Latvijas Biozinātņu un tehnoloģiju universitāte
Latvia University of Life Sciences and Technologies

Inženierzinātņu un informācijas tehnoloģiju fakultāte
Faculty of Engineering and Information Technologies

Mg. sc. ing. **Nikolajs Būmanis**

VAIRĀKU AVOTU DATU APVIENOŠANA UN ANALĪZE INTERAKTĪVAI APSTRĀDEI UN VIZUALIZĀCIJAI

MULTI-SOURCE DATA FUSION AND ANALYSIS FOR INTERACTIVE PROCESSING AND VISUALIZATION

Promocijas darba KOPSAVILKUMS

Zinātnes doktora grāda *(Ph.D.)* iegūšanai

SUMMARY
of the Doctoral thesis for the Doctoral degree of Science (Ph.D.)

N. Bumanis _____

Jelgava
2025

VISPARĪGĀ INFORMĀCIJA

Darba izpildes vieta: Latvijas Biozinātņu un tehnoloģiju universitāte (LBTU), Inženierzinātņu un informācijas tehnoloģiju fakultāte, Datoru sistēmu un datu zinātnes institūts.

Doktora studija programma: Informācijas tehnoloģijas.

Eksperimentu izpildes vieta: Latvijas Biozinātņu un tehnoloģiju universitāte (LBTU), Inženierzinātņu un informācijas tehnoloģiju fakultāte, Datoru sistēmu un datu zinātnes institūts, Lielā iela 2, Jelgava, Latvija.

Promocijas darba zinātniskā vadītāja: prof. *Dr. sc. ing.* Irina Arhipova.

Darbs akceptēts LBTU Datoru sistēmu un datu zinātnes institūta sēdē 2023. gada 19. oktobrī. Protokols Nr. 3.

Oficiālie recenzenti:

1. Dr.habil.comp. **Jānis Grundspeņķis**, RTU profesors;
2. Dr.sc.ing., **Jānis Grabis**, RTU profesors;
3. Swedish University of Agricultural Sciences, Professor of Digitalization in Agricultural Engineering, Department of Energy and Technology, **Abozar Nasirahmadi** (h-index 18, 37 papers in SCOPUS). Galvenie pētniecības virzieni ir ar mākslīgo intelektu darbināmu risinājumu izstrāde lauksaimniecības pārvaldībai, robotikas pielietojums augkopībā un lopkopībā, viedie sensori, kā arī datu analītika un mašīnmācīšanās.

Promocijas darba aizstāvēšana notiks LBTU nozares "Elektrotehnika, elektronika, informācijas un komunikāciju tehnoloģijas" promocijas padomes atklātajā sēdē 2025. gada 3. septembrī, 14:30, Jelgavā, Lielā iela 2, Inženierzinātņu un informācijas tehnoloģiju fakultātē 37. auditorijā.

Atsauksmes sūtīt Promocijas padomes sekretārei – Lielā iela 2, Jelgava, LV-3001; tālrunis: 63022584; e-pasts: tatjana.tabunova@lbtu.lv. Atsauksmes vēlams sūtīt skenētā veidā ar parakstu.

Ar promocijas darbu var iepazīties LBTU Fundamentālajā bibliotēkā, Lielā ielā 2, Jelgavā un http://llufb.llu.lv/promoc_darbi.html.

Padomes sekretāre: Mg.paed. Tatjana Tabunova.

SATURS/CONTENT

PROMOCIJAS DARBA APROBĀCIJA

Promocijas darba izstrādes laikā publicētas 11 zinātniskās publikācijas (indeksētas Web of Science un SCOPUS datubāzēs). Pētījumu rezultāti prezentēti 6 zinātniskās konferencēs.

Publikācijas vispāratzītos recenzējamos zinātniskos izdevumos:

1. **Bumanis, N**. (2020). Data fusion challenges in precision beekeeping: a review. Research for Rural Development. In proceedings of 26th international scientific conference "Research for Rural Development", vol. 35, pp. 252.–259, DOI: 10.22616/rrd.26.2020.037. **Autora devums: sagatavots visu nodaļu teksts;**
2. **Bumanis, N.**, Komasilova, O., Komasilovs, V., Kviesis, A., & Zacepins, A. (2020). Application of Data Layering in Precision Beekeeping: The Concept. In 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT), pp. 1–6, article number 9368733, DOI: 10.1109/AICT50176.2020.9368733. **Autora devums: sagatavots visu nodaļu teksts;**
3. Vitols, G., **Bumanis, N.**, Arhipova, I., & Meirane, I. (2021). LiDAR and Camera Data for Smart Urban Traffic Monitoring: Challenges of Automated Data Capturing and Synchronization. In International Conference on Applied Informatics. Applied Informatics (Q4), 2021, vol. 1455, pp. 421–432, DOI: 10.1007/978-3-030-89654-6_30. **Autora devums: darbs pie datu sinhronizācijas risinājuma, datu sinhronizācijas koncepta un ievestā risinājuma nodaļas izveide, datu sinhronizācijas risinājuma vizualizācija; darbs pie secinājumu nodaļas;**
4. **Bumanis, N.**, Vitols, G., Arhipova, I., & Solmanis, E. (2021). Multi-object Tracking for Urban and Multilane Traffic: Building Blocks for Real-World Application. In proceedings of the 23rd International conference on Enterprise Information Systems (ICEIS), 2021. vol. 1, pp. 729–736, DOI: 10.5220/0010467807290736. **Autora devums: sagatavots visu nodaļu teksts.**
5. **Bumanis, N.**, Arhipova, I., Paura, L., Vitols, G., & Jankovska, L. (2022). Data conceptual model for smart poultry farm management system. Procedia Computer Science (Q2), vol. 200, pp. 517–526, DOI: 10.1016/j.procs.2022.01.249. **Autora devums: datu kopu noteikšana, datu glabātuves arhitektūras izveide, kiberfizikālā modeļa un datu glabātuves arhitektūras nodaļu teksta sagatavošana; darbs pie secinājumu nodaļas;**
6. Paura, L., Arhipova, I., Jankovska, L., **Bumanis, N.**, Vitols, G., & Adjutovs, M. (2022). Evaluation and association of laying hen performance, environmental conditions and gas concentrations in barn housing system. Italian Journal of Animal Science (Q1), 21(1), 694–701,

DOI: 10.1080/1828051X.2022.2056528. **Autora devums: datu kopu noteikšana, analīzes rezultātu pārbaude;**

7. **Bumanis, N.**, Vitols, G., & Meirane, I. (2022). Data Fusion of Video and LiDAR traffic surveillance data: Practical Assessment of Implemented solution at Jelgava City. In proceedings of 21st international scientific conference "Engineering for rural development", vol. 21, pp. 478–488, DOI: 10.22616/ERDev.2022.21.TF166. **Autora devums: sagatavotas nodaļas – materiāli un metodes, rezultāti; darbs pie secinājumu nodaļas;**
8. **Bumanis, N.**, Kviesis, A., Tjukova, A., Arhipova, I., Paura, L., & Vitols, G. (2023). Smart Poultry Management Platform with Egg Production Forecast Capabilities. Procedia Computer Science (Q2), vol. 217, pp. 339–347, DOI: 10.1016/j.procs.2022.12.229. **Autora devums: sagatavots – nodaļas ievads, rezultāti; darbs pie secinājumu nodaļas;**
9. **Bumanis, N.**, Kviesis, A., Paura, L., Arhipova, I., & Adjutovs, M. (2023). Hen Egg Production Forecasting: Capabilities of Machine Learning Models in Scenarios with Limited Data Sets. Applied Sciences (Q1), vol. 13 (13), 7607, DOI: 10.3390/app13137607. **Autora devums: sagatavotas nodaļas – ievads, materiāli un metodes; darba pie nodaļām rezultāti, diskusija, secinājumi;**
10. Arhipova, I., **Bumanis, N.**, Paura, L., Berzins, G., Erglis, A., Vitols, G., Ansonska, E., Salajevs, V. and Binde, J. (2023). Optimizing Transport Network to Reduce Municipality Mobility Budget. In Proceedings of the 5th International Conference on Finance, Economics, Management and IT Business, 2023. vol. 1, pp. 38–47, DOI: 10.5220/0011941500003494. **Autora devums: ievada nodaļas sagatavošana, darba pie nodaļām materiāli un metodes, rezultāti;**
11. **Bumanis, N.** (2024). Overcoming Data Limitations in Precision Poultry Farming: Processing and Data Fusion Challenges. Procedia Computer Science, 232, 2302–2309. **Autora devums: sagatavots visu nodaļu teksts.**

Pētījumu rezultāti prezentēti šādās zinātniskās konferencēs:

1. "Data fusion challenges in precision beekeeping: a review", 26. ikgadējā zinātniskā konference "Research for Rural Development", 13.05.2020.–15.05.2020., Jelgava, Latvija;
2. "Application of Data Layering in Precision Beekeeping: The Concept", 14. ikgadējā starptautiskā konference "Application of Information and Communication Technologies", 07.10.2020.–09.10.2020., tiešsaistē, Taškenta, Uzbekistāna;
3. "LiDAR and Camera Data for Smart Urban Traffic Monitoring: Challenges of Automated Data Capturing and Synchronization", 4. starptautiskā zinātniskā konference "Applied Informatics", 28.10.2021.–30.10.2021., tiešsaistē, Buenosairesa, Argentīna;

4. “Data Fusion of Video and LiDAR traffic surveillance data: Practical Assessment of Implemented solution at Jelgava City”, 21. starptautiskā zinātniskā konference “Engineering for Rural Development”, 25.05.2022.–27.05.2022., Jelgava, Latvija;
5. “Smart Poultry Management Platform with Egg Production Forecast Capabilities”, 3. starptautiskā zinātniskā konference “Industry 4.0 and Smart Manufacturing”, 02.11.2022.–04.11.2022., Linca, Austrija;
6. “Overcoming Data Limitations in Precision Poultry Farming: Processing and Data Fusion Challenges”, 4. starptautiskā zinātniskā konference “Industry 4.0 and Smart Manufacturing”, 22.11.2023.–24.11.2023., Lisabona, Portugāle.

IEVADS

Mūsdienās lietu internets jeb angliski *IoT* (*Internet of Things*), kas ietver sensorus, aparatūru un programmatūru, ir kļuvis par būtisku informācijas sistēmas daļu (Nižetić et al., 2020). Tas piedāvā iespējas iegūt informāciju par notiekošajiem procesiem, izmantojot sensorus, kā arī no ārējām un/vai saistītām sistēmām. Prognozējams, ka, ja tiks pilnībā izmantots tā potenciāls, *IoT* varētu kļūt par vienu no izcilākajiem tehnoloģiskajiem sasniegumiem (Alam et al., 2016). *IoT* ir nepieciešams starpnozaru uzraudzības vai pārvaldības sistēmu izstrādei (M. Zhang et al., 2021). Daudzās nozarēs *IoT* ieviešana ir veicinājusi to pārveidošanu par viedajām nozarēm, piemēram, precīzai biškopībai (Zacepins et al., 2015) un precīzai putnkopībai (Astill et al., 2020).

Saistībā ar to, ka *IoT* specifikācijas nav vienotas, katrs autors – gan pētnieks, gan izstrādātājs – izvēlas sev atbilstošāko sensoru, aparatūras un programmatūras kombināciju. Šāda pieeja noved pie sistēmu atšķirībām, kas atklājas dažādos datu formātos un mērvienībās. *IoT* sistēmas ietvaros viens un tas pats objekts var tikt detektēts ar vairākiem sensoriem, radot dažādus datus, vai arī objekts var atrasties tikai viena sensora redzeslokā. Šādās situācijās var izvēlēties vairākus datu avotus, apvienojot skaidru skatu sniedzoša sensora datus ar iepriekš iegūto informāciju (piemēram, datiem vai pēdām) no cita sensora, lai precīzāk atjaunotu objekta stāvokli (N. E. El Faouzi & Klein, 2016).

IoT datu apvienošanas procesā tomēr bieži rodas datu kvalitātes problēmas, īpaši trūkstošo datu un noviržu gadījumā. Sensoru uzticamība ne vienmēr ir pilnīga, un var būt grūtības iegūt konsekventus datus, kas apdraud precīzu analīzi un lēmumu pieņemšanu. Šīs problēmas īpaši izpaužas sarežģītos vides apstākļos, kur sensori nespēj pilnībā fiksēt nepieciešamos datus (Kratkiewicz et al., 2019).

Pētniecībā bieži sastopamies ar ierobežotu datu problēmu, kad ne vienmēr ir iespējams iegūt pietiekamu informāciju par pētāmo objektu vai procesu. Šāds datu trūkums būtiski ierobežo izpratni par pētāmo parādību. Tas ne tikai apgrūtina padziļinātu analīzi, bet arī rada šķēršļus precīzu prognožu izveidē un modeļu apmācībā. Šādās situācijās pētnieki meklē radošus risinājumus, piemēram, apvieno datus no dažādiem avotiem, izmanto datu papildināšanas metodes vai pat rada mākslīgus datus. Tomēr katrai no šīm pieejām ir savi ierobežojumi. Mākslīgi radītie dati var neatbilst reālās sistēmas uzvedībai, tādējādi radot maldinošus secinājumus. Mūsdienās, neskatoties uz ievērojamo progresu datu vākšanas tehnoloģijās, joprojām ir nepieciešams pielāgot analīzes metodes katram konkrētam uzdevumam (Kratkiewicz et al., 2019). Viens no perspektīvākajiem risinājumiem ir vairāku avotu datu apvienošana, kas ļauj iegūt pilnīgāku priekšstatu par pētāmo sistēmu un atbildēt uz jautājumiem, kurus nav iespējams risināt ar viena avota datiem.

Galvenais datu apvienošanas mērķis ir padarīt informāciju no dažādiem avotiem un sensoriem saprotamāku un precīzāku, pat ja atsevišķi sensoru dati var šķist nepietiekami informatīvi. Datu apvienošana, kā to skaidro *Hall* un *Llinas* (Hall & Llinas, 2016), ir pieeja, kas apkopo dažādu avotu informāciju – gan

sensoru mērījumus, gan saistītos datus no datubāzēm. Šāda pieeja ļauj par pētāmo objektu iegūt pilnīgāku un precīzāku priekšstatu, nekā tas būtu iespējams, izmantojot tikai vienu datu avotu. Informācijas apvienošana no dažādiem sensoriem, kas mēra atšķirīgas fiziskās pazīmes, palīdz labāk izprast apkārtējo vidi un nodrošina izšķirošu pamatu plānošanai, lēmumu pieņemšanai un autonomu sistēmu vadībai (Alam et al., 2017). Sākotnēji datu apvienošana galvenokārt tika izmantota datu analīzei militāros nolūkos, bet tagad tā ir implementēta daudzās jomās un nozarēs (Noh, 2020; Shi et al., 2019; Sun et al., 2022).

Promocijas darbā autors datu apvienošanu analizē dažādās starpnozaru jomās, tostarp precīzajā biškopībā, precīzajā putnkopībā, kā arī objektu detektēšanā un izsekošanā transportēšanas jomā. Precīzā biškopība ir nozare, kur *IoT* tehnoloģijas ir kļuvušas neatsverami nozīmīgas efektivitātes un ražīguma sasniegšanai (Debauche et al., 2018). Datu iegūšana šeit notiek, galvenokārt izmantojot bezvadu tehnoloģijas (Huet et al., 2022) un instrumentus, kas ļauj regulāri nosūtīt lielus datu apjomus. Lai gan *IoT* tehnoloģijas precīzajā biškopībā tiek izmantotas jau vairāk nekā desmit gadu, datu apvienošana šajā kontekstā bieži ir ierobežota līdz zemākā līmeņa sensoru datiem (Rafael Braga et al., 2020). Tādēļ vidējā un augstākā līmeņa datu apvienošana joprojām ir aktuāla problēma šajā nozarē (Bumanis, 2020; Bumanis et al., 2020). Objektu detektēšana un izsekošana ir bieži sastopamas jomas, kurās plaši pielieto datu apvienošanu. Tehniskā aspektā šādus risinājumus bieži saista ar viedpilsētu (*Smart City*) konceptu, jo tie tiek izmantoti apdzīvotās un publiskās vietās (Lau et al., 2019). Daži no galvenajiem pielietojumiem ir satiksmes uzraudzība un viedo transporta sistēmu izstrāde (Kim & Jeon, 2014). Piemēram, satiksmes uzraudzības sistēmās bieži tiek kombinēti *LiDAR* un videokameru dati, lai uzlabotu informācijas precizitāti (Manogaran et al., 2021). Lai gan satiksmes uzraudzības sistēmām pilnīga datu apvienošanas ieviešana var nebūt obligāta, šīs metodes iespējams veiksmīgi izmantot arī citās, specifiskās, lietojumprogrammās (Y. Han & Hu, 2020). Salīdzinājumā ar precīzo biškopību precīzā putnkopība *IoT* jomā ir attīstījusies ievērojami tālāk (Lashari et al., 2019). Šajā nozarē tiek vākti dati par putniem, to produktiem, piemēram, gaļu un olām, kā arī par apkārtējiem apstākļiem, kas ietekmē dzīvnieku veselību, fermu produktivitāti un kopējo efektivitāti (Singh et al., 2020). Datu apvienošana šajā kontekstā tiek izmantota dažādu problēmu risināšanai, piemēram, putnu veselības uzraudzībai (Muneer et al., 2020) un gaļas kvalitātes novērtēšanai (Khulal et al., 2017). Tomēr datu apvienošana produktivitātes mērījumiem joprojām nav plaši izmantota (Bumanis, Arhipova, et al., 2022).

Īpaši datu kvalitātes un ierobežoto datu kontekstā datu apvienošanas ieviešanā ir vairāki izaicinājumi (Khaleghi et al., 2013), starp tiem:

- trūkstošie dati, kas rodas sensoru nespējas dēļ konsekventi nodrošināt pilnīgus datus;
- neprecizitātes un novirzes, ko rada sensori, kas ietekmē datu ticamību un uzticamību;

- nekonsekvences un neskaidrības datos, kas rodas sensoru darbības mainīgos vides apstākļos ietekmē;
- datu pretrunas, kas rodas, piemērojot metodes, piemēram, pierādījumu pārliecības argumentāciju un *DST* jeb Dempstera-Šafera teorijas (*Dempster-Shafer theory*) kombinācijas noteikums (Yin Liu & Zhang, 2022);
- ierobežoti un heterogēni dati, kas dažkārt var būt nepietiekami vai neviendabīgi dažādām datu modalitātēm;
- sensoru trokšņa radītās datu novirzes, kas ietekmē mērījumu precizitāti un izraisīto korelāciju;
- datu reģistrēšanas problēmas, kas rodas nepilnīgiem vai kļūdainiem datu ieguves mehānismiem.

Katrs no šiem izaicinājumiem parasti (Bakr & Lee, 2017) tiek risināts individuāli, fokusējoties uz konkrēto problēmu, nevis izmantojot vispārēju pieeju. Alternatīvi, ja pieejami vairāki sensori, kas nodrošina informāciju par vienu pētāmo objektu, var izmantot datu apvienošanu. Tā ļauj risināt tādus uzdevumus kā problemātisko datu labošanu (C. Huang et al., 2019), datu uzticamības uzlabošanu (Hong et al., 2009), datu pilnīguma palielināšanu (Consoli et al., 2015) un augstāka līmeņa informācijas iegūšanu (Jayasinghe et al., 2019).

Datu apvienošanas metodiku klasifikācija atspoguļo to spējas apstrādāt dažādus datu un informācijas veidus. Pētnieki ir izstrādājuši vairākas klasifikācijas sistēmas šo metodiku strukturēšanai (Becerra et al., 2021). Trīs būtiskākās klasifikācijas dimensijas – abstrakcijas līmeņi, datu avotu savstarpējās attiecības un ievades-izvades attiecības – veido metodiku sistematizācijas pamatu. Šāds iedalījums atspoguļo fundamentālās sakarības starp datu veidiem un to apstrādes posmiem. Abstrakcijas līmeņu klasifikācija piedāvā strukturētu ietvaru, kas sastāv no četriem līmeņiem. Signālu līmeņa apvienošana veic sensoru signālu tiešu apstrādi, kamēr pikseļu līmeņa apvienošana nodrošina attēlu datu integrāciju. Tālāk pazīmju līmeņa apvienošana pārveido signāla datus raksturīgajās pazīmēs, bet simbolu (lēmumu) līmeņa apvienošana attēlo rezultātus simboliskā formā. Otra nozīmīgā dimensija skata datu avotu savstarpējās attiecības, izdalot trīs galvenos veidus. Pirmais ir papildinošā apvienošana, kas paplašina kopējo informācijas apjomu, apvienojot datus no dažādiem avotiem. Otrais ir dublējošā apvienošana, kas uzlabo datu kvalitāti, izmantojot vairākus datu avotus vienlaikus. Trešais ir savstarpējās sadarbības apvienošana, kas ļauj radīt pilnīgi jaunu informāciju no vairākiem datu avotiem. Trešā dimensija aplūko ievades-izvades attiecības, kur galvenais princips ir datu pārveide augstāka līmeņa informācijā. Šajā procesā sākotnējie dati tiek pārveidoti noderīgākā formā, piemēram, pazīmēs vai konkrētos lēmumos.

Pētnieki (Alam et al., 2017; Becerra et al., 2021; Castanedo, 2013; Khaleghi et al., 2013; J. Liu et al., 2020) izdala dažādus datu apvienošanas arhitektūras modeļus. Starp populārākajiem modeļiem izceļas *JDL* datu apvienošanas

modelis jeb *JDL* (*Joint Directors of Laboratories*). *JDL*, būdams viens no pirmajiem datu apvienošanas modeļiem, bieži tiek izmantots kā atsauces punkts pārējo modeļu salīdzināšanai atbilstoši tā arhitektūras līmeņiem (Becerra et al., 2021). Tomēr pastāvošās arhitektūras ne vienmēr tiek tieši pielietotas praktiskajos uzdevumos. Lietojumprogrammas bieži veido savas specifiskas datu apvienošanas arhitektūras, kas tiek pielāgotas konkrētām vajadzībām, piemēram, viedo transporta sistēmu apstrādē (N. E. El Faouzi & Klein, 2016; Guerrero-Ibáñez et al., 2018). Šāda pieeja izriet no izpratnes, ka datu apvienošanas būtību nosaka ne tik daudz tās arhitektūra, cik pašas apvienošanas metodes, ar kurām tiek veikta datu apstrāde.

