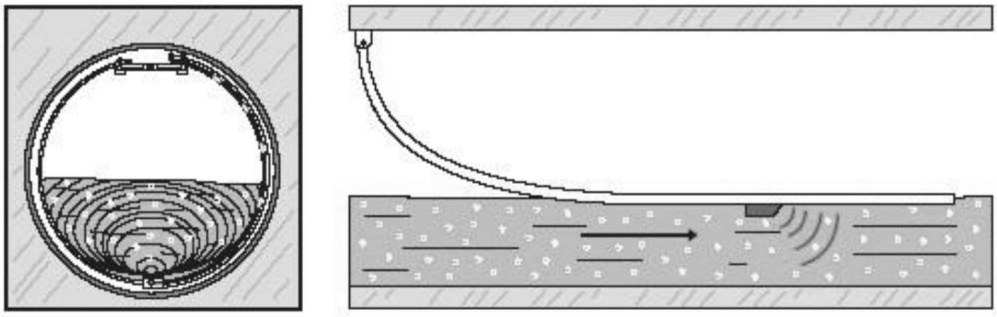


Slika 5.37: Osmatrani ispusti Beogradskog kanalizacionog sistema

Osam ispusta (slika 5.37) opremljeno je sistemima za merenje parametara količina i kvaliteta vode koja se ispušta u recipijente, reke Savu i Dunav. Parametri koji se prate su:

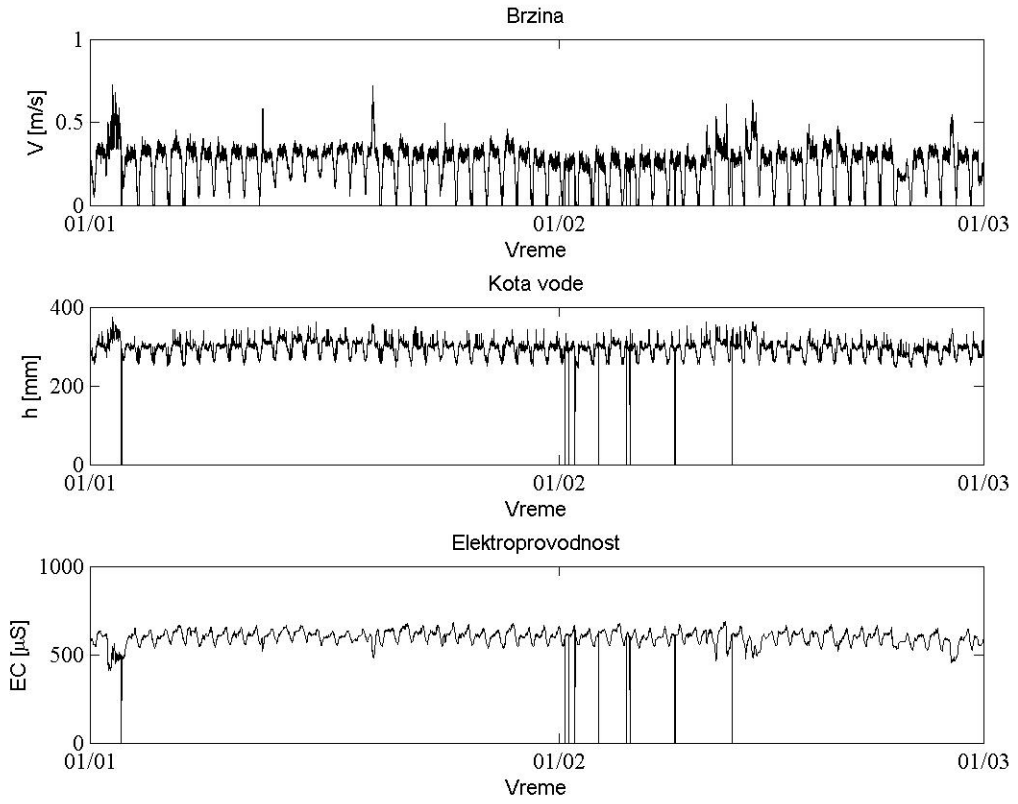
- parametri količine vode:
  - dubina vode (h);
  - brzina vode (V);
- parametri kvaliteta vode:
  - temperatura (T);
  - pH (pH);
  - elektroprovodnost (EC);
  - REDOX potencijal ($REDOX$).

Iz podataka o dubini i brzini vode (V i h) i poznate geometrije poprečnog preseka kolektora na lokaciji mernog mesta ($A(h)$) izračunava se protok Q. Podatke o dubini i brzini prati podatak o kvalitetu ultrazvučnog signala ($U(h,V)$), koji je proporcionalan snazi odbijenog ultrazvučnog signala i predstavlja stepen poverenja u merni uređaj za pojedini podatak. Pouzdanost izmerenog podatka uvećana je sistemom za registrovanje pada napona u mreži, pada napona u baterijama za podršku i sistemom za registrovanje pristupanja sistemu za merenje u cilju održavanja i provere. Način ugradnje ultrazvučnih merila na poziciji mernog mesta prikazan je na slici 5.38.



Slika 5.38: Pozicije ultrazvučnih merila

Svako merno mesto opremljeno je akvizicionom stanicom za prikupljanje podataka koji se dalje GPRS sistemom šalju u sabirni centar – bazu podataka SCADA sistema lociranu u prostorijama Beogradske kanalizacije (BK). Interval merenja je $\Delta t = 5\,\text{min}$. Moguće jeda neki podatak bude izostavljen ili da neki podatak bude dupliran u sabirnoj bazi podataka. Vremenske skale su usklađene i formiraju se uz pomoć odvojenih tajmera. Podaci se prikupljaju u relativno istim vremenskim trenucima, i u bazu podataka se beleže zajedno sa podacima o vremenu.

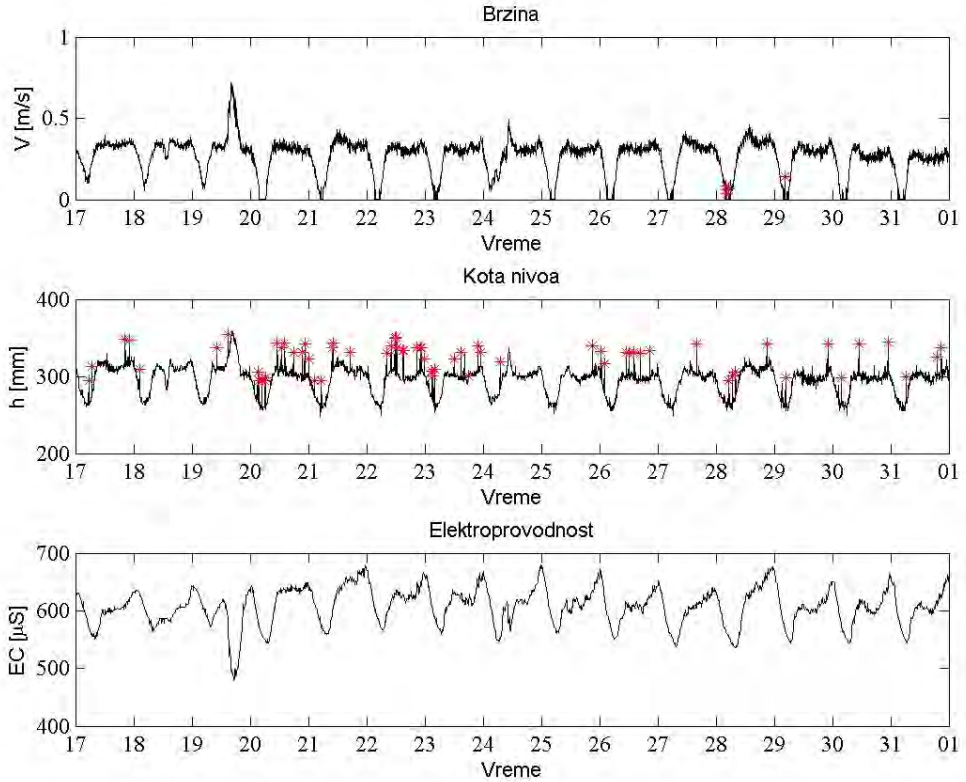


Slika 5.39: Vremenske serije brzine, dubine i elektroprovodnosti na mernom mestu Višnjica

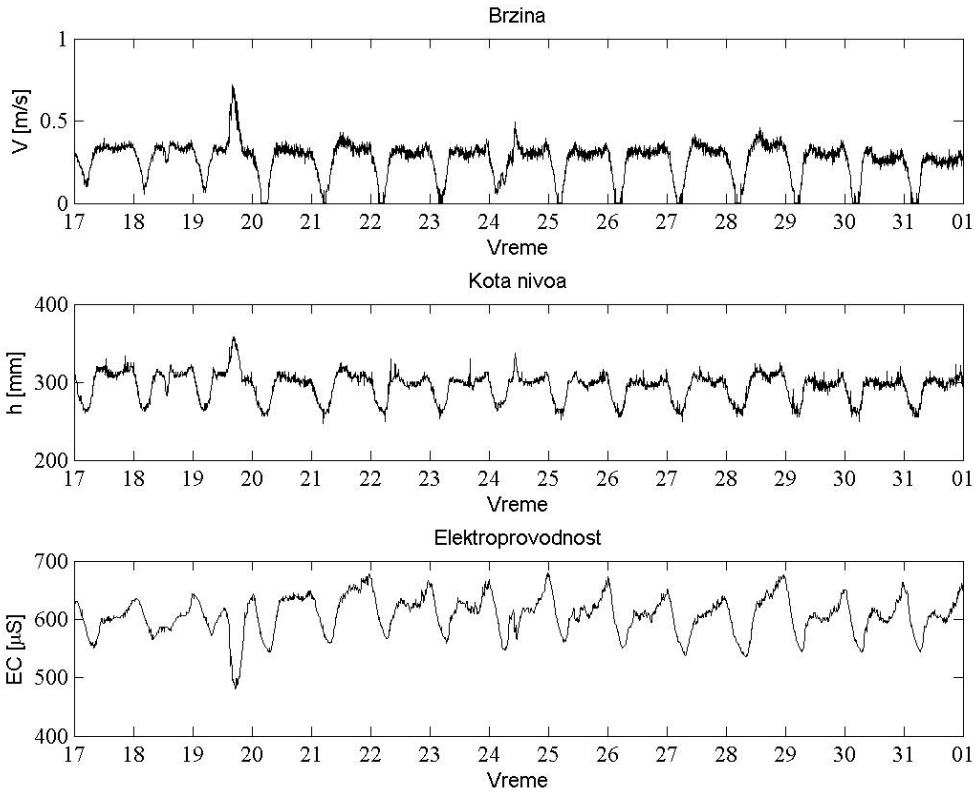
Zadatak vrednovanja ograničen je na vrednovanje u relanom vremenu, tj. za vrednovanje podatka izmerenog u vremenskom trenutku *t* dostupni su samo istorijski podaci. Kada se vrednuju podaci van realnog vremena, obično je dostupna cela vremenska serija.

S obzirom na to da je sistem BK razuĐen i da su izlivi odvojeni i pokrivaju različite podslivove, odabrano je jedno merno mesto (ispust Višnjica) da se na njemu prikaže razvoj sistema za vrednovanje podataka. TakoĐe, odabrane su dve hidrauličke veličine (dubina i brzina) i jedan parametar kvaliteta (elektroprovodnost). Na slici 5.39 prikazani su delovi (1/1/2007-28/2/2007) vremenskih serija dubine (h), brzine (V) i elektroprovodnosti (EC) na mernom mestu Višnjica na kojima se demonstrira sistem za vrednovanje podataka .

Za formiranje granica podataka i formiranje (kalibrisanje) relacija izmeĐu veličina koje se vrednuju potrebno je obezbediti podatke što boljeg kvaliteta, tj. podatke sa što manje grešaka. To se može ostvariti pažljivim jednokratnim merenjima na terenu sa pažljivo osmišljenim procedurama, ili odabirom istorijskih podataka iz kojih su uklonjene sve anomalije i sumnjive vrednosti. U ovom primeru odabran je jedan deo vremenskih serija dužine 15 dana (17/1/2007-31/1/2007), prikazan na slici 5.40. Podaci sa anomalijama (slika 5.40A) uklonjeni su i zamenjeni podacima izračunatim linearnom interpolacijom (slika 5.40B).

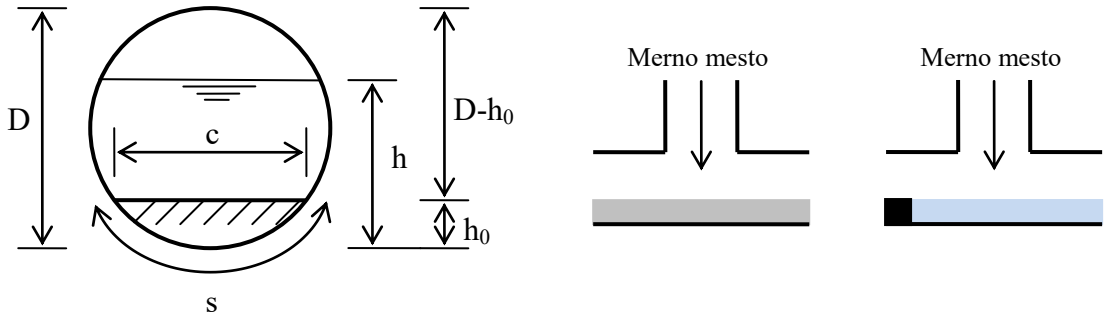


Slika 5.40A: Odabrani deo serija (17/1/2007-31/1/2007) sa ručno obeleženim anomalijama



Slika 5.40B: Odabrani deo serija (17/1/2007-31/1/2007) sa interpolovanim vrednostima za koje je iskustvom procenjeno da nisu regularne (sadrže anomalije)

Analizom podataka dolazi se do zaključaka o fenomenima koji se javljaju pri radu ispusta. Na dijagramu se može zapaziti da je nivo u noćnom periodu bez padavina relativno visok u odnosu na brzine, što se može objasniti nekim od slučajeva prikazanim na slici 5.41.



Šematizovani izgledi poprečnog profila za slučajeve A i B A: Sedimenti B: Mrtva zona

Slika 5.41: Mogući uzroci odstupanja noćnog nivoa i šematizovani izgledi poprečnog profila kolektora na mernom mestu za sve slučajeve

Zaključeno je da je u pitanju ili sediment na dnu kolektora, ili mrtva zona. Šematizovani prikaz poprečnog preseka kolektora prikazan je na istoj slici. Za prečnik kolektora usvaja se veličina $D = 0.8 \pm 0.005\,\mathrm{m}$, ukoliko se pretpostavi da je neodređenost mernog instrumenta kojim je meren prečnik (metra sa santimetarskom podelom) $u_D = 0.005\,\mathrm{m}$, gde u_D predstavlja neizvesnost podatka o prečniku kolektora. Veličina h_0 je izračunata kao najmanja registrovana kota nivoa, $h_0 = 249 \pm 8\,\mathrm{mm}$.

Veličina neodređenosti $u_h = 8$ mm određena je na osnovu kolebanja podataka istorijske vremenske serije dubine u kolektoru, prikazane na slici 5.40A.

5.3.1 Nivo šuma i neodređenost

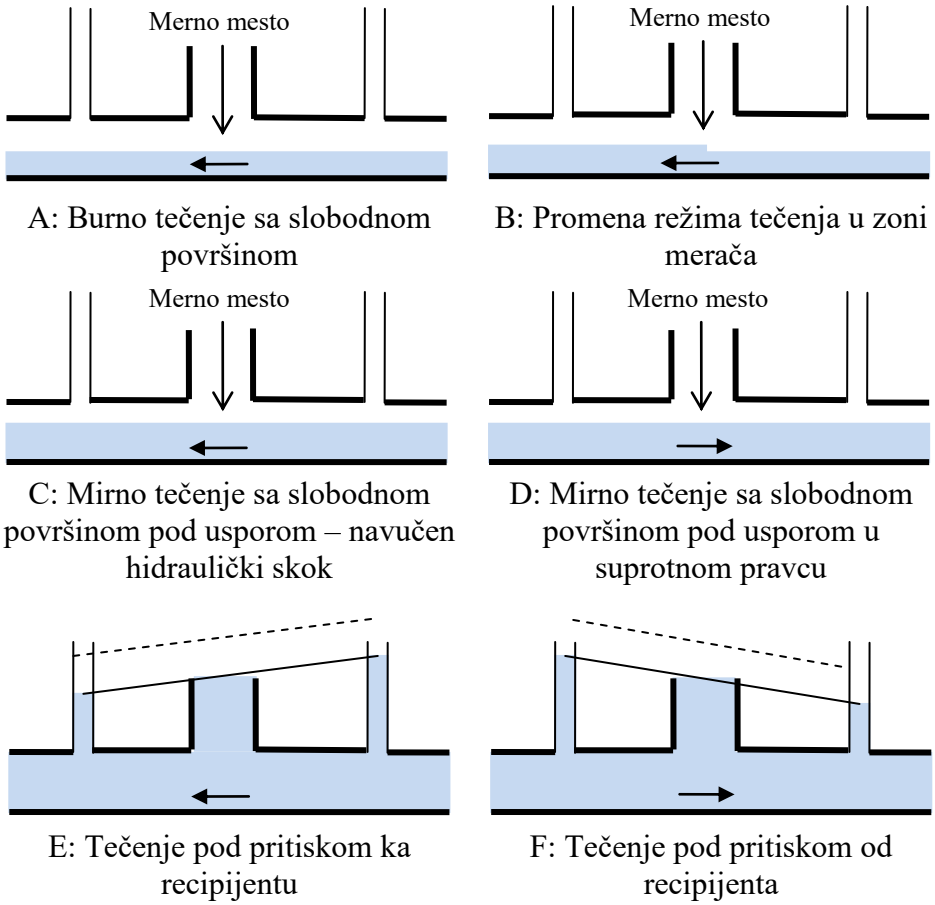
Nivo šuma koji se može očekivati u vremenskoj seriji zavisi kako od načina merenja, tako i od varijabilnosti merne veličine. Na dijagramima istorijskih vrednosti mernih veličina (slika 5.39) može se videti da neke veličine imaju veći, a neke manji šum.

Ukoliko se nivo šuma računa preko $SNR = \mu/\sigma$ odnosa (*signal to noise ratio* - odnos signala prema šumu) za odabrani deo serije, procenjene vrednosti iznose $SNR_V = 14.36$, $SNR_h = 140.2$ i $SNR_{EC} = 208.3$. Ukoliko se posmatra veličina intervala u kom se javljaju vrednosti merenih veličina u delovima serija bez trenda, procenjene vrednosti su sledeće: $u_V = \Delta V = 0.07$ m/s, $u_h = \Delta h = 8$ mm i $u_{EC} = \Delta EC = 35$ mS/cm.

U ovom primeru neodređenost je definisana kao fiksirana vrednost za ceo opseg merenja. Neodređenost je, inače, moguće definisati i u zavisnosti od izmerene veličine ili čak od veremena proteklog od poslednje kalibracije mernog uređaja.

5.3.2 Fizičke granice koje diktira sistem

Za razliku od prirodnih sistema, izgrađeni sistemi koje je formirao čovek uglavnom imaju definisanu geometriju. Ograničenja merenih veličina na izgrađenim sistemima uglavnom zavise od te geometrije i mogu se na osnovu nje i odrediti. Kanalizacioni sistem koji je sastavljen od cevi i zatvorenih kanala (kakva je mreža Beogradske kanalizacije) ima definisane prečnike cevi i geometriju poprečnih preseka kanala na osnovu kojih se mogu odrediti granične vrednosti dubina, propusne moći (protoka) i brzine. Pošto kanalizacione cevi imaju određen kapacitet, oticaj od neke ekstremne kiše (za koji ne postoje fizičke, već samo statističke granice) izliće se iz šahtova (ili neće ni ući u šahtove). Na slici 5.42 prikazano je šest konteksta tečenja u zoni mernog šahta. Tečenje sa slobodnom površinom može se ostvariti bez uticaja uspora od strane recipijenta (slika 5.42A). Promena režima tečenja u zoni mernog šahta prikazana je na slici 5.42B. Ova situacija predstavlja indikator loše odabranog mernog mesta, i zahteva njegovu promenu ili pripremu, kako bi se izbegle takve situacije. Na slikama 5.42C i 5.42D prikazano je tečenje pod usporom i to ka recipijentu (5.42C) i od recipijenta (5.42D). Tečenje prikazano na slici 5.42D predstavlja redak slučaj kada dolazi do nagle promene nivoa u recipijentu u kratkom vremenskom periodu. Sistem se može naći pod pritiskom zbog premalog kapaciteta kolektora da primi svu vodu (5.42E) ili u situaciji da zbog uticaja nivoa reke voda poteče u suprotnom smeru (slika 5.42F).



Slika 5.42: Šest mogućih konteksta tečenja u kolektoru Višnjica

Odgovarajući odnosi brzina i dubina u nekim od naznačenih konteksta prikazani su u izveštaju [31].

U slučaju merenih podataka koji su odabrani za demonstraciju sistema za vrednovanje (slika 5.40), iz raspoloživih podataka može se zaključiti da se javlja samo slučaj burnog tečenja u pravcu recipijenta prikazan na slici 5.42A.

Sa druge strane, pozicija mernih uređaja ne dozvoljava merenje dubina većih od prečnika kolektora, pa se kao fizička granica izmerene dubine može uzeti veličina bliska prečniku cevi. Takođe, osetljivost ultrazvučnih mernih uređaja je niska za male dubine i brzine, pa se često za male brzine registruje da brzine i nema.

Na osnovu gore izloženog određene su granice mogućih vrednosti merenih veličina. Fizička granica za dubinu vode u vidu minimuma je $h_0 - u_{h_0}$:

$$\left[h_{\min}, h_{\max}\right] = \left[h_0 - u_{h_0}, D + u_D\right],$$

gde $u_D = 0.5\,\text{cm}$ predstavlja neizvesnost podatka izmerenog prečnika cevi metrom sa santimetarskom podelom.

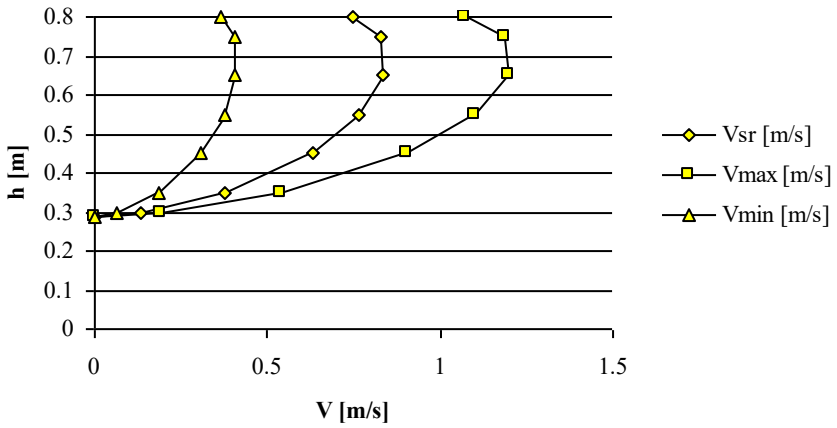
Iz fizičkih granica merenih podataka dubine vode u kolekoru mogu se odrediti i fizičke granice brzine vode. Fizička granica brzine definisana je kapacitetom cevi. Ukoliko se pretpostavi tečenje sa slobodnom površinom, sa pretpostavkom ustaljenog i jednolikog tečenja, za određivanje brzine u cevi može se upotrebiti *Šezi-Maning*-ova jednačina:

$$[V_{\min}, V_{\max}] = \left[0, \max\left(\frac{1}{n} R^{2/3} \sqrt{I_d}\right)\right].$$

Postoje dva kalibraciona parametra u Šezi-Maning-ovoj jednačini: Meningov koeficijent otpora n i pad linije dna I_d, koji se mogu sažeti u jedan - $\frac{\sqrt{I_d}}{n}$. Kalibracioni parametar moguće je dobiti u obliku intervala kalibracijom na osnovu izmerenih vrednosti. Polazne pretpostavke o geometriji sistema (D i h_0), neodređenosti ulaznih veličina (u_h i u_V) i granicama mogućih vrednosti parametara n i I_d su:

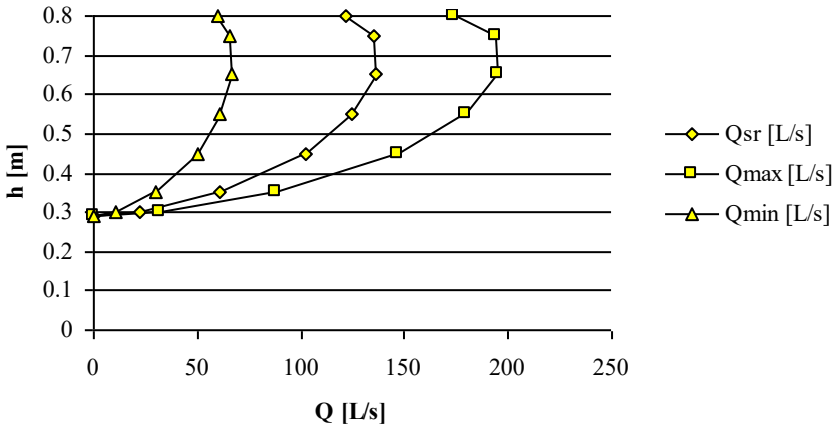
- $D = 0.8 \pm 0.005\,\text{m}$;
- $h_0 = 249 \pm 8\,\text{mm}$;
- $u_h = 8\,\text{mm}$;
- $u_V = 0.07\,\text{m/s}$;
- $n = [0.01, 0.02] m^{-1/3} s$, $I_d = [0.05, 0.15]\% - \frac{\sqrt{I_d}}{n} = [1.18, 3.873]$.

Na slici 5.43 prikazan je odnos nivoa dubine vode i brzine u kružnoj cevi.



Slika 5.43: Maksimalna brzina u cevi prema Šezi-Maning-ovoj jednačini

Prema dijagramu, opseg vrednosti u kojima se javlja brzina u kolektoru iznosi $V = [0,1.2]\,\text{m/s}$. S obzirom na to da je jedna od relacija upravo prikazana veza između brzine i dubine vode, dovoljno je voditi računa samo o ograničenju u pogledu dubine vode h.



Slika 5.44: Maksimalni protok u cevi prema Šezi-Maning-ovoj jednačini

Pomoću iste jednačine mogu se dobiti i informacije o protoku, kao što je prikazano na slici 5.44. Prema proračunu, interval u kom se nalazi protok je $Q = [0, 200]$ L/s. Kod parametara kvaliteta, fizičke granice mogu se odrediti na osnovu hipoteza, npr. da kanalizacionim sistemom nikada neće teći čista voda (tabela 5.9). Stoga je usvojeno da je $EC = [50, 2000]$ µS/cm.

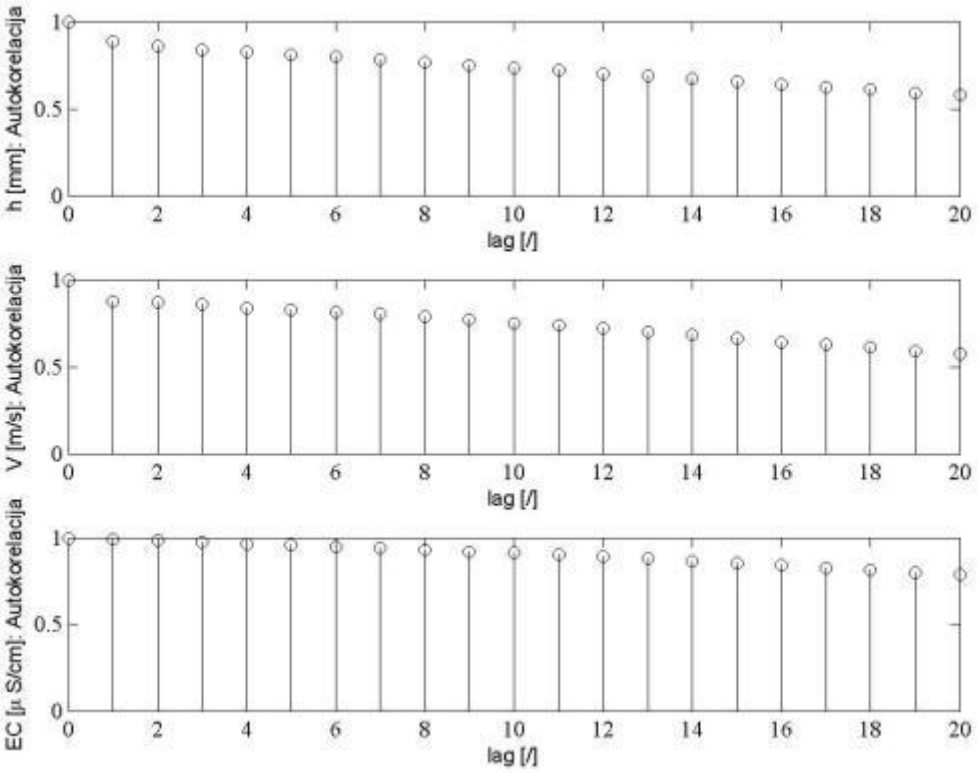
Tabela 5.9: Elektroprovodnosti nekih voda

Vrsta vode	Elektroprovodnost µS/cm
Dejonizovana voda	0.055
Zagrejana voda iz bojlera	1
Voda za piće	

5.3.3 Relacije između podataka (metode za predikciju)

Uz pretpostavku da je dubina u zoni mernog mesta normalna, merena dubina (h) i merena brzina (V)mogu se povezati Šezi-Maning-ovom jednačinom. Elektroprovodnost (EC) se može dovesti u vezu sa koncentracijom jona u kanalizacionoj vodi. Naime, koncentracija jona se menja kada se fekalna voda pomeša sa vodom atmosferskih padavina, što dovodi do smanjenja elektroprovodnosti. S obzirom na to da se protok može izračunati uz pomoć podataka o dubini i brzini, elektroprovodnost se može dovesti u vezu sa dubinom i brzinom.

Sa druge strane, autokorelacione karakteristike vremenskih serija (slika 5.45) ukazuju na postojanje značajne korelacije izmeĎu uzastopnih podataka, pa se, stoga, za svaku vrednovanu veličinu mogu formirati i auto-regresioni (AR(1)) modeli.



5.45: Autokorelacioni koeficijenti razmatranih veličina

Relacije se mogu dalje predstaviti funkcijama jedne veličine u zavisnosti od druge sa kojom su u relaciji. Tabelom 5.10 su predstavljene sve relacije u matričnoj formi.

Tabela 5.10: Matrična forma prikaza relacija i funkcionalnih zavisnosti između vrednosti

	V	h	EC	V^{t-1}	h^{t-1}	EC^{t-1}
R_1	$V^t = f_{V^t}^{R_1}\left(h^t\right)$	$h^t = f_{h^t}^{R_1}\left(V^t\right)$		$V^{t-1} = f_{V^{t-1}}^{R_1}\left(h^{t-1}\right)$	$h^{t-1} = f_{h^{t-1}}^{R_1}\left(V^{t-1}\right)$	
R_2		$h^t = f_{h^t}^{R_2}\left(EC^t\right)$	$EC^t = f_{EC^t}^{R_2}\left(h^t\right)$		$h^{t-1} = f_{h^{t-1}}^{R_2}\left(EC^{t-1}\right)$	$EC^{t-1} = f_{EC^{t-1}}^{R_2}\left(h^{t-1}\right)$
R_3	$V^t = f_{V^t}^{R_3}\left(EC^t\right)$		$EC^t = f_{EC^t}^{R_3}\left(V^t\right)$	$V^{t-1} = f_{V^{t-1}}^{R_3}\left(EC^{t-1}\right)$		$EC^{t-1} = f_{EC^{t-1}}^{R_3}\left(V^{t-1}\right)$
R_4	$V^t = f_{V^t}^{R_4}\left(V^{t-1}\right)$			$V^{t-1} = f_{V^{t-1}}^{R_4}\left(V\right)$		
R_5		$h^t = f_{h^t}^{R_5}\left(h^{t-1}\right)$			$h^{t-1} = f_{h^{t-1}}^{R_5}\left(h\right)$	
R_6			$EC^t = f_{EC^t}^{R_6}\left(EC^{t-1}\right)$			$EC^{t-1} = f_{EC^{t-1}}^{R_6}\left(EC\right)$

UvoĎenjem podataka u prethodnom vremenskom trenutku, V^{t-1}, h^{t-1} i EC^{t-1}, i prebrojavanjem i inverznih relacija, ukupan broj relacija koje se moraju definisati je šesnaest, kao što je prikazano u tabeli 5.11.

Tabela 5.11: Kratak opis relacija i funkcionalnih veza izmeĎu podataka

No	Relacija	Opis
1	$M_{R_1,V^t}: V^t = f_{V^t}^{R_1}\left(h^t\right)$	Brzina vode izražena preko izmerene dubine uz pomoć Šezi-Maning-ove formule u vremenskom trenutku *t*.
2	$M_{R_1,V^{t-1}}: V^{t-1} = f_{V^{t-1}}^{R_1}\left(h^{t-1}\right)$	Brzina izražena preko dubine uz pomoć Šezi-Maning-ove formule u prethodnom vremenskom trenutku (*t*-1).
3	$M_{R_1,h^t}: h^t = f_{h^t}^{R_1}\left(V^t\right)$	Dubina vode izražena preko izmerene brzine uz pomoć Šezi-Maning-ove -ove formule u vremenskom trenutku *t*.
4	$M_{R_1,h^{t-1}}: h^{t-1} = f_{h^{t-1}}^{R_1}\left(V^{t-1}\right)$	Dubina vode izražena preko izmerene brzine uz pomoć Šezi-Maning-ove formule u prethodnom vremenskom trenutku *t*-1.
5	$M_{R_2,EC^t}: EC^t = f_{EC^t}^{R_2}\left(h^t\right)$	Elektroprovodnost EC izražena preko izmerene dubine, fizičkim modelom promene koncentracije C jona u kanalizacionoj vodi u vremenskom trenutku *t*.
6	$M_{R_2,EC^{t-1}}: EC^{t-1} = f_{EC^{t-1}}^{R_2}\left(h^{t-1}\right)$	Elektroprovodnost EC izražena preko izmerene dubine, fizičkim modelom promene koncentracije C jona u kanalizacionoj vodi u vremenskom trenutku *t*-1.
7	$M_{R_2,h^t}: h^t = f_{h^t}^{R_2}\left(EC^t\right)$	Dubina vode izražena preko izmerene elektroprovodnosti EC na osnovu fizičkog modela promene koncentracije C jona u kanalizacionoj vodi u vremenskom trenutku *t*.
8	$M_{R_2,h^{t-1}}: h^{t-1} = f_{h^{t-1}}^{R_2}\left(EC^{t-1}\right)$	Dubina vode izražena preko izmerene elektroprovodnosti EC na osnovu fizičkog modela promene koncentracije C jona u kanalizacionoj vodi u vremenskom trenutku *t*-1.
9	$M_{R_3,EC^t}: EC^t = f_{EC^t}^{R_3}\left(V^t\right)$	Elektroprovodnost EC izražena preko izmerene brzine, fizičkim modelom promene koncentracije C jona u kanalizacionoj vodi u vremenskom trenutku *t*.
10	$M_{R_3,EC^{t-1}}: EC^{t-1} = f_{EC^{t-1}}^{R_3}\left(V^{t-1}\right)$	Elektroprovodnost EC izražena preko izmerene brzine, fizičkim modelom promene koncentracije C jona u kanalizacionoj vodi u vremenskom trenutku *t*-1.
11	$M_{R_3,V^t}: V^t = f_{V^t}^{R_3}\left(EC^t\right)$	Brzina vode izražena preko izmerene elektroprovodnosti EC na osnovu fizičkog modela promene koncentracije C jona u kanalizacionoj vodi u vremenskom trenutku *t*.
12	$M_{R_3,V^{t-1}}: V^{t-1} = f_{V^{t-1}}^{R_3}\left(EC^{t-1}\right)$	Brzina vode izražena preko izmerene elektroprovodnosti EC na osnovu fizičkog modela promene koncentracije C jona u kanalizacionoj vodi u vremenskom trenutku *t*-1.
13	$M_{R_4,V^t}: V^t = f_{V^t}^{R_4}\left(V^{t-1}\right)$	Brzina u vremenskom trenutku *t* izražena preko brzine izmerene u prethodnom vremenskom trenutku (*t*-1) pomoću AR(1) statističkog modela.
14	$M_{R_5,h^t}: h^t = f_{h^t}^{R_5}\left(h^{t-1}\right)$	Dubina u vremenskom trenutku *t* izražena preko dubine izmerene u prethodnom vremenskom trenutku (*t*-1) pomoću AR(1) statističkog modela.

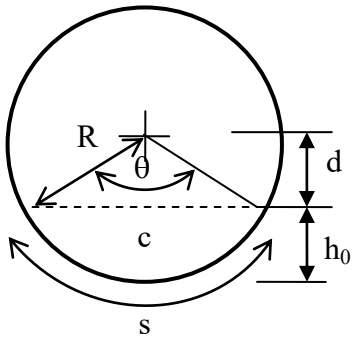
15	$M_{R_6,EC^t}: EC^t = f_{EC^t}^{R_6}\left(EC^{t-1}\right)$	Elektroprovodnost u vremenskom trenutku *t* izražena preko elektroprovodnosi u prethodnom vremenskom trenutku (*t*-1) pomoću AR(1) statističkog modela.
16	$M_{R_4,V^{t-1}}: V^{t-1} = f_{V^{t-1}}^{R_4}\left(V^t\right)$	Brzina u vremenskom trenutku (*t*-1) izražena preko brzine izmerene u vremenskom trenutku *t* pomoću AR(1) statističkog modela.
17	$M_{R_5,h^{t-1}}: h^{t-1} = f_{h^{t-1}}^{R_5}\left(h^t\right)$	Dubina u vremenskom trenutku (*t*-1) izražena preko dubine izmerene u vremenskom trenutku *t* pomoću AR(1) statističkog modela.
18	$M_{R_6,EC^{t-1}}: EC^{t-1} = f_{EC^{t-1}}^{R_6}\left(EC^t\right)$	Elektroprovodnost u vremenskom trenutku (*t*-1) izražena preko elektroprovodnosti izmerene u vremenskom trenutku *t* pomoću AR(1) statističkog modela.

Relacija R_1

Hidraulička veza dubine i brzine mogla bi da bude predstavljena u više nivoa detaljnosti. Od Šezi-Maning-ove jednačine koja povezuje brzinu i nivo pomoću geometrije kolektora, podužnog pada i Šezi-Maning-ovog koeficijenta otpora, do detaljnog CFD modela merne deonice u kom bi se ispitali i postojanje sekundarnih brzina, izdizanje nivoa u krivini, itd, različiti modeli rešavaju različite detalje u okviru hidrauličkih karakteristika sistema. Veza izmeĎu brzine i nivoa, uz pretpostavku postojanja normalne dubine, pomoću Šezi-Maning-ove formule ima sledeći oblik:

$$V(h) = \frac{1}{n} R^{2/3} \sqrt{I_d},$$

gde su: n – Maning-ov koeficijent otpora, I_d – nagib dna kolektora i $R(h) = A(h)/O(h)$ – hidraulički radijus. Površina poprečnog preseka ($A(h)$) i okvašeni obim ($O(h)$) se za poprečni presek kolektora mogu izračunati preko obrazaca prikazanih na slici 5.46.



$$R = h_0 + d$$

$$s = R\theta \quad (\theta \text{ je u radijanima})$$

$$A = 2R\sin\frac{\theta}{2} = R\sqrt{2 - 2\cos\theta}$$

$$h_0 = R\left(1 - \cos\frac{\theta}{2}\right)$$

$$\theta = \arccos\frac{d}{R}$$

Slika 5.46: Komponente kružnog odsečka

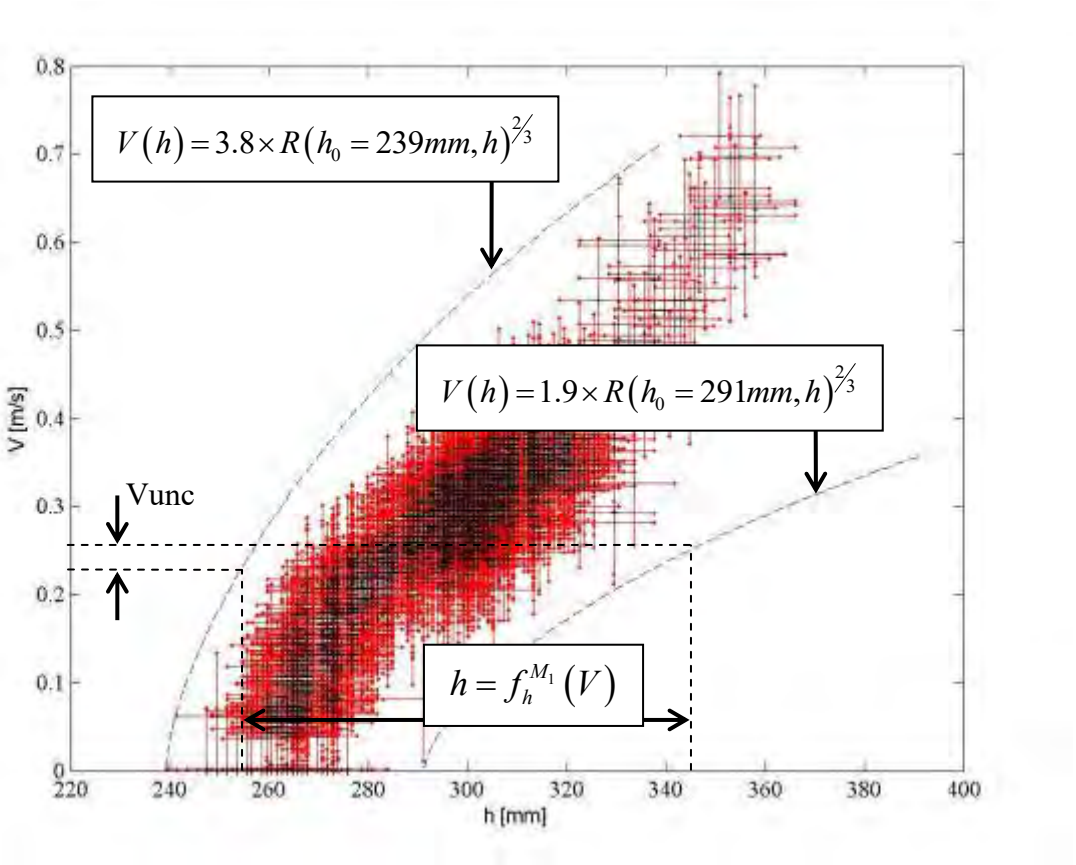
Uz pretpostavku normalne dubine, vezu izmeĎu dubinei brzine moguće je kalibrisati pomoću jednog kalibracionog parametra $C = \sqrt{I_d}/n$:

$$V = C \times R^{2/3}.$$

Pored kalibracionog parametra $C = \sqrt{I_d}/n$, utvrĐeno je da je značajan i parametar dubine ~~–~~mrtve zone" h_0 , pa se relacija može izraziti u sledećem obliku:

$$V(h_0, C, h) = C \times R(h_0, h)^{2/3}.$$

Kalibracija modela pomoću parametara $C = \sqrt{I_d}/n$ i h_0 obavljena je ručno na osnovu grafičkog prikaza izmerenih vrednosti istorijskih podataka (slika 5.47).