Datiem, pēc to analīzes un apstrādes, ir būtiska vizualizācija, jo tā ļauj pārskatāmi nodot informāciju galalietotājam. Vizualizācijas pieeja lielā mērā ir atkarīga no tā, kādus datus ir nepieciešams attēlot un cik intuitīvai jābūt vizualizācijai. Cilvēka iejaukšanās šajā procesā ir būtiska, jo kvalitatīva vizualizācija ļauj zinātniekam ne tikai labāk izprast savus datus, bet arī par saviem atklājumiem informēt citus (Lau et al., 2019; Parish & Edmondson, 2019; Weissgerber et al., 2019). Lai sasniegtu šos mērķus, tiek izmantoti dažādi rīki un paņēmieni, kas ļauj izveidot grafikus, kas balansē starp vizualizācijas efektivitāti un pielāgojamību (Waskom, 2021). Ja datu apstrādei ir ģeogrāfiskais konteksts, vizualizācijas iespējas kļūst vēl daudzveidīgākas. Šajā gadījumā var izmantot, piemēram, flīžu kartes (*Puzzle tile maps)* (Lin et al., 2019) vai telpiskos punktu mākoņus (*Point cloud*) (Schneider et al., 2020), kas ir īpaši noderīgi *LiDAR* datu attēlošanai (Deibe et al., 2019; Shirowzhan et al., 2020). Kaut arī pastāv daudz dažādu veidu, kā attēlot datus, visbiežāk tiek izmantoti tie paņēmieni, kurus ir vienkāršāk un ātrāk izveidot un kuri padara datus vieglāk saprotamus (Qin et al., 2020).

Datu apvienošanas galvenā problēma ir nepilnīgu un pretrunīgu datu pārvaldība no dažādiem avotiem, kas apgrūtina precīzas un uzticamas informācijas iegūšanu.

Mērķis un uzdevumi

Promocijas **darba mērķis** ir izveidot metodoloģiskus risinājumus datu apvienošanai un kvalitātes uzlabošanai, lai paaugstinātu datu analīzes efektivitāti un precizitāti, nodrošinot dziļāku un pamatotāku ieskatu izpētes jomā.

Mērķa sasniegšanai tika definēti šādi **darba uzdevumi:**

1. izpētīt datu kvalitātes raksturlielumus un uzlabošanas paņēmienus sensoru ģenerētiem datiem pēc to pilnīguma un precizitātes kritērijiem (sk. 1. nodaļu);
2. izpētīt pastāvošās datu apvienošanas pieejas, tai skaitā noteikt to klasifikācijas un darbības principus (sk. 1., 2., 3. nodaļas);
3. izstrādāt datu apvienošanas metodi vairāku avotu datu savstarpējās saistības vizualizācijai, balstoties uz svarīgāko periodu noteikšanu (sk. 4. nodaļu);

4. pārbaudīt izstrādāto metodi precīzās putnkopības problemātikas jomā (sk. 5. nodaļu).
5. izstrādāt datu kvalitātes (pēc pilnīguma un precizitātes kritērijiem) uzlabošanas metodi trūkstošo vērtību aizvietošanai un noviržu pielāgošanai (sk. 6. nodaļu);
6. veikt izstrādāto metožu novērtēšanu (sk. 6. nodaļu);

Pētījuma metodes

Izmantoto pētījumu metožu klāsts ietver:

- zinātniskās un citas informācijas avotu analīzi;
- salīdzināšanu, indukciju, dedukciju un slēdzienu veidošanu;
- metožu izstrādi un testēšanu *Python* programmēšanas valodā:
  - datu kvalitātes uzlabošanas metožu izveidei, ieskaitot trūkstošo vērtību aizvietošanu (metodes: *ARIMA*, modificētā standarta vidējā svērtā metode (MSVSM) un noviržu noteikšanai un pielāgošanai – multi-skalas integrēto noviržu analīzes metodi (MINA));
  - datu apvienošanas metodes izveidei;
- izstrādāto metožu novērtējumu, izmantojot validācijas kopas.

Zinātniskais jaunināj̄ums un praktiskā vērtība

- Ir izveidota datu slāņošanas metodes koncepcija, kas balstīta uz divām datu analīzes pieejām: adaptīvo svērtās interpolācijas datu apvienošanu un uz galveno komponentu analīzi (PCA) balstītu datu apvienošanu. Adaptīvā svērtās interpolācijas metode nodrošina sākotnējo svērto vērtību iegūšanu, izmantojot lietotāja definētus svarus un automātisku interpolācijas izlīdzināšanas parametru pielāgošanu, savukārt PCA tiek optimizēta vienas galvenās komponentes iegūšanai, kas atspoguļo dominējošo tendenci datos, automātiski pielāgojoties parametru savstarpējām korelācijām. Datu slāņošanas koncepcijas rezultāts ir abu metožu iegūto vērtību pārklāšanās zonu vizualizācija un kvantitatīvs novērtējums, kas, izmantojot trapeces likumu, attēlo un aprēķina kopīgo nozīmīgo reģionu proporcijas ar pielāgojamiem sliekšņiem.
- Ir izstrādātas divas kombinētās metodes datu kvalitātes uzlabošanai:
  - trūkstošo vērtību aizvietošanas modificētā standarta vidējā svērtā metode (MSVSM) ir izveidota, apvienojot vairākas pieejas: dinamisko tuvāko kaimiņu svaru noteikšanu, kas balstās uz trūkstošo vērtību attālumu un daudzumu no esošajiem punktiem, gabaliem definēto kubisko Ermita polinoma interpolāciju (*Piecewise Cubic Hermite Interpolation Polynomial, PCHIP*) datu tendences noteikšanai, kā arī trenda izlīdzināšanu ar slīdošā vidējā algoritmu. Tas nodrošina adaptīvu risinājumu dažāda veida datu anomālijām, vienlaikus saglabājot gan lokālās datu īpatnības, gan globālo tendenci;

- noviržu noteikšanas un pielāgošanas metode – multi-skalas integrēto noviržu analīzes metode (MINA), kas apvieno: adaptīvu vairāku mērogu ritošā loga analīzi ar robustām statistikām (izmantojot mediānas absolūto novirzi, MAD), vinsorizāciju ekstremālo vērtību apstrādei, un daudzlīmeņu z-vērtību analīzi ar konsensa mehānismu. Metode ietver arī trenda noviržu detektēšanu, izmantojot *Savitzky-Golay* filtru, un kontekstuālu sliekšņu pielāgošanu, balstoties uz lokālo datu variabilitāti. Tas nodrošina efektīvu gan lokālo, gan ekstremālo datu noviržu identifikāciju un pielāgošanu, vienlaikus saglabājot būtiskās datu statistiskās īpašības un sezonalitāti.

Praktiskā aprobācija

Promocijas darba praktiskā vērtība ir darba gaitā iegūto zināšanu un atziņu, kā arī izstrādāto datu apvienošanas un datu kvalitātes uzlabošanas metožu pielietošana problemorientēto uzdevumu risināšanai šādos pētniecības projektos:

- "Apvārsnis 2020" projekts "Futūristiski bišu stropi viedajai metropolei". **Autora devums**: datu slāņošanas koncepcijas izstrāde bišu stropa novietošanas pozīcijas noteikšanai, balstoties uz vairāku avotu datiem (*Futūristiski bišu stropi viedajai metropolei (HIVEOPOLIS) (HOR5)*, 2019);
- darbības programmas "Izaugsme un nodarbinātība" 1.1.1. specifiskā atbalsta mērķa "Palielināt Latvijas zinātnisko institūciju pētniecisko un inovatīvo kapacitāti un spēju piesaistīt ārējo finansējumu, ieguldot cilvēkresursos un infrastruktūrā" 1.1.1.1. pasākuma "Praktiskas ievirzes pētījumi" projekts 1.1.1.1/19/A/145 "HENCO2: Mākoņdatu vidē balstīta IT platforma putnkopības produktivitātes uzlabošanai un siltumnīcefekta gāzu emisiju samazināšanai". **Autora devums**: datu kvalitātes metožu izstrāde, balstoties uz projekta datu kvalitātes trūkumiem; izstrādāto metožu pielietošana olu īpatsvara prognozēšanas risinājuma izstrādei; mašīnmācīšanās modeļu veiktspējas novērtējums (*HENCO2: Mākoņdatu vidē balstīta IT platforma putnkopības produktivitātes uzlabošanai un siltumnīcefekta gāzu emisiju samazināšanai – ER32*, 2020);
- "Apvārsnis 2020" programmas *ERA-NET Cofund* projekts "Individuālie mobilitātes budžeti kā sociālais un ētiskais pamats oglekļa emisiju samazināšanai". **Autora devums**: sabiedriskā transporta un mobilā sakaru tīkla datu apvienošana (*Individuālie mobilitātes budžeti kā sociālais un ētiskais pamats oglekļa emisiju samazināšanai (MyFairShare) (ZV91)*, 2021);
- SIA "WeAreDots" un zinātniski tehniskās firmas "Lāsma" pētījums Nr. 1.12. "Multiobjektu detektēšana un izsekošana transportlīdzekļu satiksmes novērošanai: *3D-LiDAR* un kameras datu apvienošana". **Autora devums**: multiobjektu detektēšanas risinājumu izpēte; video plūsmas un *3D-LiDAR* datu kopu sinhronizācijas risinājuma uzlabošana

(*Multiobjektu detektēšana un izsekošana transportlīdzekļu satiksmes novērošanai: 3D-LiDAR un kameras datu apvienošana*, 2020).

Pētījuma tēzes

- Lietu interneta (*IoT*) sistēmās iegūtie datu analīzes rezultāti ir atkarīgi no datu kvalitātes uzlabošanas metožu pielietošanas, kas īpaši svarīgi ierobežota datu apjoma vai nepilnīgu datu gadījumos, nodrošinot ticamu un precīzu lēmumu pieņemšanu.
- Apkopojot vairākus datu kvalitātes uzlabošanas paņēmienus, iespējams izveidot adaptīvu datu pirmsapstrādes metodoloģiju, kas efektīvi pielāgojas dažāda veida datu anomālijām un nevienmērībai, nodrošinot stabilāku un precīzāku rezultātu turpmākajās analīzes un modelēšanas fāzēs.
- Ir iespējama nepilnīgo datu interaktīvā apstrāde un vizualizācija, izmantojot izstrādātās datu apvienošanas un datu kvalitātes uzlabošanas metodes.

Promocijas darba struktūra un apjoms

Promocijas darbs ir sarakstīts latviešu valodā, satur anotāciju, ievadu, 6 nodaļas, secinājumus, literatūras sarakstu, 58 attēlus, 15 tabulas, 3 pielikumus, kopā veidojot 140 lappuses. Darbā veiktas atsauces uz 318 literatūras avotiem.

1. PĒTIJUMA AKTUALITĀTE

IoT, kas apvieno sensorus, aparatūru un programmatūru, ir kļuvis par svarīgu informācijas sistēmu elementu (Nižetić et al., 2020). Visprogresīvākās *IoT* lietojumu jomas ir saistītas ar Industriju 4.0 (Osterrieder et al., 2020), viedās pilsētas koncepciju (Eremia et al., 2017), transportu (Porru et al., 2020) un lauksaimniecību (Villa-Henriksen et al., 2020). *IoT* sensoru tīkli darbojas trīs virzienos: uztver informāciju no vides, seko līdzi sistēmas iekšējiem procesiem un pārveido datus lēmumu pieņemšanai (Govinda & Saravanaguru, 2016; Sanyal & Zhang, 2018). Sistēmām raksturīga universāla savienojamība un dinamiskums (Patel et al., 2016; El-Mawla & Badawy, 2023), kas ir būtiski starpnozaru uzraudzības un pārvaldības sistēmu izstrādē (M. Zhang et al., 2021). Kviesis un Zacepins (2015) norāda uz sensoru sistēmu tehniskajiem ierobežojumiem – ierobežotu skaitļošanas jaudu un atmiņas resursiem. *IoT* sistēmās galvenie izaicinājumi ir trokšņa samazināšana (G. Han et al., 2022), noviržu noteikšana (Gaddam et al., 2020), trūkstošo datu aizvietošana (Yuehua Liu et al., 2020) un datu apkopošana (Sanyal & Zhang, 2018). Multimodālā datu apvienošana apvieno dažādu avotu informāciju vienotā formā (Castanedo, 2013), un to izmanto klasifikācijai, regresijai, klasteru veidošanai un dimensiju samazināšanai (Bokade et al., 2021).

IoT kontekstā "dati" primāri attiecas uz neapstrādātiem sensoru signāliem vai rādījumiem (Jifa & Lingling, 2014), bieži sajaucoties ar citu informāciju plašākā

informācijas telpā (Nasution et al., 2021). Datu apvienošanas process ietver noteiktus datu objektus, kas var būt gan neapstrādāti signāli no ierīcēm, gan apstrādāti dati, kas piesaistīti reālam objektam (Beddar-Wiesing & Bieshaar, 2020). *DIKW* (*Data, Information, Knowledge, Wisdom*) modelis (Bellinger et al., 2004) skaidro, kā datu attiecību analīze rada informāciju, informācijas modeļu izpēte veido zināšanas, un zināšanu pamatprincipu apguve noved pie gudrības. Šis modelis tiek izmantots lielo datu jautājumu risināšanai (Nasution et al., 2021) un tiek uzskatīts par datu apvienošanas modeli (Becerra et al., 2021).

JDL modelis (White & Steinberg, 1998) definē apvienošanu kā datu un informācijas asociēšanu, korelāciju un kombinēšanu no viena vai vairākiem avotiem. Sensoru sistēmu kontekstā process nodrošina precīzus pozīcijas aprēķinus, identitātes noteikšanu un situācijas novērtējumu (Hall & Llinas, 1997). Šīs definīcijas nošķir neapstrādātus datus no apstrādātas informācijas, veidojot pamatu mūsdienu datu apstrādes hierarhijai (Castanedo, 2013).

Khaleghi u. c. un M. *Kumar* u. c. (Khaleghi et al., 2013; M. Kumar et al., 2006) atzīst, ka sensoru sniegtajos datos vienmēr pastāv mērījumu neprecizitāte un nenoteiktība, radot datu nepilnības. *Lakshmanarao* un *Shashi* (Lakshmanarao & Shashi, 2020) identificē galvenos izaicinājumus: sensoru datu kvalitāte, trokšņi, vides interference un sistemātiskās kļūdas. Nopietnas problēmas rodas sistēmās, kas izmanto *Dempster-Shafer* (DST) teoriju, kad parādās pretrunīgi mērījumi. Papildu sarežģījumus rada sensoru heterogenitāte un reāllaika apstrādes nepieciešamība. Vides faktoru ietekme uz sensoru mērījumiem (M. Kumar et al., 2006) rada sistemātiskas novirzes datu kopās. *Liu* u. c. (J. Liu et al., 2020) piedāvā heterogēno datu klasifikāciju telpiskajā, temporālajā, statiskajā, dinamiskajā un atribūtu dimensijās. Bezvadu sensoru tīklos dominē decentralizētās arhitektūras pieeja (Debauche et al., 2018; Kviesis & Zacepins, 2015; Murakami et al., 2007), kas nodrošina lokālu datu apstrādi noviržu novēršanai. *Khaleghi* u. c. (Khaleghi et al., 2013) identificē reģistrācijas datu izlīdzināšanas problēmu, kas rodas, transformējot lokālo sensoru datus vienotā atskaites sistēmā. Temporālie aspekti rada īpašus izaicinājumus vairāku sensoru sistēmās, īpaši ārpus secības saņemto datu apstrādē (Besada-Portas et al., 2011).

Datu apvienošana tiek izmantota, lai risinātu šādus uzdevumus: problemātisko datu korekciju (C. Huang et al., 2019), datu uzticamības uzlabošanu (Hong et al., 2009; Kreibich et al., 2014), datu pilnīguma palielināšanu (Consoli et al., 2015) un augstāka līmeņa informācijas iegūšanu (Jayasinghe et al., 2019). *Kreibich* u. c. (Kreibich et al., 2014) norāda, ka datu ticamība būtiski samazinās nekontrolētā vidē ar augstu trokšņu līmeni. *W. Wang* u. c. (W. Wang et al., 2018) demonstrē vairāku sensoru datu apvienošanas efektivitāti vides monitoringa sistēmās, kur integrētie dati atklāj netriviālas sakarības. *Karkouch* u. c. (Karkouch et al., 2016) norāda uz faktoriem, kas ietekmē datu kvalitāti *IoT*: sistēmu mērogs, resursu ierobežojumi, tīkla arhitektūra, vides apstākļi, sensoru stāvoklis, drošības ievainojamība un datu plūsmas apstrāde. *Aboubakar* u. c. (Aboubakar et al., 2022) raksturo *IoT* kā *IP* tīklu ar paaugstinātu nestabilitāti. *Adelantado* u. c., un *Dinculeană* un *Cheng*

(Adelantado et al., 2017; Dinculeană & Cheng, 2019) uzsver *IoT* ierīču specifiku – neliela apjoma ziņojumu pārraide. *Sliwa* u. c. (Sliwa et al., 2020) identificē kritiskos šķēršļus – ierobežoto enerģiju un atmiņu, kas kavē drošības risinājumu ieviešanu.

Drošības aspektā sensoru ierīces praktiski nav iespējams pilnvērtīgi aizsargāt, jo tās nespēj veikt resursietilpīgās kriptogrāfiskās operācijas. Vides ietekme un tehniskās problēmas, tostarp zemākas kvalitātes sensoru precizitāte un kalibrācijas trūkumi, būtiski ietekmē sensoru darbību. Traucējumu novēršanai izstrādātās metodes ietver trokšņa samazināšanu, trūkstošo datu aizpildīšanu un interpolāciju, datu noviržu noteikšanu un datu apkopošanu (Karkouch et al., 2016; Souza & Amazonas, 2015). Troksnis signālu apstrādē izpaužas kā nekorelētas sastāvdaļas, ko Jcgm (Jcgm, 2008) raksturo kā mērījumu rezultātu izkliedes parametru. *Teh* u. c. (Teh et al., 2020) skaidro nenoteiktību kā kļūdu kvantitatīvo izteiksmi, savukārt *Krishnamurthi* u. c. (Krishnamurthi et al., 2020) uzsver trokšņa negatīvo ietekmi uz sistēmas resursiem. Trokšņu mazināšanas metodoloģija izmanto bīdāmā loga principu. *Vanus* u. c. (Vanus et al., 2020) norāda uz šo metožu ierobežojumiem lokālo frekvenču pielāgošanā. *M. M. Rashid* u.c. (M. M. Rashid et al., 2015) izceļ divas galvenās viļņu transformācijas pieejas: nepārtraukto transformāciju laika un frekvences dimensijās, un diskrēto transformāciju laika dimensijas analīzei.

Adhikari u. c. (Adhikari et al., 2022) norāda, ka *IoT* sistēmās datu nepilnīgums ir izplatīta parādība, kas apdraud analītiskā procesa ticamību. *Zhang* u. c. (Zhang et al., 2024) piedāvā metodisku risinājumu, ieviešot izsekošanas noņemto autoenkoderu (*TRAE, Tracking-Removed Autoencoder*) un faziklasterizācijas (*FC, Fuzzy Clustering*) metodes. *Krishnamurthi* u. c. (Krishnamurthi et al., 2020) identificē trīs fundamentālus trūkstošo datu tipus: *MCAR* (*Missing Completely at Random*), *MAR* (*Missing at Random*) un *NMAR* (*Not Missing at Random*). Izolācijas meža (*IF, Isolation Forest*) algoritms ir kļuvis par vienu no vadošajiem risinājumiem noviržu noteikšanai *IoT* datu kopās. *Muñoz* u. c. (Muñoz et al., 2024) norāda uz algoritma precizitātes uzlabošanas iespējām, izmantojot pielāgojamus parametrus. Lokālais ārpusnieku faktors (*LOF, Local Outlier Factor*) novērtē novirzes, analizējot tuvāko kaimiņu datu blīvumu, savukārt izolācijas meža algoritms identificē globālas novirzes. *Kim* u. c. (Kim et al., 2022) piedāvā adaptīvas mašīnmācīšanās metodes, kas ne tikai atpazīst novirzes, bet arī veic automātisku datu korekciju, izmantojot vēsturisko datu analīzi.

Balstoties uz nodaļā veikto analīzi, īpaši par datu kvalitātes problēmām un to risinājumiem *IoT* sistēmās, autors secina, ka tiek apstiprināta tēze:

> Lietu interneta (*IoT*) sistēmās iegūtie datu analīzes rezultāti ir atkarīgi no datu kvalitātes uzlabošanas metožu pielietošanas, kas īpaši svarīgi ierobežota datu apjoma vai nepilnīgu datu gadījumos, nodrošinot ticamu un precīzu lēmumu pieņemšanu.

Pamatojums

Veiktā analīze atklāj aspektus, kas apstiprina šo tēzi. *Karkouch* u. c. (Karkouch et al., 2016) identificē kritiskos faktorus, kas ietekmē *IoT* datu kvalitāti – sistēmu mērogu, resursu ierobežojumus, vides apstākļus un sensoru tehniskās īpatnības. Šos ierobežojumus pastiprina *IoT* ierīču specifika, ko uzsver *Adelantado* u. c., un *Dinculeană* un *Cheng* (Adelantado et al., 2017; Dinculeană & Cheng, 2019) – ierīces spēj pārraidīt tikai neliela apjoma ziņojumus, darbojas ar ierobežotu enerģiju un atmiņu.

Karkouch u. c. (Karkouch et al., 2016) norāda, ka vides faktori spēcīgi ietekmē sensoru darbību – ekstremālos apstākļos un ierobežotas apkopes situācijās datu kvalitāte var būtiski pazemināties. *Kreibich* u. c. (Kreibich et al., 2014) pierāda, ka datu ticamība ievērojami samazinās nekontrolētā vidē, kas ir tipiska *IoT* sistēmām. Šo novērojumu papildina *Consoli* u. c. (Consoli et al., 2015) secinājumi – ierobežota apjoma mērījumi no atsevišķiem sensoriem bieži vien nesniedz pietiekamu informāciju sistēmas stāvokļa novērtēšanai.

Aplūkotie risinājumi sniedz praktiskus instrumentus šo problēmu novēršanai. *C. Huang* (C. Huang et al., 2019) izstrādātās metodes problemātisko datu korekcijai, *Hong* (Hong et al., 2009) piedāvātie risinājumi datu uzticamības uzlabošanai un *Consoli* u. c. (Consoli et al., 2015) darbs datu pilnīguma palielināšanai demonstrē tiešu saikni starp datu kvalitātes uzlabošanas metodēm un precīzāku lēmumu pieņemšanu. *W. Wang* u. c. (W. Wang et al., 2018) pētījums vides monitoringa sistēmās pārliecinoši parāda, ka tieši datu kvalitātes uzlabošanas metožu pielietošana ļauj atklāt būtiskas sakarības datos.