Slika 5.47: Kalibracija modela R_1 prema Šezi-Maning-ovoj jednačini

Rezultati kalibracije prikazani su u tabeli 5.12, a odgovarajuća jednačina funkcionalne zavisnosti modela $M_{R_1,V} : V = f_V^{R_1}(h)$ ima oblik:

$$M_{R_1,V} : V(h) = [3.8, 1.9] \times R([239, 291], h)^{2/3}.$$

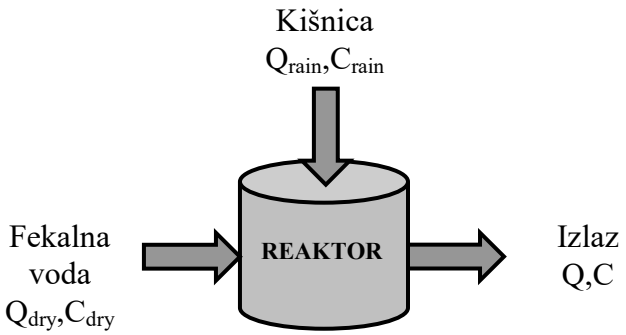
Tabela 5.12: Vrednosti parametara relacije M_1 nakon kalibracije

	C	h_0
	[$m^{1/3}/s$]	[mm]
min	1.9	239
max	3.8	291

Jednačinu funkcionalne zavisnosti $M_{R_1,h}: h = f_h^{R_1}(V)$ nije moguće eksplicitno izraziti za navedenu geometriju kolektora, već se mora numerički rešiti pomoću formule ($f_2^{R_1} = \left(f_1^{R_1}\right)^{-1}$).

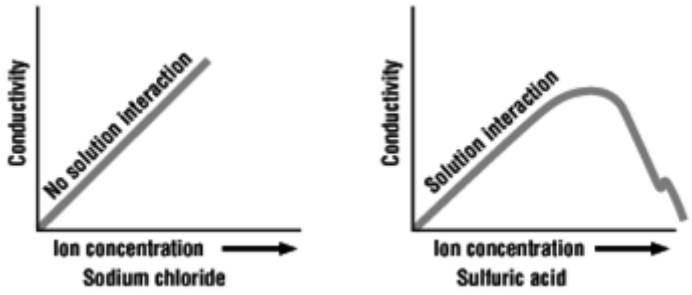
Relacije R_2 i R_3

Činjenica da elektroprovodnost zavisi od koncentracije jona i ukupnih rastvorenih materija u kanalizacionoj vodi ukazuje na mogućnost da se formira relacija između tih veličina. U vreme atmosferskih padavina fekalna voda se u kanalizacionom sistemu meša sa atmosferskom. Ove dve vrste voda imaju različite karakteristike u pogledu koncentracije jona, pa se pri njihovom mešanju smanjuje elektroprovodnost mešavine. Stoga bi se model smanjenja elektroprovodnosti za vreme atmosferskih padavina mogao opisati kao model promene koncentracije u nekom reaktoru prema šemi na slici 5.48.



Slika 5.48: Šema modela smanjenja elektroprovodnosti zbog kišnice

Pretpostavlja se da je elektroprovodnost proporcionalna koncentraciji jona (slika 5.49A), tj. ukupnim rastvorenim čvrstim materijama (*total disolved solids*, TDS)].



A **B**

Slika 5.49: Elektroprovodnost u zavisnosti od koncentracije jona

Dakle, elektroprovodnost EC se može izraziti preko linearne veze sa koncentracijom jona u kanalizacionoj vodi C:

$$EC = A_C \times C = A_C \times \frac{\Delta m}{\Delta t Q},\ \Delta m = \Delta m_{dry} + \Delta m_{rain},$$

gde se ukupna masa jona Δm može izraziti kao zbir mase jona u fekalnoj vodi (Δm_{dry}) i mase jona u vodi koja je u kolektor dospela kao kišnica (Δm_{rain}):

$$EC = \frac{A_C}{\Delta t} \times \frac{\Delta m_{dry} + \Delta m_{rain}}{Q} .$$

Konstanta A_C se može odrediti iz izdvojenih podataka za vreme perioda bez kiše:

$$EC_{dry} = \frac{A_C}{\Delta t} \times \frac{\Delta m_{dry}}{Q_{dry}} , \; \frac{A_C}{\Delta t} = EC_{dry} \times \frac{Q_{dry}}{\Delta m_{dry}} ,$$

gde je EC_{dry} elektroprovodnost za vreme perioda bez padavina. Iz navedenih jednačina može se izvesti relacija između elektroprovodnosti u toku suvog vremena i elektroprovodnosti za vreme perioda sa padavinama;

$$EC = EC_{dry} \times \frac{Q_{dry}}{\Delta m_{dry}} \frac{\Delta m_{dry} + \Delta m_{rain}}{Q} ;$$

$$EC = EC_{dry} \times \frac{Q_{dry}}{Q} \times \frac{\Delta m_{dry} + \Delta m_{rain}}{\Delta m_{dry}} ;$$

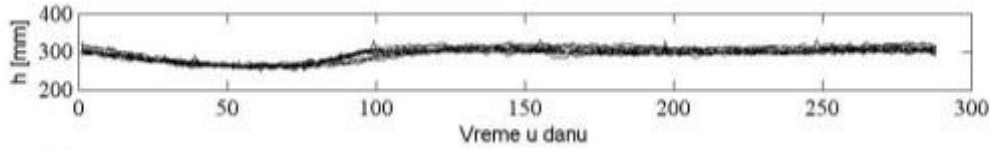
$$EC = EC_{dry} \times \frac{Q_{dry}}{Q} \times \left(1 + \frac{\Delta m_{rain}}{\Delta m_{dry}} \right) ;$$

$$EC = EC_{dry} \times \frac{Q_{dry}}{Q} \times A_C^{'} ,$$

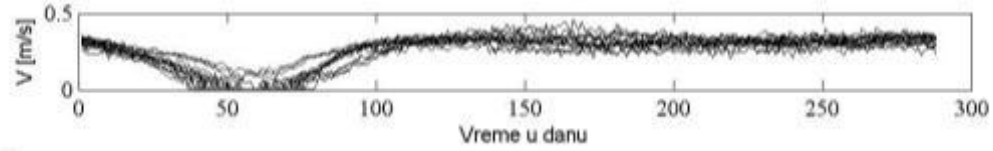
gde je $A_C^{'}$ novi koeficijent koji je jednak jedinici u periodima bez padavina, dok ga je u periodima sa padavinama potrebno kalibrisati na osnovu merenih podataka:

$$A_C^{'} = \begin{cases} 1 & , Q = Q_{dry} \\ A_C^{'} & , Q > Q_{dry} \end{cases} .$$

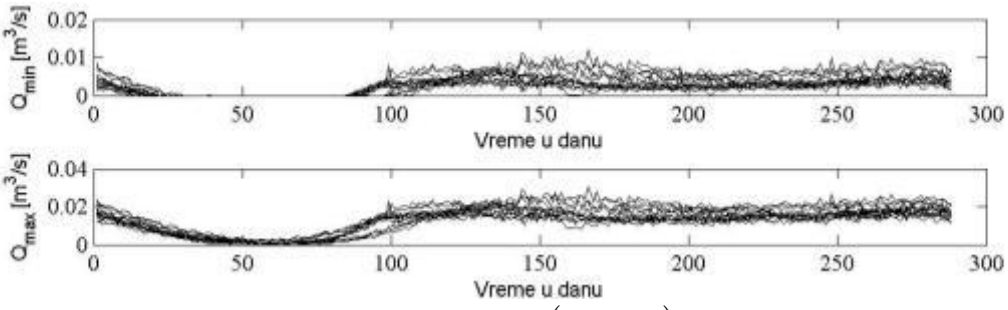
Ulazne vrednosti EC_{dry} i $Q_{dry}\left(h_{dry}, V_{dry}\right)$ mogu se proceniti i izračunati na više načina. Jedan od njih je procena intervala u kom se javljaju na nivou celog petnaestodnevnog perioda u kom postoje podaci za kalibraciju modela, tj. $EC_{dry} = \left[EC_{dry}^{\min}, EC_{dry}^{\max} \right]$, $h_{dry} = \left[h_{dry}^{\min}, h_{dry}^{\max} \right]$ i $V_{dry} = \left[V_{dry}^{\min}, V_{dry}^{\max} \right]$, na osnovu kojih se može izračunati $Q_{dry} = \left[Q_{dry}^{\min}, Q_{dry}^{\max} \right]$. Izdvojeni podaci su prikazani na slici 5.50.



A: Dubina h_{dry}



B: Brzina V_{dry}



C: Protok $Q_{dry}\left(h_{dry}, V_{dry}\right)$

D: Elektroprovodnost EC_{dry}

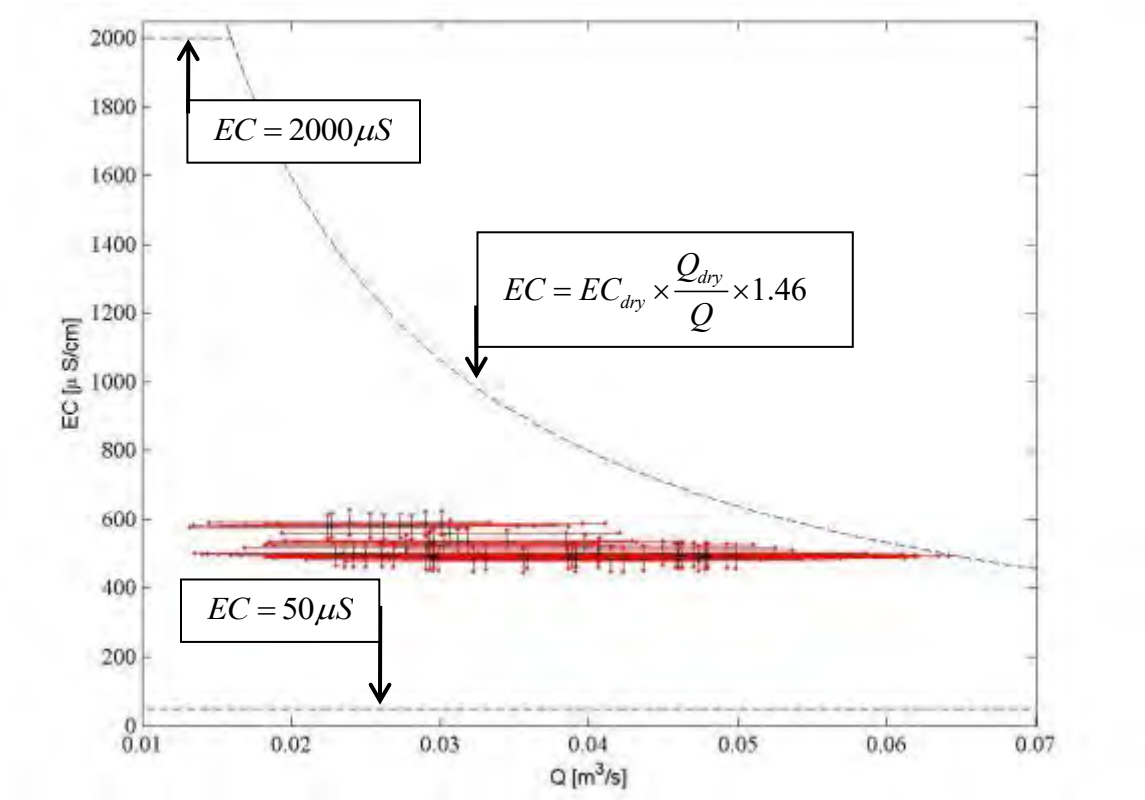
Slika 5.50: Izmerene vrednosti u toku dana za vreme bez padavina

Tabela 5.13: Granične vrednosti protoka i elektroprovodnosti u danima bez kiše

		min	max
EC_{dry}	[μS/cm]	501.503	714.721
Q_{dry}	[m^3/s]	0	0.0306

Ručnom kalibracijom može se dalje odrediti koeficijent $A_C^{'}$ (slika 5.51):

$$A_C^{'} = \begin{cases} 1 & ,Q = Q_{dry} \\ \left[A_C^{',\min}, A_C^{',\max} \right] & ,Q > Q_{dry} \end{cases} = \begin{cases} 1 & ,Q = Q_{dry} \\ 1.46 & ,Q > Q_{dry} \end{cases}.$$

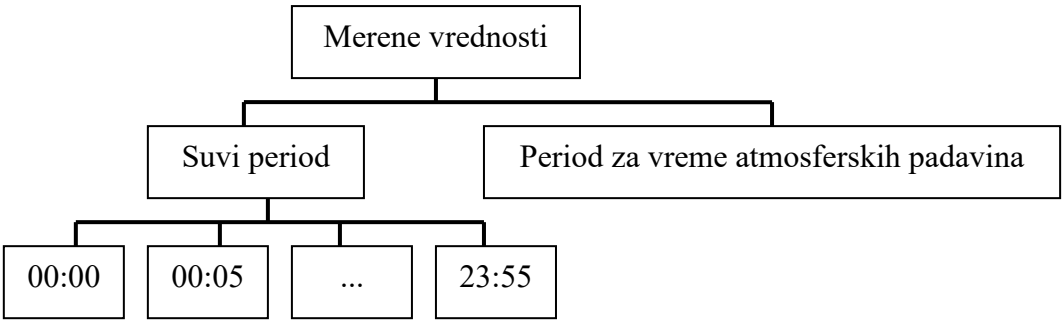


Slika 5.51: Kalibracija koeficijenta $A_C^{'}$

Sa druge strane, prepoznatljiv šablon u kom se javljaju podaci na dnevnom nivou omogućava da se intervali definišu na bazi vremenskog konteksta, tj. vremena u toku dana u kom su podaci mereni. U [15] navedeno je nekoliko konteksta koji se mogu izdvojiti za tečenje u kanalizacionom sistemu:

- klimatski: kišno/suvo;
- hidraulički: tečenje pod usporom/slobodno tečenje;
- pumpni: *on*/*off*;
- socijalni: radni dan/vikend;
- sezonski: proleće/leto/jesen/zima;
- vremenski: 0/…/23;
- u pogledu doba dana: noć/jutro/dan/veče;
- socijalni događaji: regularan dan/praznik/utakmica/...

Iz istorijskih vremenskih serija dubine vode, brzine i elektroprovodnosti (slika 4.30) i prirode samog fenomena tečenja u kombinovanoj kanalizaciji, mogu se uočiti karakteristični šablon i konteksti u kojima se javljaju mereni podaci [11]. Dva konteksta koji se iz dijagrama prvi mogu uočiti su (slika 5.52) suvi periodi i period za vreme atmosferskih padavina. Drugi kontekst (podkontekst) vezan je za tečenje u suvom periodu i za doba dana [11]: 1) jutro, 2) dan, 3) veče i 4) noć.



Slika 5.52: Konteksti u kojima se javljaju podaci

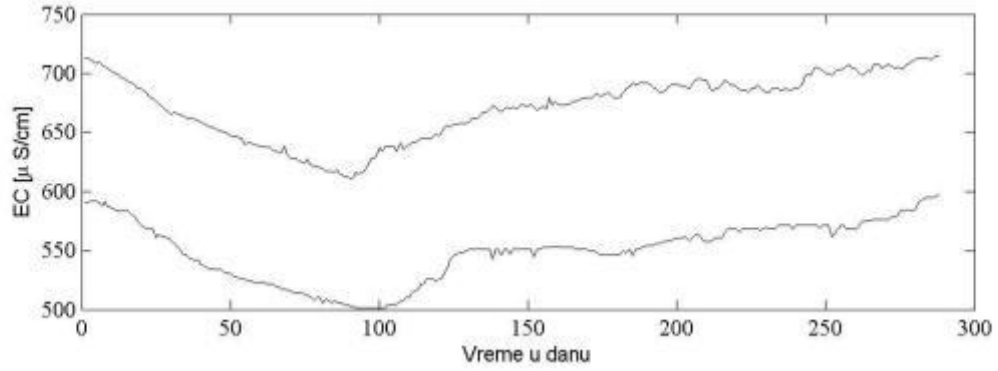
Podaci o atmosferskim padavinama su od izuzetnog značaja za proveru podataka merenih u kanalizacionom sistemu. Ukoliko su poznati podaci o atmosferskim padavinama, mereni podaci se mogu odvojiti u dve klase: klasu podataka za vreme kada nije bilo atmosferskih padavina i klasu podataka za vreme atmosferskih padavina. Kada nema padavina, navike građana formiraju karakterističan oblik hidrograma, pa samim tim i nivograma i vremenske serije brzina.

Ukoliko se izmereni podaci posmatraju prema vremenu uzorkovanja (slika 5.53) u toku dana, i predstave u matrici $Y_{n\times 288}$, gde je n broj dana u istorijskoj vremenskoj seriji sa po 288 merenja sa intervalom od $\Delta t = 5\,\text{min}$, $Y_{n\times 288}$ merene veličine V, h (Q) ili EC:

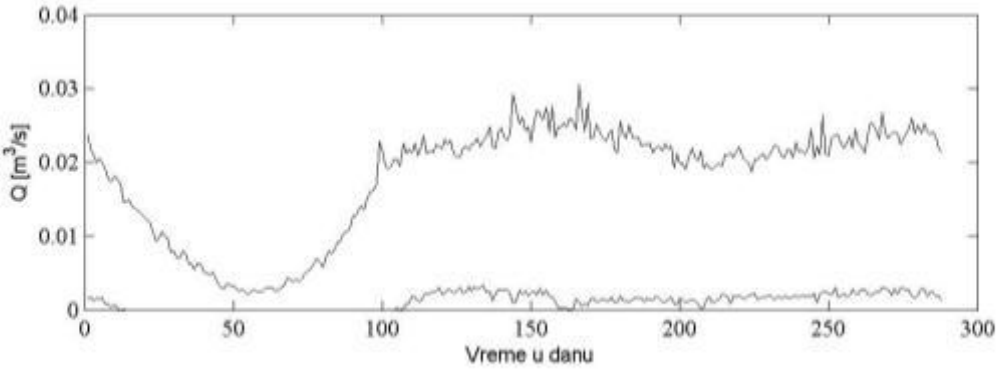
$$Y = \begin{bmatrix} y_{d/m/y\,00:00} & y_{d/m/y\,00:05} & \cdots & y_{d/m/y\,23:50} & y_{d/m/y\,23:55} \\ y_{d+1/m/y\,00:00} & y_{d+1/m/y\,00:05} & \cdots & y_{d+1/m/y\,23:50} & y_{d+1/m/y\,23:55} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ y_{d+(n-1)/m/y\,00:00} & y_{d+(n-1)/m/y\,00:05} & \cdots & y_{d+(n-1)/m/y\,23:50} & y_{d+(n-1)/m/y\,23:55} \\ y_{d+n/m/y\,00:00} & y_{d+n/m/y\,00:05} & \cdots & y_{d+n/m/y\,23:50} & y_{d+n/m/y\,23:55} \end{bmatrix},$$

podaci u kolonama mogu se modelirati intervalima:

$$Y = \left[\left[y_{00:00}^{\min}, y_{00:00}^{\max} \right] \quad \left[y_{00:05}^{\min}, y_{00:05}^{\max} \right] \quad \cdots \quad \left[y_{23:50}^{\min}, y_{23:50}^{\max} \right] \quad \left[y_{23:55}^{\min}, y_{23:55}^{\max} \right] \right].$$



A: Elektroprovodnost EC_{dry}



B: Protok $Q_{dry}\left(h_{dry}, V_{dry}\right)$

Slika 5.53: Granične vrednosti elektroprovodnosti (A) i protoka (B) u toku dana bez padavina

Protok Q može se izračunati na osnovu veze između brzine i površine poprečnog preseka:

$$Q = V \times A(h).$$

Kod relacije R_2 moguće je izraziti brzinu pomoću relacije R_1 ($V = R_1(h)$). Kod relacije R_3 moguće je izraziti dubinu h pomoću modela R_1 ($h = R_1(V)$).

Napomena: Značajno odstupanje rezultata modela od merenih vrednosti može se registrovati u situacijama kada se u kišnici nađe rastvoreni nanos, hemikalije ili kuhinjska so kojom se zimi posipaju ulice.

Relacije R_4, R_5 i R_6

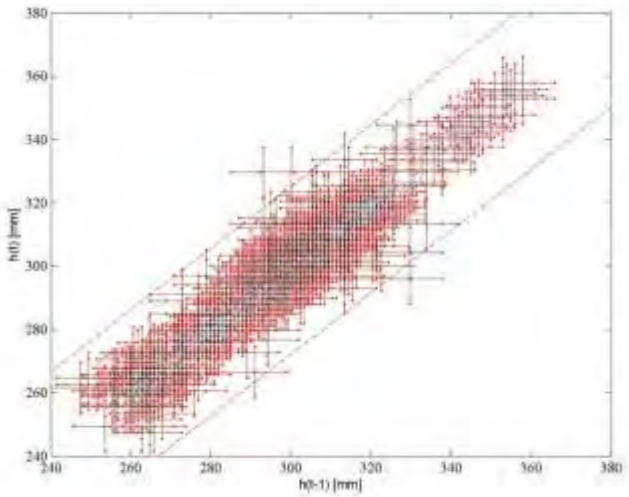
Autokorelacione karakteristike vremenskih serija (visoka autokorelacija) ukazuju na postojanje mogućnosti da se formira AR (*autoregressive*) model u formi:

$$x^t = ax^{t-1} + b,$$

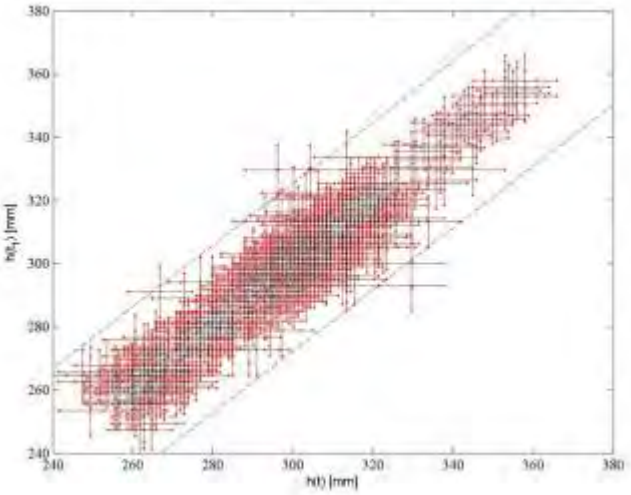
gde x predstavlja neku od merenih veličina V, h ili EC, dok eksponent t predstavlja podatak u tekućem vremenskom trenutku, a $t-1$ u prethodnom ($\Delta t = 5\,\text{min}$). Koeficijenti a i $b = \left[\underline{b}, \overline{b}\right]$ su određeni kalibracijom na osnovu istorijskih podataka.