Pētījuma gaitā atklātās datu kvalitātes problēmas un to risinājumi apstiprina tēzes pamatotību. Papildus iepriekšminētajam vēl *Gomathi u. c.* (Gomathi et al., 2018) norāda uz *IoT* tehnoloģiju radītajiem izaicinājumiem, kas ietver gan sensoru bojājumus, gan sistēmiskas problēmas. *Adhikari* u. c. (Adhikari et al., 2022), pētījumā konstatēts, ka līdz pat 40% visu datu var būt nepilnīgi, kas būtiski ietekmē analīzes rezultātus. *Martin* u. c. (Martin et al., 2023) demonstrē, ka viļņu transformācijas, adaptīvās mašīnmācīšanās un modernās trūkstošo datu aizvietošanas metodes nodrošina efektīvus risinājumus šo problēmu pārvarēšanai, uzlabojot lēmumu pieņemšanas precizitāti *IoT* sistēmās ar ierobežotu vai nepilnīgu datu apjomu.

Datu apvienošana ir process, kurā dažādas datu plūsmas tiek apvienotas, lai radītu informatīvāku un pielāgotu izvadi (Bokade et al., 2021). Datu apvienošanas veidi tiek klasificēti pēc sistēmu arhitektūras, abstrakcijas līmeņiem un apvienošanas mērķiem.

Khaleghi u. c. (Khaleghi et al., 2013) veic sistemātisku metožu izpēti, sniedzot detalizētu analīzi. *Castanedo* (Castanedo, 2013) koncentrējas uz trim algoritmiskiem aspektiem: datu asociāciju, sistēmas stāvokļa novērtējumu un lēmumu apvienošanu. Zheng (Y. Zheng, 2015) piedāvā inovatīvas metodes starpdomēnu datu integrācijai, Atluri u. c. (2018) izstrādā metodisko ietvaru telpisko un laika datu ieguvei, bet *Qin* u. c. (Qin et al., 2020) pievēršas *IoT* specifikai.

Datu apvienošanas klasifikācija balstās uz abstrakcijas līmeņiem, datu avotu un ieejas-izejas attiecībām (Becerra et al., 2021; Dasarathy, 1997). Trīs galvenās klasifikācijas metodes ir *Dasarathy* klasifikācija (Dasarathy, 1997), *Whyte* klasifikācija (Grime & Durrant-Whyte, 1994) un JDL modeļa klasifikācija (Steinberg & Bowman, 2008).

Sistēmu arhitektūrā izšķir četrus galvenos veidus. Centralizētā arhitektūra izmanto vienotu centrālo procesoru, bet saskaras ar ierobežojumiem vizuālo sensoru tīklos. Decentralizētajā arhitektūrā katrs sensors darbojas autonomi, bet rodas augošas komunikācijas izmaksas – $O(n^2)$ (Grime & Durrant-Whyte, 1994). Sadalītā arhitektūra piedāvā līdzsvarotāku pieeju, kur avota mezgli vispirms apstrādā savus mērījumus. Hierarhiskā arhitektūra apvieno iepriekšējās pieejas (Castanedo, 2013).

Beddar-Wiesing un *Bieshaar* (Beddar-Wiesing & Bieshaar, 2020) norāda, ka decentralizētā pieeja, neskatoties uz izaicinājumiem, bieži ir izvēles pamatā tās noturības dēļ. *Becerra* u. c. (Becerra et al., 2021) identificē trīs sensoru datu apvienošanas scenārijus: konkurētspējīgo (vienas modalitātes sensoru integrācija), papildinošo (dažādu sensoru papildinoši dati) un kooperatīvo (sensoru dinamiska pielāgošana).

Whyte klasifikācija piedāvā trīs sadarbības veidus. Papildinošā mijiedarbība ietver dažādu sensoru atšķirīgu datu sniegšanu, piemēram, vides monitoringa sistēmās. Datu dublēšanās gadījumā vairāki sensori vēro vienu mērķi, uzlabojot precizitāti. Kooperatīvā mijiedarbība apvieno dažādu sensoru datus jaunas informācijas iegūšanai, piemēram, 3D LiDAR un video kameru integrācijā.

Dasarathy klasifikācija (Dasarathy, 1997) strukturē apvienošanu piecos līmeņos: DAI-DAO (neapstrādātu datu apvienošana), DAI-FEO (datu transformācija pazīmēs), FEI-FEO (pazīmju apstrāde), FEI-DEO (lēmumu pieņemšana) un DEI-DEO (lēmumu sintēze). *Varshney* (Varshney, 1997) papildināja to ar DAI-DEO līmeni tiešai pārejai no datiem uz lēmumiem.

JDL modelis piedāvā piecu līmeņu hierarhisku struktūru (Llinas et al., 2004). Tas sākas ar sensoru priekšapstrādi, turpinās ar objektu stāvokļu novērtēšanu un attiecību analīzi, un noslēdzas ar darbību ietekmes analīzi un resursu pārvaldību. *Blasch* un *Plano* (Blasch & Plano, 2002) papildināja modeli ar lietotāja mijiedarbības līmeni.

Mūsdienu klasifikācija (Abdelgawad & Bayoumi, 2012; Barbedo, 2022) izdala trīs fundamentālus līmeņus. Zemā līmeņa apvienošana operē ar neapstrādātiem datiem (Di Natale et al., 2002). Vidējā līmeņa apvienošana veic pazīmju ekstrakciju (Biancolillo et al., 2014), savukārt augstā līmeņa apvienošana veido atsevišķus analīzes modeļus (L. Huang et al., 2014).

Varbūtību teorijā balstītās metodes, īpaši Beijesa secinājumi, nodrošina sistemātisku pieeju datu apvienošanai (Pires et al., 2016). Tomēr I*IoT* sistēmās tradicionālās metodes zaudē efektivitāti. *Medjahed* u. c. (Medjahed et al., 2011) identificē galvenos izaicinājumus – ierobežotu veiktspēju daudzfaktoru datos. *Rezatofighi* u. c. (Rezatofighi et al., 2015) analizē varbūtiskās datu asociācijas priekšrocības objektu izsekošanā.

Pētījums parāda, ka universāla metode nepastāv – katrs risinājums ir piemērots specifiskiem uzdevumiem. Vienkāršākiem gadījumiem pietiek ar zemā līmeņa apvienošanu, sarežģītākiem efektīvāka ir augstāka līmeņa pieeja. *IoT* sistēmās mūsdienīgās metodes spēj ievērojami uzlabot datu kvalitāti un lēmumu ticamību.

Iepriekš aplūkotās IoT datu kvalitātes problēmas un to risinājumi ir īpaši aktuāli specifiskās nozarēs, kur precīzi un savlaicīgi dati ir kritiski svarīgi lēmumu pieņemšanai. Autora pētījumi trīs dažādās jomās – precīzajā biškopībā, pilsētas satiksmes uzraudzībā un putnkopībā – demonstrē praktiskus risinājumus iepriekš identificētajiem izaicinājumiem. Katrā no šīm jomām sastopamas tipiskās IoT sistēmu problēmas – sensoru datu nepilnības, multimodālo datu integrācijas sarežģījumi un reāllaika apstrādes prasības. Šo problēmu risināšanai autors izmanto un paplašina iepriekš aprakstītās datu apvienošanas metodes, pielāgojot tās nozaru specifiskajām vajadzībām un ierobežojumiem. Īpaša uzmanība pievērsta tieši tām metodēm, kas spēj efektīvi darboties ierobežotu resursu apstākļos, vienlaikus nodrošinot augstu datu kvalitāti un ticamību.

Biškopības jomā autors veicis padziļinātu izpēti par datu vākšanas un apstrādes procesiem. Precīzā biškopība ir inovatīva dravas pārvaldības stratēģija, kas nodrošina bišu saimju monitoringu resursu optimizācijai (Zacepins et al., 2012). Datu vākšana ietver fizisko parametru mērīšanu, izmantojot sensorus stropos (Kviesis & Zacepins, 2015), kas mēra temperatūru, mitrumu, gāzu sastāvu, vibrāciju un skaņu. Dati tiek analizēti trīs līmeņos (Human et al., 2013; Zacepins & Stalidzans, 2013): dravas līmenī (meteoroloģiskie dati, video novērojumi), kolonijas līmenī (temperatūra, mitrums, svars) un individuālā līmenī (bišu uzvedība). Bezvadu tīkla tehnoloģijas (Debauche et al., 2018) bieži rada datu nepilnības, kuru risināšanai izmanto datu apvienošanas metodes. Mūsdienīgi risinājumi izmanto *IoT* ierīces un LSTM neironu tīklus spietošanas prognozēšanai (Kwon, Cho, et al., 2019), apvienojot temperatūras un skaņas datus ar Kalmana filtra algoritmu. Šāda pieeja ļauj ne tikai optimizēt resursu patēriņu, bet arī agrīni identificēt potenciālās problēmas bišu kolonijās. Datu apvienošana, kas nozarē līdz šim nav plaši pielietota, var uzlabot bišu veselības uzraudzību un prognozēt būtiskus stāvokļus – spietošanu un koloniju sabrukšanas sindromu. Autors analizē datus dravas, bišu saimju un individuālā līmenī, piedāvājot risinājumus, kas balstīti uz sensoru datiem un LSTM neironu tīkliem, lai optimizētu resursu patēriņu un produktivitāti.

Pilsētas satiksmes novērošanā *IoT* risinājumi izmanto videonovērošanas aparatūru kopā ar objektu noteikšanas un izsekošanas algoritmiem (N. Chen & Chen, 2018). Mūsdienu transporta sistēmu digitalizācija rada jaunas iespējas datu apvienošanas jomā (Neumann et al., 2016). Autors izstrādāja sinhronizācijas algoritmu *LiDAR* un videokameru datu apvienošanai, fokusējoties uz precīzu dažādu avotu sinhronizāciju ar laika zīmogiem. Algoritms ņem vērā transportlīdzekļu kustību, laikapstākļus un aparatūras konfigurāciju, lai samazinātu kļūdas ātruma noteikšanā un transportlīdzekļu klasifikācijā. Jelgavas pilsētā veiktā validācija (Bumanis et al., 2021) ilga 6 mēnešus (01.04.2021.-

30.09.2021.), izmantojot *LiDAR* sensoru un četras videokameras divu satiksmes virzienu uzraudzībai. Datu apstrāde organizēta divās paralēlās plūsmās – videokameru dati Microsoft Azure BLOB krātuvē un *LiDAR* mērījumi *XML* formātā *FTP* serverī, ar *Site-to-site VPN IKEv2* protokola aizsardzību. Datu integrācijas metodoloģija (Vitols et al., 2021) balstās uz pazīmju līmeņa datu apvienošanu, kur katrs transportlīdzeklis tiek apstrādāts kā notikums. Validācijas rezultāti uzrādīja augstu numura zīmju atpazīšanas precizitāti (>99%) un sasniedza 97% sinhronizācijas precizitāti, kas ir kritiski svarīgi transportlīdzekļu identifikācijai un pārkāpumu novērošanai. Sistēmas optimizācija ietvēra paralēlo attēlu uzņemšanu redzamajā un infrasarkanajā spektrā, datu plūsmas optimizāciju un koordinātu sistēmas konfigurāciju. Eksperiments atklāja trīs būtiskus aspektus: sensoru izvietojuma kritisko nozīmi, nepieciešamību pēc dinamiskās parametru pielāgošanas, un sistēmas dublēšanas svarīgumu precizitātes uzlabošanai.

Putnkopības nozarē autors izstrādājis kompleksu datu glabāšanas risinājumu, izveidojot datu noliktavu vides monitoringa sensoru un ražošanas ciklu datu apstrādei. Sistēma balstās uz sniegpārslu shēmas principiem, nodrošinot gan saimniecības efektīvu pārvaldību, gan atbilstību ES regulām par dzīvnieku labturību. Kibernētiski-fiziskais modelis ietver sensoru kopumu, datu apmaiņas kontrolierus un analītisko centru, koncentrējoties uz barības procesu optimizāciju un vides parametru kontroli. Centrālā datu noliktavas struktūra sastāv no [productionLog] faktu tabulas un saistītām dimensiju tabulām sensoru datiem, barības datiem un olu ražošanai, kā arī [standardValues] tabulas references vērtībām. Risinājums veiksmīgi ieviests divās Baltijas putnu fermās CO_2 un NH_3 līmeņa uzraudzībai, izmantojot Microsoft Azure platformu automātiskai datu vākšanai un apstrādei. Sistēmā uzstādīti NH_3 sensori 2,5m augstumā un CO_2 sensori 0,4m augstumā, kopumā izvietojot 6 sensoru pārus ar analizatoru un mikrokontrolieru sistēmām (2020-2021).

Visās trijās nozarēs tika konstatēts, ka nepilnīgas vai nekvalitatīvas datu ievades ietekmē prognožu veikšanu un resursu optimizāciju. Datu kvalitātes uzlabošanai tika izmantotas sensoru kalibrēšana, laika zīmogu sinhronizācija un datu apstrādes algoritmi.

Veiktā analīze atklāj būtisku nepieciešamību pēc jaunas, specializētas datu apvienošanas metodes, kas spētu efektīvi risināt identificētās *IoT* datu kvalitātes problēmas. Lai gan esošās metodes, piemēram, *TRAE* un *FC* (Zhang et al., 2024), vai *IF* un *LOF* algoritmi (Muñoz et al., 2024), piedāvā risinājumus atsevišķiem izaicinājumiem, joprojām trūkst vienota, sistemātiska risinājuma, kas vienlaicīgi aptvertu gan datu kvalitātes uzlabošanu, gan efektīvu datu apvienošanu ierobežotu resursu apstākļos. Īpaši aktuāla ir nepieciešamība pēc metodes, kas spētu pielāgoties dažādām nozarēm un datu tipiem, vienlaikus saglabājot augstu precizitāti un uzticamību. Kā norāda *Karkouch* u. c. (Karkouch et al., 2016) un *Kreibich* u. c. (Kreibich et al., 2014), datu kvalitātes problēmas ir īpaši izteiktas nekontrolētā vidē ar ierobežotiem resursiem, kas ir raksturīga *IoT* sistēmām. Šie faktori pamato nepieciešamību izstrādāt jaunu, uz datu slāņošanu balstītu metodi,

kas ne tikai risinātu datu kvalitātes problēmas, bet arī nodrošinātu efektīvu dažādu avotu datu apvienošanu, ņemot vērā gan telpiskos, gan laika aspektus.

2. DATU SLĀŅOŠANAS KONCEPTUĀLĀ METODE

Sistemātiska pieeja datu apvienošanas metožu analīzei atklāj, ka datu priekšapstrādes posms ir kritisks faktors, kas būtiski ietekmē gala rezultāta kvalitāti. Vienota modeļa izmantošana vairākām datu kopām var radīt specializācijas efektu, tādēļ metodes pielāgošanās spēja ir izšķiroša darbā ar dažādiem vēsturisko datu avotiem. Datu slāņošanas pieeja kombinē telpisko-laika analīzi ar elipsoidālo metodi. *Abu Bakr* un *Lee* (Abu Bakr & Lee, 2017) apraksta pieejas īstenošanu, organizējot datus dimensiju slāņos, kur katrs slānis reprezentē noteikta laika posma informāciju. Elipsoidālās metodes principi nosaka datu pārklāšanās zonu identificēšanu. Metode ir izmantota bišu barības meklēšanas uzvedības analīzē, apvienojot dažādu avotu datus. Pētījumi apliecina, ka veiksmīgai datu apvienošanas metodei jāietver trīs galvenie elementi: datu sagatavošanas iespējas, efektīva vēsturisko datu izmantošana un sistemātiska kvalitātes kontrole. Šo elementu integrācija ir būtiska *IoT* sistēmās un citās datu intensīvās nozarēs, kur tie tieši ietekmē gan analīzes rezultātus, gan lēmumu pieņemšanas procesu.

Bišu barības optimizācijai nepieciešamie dati ietver reģionālo informāciju (atrašanās vieta, reljefs, klimats), vietējo nektāraugu raksturojumu un vides apstākļus, kas ietekmē nektāra veidošanos (X. J. He et al., 2016; Hennessy et al., 2020). Reģionālās biškopības organizācijas apkopo šos datus ziedēšanas kalendāros, tomēr to izmantošanu ierobežo datu formātu un aptvēruma neviendabība. Metodes validācijai izvēlēti četri raksturīgi augi ar atšķirīgām īpašībām: *Grevillea robusta* (konstants ziedēšanas periods, vidēja produkcija), *Coffea arabica* (sezonāla ziedēšana, mainīga produkcija), *Eucalyptus citriodora* (resursi visa gada garumā) un *Dichrostachys cinerea* (zema intensitāte). Šī izvēle ļauj pārbaudīt metodes efektivitāti dažādos resursu pieejamības scenārijos. Datu apstrādē izmanto divas galvenās metodes: normalizāciju (nenormālam sadalījumam) un standartizāciju (normālam sadalījumam). Normalizācija nodrošina objektīvu resursu novērtējumu starp dažādiem augiem, piemēram, salīdzinot *Dichrostachys cinerea* un *Coffea arabica* datus. Standartizācija, kas pielietota *Eucalyptus citriodora* un *Grevillea robusta* datiem, sniedz precīzu resursu svārstību kvantitatīvo novērtējumu.

Efektīvai bišu dravas pārvaldībai nepieciešama kompleksa izpratne par vairākiem savstarpēji saistītiem faktoriem. Precīzās biškopības metodoloģija balstās uz trīs galveno datu slāņu mijiedarbību, kas kopā veido pamatu informētai lēmumu pieņemšanai. Pirmais un nozīmīgākais ir augu ziedēšanas slānis, kas balstīts uz detalizētiem ziedēšanas kalendāra datiem (Tree Flowering Calendar, 2020). Šis slānis sniedz būtisku informāciju par nektāra un ziedputekšņu pieejamību dažādos gada periodos, ļaujot biškopjiem precīzi plānot dravas izvietojumu un resursu izmantošanu. Otro slāni veido nokrišņu dati, kuru nozīmi

bišu barības meklēšanas efektivitātē ir apstiprinājuši vairāki pētījumi (X. J. He et al., 2016). Nokrišņu daudzums un sadalījums gada griezumā būtiski ietekmē gan nektāra izdalīšanos, gan bišu spēju piekļūt barības resursiem. Trešais slānis atspoguļo bišu koloniju aktivitātes ciklus, kas dokumentēti biškopības kalendārā (Beekeeping Calendar, n.d.). Šis slānis ir īpaši nozīmīgs stropu ikdienas pārvaldībā, jo ļauj prognozēt un plānot tādas būtiskas aktivitātes kā peru audzēšana, spietošana un intensīvie barības vākšanas periodi.

Kad ir zināmi interesējošie slāņi, ir iespējams noteikt un izvadīt derīgu informāciju. Pirmkārt, tiek pielietota interpolācija, kas palīdz palielināt datu precizitāti, padarot tos vienmērīgākus un sniedzot iespēju detalizētāk skatīt tendences. Attiecīgi, katram parametram tiek palielināta datu punktu izšķirtspēja, izmantojot lineāro interpolāciju.

Ņemot vērā datu punktu kopu $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, lineārā interpolācija ir y vērtības noteikšanas process konkrētam x , izmantojot formulu (sk. (2.1.)):

$$y = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) \tag{2.1.}$$

kur

x_0, x_1 – konkrēti x vērtības punkti, starp kuriem tiek veikta interpolācija, reālie skaitļi;

y_0, y_1 – šo x vērtību punktu atbilstošās y vērtības, reāli skaitļi;

y – aprēķinātā y vērtība, kas atbilst x vērtībai. Tas ir rezultāts no interpolācijas procesa, reālais skaitlis.

Tad katram parametram vērtības tiek reizinātas ar atbilstošajiem svariem un pēc tam normalizētas līdz diapazonam [0, 100]. Katras svērtās vērtības formula ir šāda (sk. (2.2.)):

$$\text{weighted\_value}_i = \sum_{j=1}^{n} \text{weight}_j \times \frac{\text{parameter}_j}{100} \tag{2.2.}$$

kur

n – parametru skaits, vesels skaitlis;

weight_j – svērums, kas piešķirts katrai vērtībai, reāls skaitlis;

parameter_j – konkrēts parametrs, kuram tiek piešķirts svērums, vesels vai decimālais skaitlis;

weighted_value_i – aprēķinātā svērtā vērtība kādam *i* parametram, %.

Pēc svērtās vērtības aprēķināšanas tiek pielietota galveno komponentu analīze (PCA). Šī statistiskā metode pārveido sākotnējos, iespējami savstarpēji saistītos mainīgos, jaunā koordinātu sistēmā, kur tie kļūst savstarpēji neatkarīgi. Šos jaunos, neatkarīgos mainīgos sauc par galvenajiem komponentiem. Ņemot vērā datu matricu X, pirmo galveno komponentu var atrast šādi:

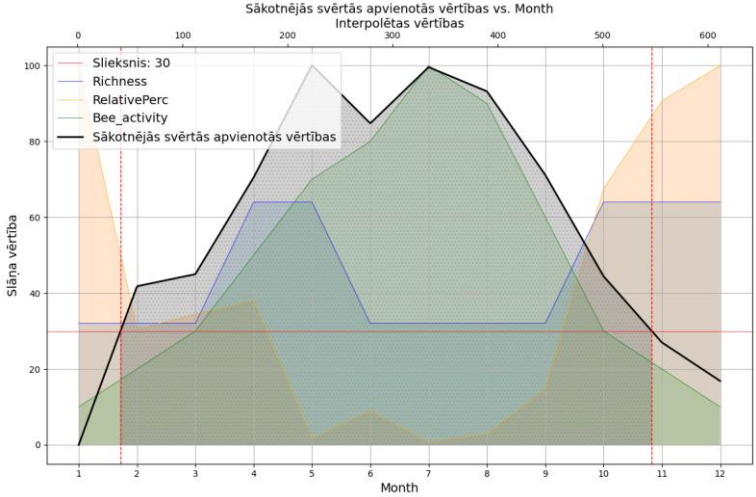
1. tiek atņemta vidējā vērtība: $X_{mean} = X - \bar{X}$;

2. tiek aprēķināta kovariācijas matrica Σ no X_{mean};
3. tiek aprēķināts Σ īpašvektors un īpašvērtība;
4. tiek izvēlēts pazīmju vektors, kas saistīts ar lielāko īpašvērtību.

Šajā kontekstā *PCA* tiek izmantots, lai no kombinētajiem parametriem iegūtu nozīmīgāko tendenci, tas ir, lai iegūtu vienu kombinētu vērtību (uz *PCA* balstīta apvienotā vērtība), kas atspoguļo vislielākās atšķirības no visiem parametriem.

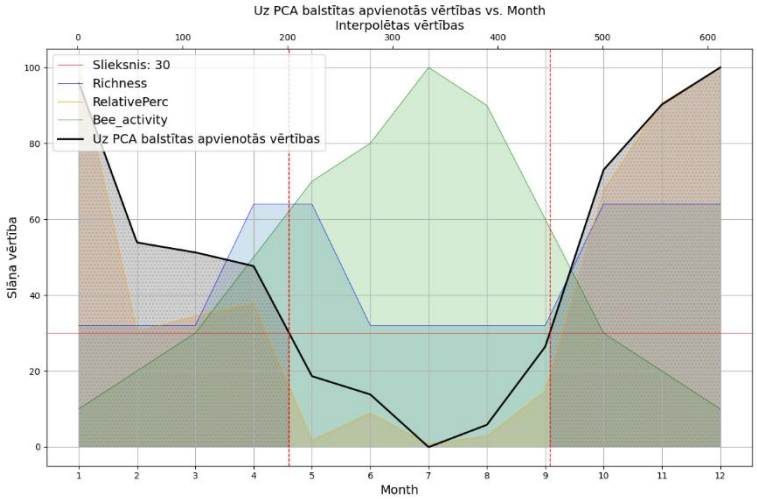
Gan svērtās, gan uz *PCA* balstītās apvienotās vērtības tiek salīdzinātas ar slieksni. Reģioni, kuros šīs vērtības pārsniedz slieksni, ir iezīmēti diagrammā. Šī ir vienkārša nosacījuma pārbaude: ja fused_value ≥ slieksnis, reģions tiek iezīmēts

Rezultātā tiek iegūtas divas diagrammas. Sākotnējo apvienoto vērtību diagramma, kur (sk. 2.1. att.) ir parādīti sākotnējie parametri, sākotnējās apvienotās vērtības un reģioni, kuros sākotnējās apvienotās vērtības pārsniedz slieksni. Tas sniedz ieskatu par to, kā vienkāršā svērtā parametru summa (sākotnējās apvienotās vērtības) darbojas dažādos mēnešos.



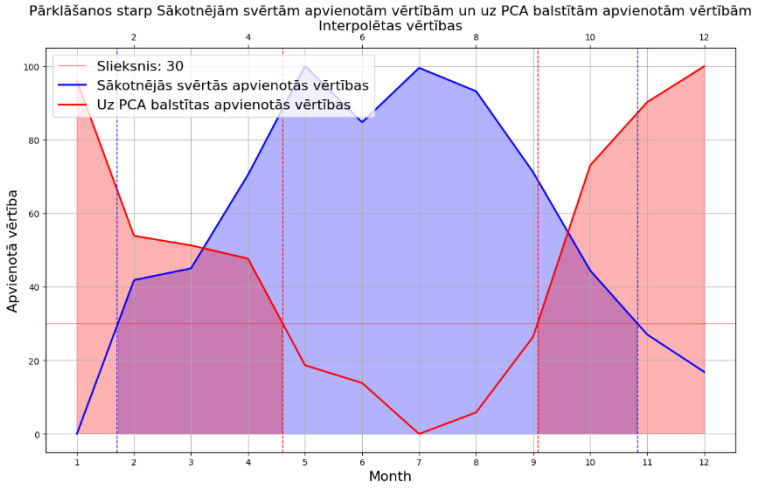
2.1. att. **Sākotnējo apvienoto vērtību diagramma.**

Uz *PCA* balstīto apvienoto vērtību diagrammā (sk. 2.2. att.) tiek parādīti sākotnējie parametri, uz *PCA* balstītās apvienotās vērtības un reģioni, kuros uz *PCA* balstītās apvienotās vērtības pārsniedz slieksni. Tas sniedz perspektīvu par to, kā uz *PCA* balstīta apvienošana, kas uztver vislielākās atšķirības no parametriem, darbojas mēnešu laikā.



2.2. att. **Uz *PCA* balstītu apvienoto vērtību diagramma.**

Izmantojot šīs diagrammas, var vizuāli salīdzināt (sk. 2.3. att.). abas datu apvienošanas metodes un izprast abus nozīmīgos reģionus (tos, kas pārsniedz slieksni).



2.3. att. **Apvienoto vērtību pārklāšanās.**

Turpmāk var pielietot trapecveida metodi. Pārklājot sākotnējo apvienoto vērtību un uz PCA balstīto apvienoto vērtību, trapecveida metode ļauj kvantitatīvi noteikt reģionus, kur šīs vērtības pārklājas (vai atšķiras). Izmērot laukumu starp šīm līknēm, var noteikt, cik lielā mērā abas apvienotās vērtības sakrīt (vai nesakrīt) noteiktā laika periodā. Jomas ar lielu pārklāšanos liecina par

spēcīgu vienošanos starp abām apvienošanas metodēm, savukārt atšķirību apgabali var norādīt uz interesējošiem reģioniem vai anomālijām.

Papildus var izcelt šādus aspektus:

- izmantojot slieksni, var identificēt un kvantitatīvi noteikt reģionus, kuros apvienotās vērtības pārsniedz noteiktu nozīmīguma līmeni. Trapecveida metode ļauj izmērīt šīs nozīmes lielumu, sniedzot skaitlisku vērtību, lai noteiktu, cik svarīgi vai ietekmīgi varētu būt noteikti laika periodi;
- trapecveida metode nodrošina iespēju kvantitatīvi salīdzināt abas apvienošanas metodes. Integrējot laukumus zem katras līknes un salīdzinot rezultātus, var pieņemt apzinātus lēmumus par to, kura apvienošanas metode varētu būt piemērotāka konkrētiem lietojumiem vai scenārijiem;
- trapecveida metode ir arī skaitļošanas ziņā efektīvi un vienkārši īstenojams paņēmiens. Ņemot vērā datu potenciāli lielo precizitāti (īpaši pēc interpolācijas), vienkārša, bet efektīva metode, nodrošina ātru analīzi bez ievērojamām skaitļošanas izmaksām;
- platības, kas aprēķinātas, izmantojot trapecveida metodi, var intuitīvi saprast. Lielāki apgabali norāda uz nozīmīgākiem notikumiem vai modeļiem datos, savukārt mazāki apgabali var norādīt uz mazāk ietekmīgiem periodiem.

Ņemot vērā divas y-vērtību kopas, *y1* un *y2*, kopējā *x* domēnā, jāaprēķina laukums starp abām līknēm. Ja viena līkne ir pilnībā virs otras, laukumu aprēķina kā starpību starp tām; ja tie krustojas, identificē pārklāšanās un nepārklāšanās sadaļas, lai aprēķinātu attiecīgos apgabalus.

Tālāk, kā tiek aprēķināti apgabali.

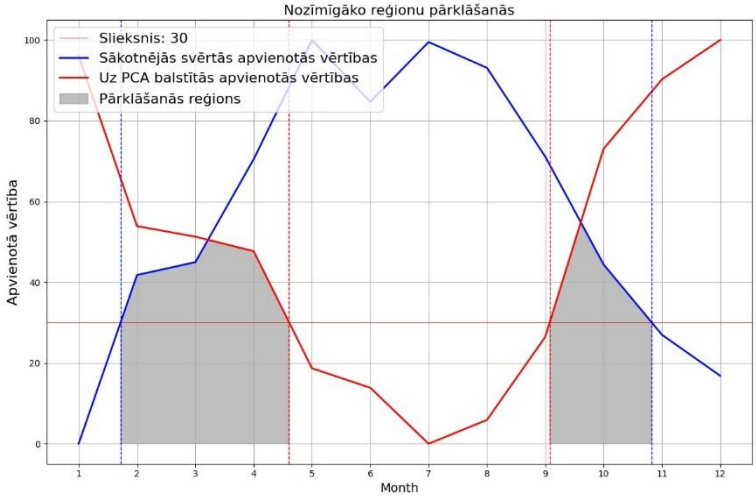
1. Tiek aprēķināta atšķirība starp divām datu kopām:
   - katram punktam x_i domēnā tiek aprēķināta atšķirība Δy_i starp abām datu kopām: $\Delta y_i = y1_i - y2_i$.
2. Tiek noteikti reģioni, kas pārklājas:
   - ja Δy_i un Δy_{i+1} ir viena un tā pati zīme, tad abas datu kopas nekrustojas starp x_i un x_{i+1};
   - ja Δy_i un Δy_{i+1} ir pretējas zīmes, tad abas datu kopas krustojas starp x_i un x_{i+1}, norādot reģionu, kas pārklājas.
3. Tiek aprēķināts laukums, izmantojot trapecveida metodi (sk. (2.3.)):
   - reģioniem, kas nepārklājas starp x_i un x_{i+1}:

$$Area = \frac{x_{i+1} - x_i}{2} \times (|y_{1i} + y_{1i+1}| - |y_{2i} + y_{2i+1}|) \qquad (2.3.)$$

   - reģioniem, kas pārklājas, šķērsošanas vietā laukums tiek sadalīts divās trapecēs, un šo trapecveida formu laukumi tiek summēti.
4. Kopējais pārklājuma laukums ir to laukumu summa, kas aprēķināti katram intervālam *x* domēnā.

Rezultātā tiek iegūta diagramma ar pārklājošiem reģioniem (sk. 2.4. att.). Veicot vizuālu analīzi, var secināt, ka:

- lielākas izmaiņas datos notiek starp 1. un 4. mēnesi un starp 9. un 12. mēnesi;
- labvēlīgākie apstākļi bišu saimes novietošanai uz analizējama auga ir starp 4. mēneša pēdējo nedēļu un 9. mēneša sākumu;
- novietošana agrāk, starp 2. un 4. mēnesi, vai vēlāk, starp 9. un 11. mēnesi, nav rekomendēta strauji mainīgo apstākļu dēļ.



2.4. att. **Apvienoto vērtību pārklājošie reģioni.**

Datu slāņošanas metode ir izstrādāta *Python* vidē, izmantojot pandas bibliotēku datu apstrādei, *numpy* skaitliskajiem aprēķiniem un *matplotlib* vizualizācijai. Metodes kodolu veido `data_fusion_main` funkcija, kas koordinē vairāku savstarpēji saistītu apakšfunkciju darbu.

Metodes darbības process sākas ar datu sagatavošanu, kur būtiska loma ir interpolācijas funkcijām. Funkcija `interpolate_data` veic lineāro interpolāciju starp datu punktiem, nodrošinot vienmērīgu datu plūsmu, savukārt `interpolate_data_auto_smoothing` automātiski meklē optimālo izlīdzināšanas pakāpi. Datu apvienošanu var veikt divos veidos – ar `weight_based_data_fusion`, kas izmanto lietotāja definētus svarus, vai ar `pca_based_data_fusion`, kas balstās uz galveno komponentu analīzi.

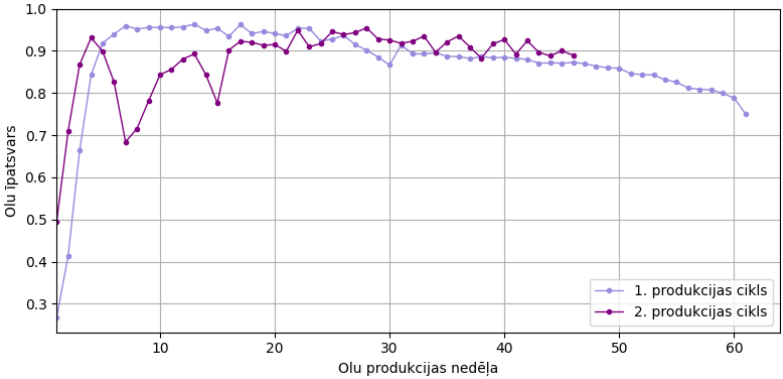
Metodes parametri ir organizēti trīs loģiskās grupās: datu avotu parametri (`DataFrames_dict_or_df`, `Data_params`, `CommonColumn`), analīzes kontroles parametri (`Weights`, `Layering_threshold`) un datu apstrādes opcijas (`Smoothing`, `AutoSmooth`, `FilterBy`, `FilterValue`). `DataFrames_dict_or_df` nosaka sākotnējo datu struktūru, nodrošinot elastīgu pieeju gan atsevišķu, gan apvienotu datu kopu apstrādei. `Data_params` precizē, kuras kolonnas tiks izmantotas analīzē, piemēram, nektāra daudzums vai

bišu aktivitāte, tādējādi fokusējot aprēķinus uz būtiskākajiem rādītājiem. `CommonColumn`, kas parasti ir laika ass (mēnesis), nodrošina datu sinhronizāciju un korektu salīdzināšanu starp dažādiem avotiem. `Weights` parametrs ļauj pielāgot katra datu slāņa ietekmi uz gala rezultātu, piemēram, piešķirot lielāku nozīmi nokrišņu datiem nekā bišu aktivitātei. `Layering_threshold` definē kritisko robežvērtību (parasti 30%), kas kalpo gan kā vizuāls atskaites punkts, gan kā automātiskās izlīdzināšanas kontroles mehānisms. `Smoothing` koeficients kontrolē datu interpolācijas precizitāti, ietekmējot līkņu gludumu un detalizācijas pakāpi. `AutoSmooth` funkcionalitāte automātiski pielāgo izlīdzināšanas intensitāti, balstoties uz noteikto slieksni, tādējādi optimizējot datu reprezentāciju. `FilterBy` un `FilterValue` parametri nodrošina iespēju fokusēties uz specifiskiem datu segmentiem, piemēram, konkrēta auga vai laika perioda analīzi, ļaujot veikt detalizētu izpēti. Šāda struktūra nodrošina elastīgu datu apstrādes procesu, ļaujot pielāgot analīzi dažādām vajadzībām. Rezultātu vizualizāciju nodrošina plot_fused_data_final funkcija, kas rada pārskatāmu datu attēlojumu ar iespēju salīdzināt gan sākotnējos datus, gan to apvienotās versijas.

Izstrādātā metodes koncepcija ietver laika datu slāņošanu, kur katrs slānis attēlo datus kā plakni. Sākotnējā koncepcija tika piemērota bišu barības meklēšanas uzdevumam, kas sniedz daudzveidīgus datus. Būtiski dati bišu barības meklēšanas optimizēšanai ietver informāciju par reģiona atrašanās vietu, topogrāfiju, klimatu, vietējiem nektāru un ziedputekšņus ražojošajiem augiem, kā arī bišu sugām un to aktivitātēm. Datu slāņošanas metode sākas ar nepieciešamo datu, piemēram, nektāra līmeņu normalizēšanu, lai izveidotu datu kopu. Pēc tam šī kopa tiek analizēta, lai noteiktu visproduktīvākās vietas bišu barošanai. Metodē tiek izmantota gan svērto vērtību pieeja, gan galveno komponentu analīze (PCA), lai apvienotu dažādus datu aspektus. Šīs pieejas ļauj salīdzināt un novērtēt dažādu parametru nozīmīgumu, piemēram, augu bagātību noteiktā laika periodā. Pārklājuma novērtēšanai starp abām pieejām tiek izmantota laukuma aprēķināšana zem līknēm. Izvēlētā pieeja ir vienkārša lietošanā un sniedz skaidri saprotamus rezultātus. Datu vizualizācija ļauj viegli identificēt nozīmīgākos periodus vai notikumus, kas ir būtiski tālākajai analīzei.

Precīzās putnkopības prognozēšanas uzdevums

Mūsdienu putnu fermās tiek nepārtraukti novēroti dažādi vides parametri, kas ietekmē olu ražošanu. Eksperti noteica būtiskos faktorus, kas ietekmē dējējvistu labklājību, piemēram, gaisa temperatūru, mitrumu, CO_2 un NH_3 līmeni. Datu pieejamības rakstura dēļ datu kopas ir ierobežotas, un olu īpatsvara dati par 61 nedēļas periodu (dati, kas savākti no 2019. gada 22. novembra līdz 2021. gada 9. februārim) tika izmantoti modeļu apmācībai, tāpat arī par 46 nedēļu periodu (dati savākti no 2021. gada 23. marta līdz 2022. gada 3. martam) (sk. 2.5. att.). Diennakts olu produkcija jeb īpatsvars tiek aprēķināts kā dienā saražoto olu skaits attiecību pret kopējo vistu skaitu attiecīgajā dienā (Paura et al., 2022).



2.5. att. **Olu īpatsvara līknes, ko izmanto apmācībai (1. cikls) un testēšanai (2. cikls).**

Testa datu kopa ievērojami atšķīrās no apmācībā izmantotās un arī no parastā olu īpatsvara modeļa. Šādu atšķirību iemesli, pēc lauksaimnieku sniegtās informācijas, varētu būt skaidrojami ar nekonsekvenci datu ievades pārvaldībā – kamēr savākto olu daudzums tiek skaitīts automātiski, galīgā vērtība tiek ievadīta manuāli. To var veikt vairākas reizes dienā vai arī neveikt vispār, piemēram, darba dienās vai tehnisku iemeslu dēļ. Apmācības un testēšanas kopas, t. i., kā tās tiek sadalītas, atšķiras nelineārajiem un *ML* balstītajiem modeļiem, un ir aprakstītas tālāk.

Precīzās putnkopības uzdevuma risināšanai Henco2 projekta ietvaros atlasītie mašīnmācīšanās modeļi tika izveidoti, izmantojot *Keras* ietvaru (Chollet, 2015) (*LSTM* un *CNN*) un *scikit-learn* bibliotēku (Pedregosa et al., 2011) (*RF* un *XGBoost*). Modeļi tika noregulēti (hiperparametru atlase), izmantojot bibliotēkas paplašinājumus, piemēram, *keras-tuner* un *sklearn.model_selection*. Atsevišķos ML modeļos netika veiktas nekādas izmaiņas attiecībā uz to bāzes arhitektūru. Modeļi tika salīdzināti, mainot katra modeļa hiperparametru vērtības. Lai atrastu labāko hiperparametru konfigurāciju, *LSTM* un *CNN* modeļi tika noskaņoti, izmantojot *Hyperband* algoritmu (Li et al., 2020), bet uz lēmumu koku balstītie modeļi – izmantojot *Random Search* (Bergstra & Bengio, 2012). LSTM un *CNN* modeļiem tika izmantota agrīna apstāšanās tehnika (ar pacietības vērtību 10), lai potenciāli samazinātu pārklāšanās problēmu. Agrīna apstāšanās tika izmantota arī *XGBoost* hiperparametru meklēšanas un apmācības fāzē, bet savstarpējās validācijas tehnika RF gadījumā.

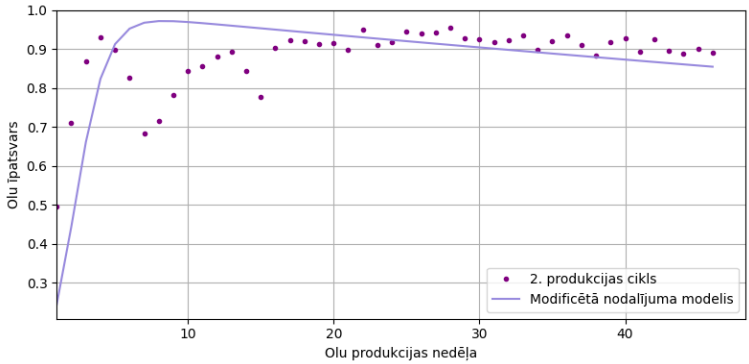
Modeļu apmācība tika veikta, izmantojot faktorus, kas putnu fermā tiek uzraudzīti ikdienā, un ar ražošanu saistītos datus ar dažādu ievades secības garumu, piemēram, izmantojot bīdāmā loga pieeju. Lietojot šādu paņēmienu, svarīgs solis bija noteikt loga izmēru, jo tas nosaka papildu prasību modeļa ievadei – iepriekšējo produktivitātes vērtību skaitu, piemēram, šajā gadījumā olu ražošanu. Ja ir izvēlēts bīdāmais logs ar izmēru 1, tas nozīmē, ka ievadei ir nepieciešami iepriekšējās dienas ražošanas dati. Sākotnējā izstrādes posmā tika apsvērtas vairākas pazīmju atlases pieejas, un, pamatojoties uz iegūtajiem datiem

no saimniecības, tika noteikts, ka perspektīvā funkciju atlase būtu piemērota izmantošanai vispārīgiem *ML* algoritmiem. Atlasītie *ML* modeļi tika apmācīti uz pirmā ražošanas cikla datiem (kas tika tālāk sadalīti 90% apmācības un 10% validācijas daļās, lai izvairītos no datu noplūdes (Hannun et al., 2021) un pārmērīgas uzstādīšanas problēmām (Ying, 2019)) un pārbaudīti otrajā ražošanas ciklā, lai prognozētu produktivitāti nākamajai dienai. Modeļu ievade tika veidota no 12 parametriem un papildus vēsturiskajiem ražošanas datiem atkarībā no bīdāmā loga izmēra. Modeļu veiktspēja tika novērtēta pēc statistikas kritērijiem. Modificētā nodalījuma modeļa parametri (sk. 2.1. tabulu) tika novērtēti, izmantojot R programmēšanas valodu, lai atbilstu olu īpatsvara līknei (1. cikls). No 2.1. tabulas visi parametri ir nozīmīgi ($p < 0,001$) un ir piemērojami šim prognozēšanas uzdevumam. Modeļa rezultāts parāda tikai olu īpatsvara tendenci, pamatojoties uz iepriekš apmācītajiem datiem, bet neietver ievades vērtības, kas varētu ietekmēt prognozi un norādīt uz iespējamām problēmām. Šis ražošanas cikls parāda, ka, lai iegūtu augstu precizitāti, nepietiek tikai ar olu produkcijas nedēļu vien, lai izdarītu secinājumus, bet tas ļauj lauksaimniekiem redzēt novirzes no tendences.

2.1. tabula. **Modificētā nodalījuma modeļa aprēķinātās parametru vērtības**

Parametrs	Aplēse	Standarta kļūda	t vērtība	Pr(>|t|)
a	0.13099	0.01316	9.954	4.45e-14 ***
b	-0.90414	0.03927	-23.024	< 2e-16 ***
d	2.24435	0.04658	48.182	< 2e-16 ***
c	-0.90766	0.03923	-23.139	< 2e-16 ***

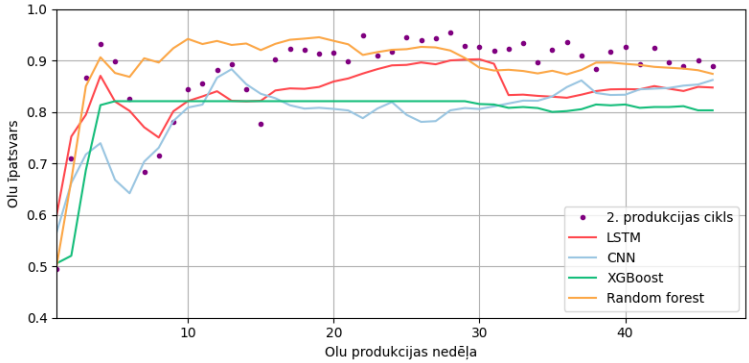
Pielāgotā līkne testa olu īpatsvara datu kopai ir šāda:



2.6. att. **Pielāgota līkne un novērotais olu īpatsvars** (Bumanis et al., 2023).