Tabela 5.14: Kalibracioni parametri AR(1) modela

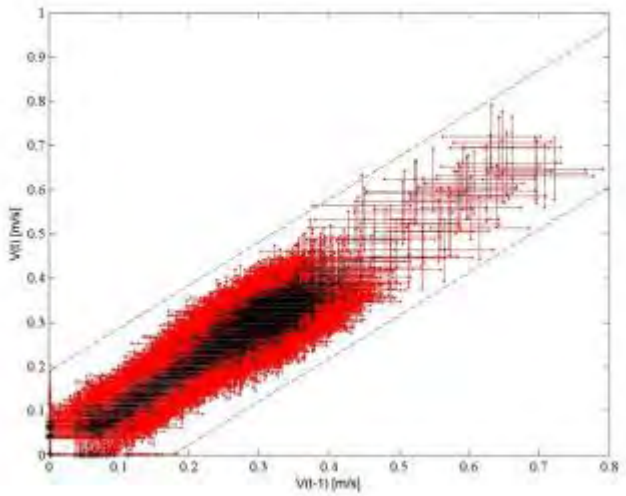
x	h	V	EC
a	0.97427	0.97258	0.99712
$b=\left[\underline{b},\bar{b}\right]$	[-33, 52]	[-0.17, 0.19]	[-59, 49]



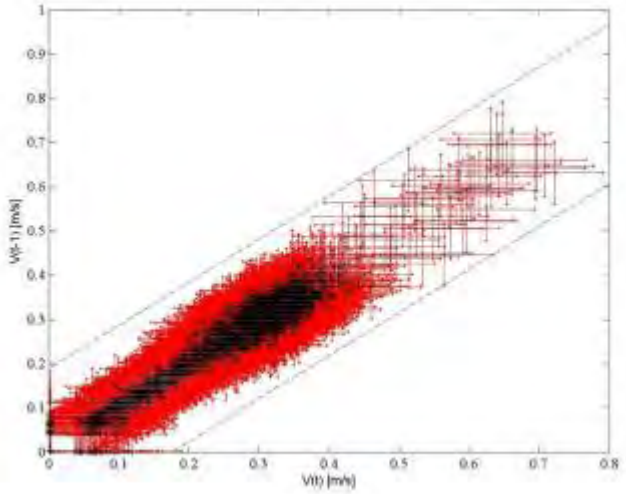
Metoda M_{R_4,h^t}: $h^t = 0.97427h^{t-1} + \left[-20,33\right]$



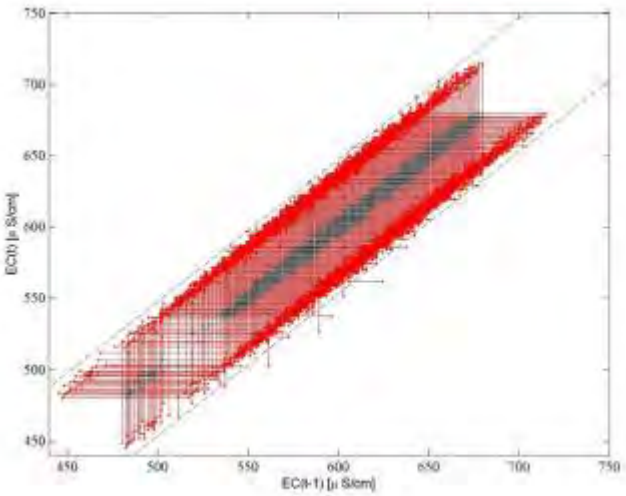
Metoda $M_{R_4,h^{t-1}}$: $h^{t-1} = 0.9745h^t + \left[-20,33\right]$



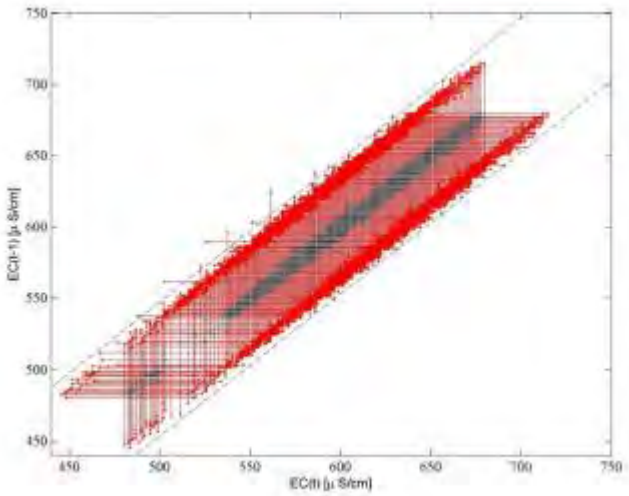
Metoda M_{R_5,V^t} : $V^t = 0.97258V^{t-1} + [-0.17, 0.19]$



Metoda $M_{R_5,V^{t-1}}$: $V^{t-1} = 0.9726V^t + [-0.17, 0.19]$



Metoda M_{R_6,EC^t} : $EC^t = 0.99712EC^{t-1} + [-45, 49]$



Metoda $M_{R_6,EC^{t-1}}: EC^{t-1} = 0.9968EC^{t} + [-45,49]$

Slika 5.54: Metode bazirane na relacijama R_4, R_5 i R_6

Definisanjem relacija izmeĎu veličina i njihovom kalibracijom završena je priprema sistema za vrednovanje.

5.3.4 Rezultati i diskusija o realnom primeru merenja u kanalizaciji

Sistem za vrednovanje, dizajniran pomoću relacija opisanih u prethodnom odeljku, testiran je na nekoliko karakterističnih delova vremenskih serija merenih u:

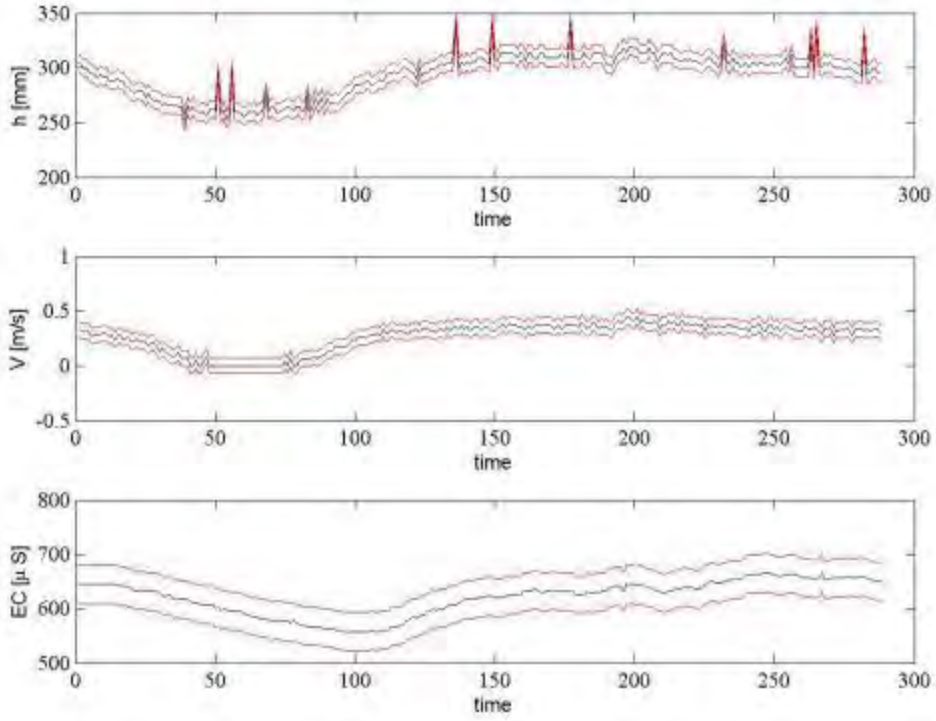
1. danu bez padavina;
2. danu sa padavinama.

Pri merenju hidrauličkih parametara i parametara kvaliteta u kanalizacionom sistemu padavine narušavaju karakteristični šablon oticaja upotrebljenih voda, a anomalije mogu biti maskirane promenom vrednosti hidrauličkih parametara. Ovim primerom pokazano je da predložena metodologija daje dobre rezultate u oba slučaja.

Treba naglasiti da su anomalije u podacima odabrane tako da ih je lako uočiti vizuelnom inspekcijom, ali da su niskog intenziteta, čime se potvrĎuje i osetljivost algoritma, kao i kvalitet matematičkih relacija izmeĎupodataka. Grube greške u primerima nisu razmatrane jer predloženi algoritam ima tu osobinu da se greške sa većim intenzitetom lakše detektuju.

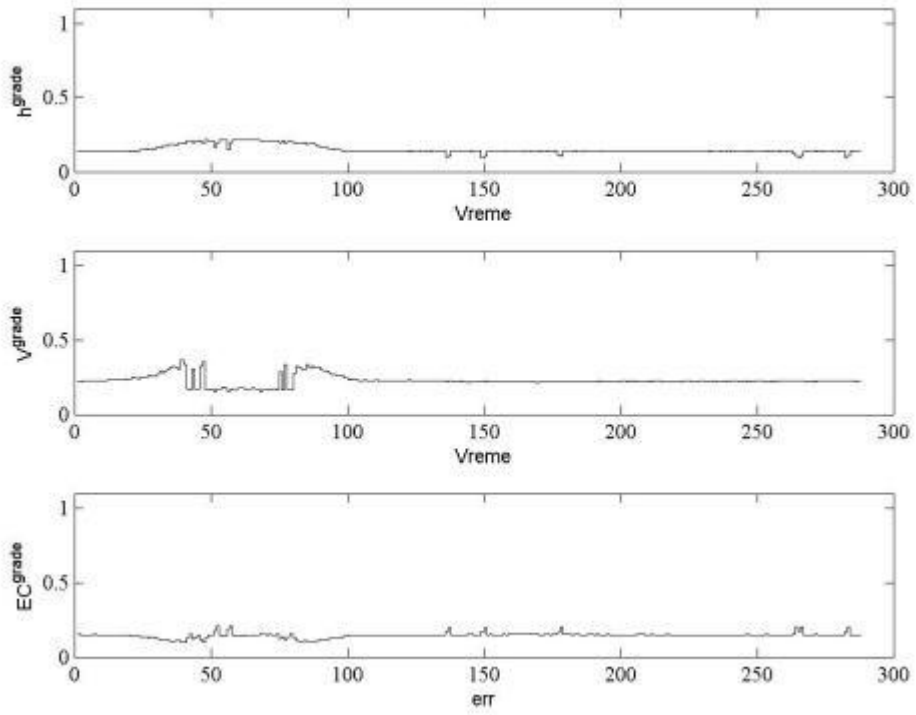
Merene vrednosti u danu bez padavina

Karakterističan šablon promene vrednosti dubine, brzine i elektroprovodnosti može se uočiti u danu 6/1/2007 u kom nije bilo atmosferskih padavina (slika 5.55).



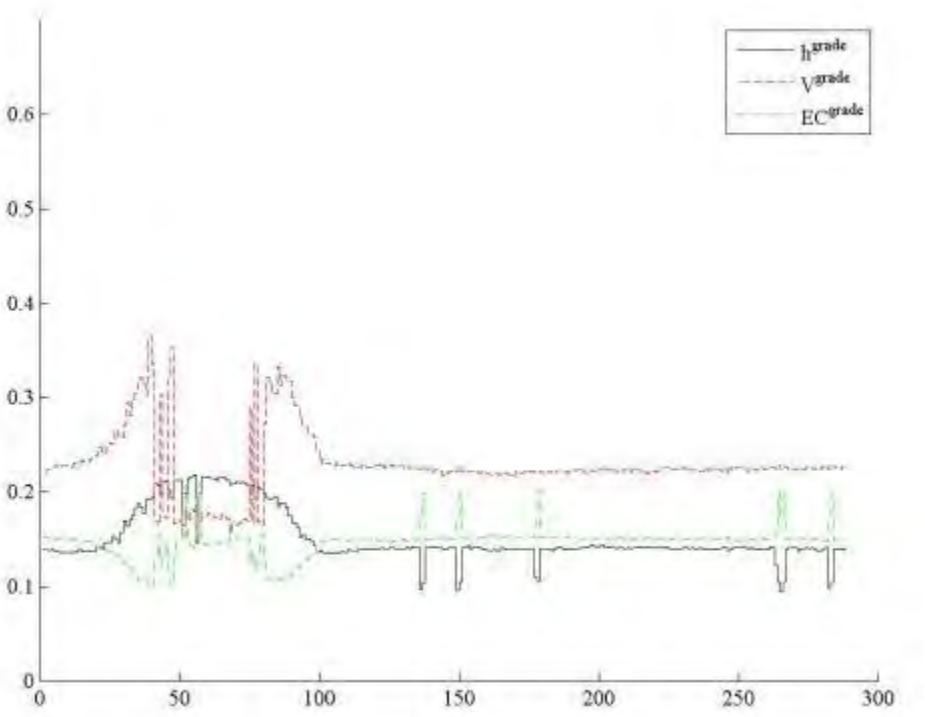
Slika 5.55: Merene vrednosti u danu bez padavina 6/1/2007

Kod vremenske serije dubine u kolektoru javljaju se pikovi koji deluju kao greške u merenjima. Može se primetiti da se pikovi javljaju kako za vreme noćnog protoka, tako i za vreme dnevnog. Na slici 5.56 prikazane su verovatnoće merenih vrednosti $p\left(x_i, X_{x_i}, MM_i\right)$ nakon primenjenog algoritma. Ove verovatnoće mogu se smatrati ocenama kvaliteta merenih vrednosti u odnosu na ostale merene vrednosti. Uzrok relativno niskih verovatnoća za većinu merenih vrednosti je svakako neizvesnost relacija kojima su povezane sa drugim merenim veličinama. Iako ne postoji jasna granica koja se može povući tako da se odvoje samo ocene koje su ispod granice, uočava se pad ocena na mestima nekih pikova kod merene dubine.



Slika 5.56: Verovatnoće merenih vrednosti

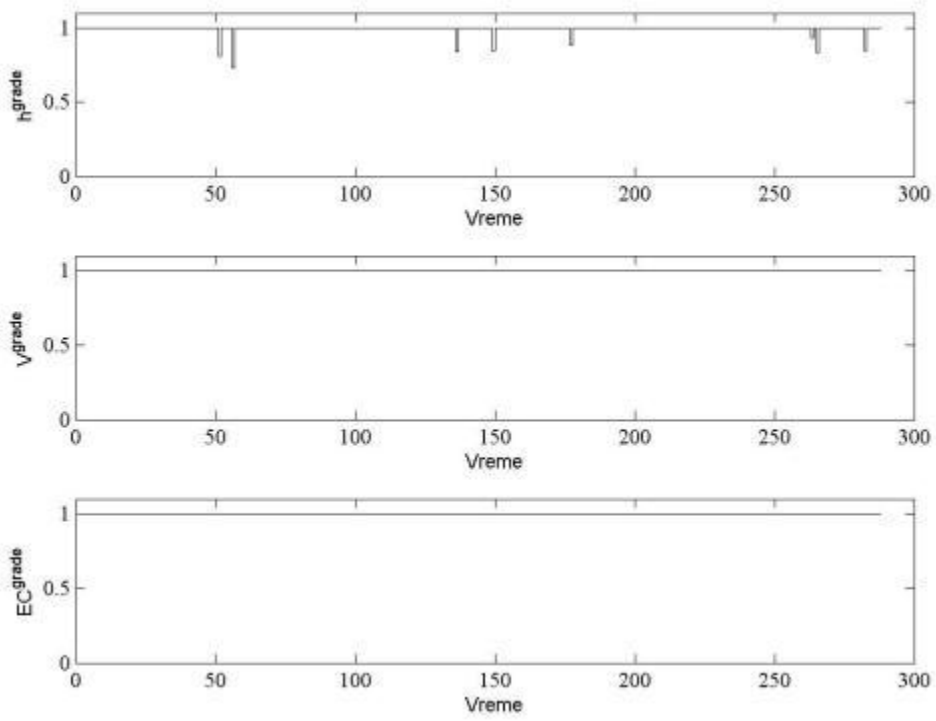
Ukoliko se verovatnoće merenih podataka prikažu na istom dijagramu (slika 5.57), smanjenje verovatnoća na mestima pikova uočava se još izraženije. Međutim, i dalje niske verovatnoće kod merenih vrednosti elektroprovodnosti onemogućavaju da se povuče jasna granica između podataka bez grešaka i podataka sa greškama. Uzrok niskih vrednosti je velika neodređenost modela kojom se izračunava elektroprovodnost.



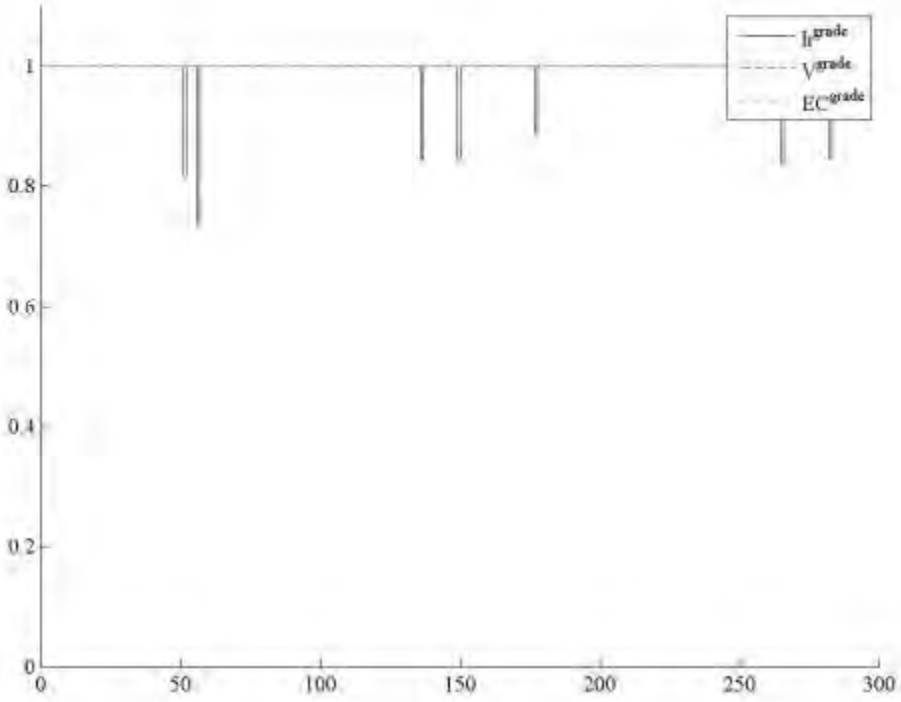
Slika 5.57: Verovatnoće na jednom dijagramu

Normiranjem verovatnoća može se izolovati greška koja potiče samo od odstupanja merene i izračunate vrednosti, pa maksimalna verovatnoća postaje jednaka jedinici. Na taj način se stvara uslov da se povuče oštra linija i da se izdvoje vrednosti sa manjim ocenama od granične (slika 5.58).

Na slici 5.58 vidi se da je veliki broj normiranih verovatnoća jednak jedinici, što znači da se merena vrednost nalazi u intervalu izračunatih vrednosti. Kada to nije slučaj, normirane verovatnoće su manje i to proporcionalno veličini razilaženja izmerenih i izračunatih vrednosti.