ML modeļi mēdz sekot nenormālam ražošanas samazinājumam (sk. 2.7. att.), tādējādi norādot uz to spēju pielāgoties šāda veida situācijām. Lai gan turpmākā datu pārbaude parādīja, ka vides faktori krasi nemainījās, lai ietekmētu ražošanas samazināšanos, modeļi prognozēja kritumu tāpēc, ka iepriekšējās dienas

(atkarībā no bīdāmā loga izmēra) ražošanas dati tika izmantoti kā ievade. ML modeļu rezultāti un novērotais olu īpatsvars ar bīdāmo logu (izmērs 2) ir šādi:



2.7. att. **Apmācīto *ML* modeļu rezultāti** (Bumanis et al., 2023).

Attiecībā uz *ML* modeļiem tika pārbaudīti vairāki (1, 2, 3, 5, 7 un 14) logu izmēri. Modeļa veiktspējas rezultāti ir parādīti 2.2 tabulā. Attiecībā uz bīdāmā loga izmēru rezultāti liecina, ka *LSTM* ir precīzāks, izmantojot bīdāmo logu ar izmēru 2, sasniedzot mazākās *MAPE* un *RMSPE* vērtības, attiecīgi 5,390% un 7,751%. Nebija arī lielas atšķirības starp modeļu veiktspēju, izmantojot loga izmērus 3 un 5, izņemot *CNN* modeli, kas darbojās vissliktāk, un tas var tikt skaidrots ar iespējamu modeļa pārmērīgu pielāgošanu.

2.2. tabula. **Mašīnmācīšanās modeļa veiktspēja** (Bumanis et al., 2023)

Bīdāmā loga izmērs	**Kļūdas metrika**	**LSTM**	**CNN**	**XGBoost**	**RF**
1	*MSE*	1.710	4.111	1.225	**0.944**
1	*MAPE*	13.909	14.224	9.994	**6.907**
1	*RMSPE*	15.439	16.258	11.708	**10.242**
2	*MSE*	**0.272**	1.884	1.060	0.726
2	*MAPE*	**5.390**	15.200	10.272	6.331
2	*RMSPE*	**7.751**	18.314	12.178	9.284
3	*MSE*	**0.203**	1.384	0.877	0.664
3	*MAPE*	6.501	39.319	9.086	**6.158**
3	*RMSPE*	**8.828**	39.993	10.875	9.110
5	*MSE*	**0.358**	0.843	0.767	0.604
5	*MAPE*	6.218	13.479	7.415	**6.077**
5	*RMSPE*	**8.781**	15.537	9.223	9.016
7	*MSE*	**0.198**	0.443	0.863	0.546
7	*MAPE*	**5.484**	13.300	9.619	6.188
7	*RMSPE*	**7.845**	14.555	11.350	9.168
14	*MSE*	**0.153**	0.308	0.719	0.453
14	*MAPE*	6.433	6.633	**6.114**	6.273
14	*RMSPE*	8.982	9.718	**8.158**	9.221

Tabulā 2.3 ir apkopoti labākie rezultāti, kas iegūti modeļa novērtēšanā. Var secināt, ka kopumā *LSTM, RF* un *XGBoost* uzrādīja vislabāko veiktspēju. Novērtēšanas rezultāti, ņemot vērā labākās metrikas vērtības dažādiem bīdāmo logu izmēriem (mašīnmācīšanās modeļiem), liecina, ka veiktspēja atšķiras. Kopumā visi modeļi nodrošina pietiekami precīzus rezultātus, lai atklātu problēmas un veiktu izmaiņas ražošanas procesā; tomēr rezultāti liecināja, ka daži modeļi, piemēram, *LSTM*, uzrādīja konkurētspējīgu veiktspēju visos bīdāmo logu izmēros, vienlaikus nodrošinot vislabākos rezultātus – izmantojot mazāku vēsturisko ražošanas datu skaitu. Var secināt, ka mašīnmācīšanās modeļi, īpaši *LSTM*, izrādās labāki par *Modified Compartmental*.

2.3. tabula. **Modeļu novērtēšanas labākie rezultāti** (Bumanis et al., 2023)

Modelis	***MSE***	***MAPE***	***RMSPE***	**Bīdāmā loga izmērs**
Modified Compartmental	0.011	9.134	14.809	n/a
LSTM	0.272	5.390	7.751	2
CNN	0.308	6.633	9.718	14
XGBoost	0.719	6.114	8.158	14
RF	0.604	6.077	9.016	5

Modeļa testēšanai izmantotais olu produkcijas cikls bija netipisks datu kvalitātes raksturlielumu, piemēram, viendabīguma un pilnīguma, ziņā, un tādējādi padarīja sarežģītāku dzīvotspējīgākā modeļa atlases un olu īpatsvara prognozēšanas procesu, pamatojoties uz šādiem datiem. Jāpiebilst, ka prognozēšana tika veikta tikai 1 dienai iepriekš, kas nosacīti ierobežo prasības precizitātes mērķim. Lai gan ir iespējams prognozēt olu īpatsvaru ilgākam periodam, rezultātos var būt strauja precizitātes samazināšanās; tādējādi atbilstošajam prognozēšanas garumam jābūt pietiekami ilgam, lai veiktu atbilstošas izmaiņas (t. i., pielāgotu ventilācijas algoritmu temperatūras izmaiņām) ražošanas procesam. Turklāt izvēli prognozēt tikai 1 dienu iepriekš noteica pieejamo apmācību datu konsekvence. Tas ietver arī atšķirības starp divu atsevišķu ciklu datiem. Testa datu kopa, kas bija ievērojami atšķirīga, parādīja ierobežojumus nelineārajam modelim, kas izmanto tikai vienu parametru (dēšanas nedēļu skaitu) un nepielāgojas izmaiņām, kuru rezultātā tika iegūtas arī *MAPE* un *RMSPE* vērtības – attiecīgi 9,134% un 14,809%. Lai gan *ML* modeļu aprēķinātās kļūdas (*MAPE* un *RMSPE*) bija robežās no 5% līdz 10%, tika novērots, ka tās var labāk pielāgoties ražošanas izmaiņām nekā pārbaudītais nelineārās regresijas modelis. Tā kā *ML* modeļos kā ievades dati tiek izmantoti arī vides dati, pēkšņas šo faktoru izmaiņas (piemēram, temperatūra, CO_2, NH_3) ietekmē produktivitāti, ko var laikus prognozēt.

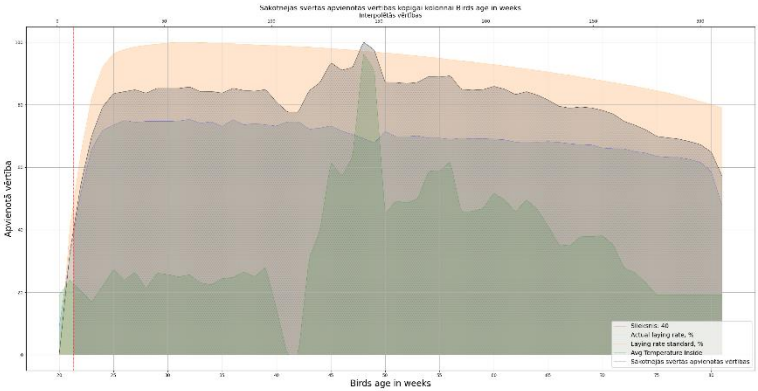
Rezultāti parādīja, ka *ML* modeļi (*LSTM, RF* un *XGBoost* ar bīdāmo logu izmērā 2) spēja prognozēt ražošanas samazināšanos (2. produkcijas cikls) apmierinošā līmenī. Rezultāti liecina, ka piedāvātie risinājumi var būt piemērojami arī saimniecībās, kurās ir ierobežotas ražošanas datu kopas un nav liela apjoma vēsturisko olu īpatsvara datu. Atkarībā no pieejamajiem

vēsturiskajiem datiem modeļu apmācībai saimniecībai iespējams izmantot arī vairāku modeļu pieeju, kur var darbināt dažādus modeļus atbilstoši lauksaimnieka vajadzībām (prognozes garumam). Turklāt tas arī saglabā iespēju izmantot nelineāro modeli situācijās, kad netiek reģistrēti dati, kas ietekmē vidi vai citus produktivitātes parametrus. Šajā gadījumā nelineāro modeli var izmantot vai nu kā atsevišķu risinājumu, vai kā papildu pierādījumu, lai sekotu ražošanas līknes dinamikai.

Datu slāņošanas metodes aprobācija

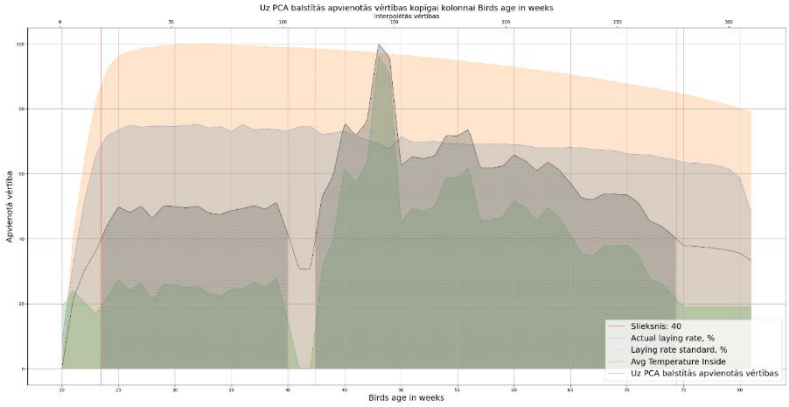
Prognozēšanas modeļu rezultātu analīzei un potenciālo problēmu identificēšanai tika izmantota datu slāņošanas metode. Šī metode nodrošina sistemātisku pieeju vairāku parametru mijiedarbības izpētei un to ietekmei uz olu ražošanas procesu. Pētījumā tika analizēti trīs fundamentāli parametri, kas raksturo olu ražošanas procesu: faktiskais olu dēšanas īpatsvars procentos, standarta dēšanas īpatsvars procentos un vidējā iekštelpu temperatūra. Parametru izvēle balstījās uz to fizioloģisko un tehnisko nozīmīgumu olu ražošanas procesā.

Atbilstoši iepriekš aprakstītajai metodikai, pēc trūkstošo vērtību interpolācijas un datu normalizācijas, tika veikta datu apvienošana, izmantojot `data_fusion_main` funkciju. Parametru svaru koeficienti tika noteikti, ņemot vērā to ietekmes pakāpi uz ražošanas procesu. Faktiskajam dēšanas īpatsvaram tika piešķirts svars 0,85, standarta dēšanas īpatsvaram 0,50, bet vidējai temperatūrai 0,35. Zemāks svars temperatūras parametram tika piešķirts, balstoties uz zinātniskajiem pētījumiem, kas norāda, ka temperatūra būtiski ietekmē produktivitāti tikai ārpus noteiktām robežvērtībām, kuras nosaka konkrētā vistu šķirne un turēšanas apstākļi. Slāņošanas slieksnis tika iestatīts uz 40 vienībām, kas ļauj efektīvi identificēt būtiskas novirzes starp faktisko un teorētisko dēšanas īpatsvaru. Datu izlīdzināšanai tika izmantota automātiskā izlīdzināšana (`AutoSmooth`=1) ar izlīdzināšanas parametru 1, kas nodrošina optimālu līdzsvaru starp īstermiņa svārstību filtrāciju un būtisko tendenču saglabāšanu datos. Metodes pielietošanas rezultāti atspoguļoti trīs attēlos (sk. 2.8. att., 2.9. att. un 2.10. att.). Sākotnējo svērto apvienoto vērtību analīze (sk. 2.8. att.) atklāj būtiskas atšķirības starp faktisko un standarta dēšanas īpatsvaru, īpaši periodos ar paaugstinātu temperatūras svārstību ietekmi. Izteikts ir periods no 40. līdz 50. nedēļai, kurā vērojams straujš kritums, kam seko pakāpeniska atgūšanās. Šīs novirzes var būt saistītas ar vairākiem faktoriem, tostarp datu kvalitāti un ražošanas procesa izmaiņām.



2.8. att. **Sākotnējās svērtās apvienotās vērtības.**

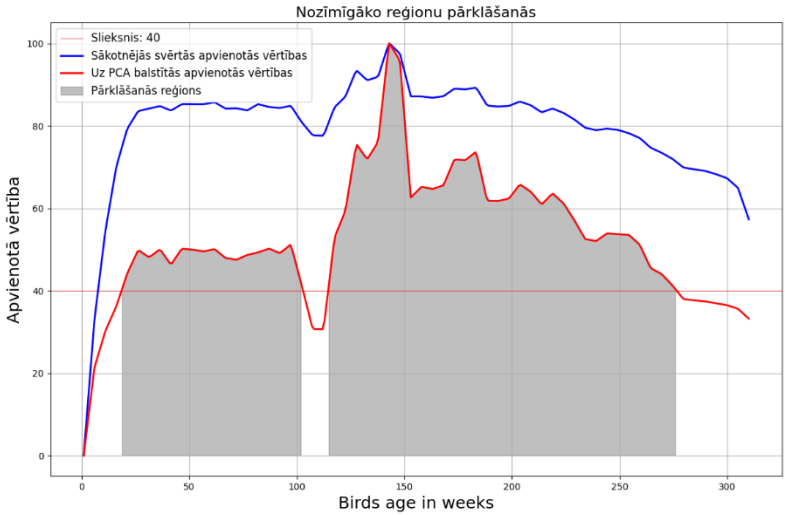
Uz PCA balstītā analīze (sk. 2.9. att.) sniedz papildu ieskatus par parametru mijiedarbību. Attēlā redzams, ka PCA komponente spēj efektīvi identificēt anomālijas datu kopā, īpaši periodā no 40. līdz 50. nedēļai, kur novērojamas būtiskas novirzes no gaidītā dēšanas īpatsvara. Šis periods sakrīt ar paaugstinātām kļūdas vērtībām *Modified Compartmental* modelī (MAPE 9,134%, RMSPE 14,809%), kas norāda uz modeļa ierobežojumiem sarežģītu situāciju analīzē.



2.9. att. **Uz PCA balstītās apvienotās vērtības.**

Nozīmīgāko reģionu pārklāšanās analīze (sk. 2.10. att.) identificē trīs raksturīgus periodus ražošanas ciklā. Ražošanas uzsākšanas periodā (20.–25. nedēļa) novērojama augsta korelācija starp abām metodēm, kas norāda uz datu kvalitātes un ražošanas procesa stabilitāti. Nestabilitātes periodā (40. –50. nedēļa) konstatētas ievērojamas atšķirības starp metodēm, kur mašīnmācīšanās modeļi, īpaši *LSTM* ar MAPE 5,390%, uzrāda ievērojami augstāku precizitāti

nekā tradicionālie modeļi. Ražošanas noslēguma periodā (75.–80. nedēļa) vērojama atkārtota metožu konverģence, kas liecina par ražošanas procesa stabilizāciju.



2.10. att. **Nozīmīgāko reģionu pārklāšanās.**

Datu slāņošanas analīze atklāj vairākas būtiskas problēmas, kas ietekmē prognozēšanas modeļu precizitāti:

1. datu kvalitātes problēmas: identificētās novirzes, īpaši 40.–50. nedēļu periodā, norāda uz potenciālām problēmām sensoru datos vai datu savākšanas procesā. Šīs problēmas var būt viens no iemesliem, kāpēc *Modified Compartmental* modelis uzrāda augstākas kļūdas vērtības (MAPE 9,134%);

2. modeļu adaptācijas spēja: ML modeļu zemākās kļūdas vērtības (MAPE 5–10%) var tikt skaidrotas ar to spēju labāk pielāgoties nelineārām izmaiņām datos, ko apstiprina PCA analīzes rezultāti. Īpaši *LSTM* modelis (MAPE 5,390%) demonstrē ievērojami labāku veiktspēju periodā ar identificētajām anomālijām;

3. sensoru sistēmu ierobežojumi: analīze norāda uz nepieciešamību pilnveidot sensoru datu kvalitātes kontroles mehānismus, īpaši attiecībā uz temperatūras mērījumiem, kur novērojamas būtiskas svārstības.

Kopumā pētījums parāda tradicionālo nelineāro modeļu un mašīnmācīšanās algoritmu veiktspējas atšķirības olu īpatsvara prognozēšanā. ML modeļi uzrāda labākus rezultātus nekā tradicionālie viena faktora modeļi, ko apliecina zemākas kļūdu vērtības. Vienlaikus rezultāti norāda uz nepieciešamību pēc specializētām metodēm datu kvalitātes problēmu risināšanai precīzās lauksaimniecības kontekstā.

Balstoties uz otrajā nodaļā izstrādāto datu slāņošanas metodi un tās praktisko pielietojumu precīzās putnkopības pētījuma rezultātu pārbaudei, autors secina, ka tiek apstiprināta tēze:

Ir iespējama nepilnīgo datu interaktīvā apstrāde un vizualizācija, izmantojot izstrādātās datu apvienošanas un datu kvalitātes uzlabošanas metodes.

Pamatojums

Otrajā nodaļā izstrādātā datu slāņošanas metode nodrošina sistemātisku pieeju dažāda rakstura datu apvienošanai un vizualizācijai. Metodes konceptuālais pamats tai dod spēju efektīvi integrēt atšķirīgus datu avotus, ko apliecina augu bagātības, nokrišņu un bišu aktivitātes datu analīze un vizualizācija vienlaikus. Metodes interaktīvais raksturs nodrošina tās spēju identificēt un vizualizēt datu pārklāšanās zonas, kas izpaužas trapecveida metodes pielietojumā reģionu analīzē (sk. 2.3. att.). Šī pieeja ļauj kvantitatīvi novērtēt gan pārklāšanās zonas, gan atšķirību reģionus, tādējādi sniedzot skaidru priekšstatu par datu mijiedarbību.

Metodes praktiskā pielietojamība tiek apstiprināta precīzās putnkopības pētījumā, kur attēlos 2.8., 2.9. un 2.10. redzamā vizualizācija parāda metodes daudzpusīgo analītisko potenciālu. Attēlos 2.8. un 2.9. atspoguļotā vizualizācija parāda metodes spēju apvienot un analizēt trīs būtiskus parametrus – faktisko dēšanas īpatsvaru, teorētisko dēšanas īpatsvaru un vidējo iekštelpu temperatūru. Īpaši nozīmīga ir metodes spēja identificēt kritiskos periodus, izmantojot gan svērto summu, gan PCA balstīto pieeju. Attēlā 2.10. paradīta pārklāšanās reģionu analīze sniedz papildu kvantitatīvu novērtējumu par metožu konverģenci un diverģenci trīs raksturīgos periodos – ražošanas uzsākšanas (20.–25. nedēļa), nestabilitātes (40.–50. nedēļa) un noslēguma (75.–80. nedēļa) posmos.

Metodes efektivitāte datu kvalitātes problēmu identificēšanā izpaužas tās spējā atklāt būtiskas novirzes un neparastus datu periodus. Trūkstošo vērtību aizpildīšanai tika izmantota interpolācija, kas nodrošināja datu nepārtrauktību un normalizāciju, kas bija būtisks priekšnosacījums PCA komponenta izmantošanai. Šī pieeja nodrošināja efektīvu datu vizualizāciju un palīdzēja identificēt kritiskās datu kvalitātes problēmas.

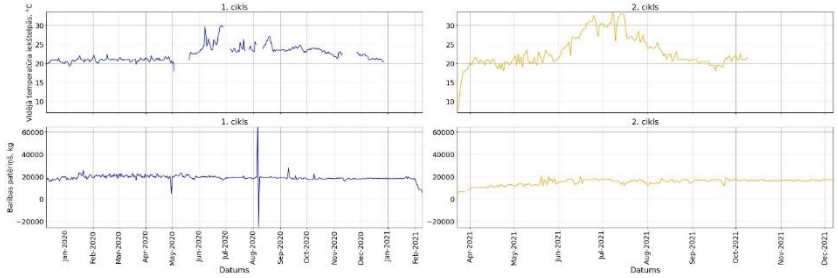
3. DATU KVALITĀTES UZLABOŠANA

Datu kvalitātes kritēriji – pilnīgums un precizitāte – ir fundamentāli datu apstrādes procesos, jo tie tieši ietekmē analīzes, prognozēšanas un lēmumu pieņemšanas kvalitāti. Datu pilnīgums raksturo pieejamo mērījumu īpatsvaru no visiem nepieciešamajiem mērījumiem, savukārt precizitāte nosaka mērījumu atbilstību patiesajām vērtībām. Datu pilnīgumu var novērtēt divos aspektos. Pirmais ir nepieciešamo datu kopu esamība – vai visas vajadzīgās datu grupas ir pieejamas. Piemēram, vides monitoringa sistēmā būtiski ir gan temperatūras, gan mitruma, gan CO_2 līmeņa mērījumi. Otrais aspekts ir datu ierakstu pilnīgums katrā kopā – vai nav iztrūkstošu mērījumu konkrētā laika periodā. Trūkstošo datu problēmu var risināt ar interpolāciju vai līdzīgu ierakstu izmantošanu. Datu

precizitāti ietekmē gan tehniskās neprecizitātes, kas saistītas ar mēraparatūras un sensoru darbību (kalibrācija, ierīču ierobežojumi), gan cilvēciskais faktors – kļūdas manuālā datu apstrādē. Īpaši kritiska ir datu precizitāte *IoT* risinājumos un mākslīgā intelekta lietojumos, kur neprecīzi ievaddati var būtiski ietekmēt modeļu apmācības procesu. Efektīvai datu kvalitātes nodrošināšanai nepieciešama sistemātiska pieeja. Trūkstošo vērtību aizpildīšanai izmanto matemātiskās metodes, piemēram, interpolāciju, kas balstās uz esošajiem precīzajiem datiem. Precizitātes nodrošināšanai veic regulāras pārbaudes, kas ļauj identificēt un novērst problēmu cēloņus – gan tehniskās kļūmes, gan cilvēciskā faktora radītās novirzes. Šāda kompleksa pieeja ļauj izveidot uzticamu pamatu tālākai datu analīzei.

Datu kvalitātes metodes tika izveidotas, risinot precīzās putnkopības olu dēšanas īpatsvara prognozēšanas uzdevumu. Metožu izstrādei un testēšanai tika izmantota datu kopa par diviem (vienu pilnu un otru daļēju) olu dēšanas cikliem – par 61 nedēļas periodu (dati, kas savākti no 2019. gada 22. novembra līdz 2021. gada 9. februārim) un par 46 nedēļu periodu (dati savākti no 2021. gada 23. marta līdz 2022. gada 3. martam). Olu dēšanas periodos tika reģistrēti dažāda veida mājputnu un vides dati, kas sniedz informāciju par mikroklimatu (temperatūra, mitrums, CO_2, NH_3), kā arī dati par putnu barošanu (ūdens un barības patēriņš un tā sastāvs, t. i., makro/mikro barības vielas un mikroelementi). Temperatūras un mitruma uzraudzības sensori tika novietoti vistu kūts centrā. CO_2 (*IR-2* sensors, *GDS Technologies Garforth*) un NH_3 (NH_3/*MR-100* sensors, *Membrapor AG*) koncentrācijas tika mērītas nepārtraukti ik pēc 10 minūtēm, bet vidējās vērtības tika aprēķinātas katru stundu pēc tam.

Sakarā ar uzraudzības sistēmas agrīnu ieviešanas fāzi datu kvalitāte savāktajiem datiem nav ideāla. Tā, piemēram (sk. 3.1. att.), vidējiem temperatūras datiem ir nepilnības, bet barības patēriņam – novirzes.



3.1. att. **Divu ražošanas ciklu datu kvalitātes reprezentācija.**

Apraksts: pirmā cikla (1. kolonna) un otrā cikla (2. kolonna) vidējo iekštelpu temperatūras (1. rinda) un barības patēriņi (2. rinda).

Turpmākai analīzei tika izmantoti pirmā cikla vidējās iekštelpu temperatūras dati. Sakarā ar to, ka datu failā ir trīs kolonnas, analizējama ir šo kolonnu vidējā vērtība: temperatūra no sensora, kas atrodas 1. būru stāvā, temperatūra no sensora, kas atrodas 8. būru stāvā, un temperatūra, kas ir ievadīta manuāli. Gala

vērtība, gadījumā, kad nav pieejamas visu kolonnu vērtības, tika iegūta, izmantojot esošās.

Vidējai temperatūrai ir nedaudz ciklisks modelis ar atkārtotiem maksimumiem un zemākajām vietām. Tas varētu liecināt par sezonālām izmaiņām vai citu periodisku faktoru, kas ietekmē temperatūru. Turklāt sērijas pēdējā daļā ir vērojama ievērojama augšupejoša tendence, kas norāda, ka temperatūra šajā periodā kopumā paaugstinājās.

Kopā, var izdalīt šādas statistiskās vērtības:

- datu punktu skaits: 335;
- vidējais: 22,37 °C;
- standarta novirze: ≈1,96 °C;
- minimālā vērtība: 18,00 °C;
- 25. procentile (Q1): ≈20,93 °C;
- mediāna (50. procentile): 21,90 °C;
- 75. procentile (Q3): ≈23,55 °C;
- maksimālā vērtība: 29,90 °C;
- trūkstošo vērtību skaits: 93.

ARIMA modelis

Viena no metodēm, precīzāk – statistiskiem modeļiem, kas apvieno automātisko regresīvo funkciju (*AR*), integrāciju (I, kas attiecas uz datu diferencēšanu, lai padarītu tos nekustīgus) un mainīgā vidējā (*MA*) komponentus, ir *ARIMA* modelis. Modelis var tikt pielietots trūkstošo datu noteikšanai un aizvietošanai. Tālāk aprakstīti atsevišķi komponenti.

- *AutoRegressive* (*AR*): automātiskās regresijas parametrs. Modelis, kas izmanto atkarīgo attiecību starp novērojumu un vairākiem novēlotiem novērojumiem (iepriekšējie laika posmi) (sk. (3.1.)):

$$\mathrm{AR(p)}: \mathrm{Y_t} = \mathrm{c} + \phi_1 \mathrm{Y_{t-1}} + \phi_2 \mathrm{Y_{t-2}} + \cdots + \phi_\mathrm{p} \mathrm{Y_{t-p}} + \epsilon_\mathrm{t} \qquad (3.1.)$$

kur

(Y_t) – sērijas vērtība laikā, t;
c – konstante, vesels vai decimālais skaitlis;
$\phi_1, \phi_2, \dots, \phi_p$ – modeļa parametri;
p – *AR* termina secība;
ϵ_t – balts troksnis (kļūdas termins) laikā t.

- I (d): integrēts parametrs. Novērojumu diferencēšana, lai laiku rindas būtu stacionāras; d ir nesezonālu atšķirību skaits (sk. (3.2.)):

$$\mathrm{I(d)}: \nabla^\mathrm{d} \mathrm{Y_t} \qquad (3.2.)$$

- *MA* (q): mainīgais vidējais parametrs. Modelis, kas izmanto atkarību starp novērojumu un atlikušo kļūdu no slīdošā vidējā modeļa, ko izmanto novēlotiem novērojumiem (sk. (3.3.)):

$$\mathrm{MA(q)}: \mathrm{Y_t} = \mathrm{c} + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q} \quad (3.3.)$$

Apvienojot *AR*, *I* un *MA* parametrus, *ARIMA* modelis tiek izteikts šādi (sk. (3.4.)):

$$\mathrm{ARIMA(p,d,q)}: \nabla^d \mathrm{Y_t} = \mathrm{c} + \phi_1 \mathrm{Y}_{t-1} + \cdots + \phi_p \mathrm{Y}_{t-p} + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q} + \epsilon_t \quad (3.4.)$$

kur

p – modelī iekļauto nobīdes novērojumu skaits, vesels skaitlis;

d – reižu skaits, kad neapstrādātie novērojumi ir mainīti, lai sērija būtu nekustīga, vesels skaitlis;

q – slīdošā vidējā loga izmērs, vesels skaitlis.

ARIMA modeļa pielietošana prasa stacionāru laika rindu datus. Lai pārbaudītu šo priekšnosacījumu, izmanto papildināto Dikija-Fullera (*Augmented Dickey-Fuller test, ADF*) testu. Šis tests nosaka, vai laika rinda ir stacionāra. *ADF* testa nulles hipotēze apgalvo, ka laika rindai piemīt vienības sakne. Testa rezultāti balstās uz p vērtības analīzi – ja tā ir zemāka par izvēlēto nozīmības līmeni (parasti 0,05), tas norāda uz laika rindas stacionaritāti un no laika atkarīgas struktūras esamību. Šāds rezultāts apstiprina datu piemērotību ARIMA modeļa izmantošanai.

Attiecīgi, laika rindu sērijai $\mathrm{y_t}$ *ADF* pārbauda nulles hipotēzi (sk. (3.5.)):

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \cdots + \epsilon_t \quad (3.5.)$$

kur

Δ – atšķirības operators. Šis operators tiek izmantots, lai aprēķinātu pašreizējās vērtības y_t un tās iepriekšējās vērtības y_{t-1} starpību laika rindā, veidojot Δy_t;

α – konstante;

β – fiksē potenciālu lineāro laika tendenci;

γ – koeficients sērijas nobīdes līmenī, kas fiksē vienības saknes klātbūtni. Tas ir būtisks, lai noteiktu, vai laika rinda ir stacionāra vai ne;

δ_i – atkarīgā mainīgā aizkavēto pirmo starpību koeficienti;

ϵ_t – gadījuma kļūdas (vai traucējumu) komponents, kas atspoguļo neizskaidrojamo daļu laika rindā.

ADF testa mērvienības tiek noteiktas pēc laika rindas datu veida, ar kuru tiek strādāts, un to mērvienībām. Piemēram, ja laika rindas dati ir eiro izteikti finanšu dati, tad lielākā daļa parametru būs izteikti eiro vai eiro laika vienībā.

Tests koncentrējas uz γ. Ja γ būtiski atšķiras no nulles, tad sērijai nav vienības saknes un tā ir stacionāra.

Šajā gadījumā ADF statistika ir negatīvs skaitlis −1,6112. Jo negatīvāks tas ir, jo spēcīgāk tiek noraidīta hipotēze, ka pastāv vienības sakne, un līdz ar to, jo spēcīgāki pierādījumi par stacionaritāti.

P vērtība atspoguļo varbūtību, ka datiem būtu novērotā struktūra (vai mazāk ticams), ja nulles hipotēze būtu patiesa. Ņemot vērā p-vērtību 0,4773, kas ir lielāka par parasti izmantoto nozīmīguma līmeni 0,05, nevar noraidīt nulles hipotēzi. Tas nozīmē, ka sērija nav stacionāra.

ARIMA modelēšanā sērijai jābūt stacionārai. Sērijas nestacionaritāte nozīmē, ka pirms *ARIMA* lietošanas, iespējams, vajadzēs atšķirt sēriju, lai tā būtu nekustīga. To norāda ar "I" *ARIMA*, kas apzīmē integrēto parametru. Atšķirību skaits, kas nepieciešams, lai sērija būtu stacionāra, nosaka d parametrs.

Atšķirīgās sērijas šķita nekustīgākas, kas liecina, ka $d = 1$ varētu būt labs sākumpunkts *ARIMA*. Tomēr d izvēlei jābūt balstītai uz stacionaritātes sasniegšanu un *ARIMA* modeļa Akaike informācijas kritērija *(AIC*) samazināšanu.

Attiecīgi šajā gadījumā, lai noteiktu atbilstošo atšķirības līmeni (*d ARIMA*), tiek veikts:

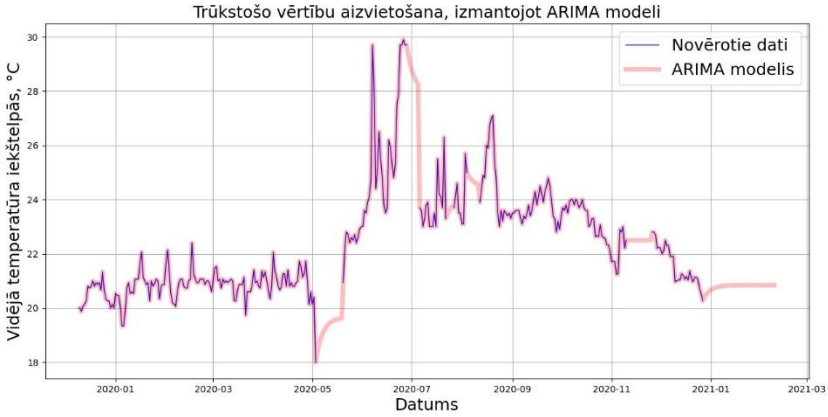
- tiek sākts ar $d = 1$;
- d tiek palielināts un atšķirīgo sēriju stacionaritāte tiek pārbaudīta, izmantojot paplašinātā Dikija-Fullera (*ADF*) testu;
- katrai d vērtībai tiek pielāgots *ARIMA* modelis un tiek salīdzinātas *AIC* vērtības;
- optimālais d ir tas, kas padara sēriju nekustīgu un samazina *AIC*.

Veicot pārbaudi, tiek iegūts:

- optimālais atšķirības līmenis, d: 1;
- saistītā minimālā *AIC* vērtība: 707,13.

Rezultātā viena ($d = 1$) atšķirība ir pietiekama, lai sērija būtu stacionāra un nodrošinātu vislabāko līdzsvaru (atbilstoši *AIC) ARIMA* modelim.

Attiecīgi izmantojot noteikto d parametra vērtību, tiek iegūts šāds rezultāts (sk. 3.2. att.):



3.2. att. **Trūkstošo datu aizvietošana, izmantojot *ARIMA* modeli.**

Modificētā standarta vidējā svērtā metode

Alternatīvi pielietojot datu apvienošanas principus trūkstošo datu pievienošanai var izmantot gan vietējos novērotos datus (izmantojot, piemēram, standarta vidējo svērto metodi), gan pamatā esošo tendenci, kā arī korekcijas, kuru bāze ir datu raksturlielumi, piemēram, šķībums.

Šajā gadījumā vietējā informācija sniedz izpratni par tiešo kontekstu ap trūkstošo vērtību. Globālā informācija vai tendence palīdz izprast plašākus datu modeļus. Šķībums var sniegt ieskatu datu vispārējā sadalījumā un būtībā.

Attiecīgi, ja laika rindā indeksā j trūkst datu punkta, tad vērtība y_j tiek aprēķināta šādi:

1. tiek papildināta tendence, izmantojot lineāras regresijas modeli, lai noteiktu optimālo kaimiņu skaitu, ko izmantot vietējam svērtam vidējam (sk. (3.6.)):

$$y = \beta_0 + \beta_1 x \tag{3.6.}$$

kur

N – aplūkojamo datu punktu skaits, vesels skaitlis;
y_i – apzīmē faktisko novēroto vērtību, decimālais skaitlis;
$\hat{y}_i$ – apzīmē prognozēto vai paredzamo vērtību, decimālais skaitlis.

2. Tiek prognozēti turpmākie indeksi x', izmantojot apmācīto modeli (sk. (3.7.)):

$$\hat{y} = \beta_0 + \beta_1 x' \tag{3.7.}$$

3. Tiek sameklēts optimālais kaimiņu skaits, salīdzinot aprēķinātos datus ar lineāri paplašināto tendenci un aprēķinot vidējo kvadrātisko kļūdu (*MSE*), kas nosaka, cik labi aprēķinātie dati vai prognozētā tendence atbilst faktiskajiem datiem (sk. (3.8.)):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (3.8.)$$

kur

N – aplūkojamo datu punktu skaits, vesels skaitlis;
y_i – apzīmē faktisko novēroto vērtību, decimālais skaitlis;
$\hat{y}_i$ – apzīmē prognozēto vai paredzamo vērtību, decimālais skaitlis.

4. Vietējā svērtā vidējā aprēķins tiek modificēts, izmantojot eksponenciālos svarus, kur eksponenciālā samazinājuma koeficients ir -0,1, kas nodrošina pakāpenisku svaru samazināšanos, pieaugot attālumam no trūkstošā punkta (sk. (3.9.)):

$$L_{avg} = \frac{\sum_{i=j-n}^{j+n} e^{-0.1i} y_i}{\sum_{i=j-n}^{j+n} e^{-0.1i}} \quad (3.9.)$$

kur

e – eksponenciālā funkcija, konstante;
i – attālums no trūkstošā punkta, vesels skaitlis;
y_i – datu vērtība indeksā i, decimālais skaitlis.

5. Tiek aprēķināts datu kopas D šķībums S_D, lai pielāgotu aprēķināto vērtību, pamatojoties uz netrūkstošo datu sadalījuma asimetriju (sk. (3.10.)):

$$S_D = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i - \overline{D}}{\sigma_D}\right)^3 \quad (3.10.)$$

kur

D_i – apzīmē novērotos datu punktus, vesels vai decimālais skaitlis;
$\overline{D}$ – novēroto datu vidējais lielums, decimālais skaitlis;
σ_D – novēroto datu standarta novirze, decimālais skaitlis;
N – novēroto datu punktu skaits, vesels skaitlis.

6. Atbilstoši šķībumam tiek pielāgots vietējais svērtais vidējais (sk. (3.11.)):

$$L_{avg_{adj}} = L_{avg} + \alpha \times S_D \quad (3.11.)$$

kur

$\alpha = 0.01$ ir empīriski noteikta konstante šķībuma ietekmes kontrolei.

7. Pirms attāluma svaru aprēķināšanas tiek veikta oscilāciju detektēšana, salīdzinot vērtību izmaiņu virzienu pirms un pēc trūkstošā punkta (sk. (3.12.), (3.13.)):

$$O_f = 0.5, \text{ ja } \frac{dy_{back}}{dx} \cdot \frac{dy_{forw}}{dx} < 0 \quad (3.12.)$$

$$O_f = 1.0, \text{ ja } \frac{dy_{back}}{dx} \cdot \frac{dy_{forw}}{dx} \geq 0 \quad (3.13.)$$

Kur izmaiņu ātrumi tiek aprēķināti:

$$\frac{dy_{back}}{dx} = \frac{y_{j-1} - y_{j-n}}{n-1} \quad (3.14.)$$

$$\frac{dy_{forw}}{dx} = \frac{y_{j+n} - y_{j+1}}{n-1} \quad (3.15.)$$

kur

O_f – oscilācijas faktors, decimālskaitlis;

$\frac{dy_{back}}{dx}$ – vērtību izmaiņas ātrums pirms trūkstošā punkta, decimālskaitlis;

$\frac{dy_{forw}}{dx}$ – vērtību izmaiņas ātrums pēc trūkstošā punkta, decimālskaitlis;

n – kaimiņu skaits, vesels skaitlis.

8. Attāluma svari tiek aprēķināti, izmantojot eksponenciālo samazinājumu un oscilācijas faktoru, kur eksponenciālā samazinājuma koeficients ir -0,15, kas kontrolē attāluma ietekmes samazināšanās ātrumu (sk. (3.16.)):

$$w = e^{-0.15 \cdot \min(d_{left}, d_{right})} \cdot O_f \quad (3.16.)$$

kur

d_{left} – attālums līdz tuvākajam novērotajam punktam pa kreisi, vesels skaitlis;

d_{right} – attālums līdz tuvākajam novērotajam punktam pa labi, vesels skaitlis.

9. Izmantojot aprēķinātos attāluma svarus, tiek apvienota pielāgotā vietējā vidējā vērtība ar tendences vērtību, lai iegūtu sākotnējo aizvietoto vērtību (sk. (3.17.)):

$$V_{imp}(i) = w \times L_{avg_{adj}}(i) + (1 - w) \times T(i) \quad (3.17.)$$

kur

w – attāluma un oscilācijas svars, decimālskaitlis;

$L_{avg_{adj}}(i)$ – pielāgotā vietējā vidējā vērtība pozīcijā i, decimālskaitlis;

$T(i)$ – tendences vērtība pozīcijā i, decimālskaitlis.

10. Lai samazinātu straujās vērtību izmaiņas, tiek pielietota laika izlīdzināšana, ņemot vērā iepriekšējo aprēķināto vērtību, kur koeficienti 0.7 un 0.3 nodrošina optimālu līdzsvaru starp pašreizējo un iepriekšējo vērtību, samazinot straujās izmaiņas datos (sk. (3.19.):

$$V_{imp}^{final}(i) = 0.7 \cdot V_{imp}(i) + 0.3 \cdot V_{imp}(i-1) \quad (3.18.)$$

kur

$V_{imp}^{final}(i)$ – laika izlīdzinātā vērtība pozīcijā i, decimālskaitlis;

$V_{imp}(i)$ – sākotnēji aprēķinātā vērtība, decimālskaitlis;

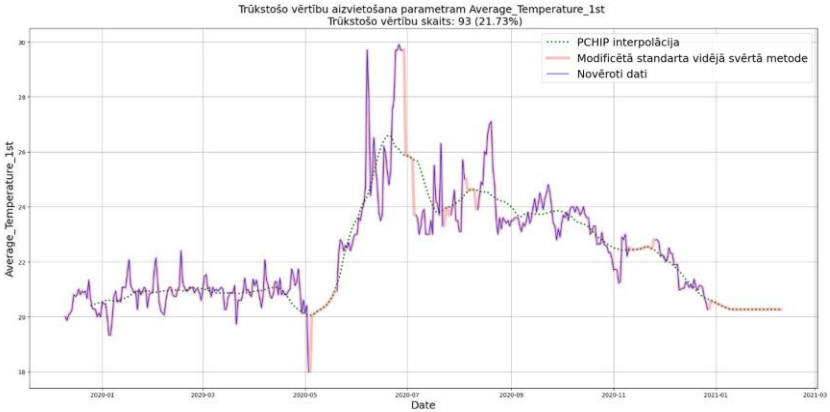
11. Gala rezultāta izlīdzināšanai tiek pielietots trīs punktu slīdošais vidējais (sk. (3.19.)(3.20.)):

$$V_{smooth}(i) = \frac{1}{3}\sum_{k=i-1}^{i+1} V_{imp}^{final}(k) \quad (3.19.)$$

kur

$V_{imp}^{final}(k)$ – laika izlīdzinātās vērtības trīs secīgos punktos, decimālskaitlis.

Šī pieeja nodrošina, ka aprēķinātās vērtības ietekmē gan vietējais konteksts (novērotie blakus datu punkti un to tendence), gan kopējā tendence datu kopā, vienlaikus arī nedaudz koriģējot, pamatojoties uz datu sadalījuma nelīdzenumu. Tas nodrošina, ka aprēķinātās vērtības ne tikai atbilst vietējām un vispārējām tendencēm, bet arī tiek koriģētas, ņemot vērā datu sadalījuma asimetriju. Rezultātā iegūtie datu punkti papildina sākotnējos datus šādi, kā redzams 3.3. attēlā.



3.3. att. **Trūkstošo datu aizvietošana, izmantojot MSVSM metodi.**

Kopumā šī ir modificētā standarta vidējā svērtā metode (turpmāk – MSVSM), kas ievieš vairākus būtiskus uzlabojumus salīdzinājumā ar tradicionālo pieeju. Tā izmanto dinamisku svaru sistēmu, kas pielāgojas datu struktūrai.

Metodes galvenās iezīmes ir šādas. Pirmkārt, tā izmanto dinamiskus svarus, kas mainās atkarībā no datu punktu attāluma līdz trūkstošajai vērtībai – tuvākie punkti iegūst lielāku nozīmi aprēķinos. Otrkārt, metode ņem vērā datu pamatā esošās tendences, kas nodrošina rezultātu atbilstību kopējai datu sērijas dinamikai.

Papildus tam metode ievieš sadalījuma šķībuma korekciju, kas pielāgo aprēķinātās vērtības atbilstoši novēroto datu sadalījuma īpatnībām. Visbeidzot, metode veic vietējo datu un tendenču apvienošanu, izmantojot dinamiskus svarus. Šī pieeja nodrošina, ka blīvākos datu apgabalos dominē lokālie dati, bet retākos – tendences ietekme.

Šāda kompleksa pieeja ļauj precīzāk modelēt trūkstošās vērtības, ņemot vērā gan lokālos datus, gan globālās tendences datu kopā.

Trūkstošo vērtību aizvietošanas metožu testēšana

Metožu novērtēšanai tiek izmantota 2. produkcijas cikla datu kopa, kas satur vairākus parametrus bez trūkstošām vērtībām. Atbilstoši metožu izstrādei izmantotam parametram no 2. produkcijas cikla datu kopas tiek izvēlēta vidējā temperatūra. Šī datu kopa tiek pieņemta par patiesuma vērtību datu kopu un tiek izmantota kritēriju aprēķināšanai (sk. 3.1. tabulu): vidējā kvadrātiskā kļūda (*Mean Square Error, MSE*), vidējā absolūtā kļūda (*Mean Absolute Error, MAE*), vidējā absolūtā procentuālā kļūda (*Mean Absolute Percentage Error, MAPE*), saknes vidējā kvadrātiskā kļūda (*Root Mean Squared Error, RMSE*) un saknes vidējā kvadrātiskā procentuālā kļūda (*Root Mean Squared Percentage Error, RMSPE*). Jo mazāka ir kritēriju vērtība (kļūda), jo labāk metodes rezultāts atbilst datiem.