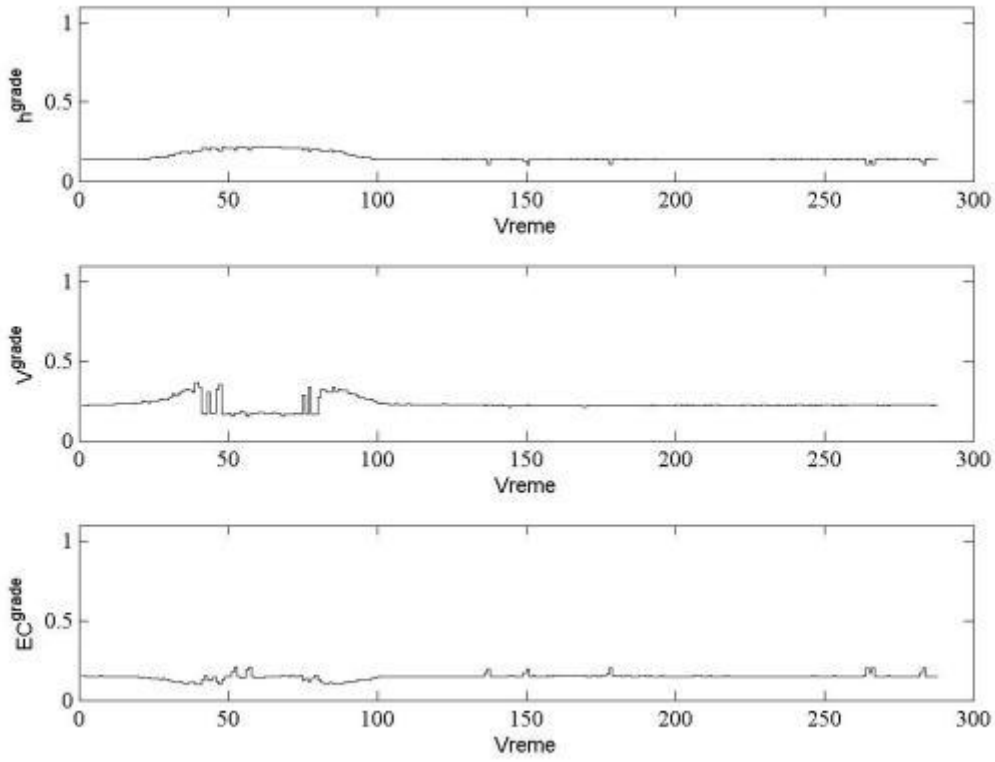
Slika 5.58: Normirane verovatnoće [0,1]



Slika 5.59: Normirane verovatnoće na jednom dijagramu

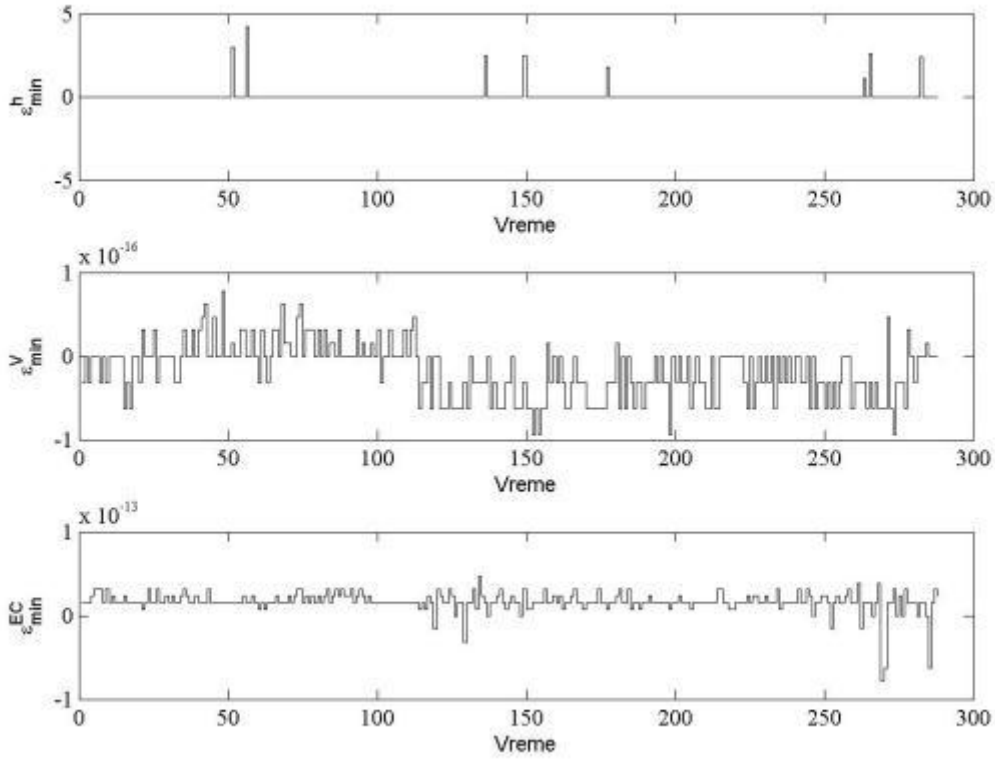
Na ovakvim dijagramima moguće je povući liniju koja označava granicu do koje je normirana verovatnoća prihvatljiva. Ta granica zavisi od načina upotrebe podatka i grešaka koje od takve upotrebe mogu nastati.

Granice u kojima se kreće neizvesnost metoda koje se koriste za predikciju su između 0 (za veliku neodređenost) i 1 (kada nema neodređenosti). Na slici 5.60 prikazane su neizvesnosti relacija koje se koriste u sistemu za vrednovanje. Vidi se da su neizvesnosti relativno visoke, pri čemu je najmanja neizvesnost kod brzine, dok je najveća kod elektroprovodnosti u slučajevima kada nema protoka. Uvidom u relacije kojima su vezani podaci može se zaključiti da su izračunate neizvesnosti dobro procenjene.

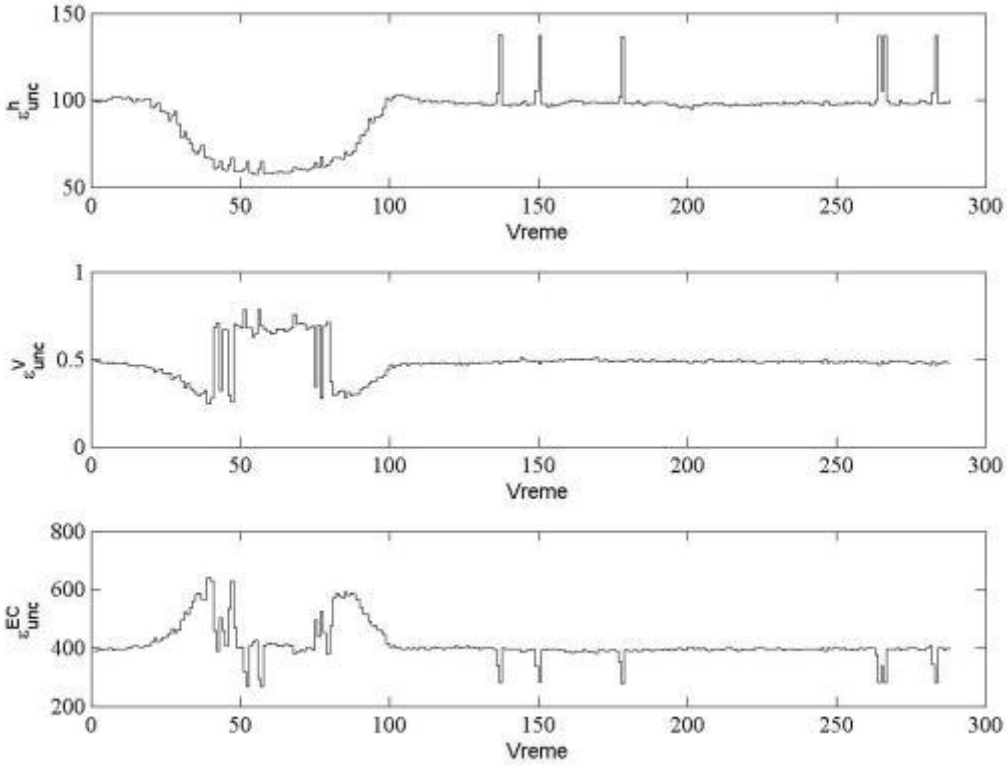


Slika 5.60: Mera neizvesnosti sistema za vrednovanje

Na slikama 5.61 i 5.62 prikazane su minimalna greška (izračunata izrazom 4.7) i neizvesnost greške merenih veličina (izračunata izrazom 4.8).

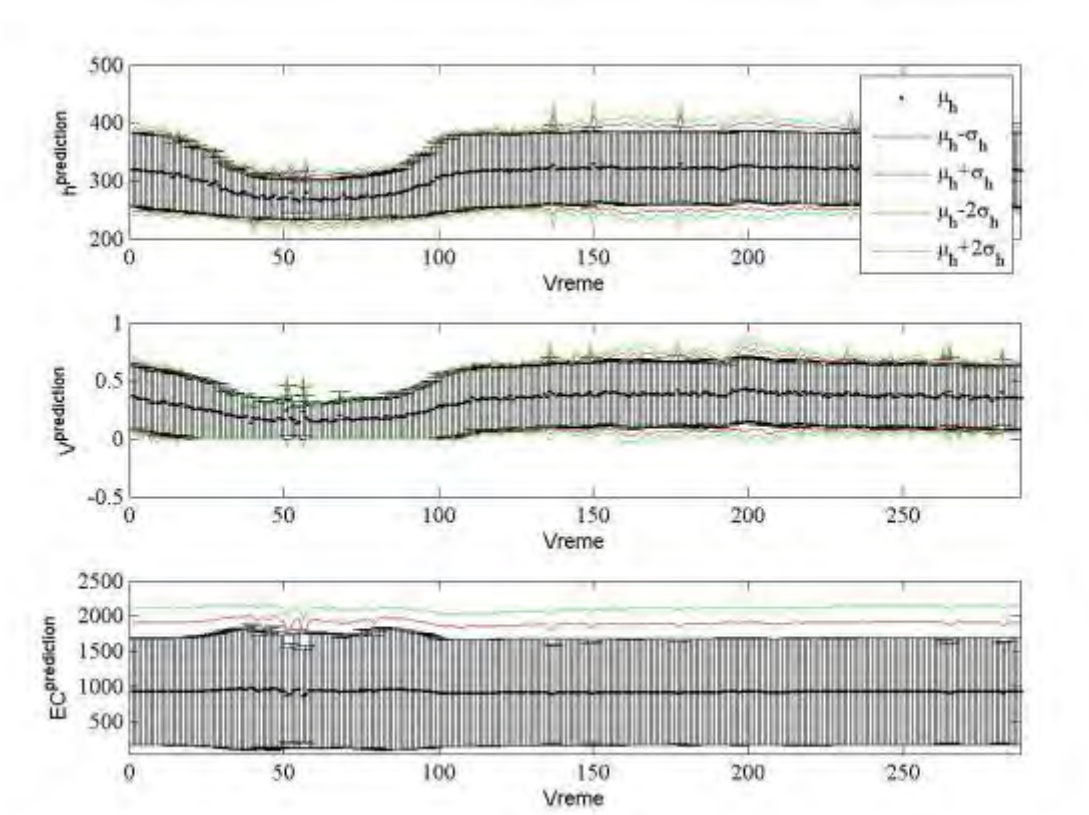


Slika 5.61: Minimalna greška



Slika 5.62: Neizvesnost greške

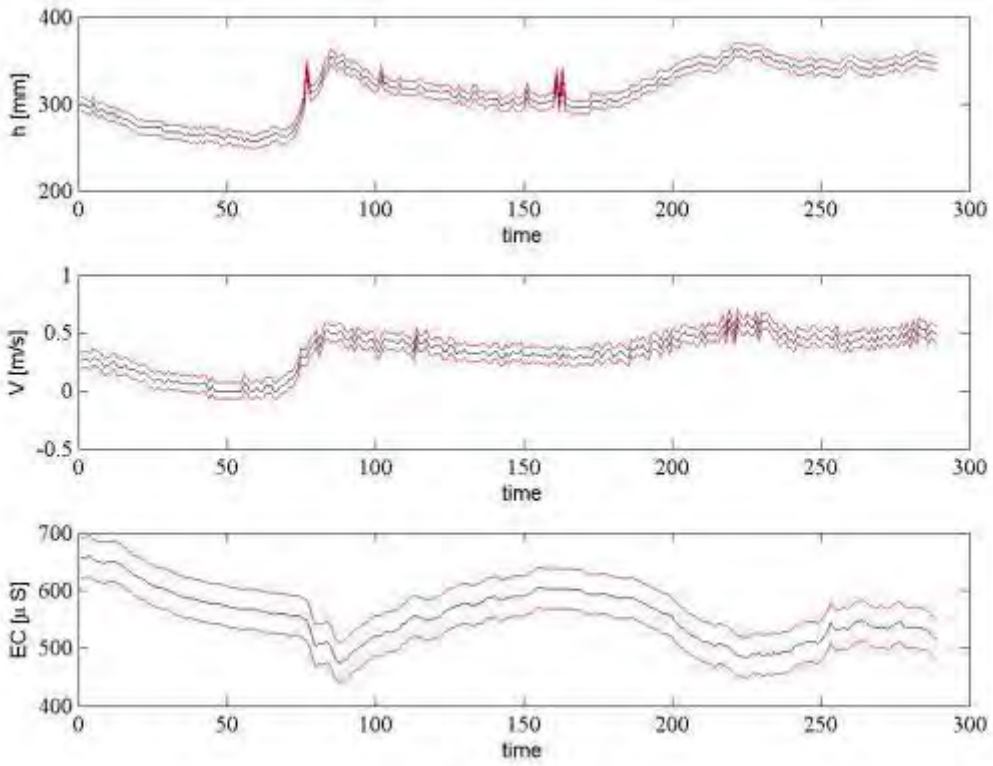
Na samom kraju prikazani su intervali očekivanih vrednosti na osnovu izračunatih vrednosti, i prateći intervali varijansi (slika 5.63). Vidi se da su varijanse veće na mestima na kojima se nalaze anomalije u podacima.



Slika 5.63: Intervali u kojima se pretpostavlja da se nalazi tačna vrednost

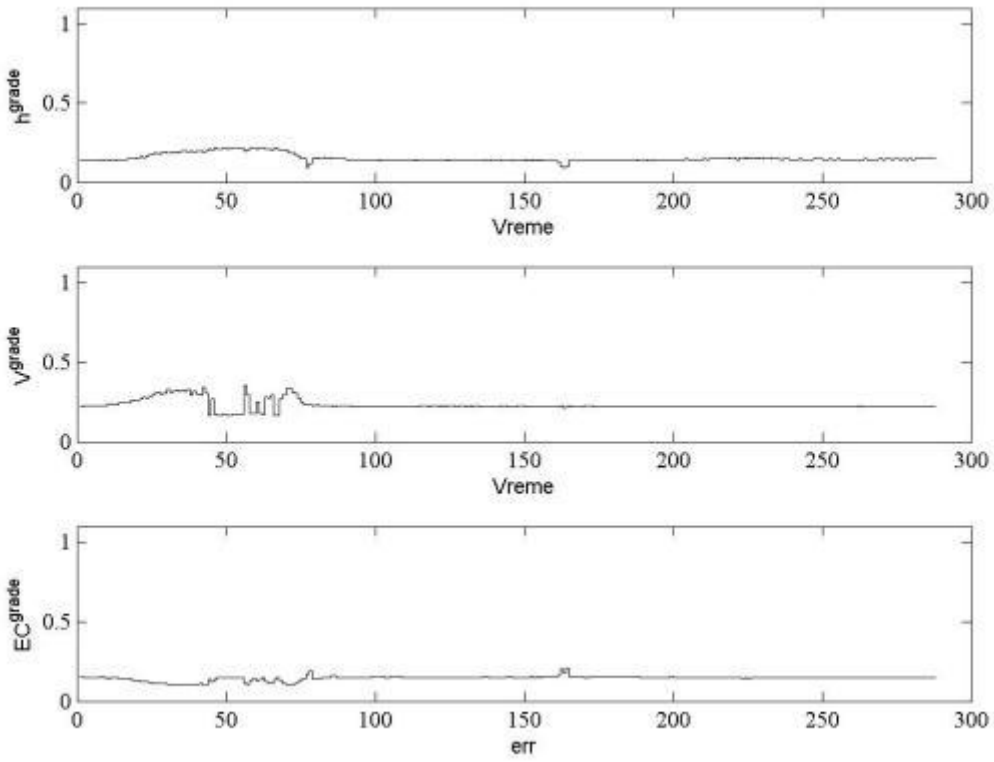
Merene vrednosti u danu sa padavinama

Dan sa padavinama (13/2/2007) odabran je kako bi se demonstrirala adaptivnost sistema za vrednovanje na povećane vrednosti dubina i brzina, kao i na smanjene vrednosti elektroprovodnosti. Posebno je interesantno videti kako će predloženi algoritam rešiti problem anomalija na uzlaznoj grani hidrograma (slika 5.64), s obzirom da je to mesto na kome je izuzetno teško razlučiti da li se radi o anomaliji ili je u pitanju deo hidrograma.

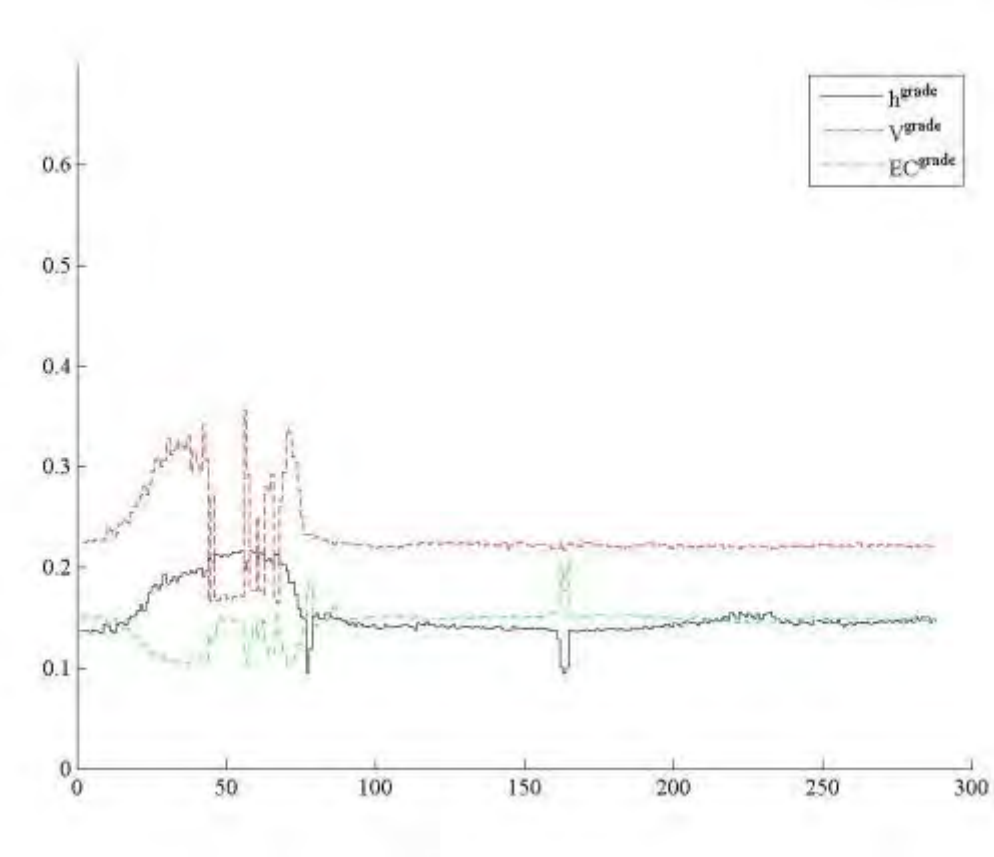


Slika 5.64: Merene vrednosti u danu bez padavina 13/2/2007

Na slikama 5.65 i 5.66 prikazane su ukupne verovatnoće merenih vrednosti koje uključuju kako slaganje merenih i izračunatih vrednosti, tako i neizvesnosti koje se unose u sistem neizvesnim relacijama koje povezuju podatke. Kao i u prethodnom primeru, izuzetno je teško, ako ne i nemoguće, povući granicu između regularnih podataka i podataka sa anomalijama.

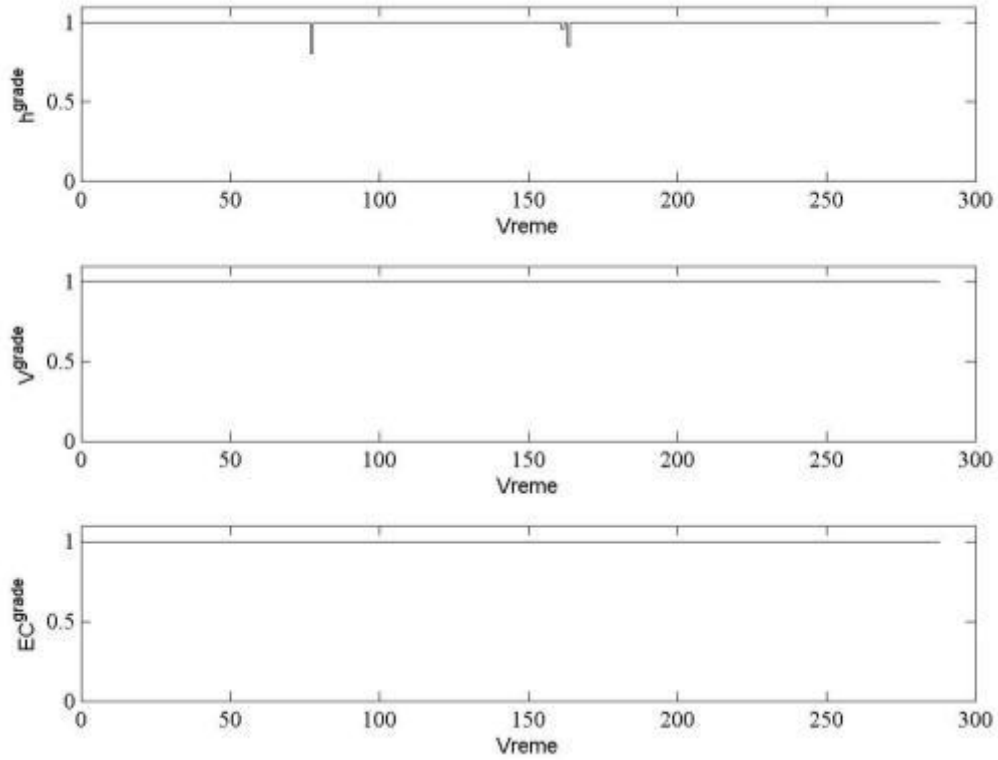


Slika 5.65: Verovatnoće merenih vrednosti (ukupne ocene)

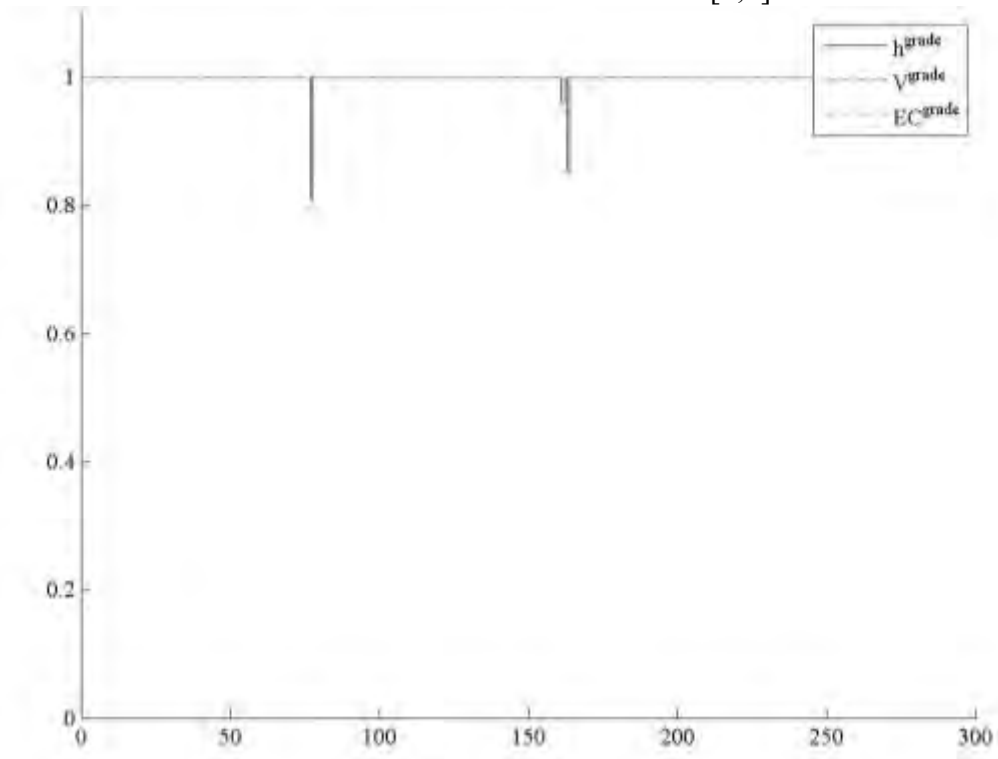


Slika 5.66: Verovatnoće na jednom dijagramu

Normirane verovatnoće pružaju mogućnost da se napravi granica izmeĎu regularnih i neregularnih podataka. Ta granica zavisi, pre svega, od upotrebe podatka, tj. od greške koju bi na rezultat upotrebe preneo neregularni podatak. Na slikama 5.67 i 5.68 uočavaju se tri podatka za koje su normirane verovatnoće manje od maksimalnih. Ta tri podatka se, po istaknutim anomalijama, mogu uočiti i na slici 5.68 na kojoj su podaci i prikazani.