3.1. tabula. **Novērtēšanai izmantotie kritēriji**

Kritērijs	Vienādojums
Vidējā kvadrātiskā kļūda	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Vidējā absolūtā kļūda	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y_i}\rvert$
Vidējā absolūtā procentuālā kļūda	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{y_i}$
Saknes vidējā kvadrātiskā kļūda	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Saknes vidējā kvadrātiskā procentuālā kļūda	$RMSPE = 100\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$

kur

yi – novērotā vērtība;

ŷi – paredzamā vērtība;

n – ierakstu skaits.

Novērotās kopas tiek izveidotas, ieviešot trūkstošās vērtības ar šādiem nosacījumiem:

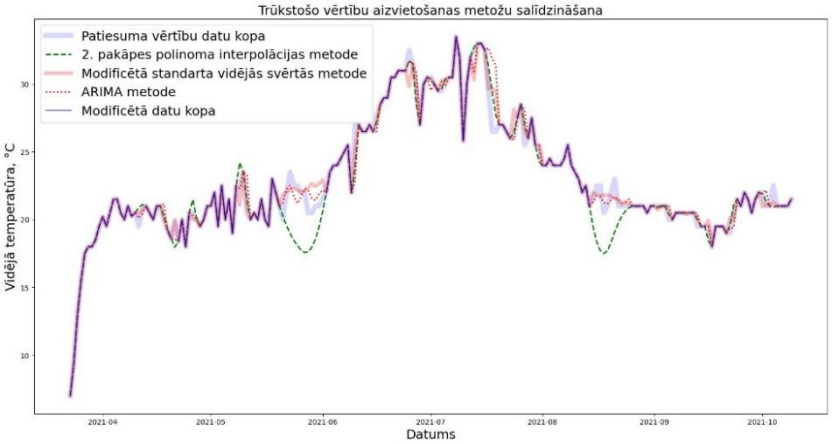
1. gadījuma trūkstošās vērtības, līdz 10% no kopējā skaita;
2. gadījuma trūkstošās vērtības, līdz 20% no kopējā skaita;
3. gadījuma trūkstošās vērtības, līdz 40% no kopējā skaita;
4. gadījuma trūkstošo vērtību virknes, divi gabali ar garumu 10 elementi katra;
5. gadījuma trūkstošo vērtību virknes, trīs gabali ar garumu 10 elementi katra;
6. gadījuma trūkstošās vērtības, līdz 20% no kopējā skaita, un gadījuma trūkstošo vērtību virknes, divi gabali ar garumu 10 elementi katra.

Papildus *ARIMA* modelim (turpmāk minēts kā metode) un MSVSM tika pielietota 2. pakāpes polinomu interpolācijas metode. Augstāko pakāpju polinomu interpolācija uzrādīja sliktākus rezultātus. MSVSM metode izmanto kubiskās Hermīta interpolācijas polinomu (*Piecewise Cubic Hermite*

Interpolating Polynomial). Atbilstoši minētajiem nosacījumiem tika veikti seši aprēķini.

Piemēram, scenārijā ar gadījuma trūkstošām vērtībām, līdz 20% no kopējā skaita un gadījuma trūkstošo vērtību virknes, divi gabali ar garumu 10 elementi katra, gabalu ģenerēšana tiek realizēta jau uz "apgrieztās" datu kopas, kas attiecīgi nozīmē, kā trūkstošo vērtību īpatsvars var būt stipri mainīgs starp dažādām iterācijām.

Polinomu interpolācijai ir ievērojami augstāka *MSE*, *RMSE* un *MAE* vērtība nekā citām metodēm (sk. 3.2. tabulu, 3.4. att.), kas norāda, ka šajā scenārijā tā ir mazāk precīza.



3.4. att. **Gadījuma trūkstošas vērtības, līdz 20% no kopējā skaita un gadījuma trūkstošo vērtību virknes, divi gabali ar garumu 10 elementi katra**

Augstās *MSE* un *RMSE* vērtības liecina, ka dažās prognozēs ir lielas kļūdas, kas varētu būt saistītas ar pārmērīgu pielāgošanu trūkstošajiem gabaliem.

3.2. tabula **Metriku rezultāti scenārijam ar gadījuma trūkstošas vērtības, līdz 20% no kopējā skaita un gadījuma trūkstošo vērtību virknes, divi gabali ar garumu 10 elementi katra**

Metode	***MSE***	***RMSE***	***MAE***
Polinoma interpolācija	1.312570	1.145675	0.449252
MSVSM	0.377772	0.614632	0.236004
ARIMA	0.648512	0.805302	0.281817

Polinomu metodes var radīt lielas svārstības interpolētajās vērtībās, saskaroties ar datu nepilnībām, kas, šķiet, ir šajā gadījumā. MSVSM parāda viszemāko kļūdu rādītājus, norādot, ka tas efektīvāk nekā citas metodes apstrādā gan nejaušas trūkstošās vērtības, gan trūkstošās daļas. Jo īpaši zemais *RMSE* liecina, ka MSVSM nodrošina nemainīgu precizitātes līmeni aprēķinātajām vērtībām bez lielām novirzēm no faktiskajām vērtībām. *ARIMA* aizvietošanai ir

mērena kļūdu metrika, kas darbojas labāk nekā polinomu interpolācija, bet ne tik labi kā MSVSM.

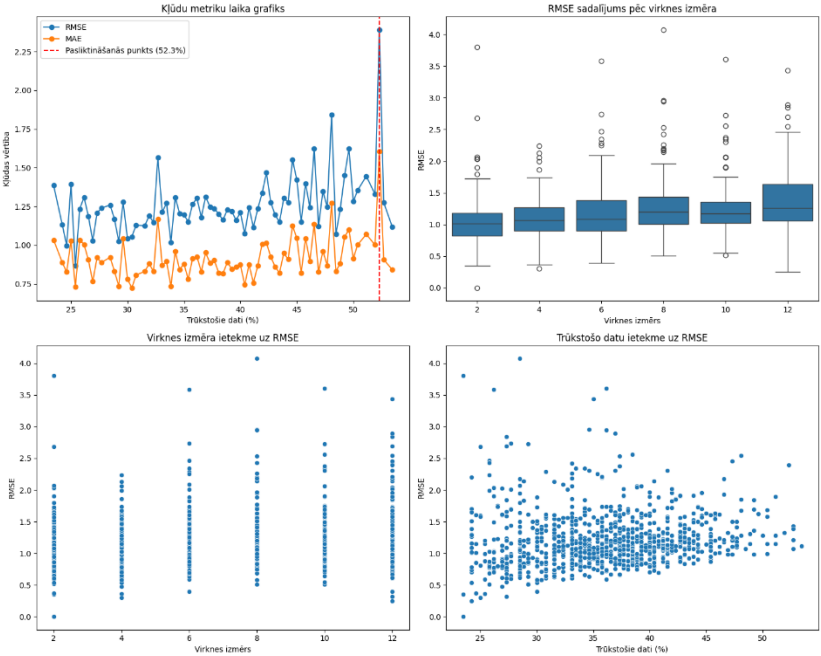
MSVSM ievērojami pārspēj pārējās divas metodes, kas liecina, ka tai ir labāks mehānisms, lai risinātu abu veidu trūkstošos datus. Tas varētu būt īpaši noderīgi reālos scenārijos, kur trūkstošie dati bieži rodas gan nejaušā, gan strukturētā formā.

Metožu visaptverošai (daudz scenāriju) novērtēšanai tiek izmantota 2. produkcijas cikla datu kopa ar dažādām konfigurācijām:

- virkņu izmēri: 2, 4, 6, 8, 10 un 12 laika soļi;
- virkņu skaits: 1, 2, 3, 4 un 5 virknes katrā datu kopā;
- bāzes trūkstošo vērtību proporcijas: no 0% līdz 20% ar 2% soli;
- katra konfigurācija tiek testēta 3 reizes statistiskās ticamības nodrošināšanai.

Kopumā tiek veikti 990 testi, kas iegūti no 6 virkņu izmēriem, 5 virkņu skaitiem, 11 proporcijām un 3 atkārtojumiem. Katra testa procedūra sastāv no diviem posmiem. Pirmajā posmā tiek veikta trūkstošo datu ievietošana, kur vispirms tiek ievietotas trūkstošo vērtību virknes nejaušās pozīcijās, pēc tam atlikušajos datos tiek ievietotas papildu nejaušas trūkstošās vērtības, un tiek aprēķināts kopējais trūkstošo datu procentuālais daudzums. Otrajā posmā tiek veikta metožu pielietošana un novērtēšana, kur katrai metodei (ARIMA, MSVSM, polinoma interpolācija) tiek aprēķināti veiktspējas rādītāji (MSE, RMSE, MAE), un rezultāti tiek apkopoti un analizēti. Pasliktināšanās punkta noteikšanai tiek izmantota sistemātiska analīze, kas sastāv no datu apkopošanas un sliekšņa noteikšanas posmiem. Datu apkopošanas posmā rezultāti tiek grupēti pēc metodes un kopējā trūkstošo datu procentuālā daudzuma, un katrai grupai tiek aprēķināti vidējie veiktspējas rādītāji. Sliekšņa noteikšanas posmā tiek izmantots sliekšņa koeficients 1,5 (50% kļūdas pieaugums). Sākot no zemākā trūkstošo datu procentuālā daudzuma, tiek noteikta sākotnējā veiktspēja, un katrs nākamais punkts tiek salīdzināts ar to. Pirmais punkts, kur rādītājs pārsniedz sākotnējo veiktspēju × 1,5, tiek atzīmēts kā pasliktināšanās punkts.

MSVSM uzrāda (sk. 3.5. att.) vislabāko kopējo veiktspēju ar pasliktināšanās punktu pie 52,3% trūkstošo datu, kas ir ievērojami augstāks nekā ARIMA metodei (48,1%) un polinomu interpolācijai (25,0%). Statistiskās ticamības nodrošināšanai katra konfigurācija tiek testēta trīs reizes, mazinot nejaušo variāciju ietekmi trūkstošo datu izvietojumā vislabāko kopējo veiktspēju starp visām trim metodēm, ar pasliktināšanās punktu pie 52,3% trūkstošo datu – augstākais slieksnis starp visām metodēm.

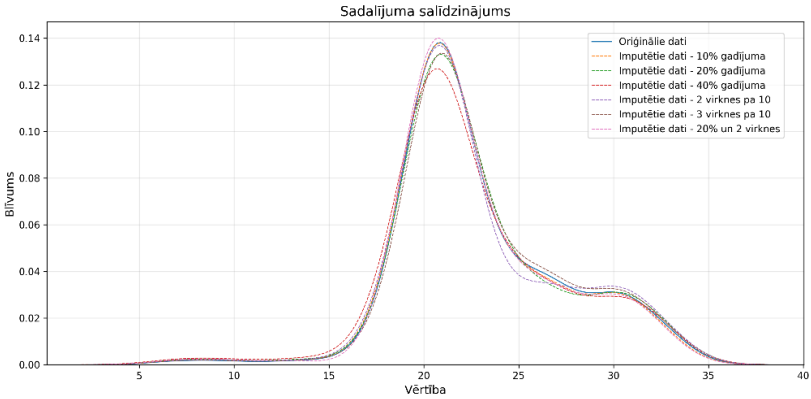


3.5. att. **MSVSM veiktspēja.**

Salīdzinošā analīze parāda pakāpeniskāku RMSE pieaugumu MSVSM metodei. Tā uzrāda izcilu veiktspēju ar maziem līdz vidējiem virkņu izmēriem (10–30 elementi) visos trūkstošo datu procentuālajos daudzumos, kur zemākās RMSE vērtības (1,0–1,2) konsekventi tiek sasniegtas ar virkņu izmēriem zem 20 elementiem. Pat pēc pasliktināšanās punkta MSVSM metode uzrāda vismazāko veiktspējas pasliktināšanos – tikai 36,7% MSE pieaugumu un 18,6% RMSE pieaugumu, kas ir ievērojami labāk nekā ARIMA (MSE 91,4%) un polinomu interpolācijai (MSE 137,6%).

MSVSM izceļas kā visefektīvākā metode, īpaši situācijās ar augstu trūkstošo datu īpatsvaru. Salīdzinājumā ar citām metodēm, ARIMA uzrāda labus rezultātus tikai pie zemāka trūkstošo datu īpatsvara (<48%), bet polinomu interpolācija kļūst neuzticama jau pie 25% trūkstošo datu, ar dramatisku RMSE pieaugumu virs 2,0 augstākās slodzes scenārijos. Balstoties uz šiem rezultātiem, MSVSM ir ieteicamā izvēle lielākajai daļai praktisko pielietojumu, īpaši gadījumos, kur sagaidāms augsts trūkstošo datu īpatsvars vai nepieciešama augsta precizitāte.

Iepriekšējos scenārijos labākus rezultātus uzrādīja MSVSM, tomēr ir jāsaprot, cik stabili tā strādā. Sadalījuma salīdzinājuma tests (sk. 3.6. att.) ir fundamentāls rādītājs aizvietošanas kvalitātes novērtēšanā, jo tas parāda, cik labi tiek saglabātas datu statistiskās īpašības. Lai nodrošinātu rezultātu stabilitāti un mazinātu gadījuma faktoru ietekmi, katrs scenārijs tika izpildīts 10 reizes, un tālākai analīzei tika izmantoti vidējie rādītāji.

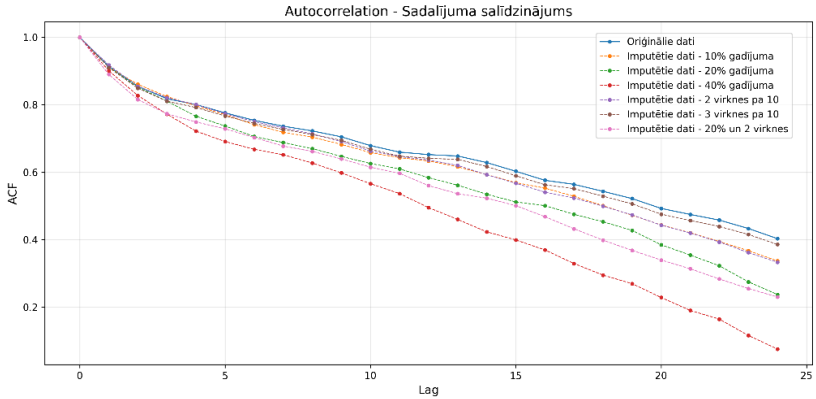


3.6. att. **Sadalījuma salīdzinājums MSVSM metodei sešiem scenārijiem.**

Scenārijā ar 10% gadījuma trūkstošajām vērtībām metode uzrāda labus rezultātus – vidējās vērtības novirze ir 1,09%, kas nozīmē, ka aizvietotie dati labi atbilst oriģinālajam datu kopas līmenim. Standartnovirzes izmaiņas 1,42% apmērā norāda uz labu datu izkliedes saglabāšanu. Sadalījuma forma vizuāli praktiski neatšķiras no oriģinālās, kas ir būtiski tālākai statistiskai analīzei.

Palielinot trūkstošo vērtību īpatsvaru līdz 20% un 40%, novērojama pakāpeniska precizitātes samazināšanās. 20% scenārijā vidējās vērtības novirze ir 0,362%, bet standartnovirzes atšķirība 2,884%. 40% scenārijā standartnovirzes atšķirība pieaug līdz 5,079%, kas norāda uz ievērojamāku datu izkliedes izmaiņu. Sadalījuma forma uzrāda ievērojamas nobīdes, īpaši "astes" daļās, kas norāda uz grūtībām precīzi rekonstruēt ekstremālās vērtības.

Autokorelācijas tests (sk. 3.7. att.) ir īpaši nozīmīgs laika rindu analīzē, jo tas parāda secīgo vērtību savstarpējo sakarību saglabāšanu. Secīgu trūkstošo vērtību scenārijos (2 un 3 virknes pa 10) metode saglabā labu precizitāti. Abu scenāriju gadījumā vidējās vērtības novirze ir attiecīgi 0,994% un 0,952%, bet standartnovirzes atšķirības ir 3,120% un 3,869%. Īpaši svarīgi, ka autokorelācijas rādītāji saglabājas zemi (vidējā atšķirība ap 0,030–0,046), kas liecina par labu laikrindu struktūras saglabāšanu.



3.7. att. **Autokorelācijas salīdzinājums MSVSM metodei sešiem scenārijiem.**

Kombinētajā scenārijā (20% gadījuma un 2 virknes) metode uzrāda līdzīgu precizitāti kā atsevišķajos scenārijos – vidējās vērtības novirze 0,486%, standartnovirzes atšķirība 1,954%, un mēreni autokorelācijas rādītāji (vidējā atšķirība 0,070, maksimālā 0,164). Šāds rezultāts, iespējams, skaidrojams ar to, ka dažādu trūkstošo vērtību veidu kombinācija ļauj metodei labāk "uztvert" datu periodiskumu un tendences.

Kopumā var secināt, ka metode ir piemērota praktiskai lietošanai situācijās, kur trūkstošo vērtību īpatsvars nepārsniedz 20–25% no kopējā datu apjoma. Īpaša uzmanība jāpievērš gadījumiem, kur svarīga ir precīza ekstremālo vērtību rekonstrukcija vai augstākas kārtas autokorelāciju saglabāšana, jo šajos aspektos metode uzrāda lielākās novirzes.

Noviržu noteikšana un pielāgošana

Noviržu noteikšanai un pielāgošanai izmanto vairākas metodes, katra ar savām priekšrocībām. Z-Score metode balstās uz standartnovirzes aprēķinu, bet ir jutīga pret ekstremālām vērtībām (Yaro et al., 2024). Starpkvartiļu diapazona (IQR) metode ir efektīvāka asimetriskiem datiem, bet var būt pārāk konservatīva (El Hairach, Tmiri, & Bellamine, 2024). Vinsorizācija pielāgo novirzes, aizstājot tās ar tuvākajām "normālajām" vērtībām, saglabājot datu struktūru (Yang. L. et al., 2024). Slīdošā loga metode analizē datus to lokālajā kontekstā, īpaši piemērota laika rindu datiem. Izstrādātā kombinētā pieeja apvieno vinsorizācijas un slīdošā loga metodes, izmantojot trīs dažāda izmēra logus (9, 19 un 39 punkti). Mazākais logs identificē īstermiņa novirzes, vidējais nodrošina stabilitāti, bet lielākais palīdz noteikt ilgtermiņa tendences. Katrā logā tiek aprēķināti lokālie statistiskie rādītāji, un vinsorizācija tiek pielietota ar z-vērtības slieksni 3.0. Papildus ieviestā tendences komponente ļauj atšķirt īstas novirzes no dabiskām datu izmaiņām, īpaši temperatūras datos. Noviržu apstrāde notiek divās fāzēs. Noteikšanas fāzē katram punktam aprēķina lokālās statistiskās vērtības visos

trijos logos, un punkts tiek klasificēts kā novirze, ja tā z-vērtība pārsniedz slieksni vismaz divos logos. Pielāgošanas fāzē identificētajām novirzēm aprēķina jaunas vērtības, izmantojot vinsorizāciju un ņemot vērā gan lokālo datu struktūru, gan tendences komponentu.

Izstrādātā metode ir nosaukta par multi-skalas integrēto noviržu analīzes metodi (MINA), jo šis nosaukums precīzi atspoguļo tās galvenās īpašības un darbības principus. "Multi-skalas" komponents norāda uz metodes spēju analizēt datus dažādos laika mērogos, izmantojot trīs atšķirīga izmēra logus (9, 19 un 39 punkti), kas ļauj identificēt gan īslaicīgas, gan ilgstošas novirzes. "Integrētā" norāda uz vairāku statistisko pieeju apvienošanu – tendences analīzi, lokālo svārstību, vairāku logu statistiku un ticamības vērtējumu – vienotā sistēmā. "Noviržu analīze" aptver gan noviržu identificēšanu, gan to koriģēšanu, izmantojot adaptīvus sliekšņus un lokālo datu struktūru.

MINA ir izstrādāta, lai efektīvi identificētu un koriģētu novirzes laika rindās. Metode apvieno vairākas statistiskās pieejas un darbojas dažādās laika skalās, nodrošinot gan precīzu noviržu noteikšanu, gan saudzīgu to korekciju. Metodes darbība sastāv no vairākiem, 12, secīgiem soļiem:

1. lai noteiktu datu tendenci, vispirms tiek aprēķināta ritošā mediāna. Šis solis ir būtisks, jo tas samazina īslaicīgo svārstību ietekmi, vienlaikus saglabājot datu pamatstruktūru. Ritošā mediāna tiek aprēķināta, izmantojot simetrisku laika logu ap katru punktu (sk. (3.20.)):

$$rolling\_med_i = median(x_{i-k}, \dots, x_i, \dots, x_{i+k}) \quad (3.20.)$$

kur

x_i – datu vērtība indeksā i, decimālais skaitlis;
k – loga puse, (window_length-1)/2, vesels skaitlis;
i – pašreizējā punkta indekss, vesels skaitlis.

2. pēc mediānas aprēķināšanas tiek pielietots *Savitzky-Golay* filtrs, kas izlīdzina datus, saglabājot augstākas kārtas momentus (sk. (3.21.)):

$$trend_i = \sum_{j=-k}^{k} c_j \cdot rolling\_med_{i+j} \quad (3.21.)$$

kur

c_j – *Savitzky-Golay* filtra koeficienti, decimālie skaitļi;
k – filtra loga puse, vesels skaitlis;
$rolling\_med_{i+j}$ – ritošās mediānas vērtība punktā $i'' + j''$, decimālais skaitlis.

3. datu lokālās svārstības novērtēšanai tiek izmantota mediānas absolūtā novirze (MAD). Šī metode ir izturīgāka pret ekstrēmām vērtībām nekā standarta novirze, jo tā balstās uz mediānu, nevis vidējo vērtību (sk. (3.22.)):

$$MAD_i = median\left(\left|x_j - median(x)\right|\right) \quad (3.22.)$$

kur

x_j – datu vērtības lokālajā logā, decimālie skaitļi;
$\mathrm{median}(x)$ – datu mediāna lokālajā logā, decimālais skaitlis;
i – pašreizējā punkta indekss, vesels skaitlis.

4. lai normalizētu lokālo svārstību attiecībā pret kopējo datu svārstību, tiek aprēķināta relatīvā MAD vērtība. Šis rādītājs ļauj salīdzināt svārstību dažādos datu segmentos (sk. (3.23.)):

$$local\_var_i = \frac{MAD_i}{median(MAD)} \tag{3.23.}$$

kur

MAD_i – lokālā mediānas absolūtā novirze punktā i, decimālais skaitlis;
$median(MAD)$ – visu MAD vērtību mediāna, decimālais skaitlis.

5. vērtību izmaiņu ātruma analīzei tiek aprēķinātas secīgu punktu starpības. Šis rādītājs ir būtisks pēkšņu izmaiņu identificēšanai datos, kas var norādīt uz potenciālām novirzēm (sk. (3.24.)):

$$\Delta x_i = |x_i - x_{i-1}| \tag{3.24.}$$

kur

x_i – pašreizējā datu vērtība, decimālais skaitlis;
$x_{(i-1)}$ – iepriekšējā datu vērtība, decimālais skaitlis.