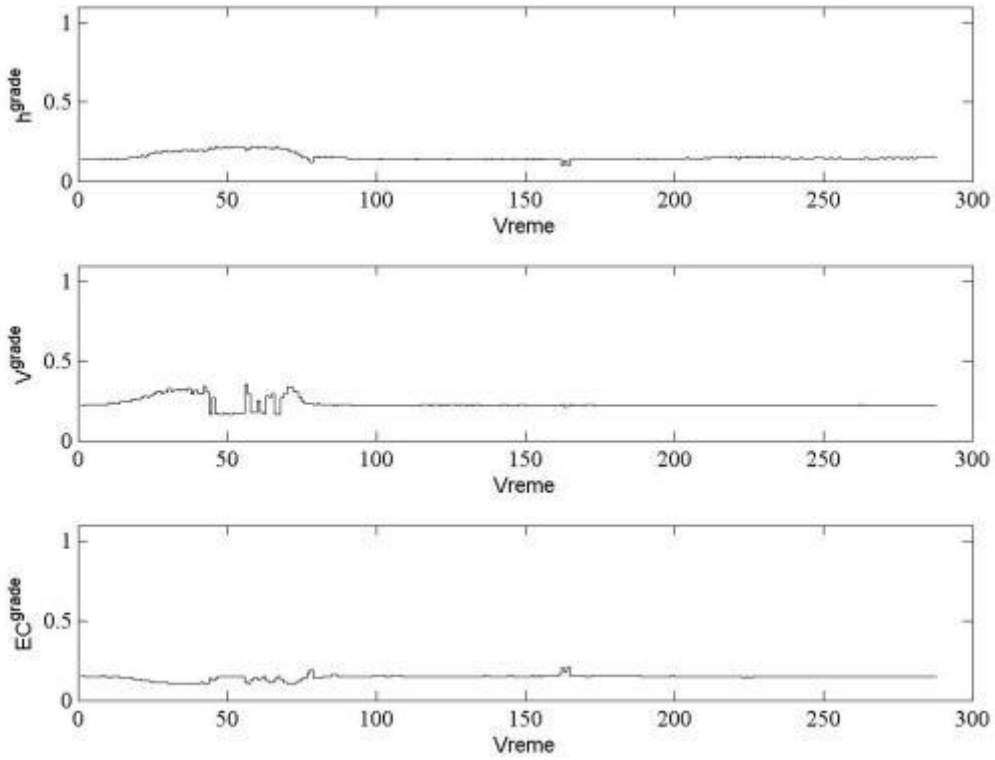


Slika 5.67: Normirane ocene u intervalu [0,1]

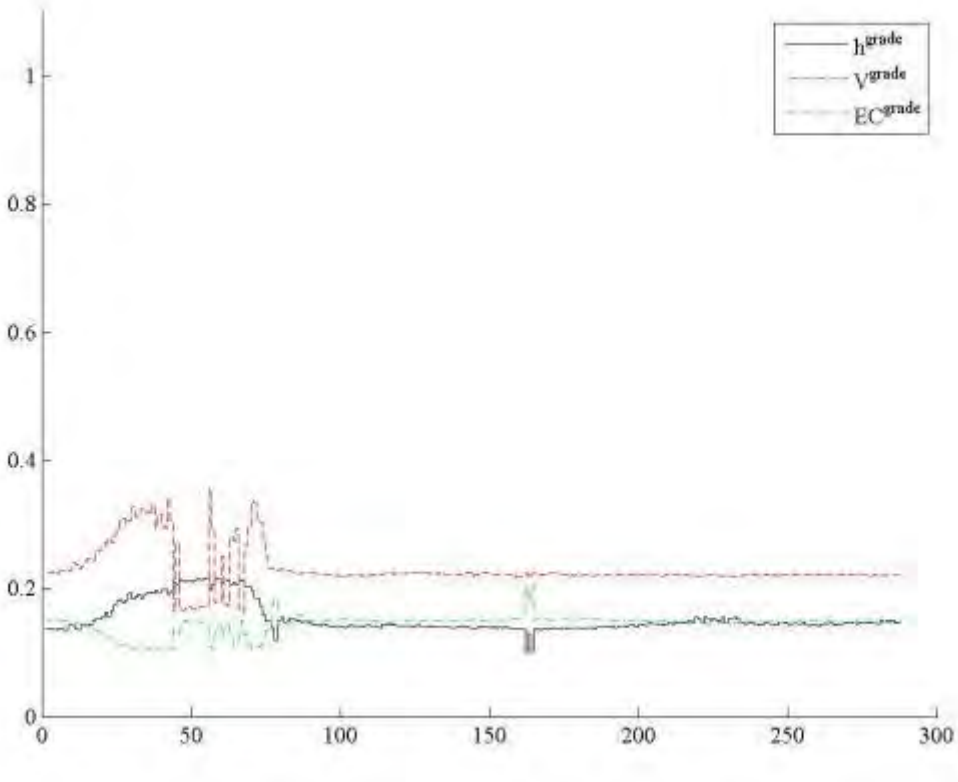


Slika 5.68: Ocene normirane na jednom dijagramu

Mera neizvesnosti relacija koje su korišćene za potrebe vrednovanja podataka predstavljena je na slikama 5.69 i 5.70. Može se uočiti da je neizvesnost relativno visoka (manji broj označava veću neizvesnost), a, kao što se i očekivalo, najveća je za relacije vezane za elektroprovodnost.

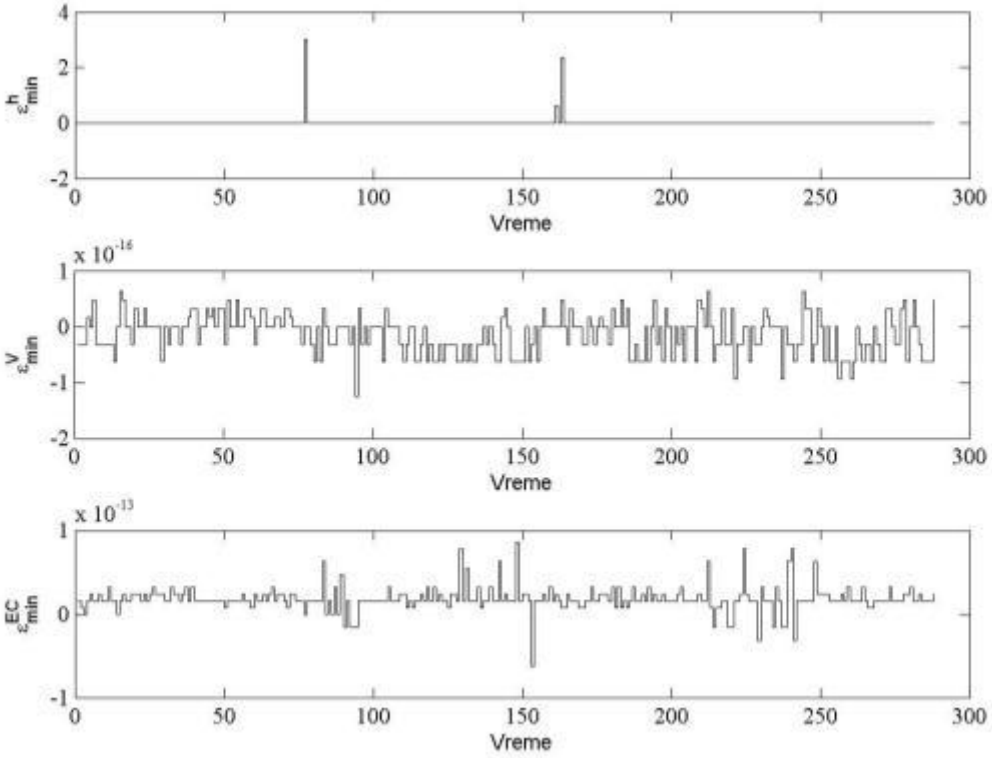


Slika 5.69: Mera neizvesnosti sistema za vrednovanje



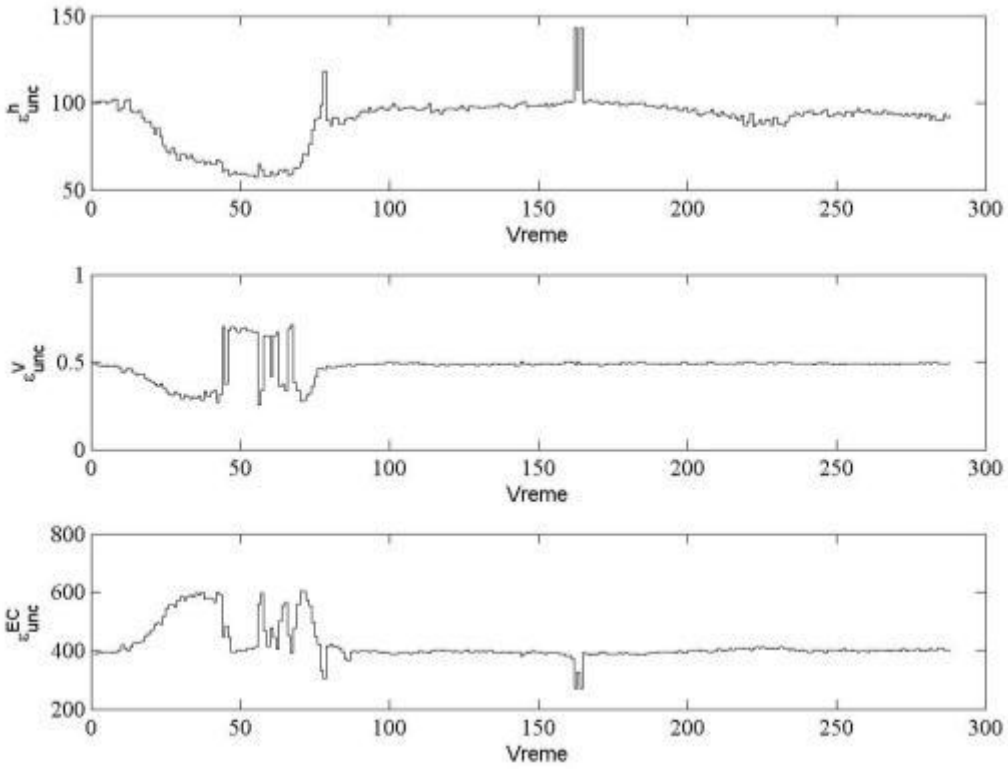
Slika 5.70: Mera neizvesnosti sistema za vrednovanje na jednom dijagramu

Na slikama 5.71 i 5.72 prikazane su minimalna greška i neizvesnost greške merenih veličina. Minimalna greška kod brzine i elektroprovodnosti izuzetno je mala i kreće se oko nule. Ovakvi rezultati posledica su greške zaokruživanja pri računanju.



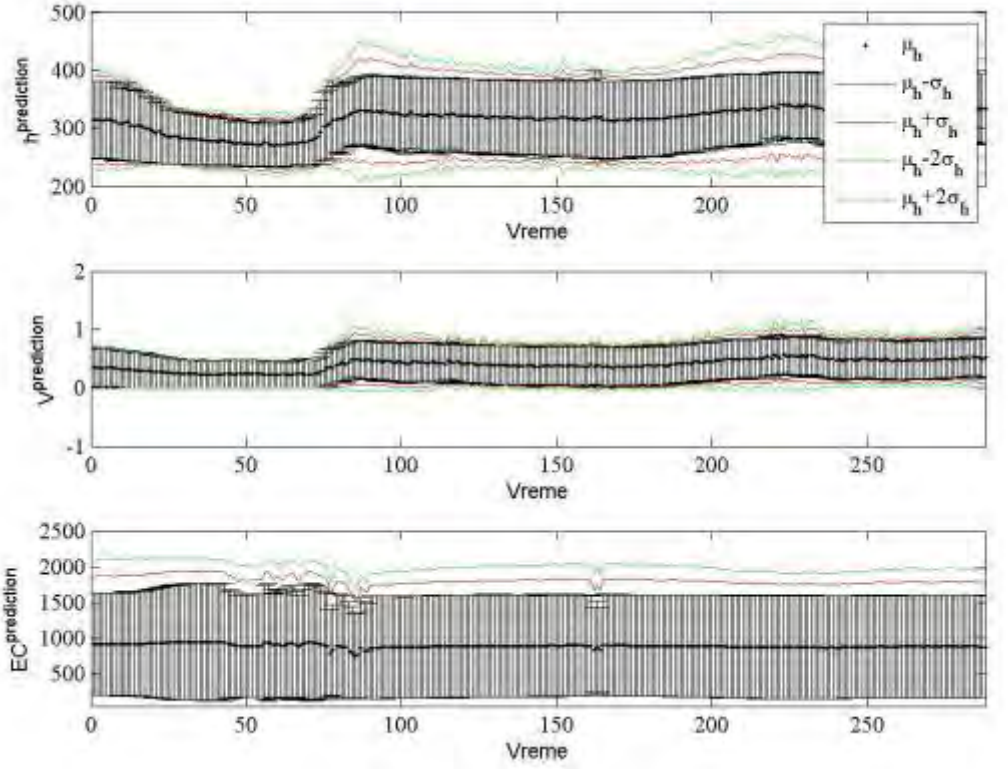
Slika 5.71: Minimalna greška

UporeĎivanjem minimalnih grešaka podataka na slici 5.71 i grešaka koje se mogu vizuelno uočiti na slici 5.64 primećuje se da veličina minimalne greške odgovara veličini pika bez obzira da li se on nalazi na uzlaznoj grani hidrograma ili ne. Ovaj efekat je izuzetno važan jer se kod mnogih metoda ili ne detektuju upravo greške koje se javljaju na delovima vremenske serije sa izraženim trendom ili se sa greškama detektuju i regularni podaci.



Slika 5.72: Neizvesnost greške

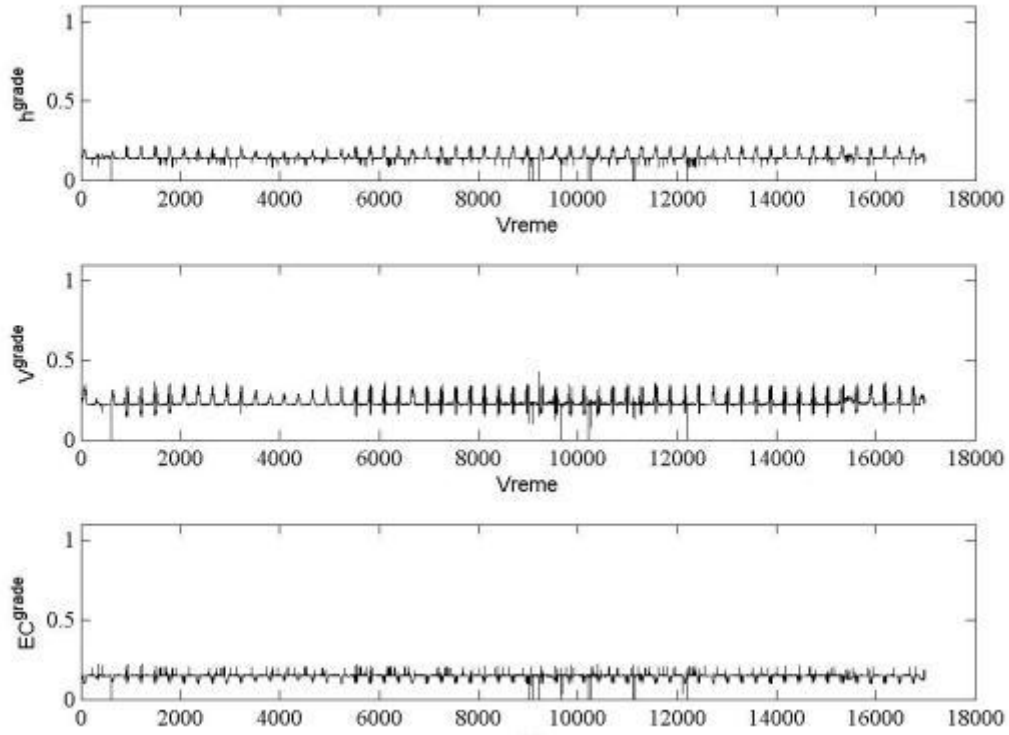
Matematička očekivanja i varijanse intervala u kojima se očekuju tačne vrednosti merenih veličina prikazane su na slici 5.73.



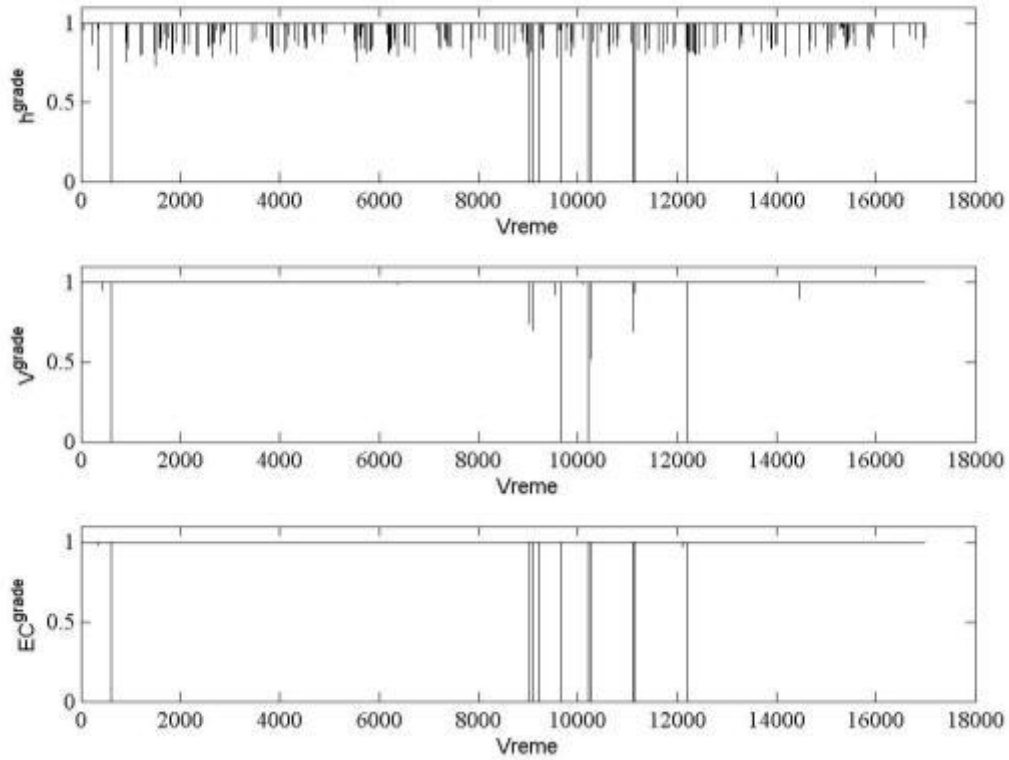
Slika 5.73: Intervali u kojima se pretpostavlja da se nalazi tačna vrednost

Cela raspoloživa serija podataka

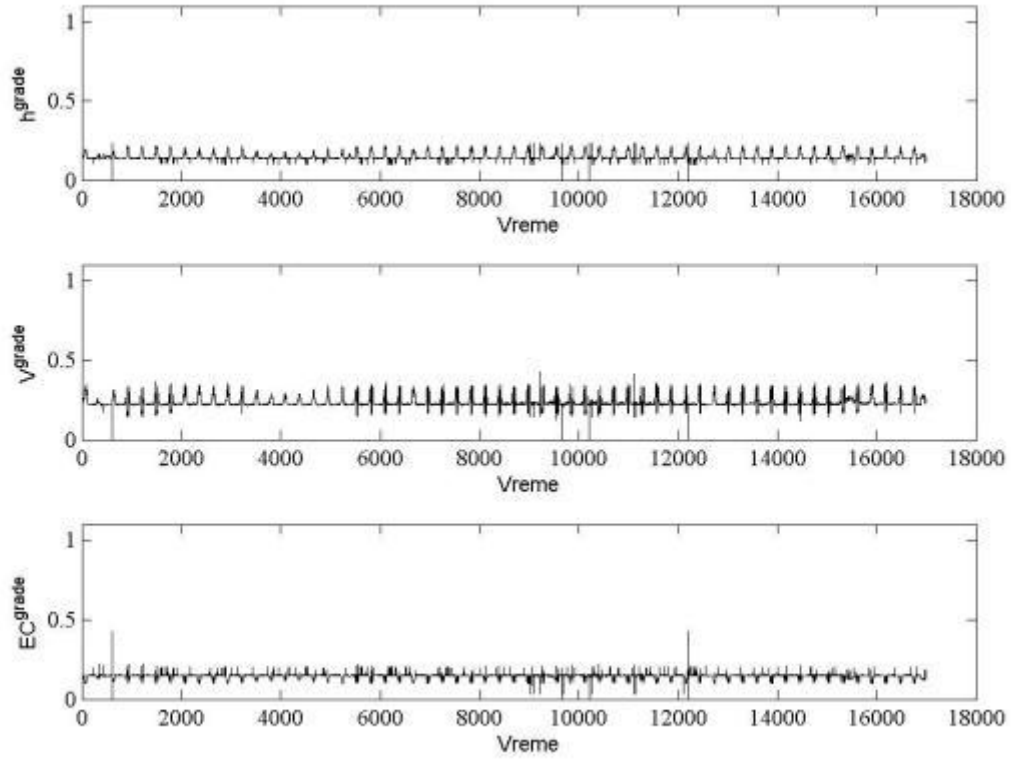
Cela raspoloživa istorijska serija podataka obuhvata prva dva meseca 2007. godine i to od 1/1/2007 00:00:00 do 28/2/2007 59:55:00. Rezultati su prikazani na slikama 5.74-5.78.



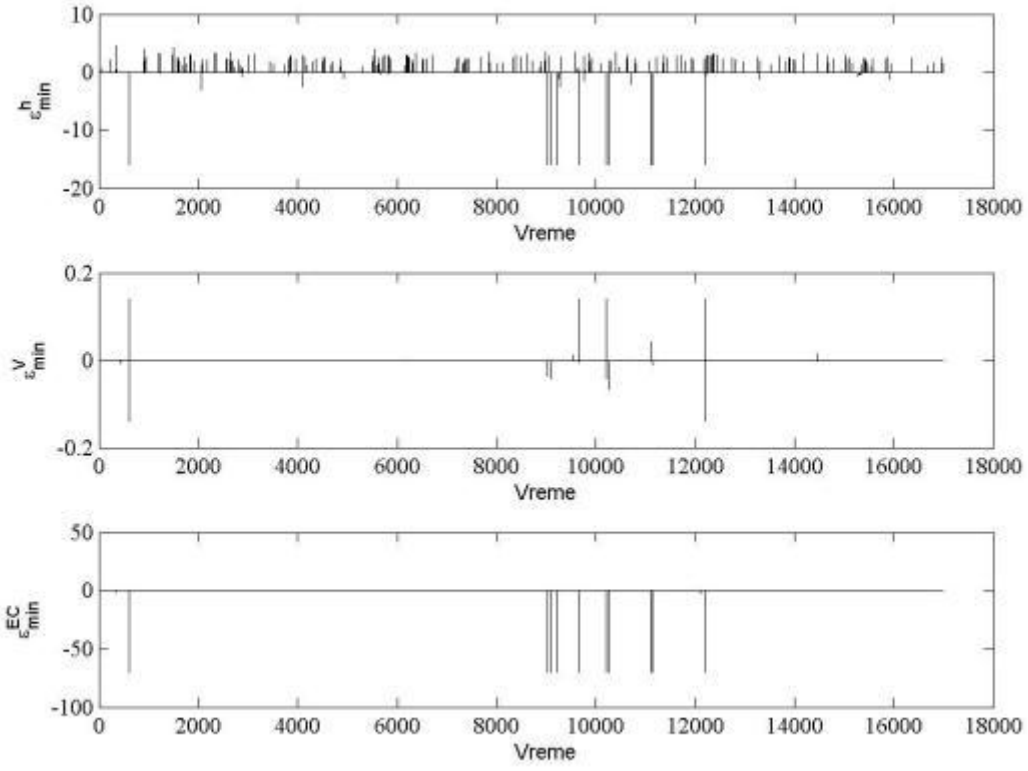
Slika 5.74: Ukupne verovatnoće podataka



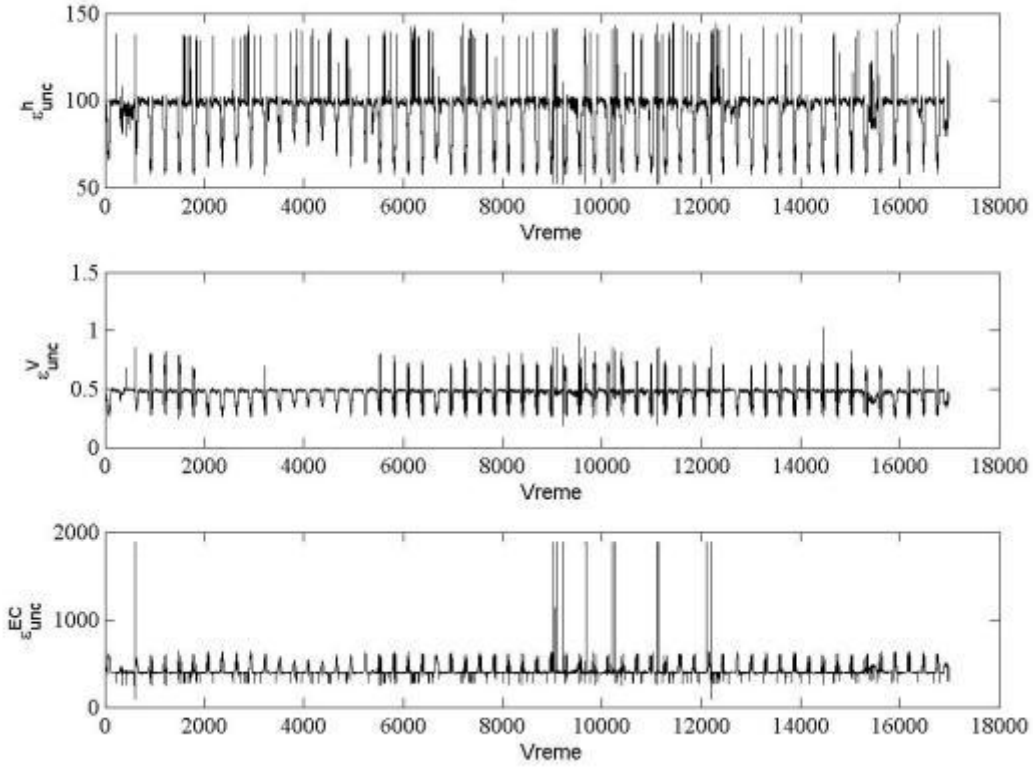
Slika 5.75: Normirane verovatnoće vrednovanja



Slika 5.76: Neizvesnost metoda vrednovanja



Slika 5.77: Minimalne greške

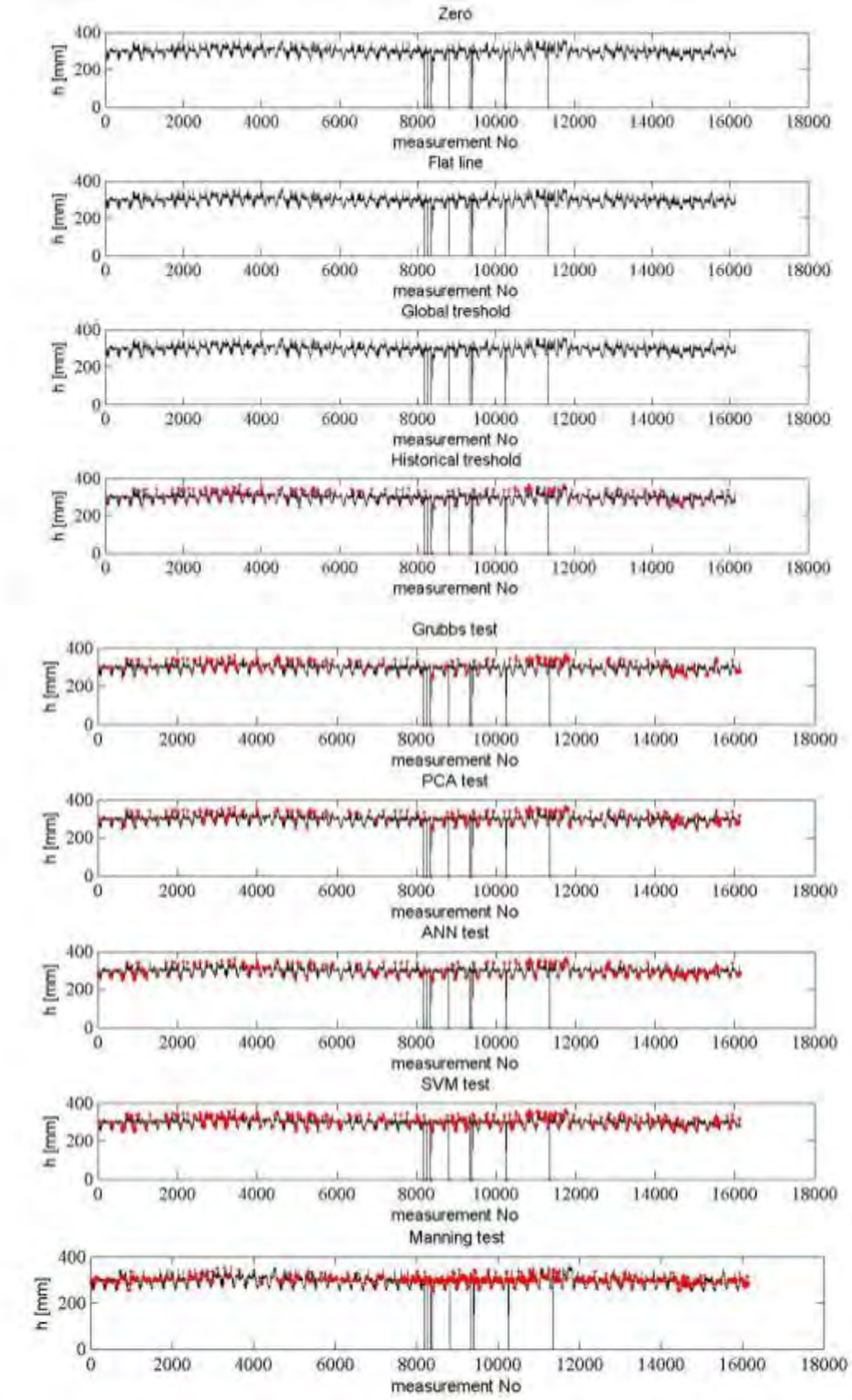


Slika 5.78: Neizvesnost greške

Rezultati vrednovanja pomoću algoritma predloženog u ovoj disertaciji upoređeni su sa rezultatima prikazanim u radovima [11] i [15], gde je sprovedeno vrednovanje dubine iste vremenske serije pomoću devet metoda vrednovanja. Rezultati iz navedenih radova iskorišćeni su za procenukvaliteta predloženog algoritma pomoću izraza 4.9, 4.10 i 4.11.

5.3.5 Ocena rezultata realnog primera merenja u kanalizaciji

U radu [11] razvijeno je devet metoda za vrednovanje podataka koje su testirane na vremenskoj seriji merene dubine u kolektoru. Dobijeni rezultati prikazani su na slici 5.79 i u tabeli 5.15.



Slika 5.79: Rezultati detekcije anomalija uz pomoć devet metoda (pozajmljeno iz rada [11])

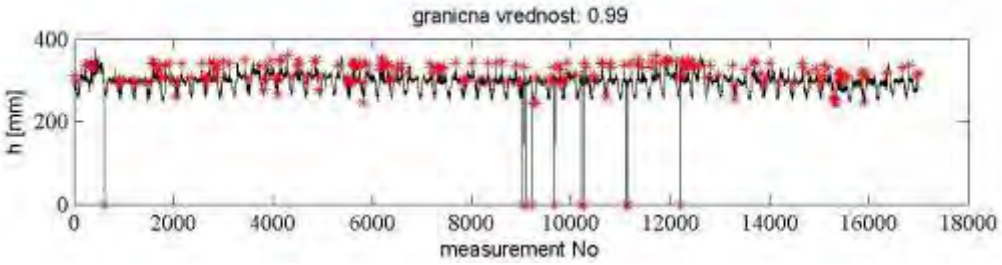
Tabela 5.15: Ocene kvaliteta devet metoda za vrednovanje uporeĎene sa rezultatima iskustvenog vrednovanja na osnovu vizuelne inspekcije (Pozajmljeno iz rada [11])

	No anomalies	No detected	No missed	No false	p	$p_{false\ tolerant}$	$p_{false\ sensitive}$
M_1	244	24	220	0	0.052	0.05	0.10
M_2	244	0	244	28	0.000	0.00	0.00
M_3	244	0	244	0	0.000	0.00	0.00
M_4	244	202	42	220	0.399	0.71	0.44
M_5	244	187	57	334	0.294	0.62	0.32
M_6	244	224	20	359	0.360	0.85	0.37
M_7	244	109	135	896	0.085	0.29	0.10
M_8	244	237	11	483	0.321	0.93	0.33
M_9	244	71	173	2725	0.023	0.17	0.02

Na osnovu prikazanih rezultata vidi se da je metoda M_4 dala najbolje rezultate kada je upoređena sa ručnom vizuelnom validacijom. Po izrazu 4.9, verovatnoća uspešnosti metode iznosi p=0.399. U radu [15] osrednjavanjem rezultata nekoliko metoda poboljšana je ocena procedure vrednovanja:

$$p = \frac{N_{registered}}{N_{anomalies} + N_{missed} + N_{registered\ nonanomalies}} = \frac{195}{244 + 49 + 58} = 0.56 .$$

Rezultati predložene metode pokazuju neuporedivo bolje rezultate, što se može videti na slici 5.80 i u tabeli 5.16. Za ocenu da li je podatak regularan ili ne (da li ima anomaliju) iskorišćena je informacija o normiranoj verovatnoći sa graničnom vrednosti $p^{norm} \leq 0.99$:



Slika 5.80: Anomalije u podacima o dubini vode u kolektoru

Tabela 5.16: Rezultati predložene metode

No anomalies	*No detected*	*No missed*	*No false*	*p*	$P_{false\ tolerant}$	$P_{false\ sensitive}$
244	219	25	73	0.64	0.81	0.69

Nažalost, nijedna od navedenih ocena ne reflektuje prirodu metoda koje se u procesu vrednovanja koriste, već se kao rezultat upotrebljava samo konačna vrednost pouzdanosti. Priroda metoda bi se mogla iskazati osetljivošću metoda, kvalitetom kalibracije ili nekom drugom osobinom. Ukoliko se rezultat vrednovanja podataka prikaže pomoću skupa rezultata korišćenih metoda u sistemu, može se doći do informacija i o navedenim osobinama metoda koje se pri vrednovanju koriste. Takođe veliki uticaj kod izračunavanja navedenih ocena imaju sami podaci, tj. rezultati neke druge metode, čime se umanjuje opštost ocene. Treba imati na umu da ni referentne ocene uglavnom nisu apsolutno tačne i da se često zasnivaju na subjektivnom osećaju, kao u slučaju iskustvenog vrednovanja vizuelnom inspekcijom.

5.3.6 Zaključak realnog primera merenja u kanalizaciji

U navedenom primeru prikazani su rezultati primene predloženog algoritma na realnim podacima prikupljenim na odabranom ispustu Beogradske kanalizacije. Relacije između merenih veličina formirane su na osnovu istorijskih podataka, a sam algoritam testiran je na dve vremenske serije – danu bez padavina i danu sa padavinama. Sistem je pokazao izuzetne rezultate u detekciji grešaka u podacima kada je upoređen sa vizuelnom inspekcijom.

Kvalitet vrednovanja predloženim algoritmom svodi se na kvalitet (greške i neodređenost) relacija koje se između podataka mogu uspostaviti i formirati. Ukoliko se želi bolji sistem za vrednovanje, uz manje neizvesnosti, može se delovati u pravcu formiranja relacija između podataka sa manje grešaka i neizvesnosti, kao i pribavljanju dodatnih merenih podataka sa manje grešaka i neizvesnosti za potrebe kalibracije relacija između podataka.

6. Zaključak

Vrednovanje podatka predstavlja proces ocene kvaliteta tog podatka i detekcije grešaka koje mogu da daju negativne efekte pri njegovoj kasnijoj upotrebi. Proces vrednovanja za rezultat ima uvid u pouzdanost podatka, koji prati i informacija o kvalitetu samog vrednovanja. Predložena metodologija predviđa formiranje sistema za vrednovanje podataka u vidu niza relacija pomoću kojih je vrednovani podatak moguće izračunati i statističkog okvira za tumačenje rezultata upoređivanja merenih i izračunatih vrednosti.

Mereni i izračunati podaci koriste se u obliku intervala, čime je obuhvaćena i njihova neodređenost. (na sličan način bi se mogli predstaviti i u obliku rasplinutog skupa ili statističke raspodele). Relacije između podataka se mogu formirati u bilo kom obliku, s tim što rezultat relacija mora biti u obliku intervala.

Pored neodređenosti ulaznih vrednosti, u predloženom algoritmu vodi se računa i o neodređenosti samog modela, tj. o neizvesnosti koja potiče od konceptualne predstave matematičkog modela i neizvesnosti podataka koji se koriste za kalibraciju. U procesu pretprocesiranja neophodno je iz serije istorijskih podataka koja se koristi za kalibraciju i podešavanje relacija otkloniti greške.

Izbor relacija i metoda vezuje se za najbolju modelarsku praksu. Ukoliko ekspert ima mogućnost da bira između različitih relacija koje opisuju vezu između merenih podataka, potrebno je odabrati onu koja pruža najmanju neodređenost. Time se i izbor metode svodi na izbor one koja ima najmanju neodređenost.

S obzirom na to da su uvedene neizvesnosti relacija, greške u izračunatim vrednostima vrednovanih veličina potiču isključivo od ulaznih veličina, koje se takođe vrednuju. Statističkom obradom izmerenih i izračunatih vrednosti izračunavaju se težinski koeficijenti uz svaku izračunatu vrednost, koji u sebi nose informaciju o postojanju potencijalne greške u izmerenoj vrednosti, koja je ujedno i ulazna veličina u relaciji između podataka.

Nakon izračunavanja težinskih koeficijenata moguće je odrediti ukupne verovatnoće izmerenih vrednosti koje u sebi sadrže i informaciju o relacijama između vrednovanih podataka. Te verovatnoće mogu se upotrebiti kao ocene pouzdanosti merenih vrednosti u skladu sa neizvesnošću samog sistema vrednovanja. Takođe, moguće je izračunati matematička očekivanja i varijanse intervala u kojima se očekuje tačna vrednost merene veličine.

Detektovano je da se kod ukupnih verovatnoća podataka koji se vrednuju teško može povući jasna granica koja bi razdvajala regularne podatke od neregularnih. Razlog za to je upravo neizvesnost relacija između podataka. Da bi se napravila granica između regularnih i neregularnih podataka, verovatnoće su normirane tako da se iz normiranih verovatnoća (ocena) gubi informacija o neizvesnosti samog sistema za vrednovanje. Ocena neizvesnosti samog sistema za vrednovanje se dobija preko maksimalnih uslovnih verovatnoća merenih vrednosti u odnosu na izračunate.

Metodologija je u disertaciji primenjena na tri primera, jednom hipotetičkom opštem primeru, drugom hipotetičkom hidrotehničkom primeru, i na trećem primeru koji je baziran na realnim podacima, a rezultati su upoređeni sa rezultatima iz literature. Pokazano je da algoritam daje pouzdane rezultate i da se mogu izračunati potencijalne greške u podacima. Testiranje je obavljeno u MatLab programskom okruženju.

Izdvajaju se dva pravca daljeg istraživanja ove tematike. Prvi se odnosi na proširenje algoritma na druge dve matematičke formulacije neodreĎenih veličina: rasplinute skupove i statističke raspodele. Drugi se odnosi na razvoj automatskih metoda za izbor i kalibraciju metoda za predikciju.

7. Literatura

1. Abbott, M. B.: Hydroinformatics: Information Technology and the Aquatic Environment, 1991, Avebury, ISBN: 978-1856288323
2. Aha, D. W. and R. B. Bankert: Feature Selection for Case-Based Classification of Cloud Types: An Empirical Comparison, 1994, Proceedings of the AAAI-94 Workshop on Case-Based Reasoning
3. Allan, J., J. Carbonell, G. Doddington, J. Yamron and Y. Yang: Topic Detection and Tracking Pilot Study, Final Report, 1998, Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop
4. AWS Scientific, Inc.: WIND RESOURCE ASSESSMENT HANDBOOK, Fundamentals for Conducting a Successful Monitoring Program, 1997
5. Babović, V.: Hydroinformatics: Emergence, Evolution, Intelligence (IHE thesis series), 1996, Taylor & Francis, 1st edition, ISBN-13: 978-9054104049
6. Bakshi, B. R. and G. Stephanopoulos: Wave-net: A multiresolution, hierarchical neural network with localized learning, 1993, AICHE J. 39(1), pp: 57-81
7. Bakshi, B. R: Multiscale PCA with application to multivariate statistical process monitoring, 1998, AIChE J. 44, pp: 1596-1610
8. Barnett, V. and T. Lewis: Outliers in Statistical Data, 1994, 3rd edn. John Wiley & Sons:
9. Becraft W. and P. Lee: An integrated neural Network/Expert system approach for fault diagnosis, 1993, Computers and Chem. Eng. 17(10), pp:1001-1014
10. Boukhris, A., S. Giuliani and G. Mourot: Rainfall-runoff multi-modelling for sensor fault diagnosis, 2001, Control Engineering Practice 9, pp: 659-671
11. Branisavljević, N., Z. Kapelan and D. Prodanović: Improved real-time data anomaly detection using context classification, 2011, Journal of Hydroinformatics, Vol 13, No 3, pp: 307–323, IWA Publishing
12. Branisavljević, N., D. Prodanović, M. Arsić, Z. Simić, J. Borota: Hydro-Meteorological Data Quality Assurance and Improvement, 2009, Journal of the Serbian Society for Computational Mechanics, Vol. 3, No. 1, pp: 228-249
13. Branisavljević, N., D. Prodanović: Razvoj funkcija za merenje veličina u hidrotehnici u MatLab okruženju, 2003, 13. Savetovanje JDHI, Soko Banja
14. Branisavljević, N., Z. Kapelan and D. Prodanovć: Bayesian-Based Detection of Measurement Anomalies in Environmental Data Series, 2009, The 8th International Conference onHydroinformatics, HIC 2009, Concepción, Chile
15. Branisavljević, N., Z. Kapelan and D. Prodanović: Automated Validation Of Real-Time Sewer Monitoring Data, 2010, 9th International Conference on Hydroinformatics, HIC 2010, Tianjin, China
16. Branisavljević, N.: Propagacija neodreĎenosti kod zatvorenih hidrotehničkih modela, 2008, Magistarska teza, GraĎevinski fakultet u Beogradu
17. Brodley, C. E. And M. A. Friedl: Identifying and Eliminating Mislabeled Training Instances, 1996, Proceedings of the 13th National Conference on Artificial Intelligence, pp: 799– 805, AAAI Press.
18. Byers, S. and A. E. Raftery: Nearest Neighbor Clutter Removal for Estimating Features in Spatial Point Processes, 1998, Journal of the American Statistical Association 93(442), pp: 577–584
19. Chandola, V., A. Banerjee and V. Kumar: Anomaly Detection: A Survey, 2009, ACM Computing Surveys, Vol. 41(3)
20. Chen, J. and R. J. Patton: Robust Model-Based Fault Diagnosis for Dynamic Systems, 1999, Kluwer Academic Publishers, Boston/Dordrecht/London
21. Cheung, J. T. and G. Stephanopoulos: Representation of process trends part I: a formal representation framework, 1990, Computers and Chem. Eng. 14(4-5), pp: 495-510

22. Clarke, D. W. and P. M. A. Fraher: Model-based validation of a DOx sensor, 1996, Control Engineering Practice, Volume 4, Issue 9, pp: 1313-1320
23. Conejo, R., E. Guzman, and J. Perez-de-la-Cruz: Knowledge-Based Validation For Hydrological Information Systems, 2007, Applied Artificial Intelligence, 21, pp: 803–830, Copyright 2007 Taylor & Francis Group
24. Crumbling, D. M.: In Search Of Representativeness: Evolving The Environmental Data Quality Model, 2001, Quality Assurance, 9, pp: 179–190
25. Datta, P. and D. Kibler: Learning prototypical concept descriptions,1995, Proceedings of the 12th International Conference on Machine Learning, pp: 158–166, Morgan Kaufmann.
26. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, 1977, Journal of the Royal Statistical Society, Series B (Methodological) 39 (1), pp: 1–38
27. Dereszynski and Dietterich: Probabilistic Models for Anomaly Detection in Remote Sensor Data Streams, 2007, Proceedings of the Twentysecond Conference on Uncertainty in AI UAI
28. Ding, X. and P. M. Frank: Fault detection via factorization approach, 1990, Syst. Contr. Lett. 14(5), 431—436
29. ĐorĎević, S.: Matematički model oticanja sa urbanih slivova interaktivnim tečenjem po površini i kroz mrežu podzemnih kolektora, 2002, doktorska disertacija, GraĎevinski fakultet Univerziteta u Beogradu)
30. Einfalt, T., B. Maul-Kiitte and S. Spies: A Radar Data Quality Control Scheme Used in Hydrology, 2000, Phys. f&m. Earfh (B), Vol. 25, No. 10-12, pp: 1141-l 146, Published by Elsevier Science Ltd
31. Enfinger, K. and P. Stevens: Scattergraph Principles and Practice Characterization of Sanitary Sewer and Combined Sewer Overflows, ADS Environmental Services 4940 Research Drive, Huntsville, Alabama 35805, www.adsenv.com/scattergraph.
32. Eskin, E.: Anomaly detection over noisy data using learned probability distributions, 2000, Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc.
33. Faloutsos, C., F. Korn, A. Labrinidis, Y. Kotidis, A. Kaplunovich and D. Perkovic: Quantifiable Data Mining Using Principal Component Analysis, 1997, Technical Report CS-TR-3754, Institute for Systems Research, University of Maryland, College Park
34. Faramand, T.: Automated Validation and Grading of Aquatic Time Series Using a Probabilistic Parity Space Method, 2006, NWQMC, National Monitoring Conference
35. Figliola, R. S., D. E. Beasley: Theory and Design for Mechanical Measurements, 2005, Wiley, 4 edition, ISBN-13: 978-0471445937
36. Fletcher, T. D. and A. Deletić (Editors): Data Requirements For Integrated Urban Water Management, 2007 Taylor & Francis Group, London, UK
37. Frank, P.: Fault Diagnosis in Dynamic Systems Using Analytical and Knowledge-based Redundancy A Survey and Some New Results, 1990, Automatica, Vol. 26, No. 3, pp. 459-474
38. Friendly, M.: Milestones in the history of thematic cartography, statistical graphics, and data visualization, 2008
39. Fuller, W. A.: Measurement Error Models, 1987, John Wiley & Sons, ISBN 0-471-86187- 1)
40. Fussell, J. B.: Fault tree analysis - state of the art,1974, IEEE Trans. on Reliability 23(1), pp:51-53
41. Gertler, J. J.: Analytical Redundancy Methods in Fault Detection and Isolation, 1991, Survey and Synthesis IFAC Safeprocess, Volume: 1, pp: 9-21
42. Golz, Einfalt, Gabella and Germann: Quality control algorithms for rainfall measurements, 2005, Atmospheric Research 77, pp: 247–255
43. Grubbs, F. E.: Procedures for Detecting Outlying Observations in Samples, 1969, Technometrics 11: 1–21

44. Guillet, Fabrice; Hamilton, Howard J. (Eds.): Quality Measures in Data Mining, Series: Studies in Computational Intelligence, 2007, ISBN: 978-3-540-44911-9
45. Hajdin, G.: Mehanika Fluida, 1992, GraĐevinski fakultet, Beograd
46. Hamelin, F. and D. Sauter: Robust residual generation for fdi in uncertain dynamic systems, 1995, Proc. 34th IEEE Conf. on Decision & Contr., New Orleans, USA
47. Hamioud, F., C. Joannis and J. Ragot: Statistical modeling for validation of hydrometric data issued from the sewer networks, 2004, 19th European Junior Scientists Workshop, Lyon, France, March 11-14
48. Han, C., R. Shih, and L. Lee: Quantifying signed directed graphs with the fuzzy set for fault diagnosis resolution improvement, 1994, Indust. and Eng. Chemistry Research 33(8), pp:1943-1954
49. Henley, E.J.: Application of expert systems to fault diagnosis, 1984, AIChE Annual Meeting, San Francisco, CA
50. Hickinbotham, S. and J. Austin: Novelty Detection in Airframe Strain Data, 2000, Proceedings of 15th International Conference on Pattern Recognition, Barcelona, pp: 536–539
51. Himberg, J., A. Jussi, E. Alhoniemi, J. Vesanto and O. Simula: The Self- Organizing Map as a Tool in Knowledge Engineering, 2001, Pattern Recognition in Soft Computing Paradigm, pp: 38–65. Soft Computing. World Scientific Publishing
52. Hodge, V. J. and J. Austin: A Survey of Outlier Detection Methodologies, 2004, Artificial Intelligence Review, 22, Kluwer Academic Publishers, pp. 85–126
53. Hoeting, J. A., D. Madigan, A. E. Raftery and C. T. Volinsky: Bayesian Model Averaging: A Tutorial, 1999, Statistical Science, Vol. 14, No. 4, pp: 382–417
54. Hudson, H., D. Mcmillan and C. Pearson: Quality assurance in hydrological measurement, 1999, Hydrological Sciences-Journal-des Sciences Hydrologiques, 44(5)
55. Iri, M., K. Aoki, E. OShima, and H. Matsuyama: An algorithm for diagnosis of system failures in the chemical process, 1979, Computers and Chem. Eng. 3(1-4), 489—493
56. Isermann, R: Model-based fault-detection and diagnosis – status and applications, 2005, Annual Reviews in Control 29, pp: 71–85
57. Isermann, R.: Fault-diagnosis systems: an introduction from fault detection to fault tolerance, 2006, Berlin, Springer
58. Janusz, M. and V. Venkatasubramanian: Automatic generation of qualitative description of Process Trends for fault detection and diagnosis, 1991, Eng. Applications of Artificial Intelligence 4(5), pp: 329-339
59. Japkowicz, N., C. Myers and M. A. Gluck: A Novelty Detection Approach to Classification, 1995, Proceedings of the 14th International Conference on Artificial Intelligence (IJCAI-95), pp: 518–523
60. John, G. H.: Robust Decision Trees: Removing Outliers from Databases, 1995, Proceedings of the First International Conference on Knowledge Discovery and Data Mining, pp: 174–179. Menlo Park, CA: AAAI Press.
61. Jørgensen, H.K., S. Rosenørn, H. Madsen and P.S. Mikkelsen**:** Quality control of rain data used for urban runoff system, 1998, Water Science and Technology, pp: 113-120
62. Karr, A. F., A. P. Sanil, D. L. Banks: Data quality: A statistical perspective Statistical Methodology, 2006, Volume: 3, Pages: 137-173
63. Katipamula, S., M. R. Brambley: Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems - A Review, Part I, 2005, Hvac&R Research, Volume 11, Number 1
64. Katipamula, S., M. R. Brambley: Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems— A Review, Part II, 2005, Hvac&R Research, Volume 11, Number 2
65. Kester, Walt, (Ed.): *The Data Conversion Handbook*, 2005, Elsevier: Newnes, ISBN 0-7506-7841-0
66. Kinosita, T.: Quality Assurance in Hydrological Data, 2003, HW03, Abstracts week B, p.B.349, IUGG 2003. Sapporo

67. Knorr, E. M. and R. T. Ng: Algorithms for Mining Distance-Based Outliers in Large Datasets, 1998, Proceedings of the VLDB Conference, pp: 392–403, New York, USA.
68. Kohonen, T.: Self-Organizing Maps. Springer, 1995, Berlin, Heidelberg.
69. Kresta, J. V., J. F. MacGregor, and T. E. Marlin: Multivariate statistical monitoring of process operating performance, 1991, Can. J. of Chem. Eng. 69(1), 35
70. Laurikkala, J., M. Juhola and E. Kentala: Informal Identification of Outliers in Medical Data, 2000, Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology IDAMAP-2000 Berlin, 22 August, Organized as a workshop of the 14th European Conference on Artificial Intelligence ECAI-2000, pp: 20-24
71. Leeuwenberg, E. L. J.: Structural information of visual patterns: an efficient coding system in perception, 1968, The Hague: Mouton
72. Lempio, G, C. Podlasly, T. Einfalt: NIKLAS - Automatical quality control of time series data, 2010, Erad 2010 - The Sixth European Conference On Radar In Meteorology And Hydrology, Sibiu, Romania
73. Leonard, J. A. and M. A. Kramer: Diagnosing dynamic faults using modular neural nets, 1993, IEEE Expert 8(2), pp: 44-53
74. Liu, J. P. and C. S. Weng: Detection of outlying data in bioavailability/bioequivalence studies, 1991, Statistics Medicine 10
75. Maksimović, Č.: Merenja u hidrotehnici, 1993, GraĐevinski fakultet, Beograd
76. Metzger, W.: Laws of Seeing, 1936, The MIT Press, Cambridge, Massachusetts, London, England
77. Milne, R.: Strategies for diagnosis, 1987, IEEE Trans. on Syst., Man and Cyber. 17(3), 333-339
78. Moran, A.W., P.G. O'Reilly and G.W. Irwin: Probability estimation algorithms for self-validating sensors, 2001. Control Engineering Practice, Volume 9, Issue 4, pp: 425-438
79. Mourad, M. and J.-L. Bertrand-Krajewski: A method for automatic validation of long time series of data in urban hydrology, 2002, Water Science and Technology, Vol 45, No 4–5, pp: 263–270, IWA Publishing
80. Mulacova, J.: Failure Detection Expert Software, 2007, Luleå University of Technology Master Thesis, Continuation Courses Space Science
81. Musa, M., E. Grüter, M. Abbt, C. Häberli, E. Häller, U. Küng, T. Konzelmann, R. Dössegger: Quality Control Tools for Meteorological Data in the MeteoSwiss Data Warehouse System, 2003, Proc. ICAM/MAP 2003., Brig, Switzerland, 19.-23. May
82. Nairac, A., N. Townsend, Carr, S. King, P. Cowley and L. Tarassenko: A System for the Analysis of Jet System Vibration Data, 1999, Integrated ComputerAided Engineering 6(1), pp: 53–65
83. Niida, K.: Expert system experiments in processing engineering,1985, Inst. Of Chem. Eng. Symposium Series, pp: 529-583
84. Nomikos, P. and J. MacGregor: Monitoring batch processes using multiway principal component analysis, 1994, AICHE J. 40(8), pp: 1361-1375
85. Organtinia, G., A. Biagiob and M. Maggiorib: A universal framework for online data validation, 2004, Nuclear Instruments and Methods in Physics Research A 534 (2004), pp: 120–124
86. Parra, L., G. Deco and S. Miesbach: Statistical Independence and Novelty Detection with Information Preserving Nonlinear Maps, 1996, Neural Computation 8 (2), pp: 260–269
87. Patcha, J. Park: An overview of anomaly detection techniques: Existing solutions and latest technological trends, 2007, Computer Networks, Vol. 51, No. 12., pp. 3448-3470
88. Patton, R. J. and J. Chen: A review of parity space approaches to fault diagnosis, 1991, Proc. IFAC/IMACS Sympo. SAFEPROCESS'91, volume 1, pages 239—255, Baden-Baden
89. Piatyszek, E., P. Voignier, D. Graillot: Fault detection on a sewer network by a combination of a Kalman filter and a binary sequential probability ratio test, 2000, Journal of Hydrology 230 (2000) , pp: 258–268

90. Piatyszek, Joannis and Aumond: Using typical daily flow patterns and dry-weather scenarios for screening flow rate measurements in sewers, 2002, Water Science and Technology, Vol 45, No 7, pp: 75–82, IWA Publishing
91. Prodanović, D.: Merenja u hidrotehnici, kurs na GraĎevinskom fakultetu u Beogradu
92. Quantrille, T. E. and Y. A. Liu: Artificial Intelligence in Chemical Engineering, 1991, Academic Press, San Diego, LA
93. Rabinovich, S. G: Measurement Errors and Uncertainties, 2005, Springer, ISBN-13: 978-0387-25368-9
94. Ramaswamy, S., R. Rastogi and K. Shim: Efficient Algorithms for Mining Outliers from Large Data Sets, 2000, Proceedings of the ACM SIGMOD Conference on Management of Data, Dallas, TX, pp: 427–438
95. Roberts, S. and L. Tarassenko: A Probabilistic Resource Allocating Network for Novelty Detection, 1995, Neural Computation 6: pp: 270–284.
96. Rosner, B.: Percentage points for a generalized many-outlier procedure, 1983, Technometrics 25, 2 (may)
97. Rossman, L.: Storm Water Management Model, 2004, User's Manual, Version 5.0, U.S. Environmental Protection Agency, USA)
98. Rousseeuw, P. and A. Leroy: Robust Regression and Outlier Detection, 1996, 3rd edn. John Wiley & Sons
99. Saunders, R. And J. S. Gero: A Curious Design Agent: A Computational Model of Novelty-Seeking Behaviour in Design, 2001, Proceedings of the Sixth Conference on Computer Aided Architectural Design Research in Asia (CAADRIA 2001), Sydney
100. Schlaeger, F., M. Natschke and D. Witham: Quality Assurance for Hydrometric Network Data as a Basis for Integrated River Basin Management, 2005, , 2007, IAHS Publication, NUMB 310, pp: 327-337
101. Scholz, M., M. Fraunholz, and J. Selbig: Nonlinear principal component analysis: neural network models and applications, 2007, Principal Manifolds for Data Visualization and Dimension Reduction, edited by Alexander N. Gorban, Balázs Kégl, Donald C. Wunsch, and Andrei Zinovyev. Volume 58 of LNCSE, pp: 44-67. Springer Berlin Heidelberg
102. Shannon, C.E.: A Mathematical Theory of Communication, 1948, *Bell System Technical Journal*, 27, pp: 379–423 & 623–656
103. Shekhar, S., Lu, C. and Zhang P.: Detecting Graph-Based Spatial Outliers: Algorithms and Applications, 2001, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
104. Shewhart, W. A.: Economic Control of Quality of Manufactured Product, 1931, D. Van Nostrand Company, New York NY
105. Skalak, D. B. and E. L. Rissland: Inductive Learning in a Mixed Paradigm Setting, 1990, Proceedings of the Eighth National Conference on Artificial Intelligence, Boston, MA, pp: 840–847
106. Solomonoff, R.: A Preliminary Report on a General Theory of Inductive Inference, 1960, Report V-131, Zator Co., Cambridge, Ma.
107. Steiner, M., J. A. Smith, S. J. Burges, C. V. Alonso R. W. and Darden: Effect of bias adjustment and rain gauge data quality control on radar rainfall estimation, 1999, Wat. Resour. Res., 35 (8), pp: 2487-2503
108. Stolfo, S. J., A. L. Prodromidis, S. Tselepis, W. Lee, D. W. Fan and P. K. Chan: JAM: Java Agents for Meta-Learning over Distributed Databases, 1997, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp: 74–81
109. Tang, J., Z. Chen, A. Fu and D. Cheung: A Robust Outlier Detection Scheme in Large Data Sets, 2002, 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Taipei, Taiwan
110. Tarifa, E. and N. Scenna: Fault diagnosis, directed graphs, and fuzzy logic, 1997, Computers and Chem. Eng. 21, pp:649-654

111. Tax, D. M. J., A. Ypma and R. P. Duin: Support Vector Data Description Applied to Machine Vibration Analysis, 1999, Proceedings of ASCI'99, Heijen, Netherlands
112. Terakawa, A.: WORLD METEOROLOGICAL ORGANIZATION OPERATIONAL HYDROLOGY REPORT No. 48, HYDROLOGICAL DATA MANAGEMENT: PRESENT STATE AND TRENDS, 2003, Secretariat of the World Meteorological Organization, Geneva, Switzerland
113. Torr, P. H. S. and D. W. Murray: Outlier Detection and Motion Segmentation, 1993, Proceedings of SPIE
114. Umeda, T., T. Kuriyama, E. Oshima, and H. Matsuyama: A graphical approach to cause and effect analysis of chemical processing systems,1980, Chem. Eng. Science 35(12), pp:2379-2388
115. Upton, and Rahimi: On-line detection of errors in tipping-bucket raingauges, 2003, Journal of Hydrology 278, pp: 197–212
116. USEPA: Guidance for Data Quality Assessment: Practical Methods for Data Analysis, 2000, http://www.epa.gov//quality/qs-docs/g9-final.pdf
117. USEPA: OEI Quality System 2000: Office of Environmental Information Management System for Quality, 2000, http://www.epa.gov/oei/quality.htm
118. USEPA: QUAL2E Windows interface user's guide, 1995, United States Environmental Protection Agency)
119. Vaitl, W.: Beschreibung der Prüfkriterien für die Qualitätskontrolle stündlicher bzw. 10-minütiger Daten von automatischen agrarmeteorologischen Stationen der Bayerischen Landesanstalt für Bodenkultur und Pflanzenbau, 1988, p 16, München-Freising
120. Van Der Helm, P. A. and L. J. Leeuwenberg: Accessibility: a Criterion for Regularity and Hierarchy in Visual Pattern Codes, 1991, Journal of Mathematical Psychology 35, pp: 151-213
121. Vedam, H. and V. Venkatasubramanian: A wavelet theory-based adaptive trend analysis system for process monitoring and diagnosis, 1997, American Contr. Conf., pp: 309-313
122. Vedam, H., V. Venkatasubramanian, and M. Bhalodia: A b-spline based method for data compression, process monitoring and diagnosis, 1998, Computers and Chem. Eng. 22, pp: 827-830
123. Venkatasubramanian, V., R. Rengaswamy, K. Yin, and S. N. Kavuri: A review of process fault detection and diagnosis part I: Quantitative modelbased methods, 2003, Computers and Chem. Eng. 27, 293—311
124. Venkatasubramanian, V., R. Rengaswamy, K. Yin, and S. N. Kavuri: A review of process fault detection and diagnosis part II: Qualitative models and search strategies, 2003, Computers and Chem. Eng. 27, 313—326
125. Venkatasubramanian, V., R. Rengaswamy, K. Yin, and S. N. Kavuri: A review of process fault detection and diagnosis part III: Process history based methods, 2003, Computers and Chem. Eng. 27, 327—346
126. Vesanto, J., J. Himberg, M. Siponen and O. Simula: Enhancing SOM Based Data Visualization, 1998, Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems - Methodologies for the Conception, Design and Application of Soft Computing, Vol. 1, pp: 64–67, Singapore: World Scientific.
127. Wald, A.: Sequential Analysis, 1947, Wiley, New York
128. Wang, J.: Encyclopedia of Data Warehousing and Mining, 2006, Idea Group Reference, ISBN 1-59140-559-9
129. Wettschereck, D.: A Study of Distance-based Machine Learning Algorithms, 1994, Ph.D. Thesis, Department of Computer Science, Oregon State University, Corvallis
130. Willsky, S.: A survey of design methods for failure detection in dynamic systems, 1976, Automatica 12(6), pp. 601—611
131. Witten, H. Ian, F. Eibe: Data Mining, Practical Machine Learning Tools and Techniques, 2005, Elsevier Inc., ISBN: 0-12-088407-0

132. WMO: WMO Guide To Meteorological Instruments And Methods Of Observation, 2008, Secretariat of the World Meteorological Organization, Geneva, Switzerland, WMO-No. 8 (Seventh edition)
133. WMO-No. 8: Guide to Meteorological Instruments and Methods of Observation, Preliminary seventh edition, 2006, Secretariat of the World Meteorological Organization, Geneva, Switzerland
134. WWW: Data Storage Validation and Retrieval System (DSVRS) - http://www.eicinformation.org/internal.asp?id=7&type=normal&title=Data,+Storage,+Validation+&+Retrieval+System
135. Wyatt, D. W., H. T. Castrup: Managing Calibration Intervals, 1991, Presented at the NCSL 1991 Annual Workshop & Symposium, Albuquerque, August 1991
136. Zhao, J., B. Chen, and J. Shen: A hybrid ANN-ES system for dynamic fault diagnosis of hydrocracking process, 1997, Computers and Chem. Eng. 21, pp: 929-933