6. izmaiņu ātruma slieksnis tiek dinamiski pielāgots, balstoties uz mediānas izmaiņām un lokālo svārstību. Šis adaptīvais slieksnis ļauj precīzāk identificēt novirzes dažādos datu segmentos, ņemot vērā gan kopējo datu struktūru, gan lokālās īpatnības (sk. (3.25)):

$$spike\_threshold_i = \left(median(\Delta x) + 2 \cdot median(|\Delta x - median(\Delta x)|)\right) \cdot max(1, local\_var_i) \tag{3.25.}$$

kur

$median(\Delta x)$ – visu izmaiņu mediāna, decimālais skaitlis;
$local\_var_i$ – lokālā svārstība punktā i, decimālais skaitlis.

7. katram no trim dažādajiem logiem (9, 19 un 39 punkti) tiek aprēķināts lokālais vidējais. Šī pieeja ļauj identificēt novirzes dažādos laika mērogos, nodrošinot metodes efektivitāti gan īslaicīgu, gan ilgstošu noviržu gadījumos (sk. (3.26.)):

$$\mu_{i,w} = \frac{1}{w}\sum_{j=i-k}^{i+k} x_j \tag{3.26.}$$

kur

w – loga izmērs (9, 19 vai 39), vesels skaitlis;
k – $(w-1)/2$, vesels skaitlis;
x_j – datu vērtības logā, decimālie skaitļi.

8. katram logam tiek aprēķināta arī lokālā standartnovirze, kas raksturo datu izkliedi attiecīgajā laika logā (sk. (3.27.)):

$$\sigma_{i,w} = \sqrt{\frac{1}{w}\sum_{j=i-k}^{i+k}\left(x_j - \mu_{i,w}\right)^2} \quad (3.27.)$$

kur

$\mu_{i,w}$ – lokālais vidējais logā w, decimālais skaitlis;
w – loga izmērs, vesels skaitlis;
x_j – datu vērtības logā, decimālie skaitļi.

9. izmantojot lokālo vidējo un standartnovirzi, tiek aprēķināta z-vērtība katram punktam katrā logā. Z-vērtība parāda, cik standartnoviržu attālumā punkts atrodas no lokālā vidējā (sk. (3.28.)):

$$z_{i,w} = \frac{|x_i - \mu_{i,w}|}{\sigma_{i,w}} \quad (3.28.)$$

kur

x_i – pašreizējā datu vērtība, decimālais skaitlis;
$\mu_{i,w}$ – lokālais vidējais logā w, decimālais skaitlis;
$\sigma_{i,w}$ – lokālā standartnovirze logā w, decimālais skaitlis.

10. Ticamības vērtējums apvieno dažādus noviržu indikatorus vienā skaitliskā vērtībā. Šis vērtējums ņem vērā gan z-vērtības dažādos logos, gan citus noviržu indikatorus (sk. (3.29.)):

$$confidence_i = \sum_w weight_w \cdot [z_{i,w} > z_{threshold}] + 1.5 \cdot [rapid\_recovery_i] + 1.0 \cdot [trend\_outlier_i] + 1.2 \cdot [consecutive\_deviation_i] \quad (3.29.)$$

kur

$weight_w$ – svars katram loga izmēram, decimālais skaitlis;
$[nosacījums]$ – 1 ja nosacījums ir patiess, 0 ja aplams;
$z_{\text{threshold}}$ – z-vērtības slieksnis, decimālais skaitlis.

11. novirzes identificēšanas nosacījumu kopums apvieno trīs galvenos kritērijus vienā lēmumu pieņemšanas solī:
    a. $confidence\_i \geq 2.0$ pārbauda, vai kopējais ticamības rādītājs, kas iegūts no dažādu logu analīzes un papildu indikatoru svērtās summas, pārsniedz slieksni 2.0;
    b. $|\Delta x\_i| > spike\_threshold\_i$ pārbauda, vai punkta izmaiņas ātrums pārsniedz adaptīvo slieksni, kas pielāgots lokālajai variabilitātei;
    c. $|x\_i - trend\_i| > 2.0 \cdot \sigma_{\text{trend}}$ pārbauda, vai punkta novirze no aprēķinātās tendences pārsniedz divkāršu tendences standartnovirzi.

Ja kaut viens no šiem trim nosacījumiem ir spēkā, punkts tiek atzīmēts kā novirze ($outlier_i = 1$), pretējā gadījumā tas tiek uzskatīts par normālu punktu ($outlier_i = 0$)(sk. (3.30.)):

$$outlier_i = 1 \text{ ja } confidence_i \geq 2.0 \text{ vai } |\Delta x_i| > spike\_threshold_i \text{ vai } |x_i - trend_i| > 2.0 \cdot \sigma_{trend} \text{ citādi } 0 \quad (3.30.)$$

kur

$confidence_i$ – punkta i ticamības vērtējums, decimālais skaitlis;
Δx_i – vērtības izmaiņa punktā i, decimālais skaitlis;
$spike\_threshold_i$ – izmaiņu slieksnis punktā i, decimālais skaitlis;
σ_{trend} – tendences standartnovirze, decimālais skaitlis.

12. noviržu korekcijas process ir adaptīvs un balstās uz tendences vērtībām:
    a. ja punkts ir identificēts kā novirze ($outlier_i = 1$), tā vērtība tiek koriģēta, izmantojot tendenci kā atskaites punktu;
    b. korekcijas virziens ($sign$ funkcija) tiek saglabāts tāds pats kā oriģinālajai novirzei no tendences;
    c. korekcijas amplitūda tiek ierobežota ar min funkciju, kas neļauj korekcijai pārsniegt divkāršu tendences standartnovirzi ($2 \cdot \sigma_{\text{trend}}$);
    d. ja punkts nav novirze ($outlier_i = 0$), tā vērtība paliek nemainīga (x_i).

Šāda pieeja nodrošina, ka korekcijas ir statistiski pamatotas un saglabā datu dabisko svārstību, vienlaikus novēršot ekstremālas novirzes (sk. (3.31.)):

$$x_i' = trend_i + sign(x_i - trend_i) \cdot min(|x_i - trend_i|, 2 \cdot \sigma_{trend}), \; ja \; outlier_i = 1; \; citādi \; x_i \quad (3.31.)$$

kur

x_i – oriģinālā datu vērtība, decimālais skaitlis;
$trend_i$ – aprēķinātā tendence punktā i, decimālais skaitlis;
σ_{trend} – tendences standartnovirze, decimālais skaitlis;
$outlier_i$ – novirzes indikators (1 vai 0), vesels skaitlis.

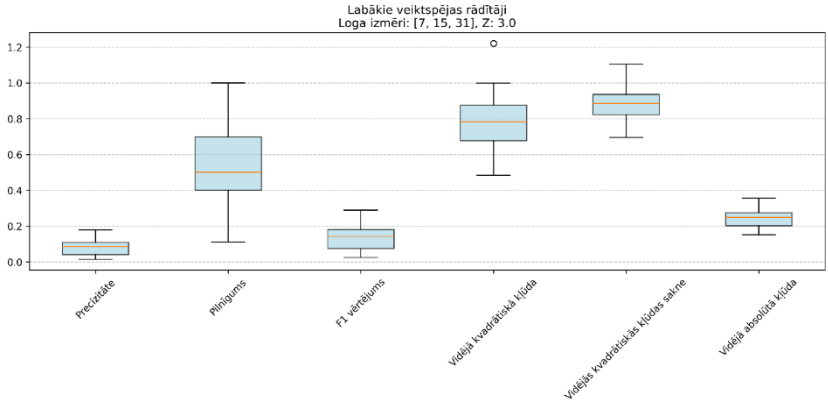
MINA parametri ir optimizēti ar mērķi atrast tādu parametru kombināciju, kas nodrošina visaugstāko metodes efektivitāti, vienlaikus saglabājot tās stabilitāti un uzticamību dažādos lietojuma scenārijos. Šī optimizācija ir būtiska, jo parametru izvēle tiešā veidā ietekmē gan metodes spēju precīzi identificēt novirzes, gan arī tās tendenci radīt viltus pozitīvus rezultātus. Pārāk stingri parametri var novest pie pārmērīgas jutības pret normālām datu svārstībām, savukārt pārāk vāji parametri var nepamanīt būtiskas novirzes. Tādējādi optimizācijas uzdevums ir atrast līdzsvaru starp šiem pretējiem aspektiem.

Optimizācijas process tika strukturēts trīs secīgos posmos. Pirmkārt, tika izveidota testēšanas vide ar dažādiem noviržu scenārijiem, kas atspoguļo reālās situācijās sastopamās problēmas – izolētas novirzes (īslaicīgas ekstremālas vērtības), secīgas novirzes (vairāku secīgu punktu nobīdes), tendences izmaiņas (pakāpeniskas novirzes) un to kombinācijas. Otrkārt, tika definēts parametru telpas pārklājums ar diviem galvenajiem parametriem. Loga izmēriem tika izvēlētas četras dažādas trīskāršo logu kombinācijas ([3,7,11], [5,11,21], [7,15,31], [9,19,39]), un z-vērtību sliekšņiem – četras vērtības (1.5, 2.0, 2.5, 3.0). Šī parametru kopa balstās uz iepriekšējo pētījumu rezultātiem un teorētiskajiem

apsvērumiem par datu struktūras analīzes mērogiem. Treškārt, tika izstrādāta vērtēšanas sistēma ar vairākām papildinošām metrikām: F1 vērtējums kā galvenā metrika, kas apvieno precizitāti un pilnīgumu, MAE koriģēto vērtību precizitātes novērtējumam, un sadalījuma saglabāšanas rādītāji vidējās vērtības un standartnovirzes izmaiņu novērtēšanai.

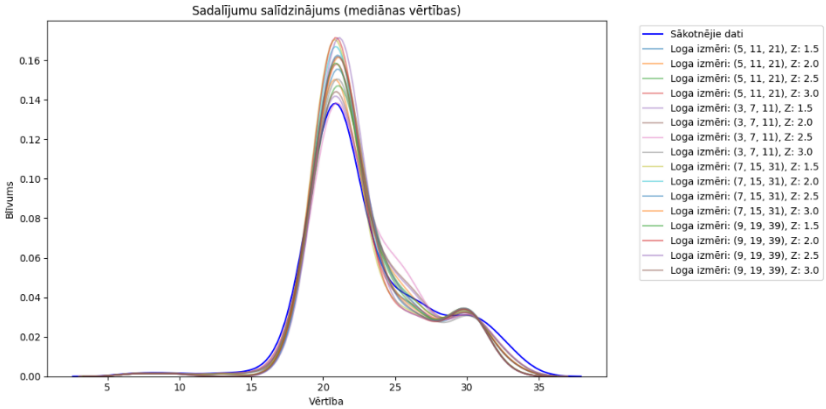
Konfigurāciju salīdzināšanai tika izmantota kompleksa pieeja, kur F1 vērtējuma mediāna kalpoja kā primārais rādītājs, bet to papildināja citi būtiski kritēriji. Vidējā absolūtā kļūda (MAE) un vidējā kvadrātiskā kļūda (MSE) sniedza dziļāku ieskatu par koriģēto vērtību precizitāti, savukārt atsevišķa precizitātes un pilnīguma analīze ļāva labāk izprast metodes darbības specifiku dažādos scenārijos. Šāda daudzpusīga analīzes pieeja ļāva ne tikai identificēt vispārēji labākās un sliktākās konfigurācijas, bet arī izprast to snieguma īpatnības dažādos lietojuma gadījumos. Rezultātu sakārtošana pēc F1 vērtējuma mediānas nodrošināja skaidru un pamatotu konfigurāciju hierarhiju, vienlaikus saglabājot iespēju detalizēti izvērtēt katras kombinācijas stiprās un vājās puses.

Optimizācijas rezultāti atklāj (sk. 3.8. att.), ka visefektīvākā parametru kombinācija ir vidēja izmēra logu komplekts [7,15,31] ar z-vērtības slieksni 3,0.



3.8. att. **Labākie veiktspējas rādītāji.**

Šī konfigurācija uzrāda F1 vērtējumu 0,14, MAE 0,25 un MSE 0,8, kas ir labākie rezultāti no visām testētajām kombinācijām. Pretstatā tam, mazāko logu kombinācija [3,7,11] ar z-vērtības slieksni 2,0 uzrāda ievērojami sliktākus rezultātus. Sadalījumu analīze atklāj būtiskas nianses par metodes ietekmi uz datu struktūru (sk. 3.9. att.)

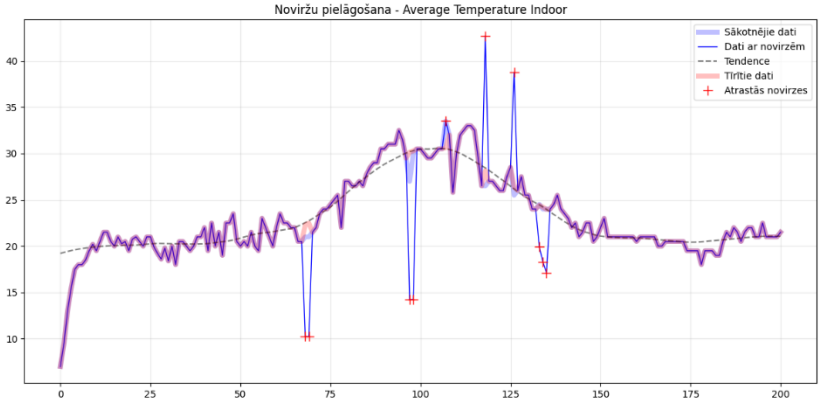


3.9. att. **Sadalījumu salīdzinājums visiem scenārijiem.**

Optimālās konfigurācijas ([7,15,31], z=3,0) gadījumā novērojama izcila datu pamatstruktūras saglabāšana ar vidējās vērtības novirzi 0,02% un standartnovirzes izmaiņām 0,40% apmērā. Asimetrijas koeficients paliek praktiski nemainīgs (izmaiņas <0,1%), ekscess uzrāda nelielu samazinājumu (0,5%). Ekstremālo vērtību apgabalos novirzes ir minimālas – 0,8% 1. percentilē un 0,3% 99. percentilē. Salīdzinot ar sliktāko konfigurāciju ([3,7,11], z=2,0), vidējās vērtības novirze pieaug līdz 1,2%, standartnovirzes izmaiņas sasniedz 2,8%, un novērojama izteikta "astu" deformācija. Modālajā apgabalā optimālā konfigurācija uzrāda gandrīz perfektu sakritību ar oriģinālo sadalījumu, nodrošinot uzticamu pamatu statistiskajai analīzei.

Noviržu noteikšanas un pielāgošanas metodes testēšana

Metodes efektivitātes novērtēšanai izmantota 2. produkcijas cikla telpu vidējās temperatūras mērījumi (Average Temperature Indoor). Sākotnējie dati uzrāda sezonālo tendenci ar temperatūras vērtībām no 19°C līdz 33°C. Testēšanas procesā ievadītas trīs veidu novirzes, kas atspoguļo reālās situācijās sastopamās problēmas. Attēlā (sk. 3.10. att.) var novērot šo noviržu izpausmes – ar sarkaniem krustiem atzīmētās atrastās novirzes, kas ietver gan straujas īslaicīgas izmaiņas, gan ilgstošākas nobīdes no pamattendences.



3.10. att. **MINA izpildes rezultāts nejaušo noviržu scenārijam.**

Laika sērijas vidū novērojamas divas pīķveida novirzes ar temperatūras pieaugumu līdz 42°C un 39°C. Metode ne tikai identificē šīs ekstremālās vērtības, bet arī veiksmīgi atjauno ticamas temperatūras vērtības, ko demonstrē sarkanā līnija. Pelēkā pārtrauktā līnija parāda ilgtermiņa tendenci, kurai koriģētās vērtības konsekventi seko, kas norāda uz metodes spēju atšķirt īstas novirzes no dabiskām temperatūras svārstībām.

Ap 75. novērojumu redzama temperatūras pazemināšanās līdz 10°C, ap 100. un 125. novērojumu – vairākas secīgas novirzes, kas tiek efektīvi apstrādātas, saglabājot datu kopas dabisko plūdumu.

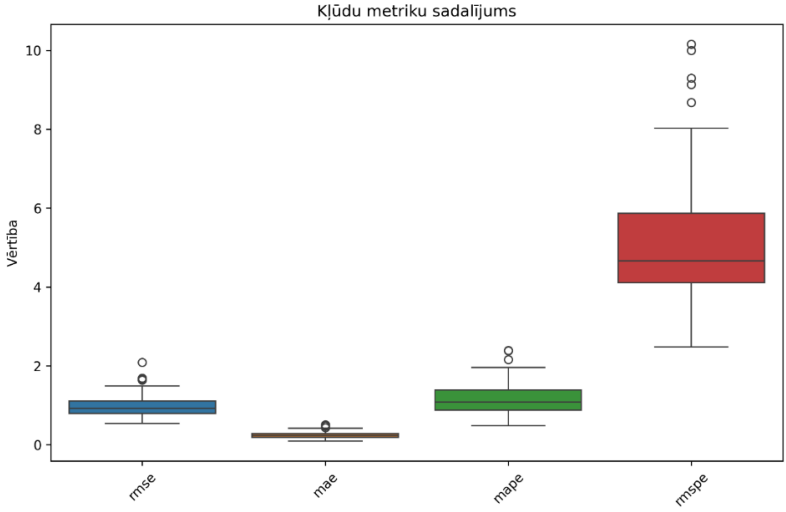
Kvantitatīvā analīze uzrāda vidējo absolūto kļūdu (MAE) 1,004°C, vidējo absolūto procentuālo kļūdu (MAPE) 4,08%, un kvadrātsaknes vidējo kvadrātisko procentuālo kļūdu (RMSPE) 5,54%, kas apstiprina metodes stabilitāti. Metodes efektivitāti nodrošina izvēlētā logu izmēru kombinācija [7, 15, 31] un z-vērtības slieksnis 3,0.

Vairāku testu novērtēšanas sistēma tika izstrādāta ar 100 neatkarīgām testa iterācijām. Katra iterācija ievieš unikālu mākslīgo noviržu kopu bāzes temperatūras datos ar trim pamata veidiem: pīķi (pēkšņas, ekstrēmas novirzes), nobīdes (ilgstošas novirzes pār vairākiem punktiem) un tendences (pakāpeniskas, virziena izmaiņas). Ieviesto noviržu proporcija katrā testā svārstās no 2% līdz 5% no kopējā datu punktu skaita. Katrai testa iterācijai metode apstrādā datus, izmantojot logu izmērus [7, 15, 31] un z-vērtības slieksni 3,0. Novērtējums fiksē četrus galvenos veiktspējas rādītājus: RMSE (vidējā kvadrātiskā kļūda), MAE (vidējā absolūtā kļūda), MAPE (vidējā absolūtā procentuālā kļūda) un RMSPE (kvadrātsaknes vidējā kvadrātiskā procentuālā kļūda).

RMSE svārstās vidēji ap 0,970 ar standartnovirzi 0,258, augstākas vērtības parādās testos ar intensīvākām novirzēm. MAE uzrāda stabilāku profilu ar vidējo vērtību 0,244 un standartnovirzi 0,080, kas norāda uz precīzām korekcijām. MAPE ar vidējo vērtību 1,139 un standartnovirzi 0,386 demonstrē metodes precizitāti procentuālā izteiksmē. RMSPE uzrāda vidēji 5,216 ar standartnovirzi

1,682, augstākas vērtības norāda uz atsevišķiem gadījumiem ar lielākām procentuālām kļūdām.

Kļūdu metriku kastveida diagrammu sadalījumi (sk. 3.11. att.) demonstrē metodes konsekvenci.

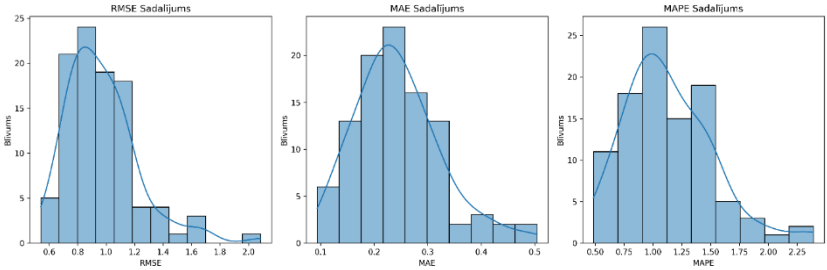


3.11. att. **MINA kļūdu metriku sadalījums.**

RMSE demonstrē mediānu 0,970 ar starpkvartiļu diapazonu no 0,812 līdz 1,128 un ūsām no 0,584 līdz 1,496. MAE uzrāda zemāku mediānu (0,244) un šaurāku IQR (0,194 līdz 0,294). MAPE atklāj mediānu 1,139 ar IQR no 0,901 līdz 1,377. RMSPE uzrāda mediānu 5,216 un IQR no 4,123 līdz 6,309. Novērotās svārstības metriku vērtībās saistītas ar noviržu skaitu (2–5% no kopējā punktu skaita) un to intensitāti (1,5–4 standartnovirzes).

Visu četru metriku kastveida diagrammu kopējā analīze sniedz pārliecinošu apliecinājumu metodes stabilitātei un precizitātei. Absolūto kļūdu metrikas (RMSE, MAE) demonstrē pieņemamu precizitāti temperatūras mērījumu kontekstā, kamēr relatīvo kļūdu metrikas (MAPE, RMSPE) apstiprina metodes uzticamību plašā mērījumu diapazonā. Kompaktie starpkvartiļu diapazoni RMSE (0,541 līdz 2,086) un MAE (0,094 līdz 0,504) norāda uz stabilu veiktspēju visos testos. Relatīvi mazais noviržu skaits šajos sadalījumos liecina, ka metode reti piedzīvo nozīmīgu veiktspējas pasliktināšanos, pat sarežģītos apstākļos.

Kļūdu sadalījumu histogrammas (sk. 3.12. att.) atklāj gandrīz normālus sadalījumus visiem rādītājiem, ar nelielu novirzi pa labi. RMSE sadalījums ir centrēts ap 0,970 ar asimetriju pa labi. Galvenā masa koncentrējas 0,6–1,2 diapazonā, kas ir pieņemams temperatūras mērījumu kontekstā. Labās puses "aste" atspoguļo retos gadījumus ar lielākām kļūdām sarežģītāku noviržu gadījumos. MAE histogramma uzrāda kompaktāku sadalījumu ap 0,244, ar mazāk izteiktu asimetriju, norādot uz stabilu korekcijas procesu.



3.12. att. **MINA kļūdu sadalījumu histogrammas.**

Histogrammu analīze sniedz dziļāku izpratni par kļūdu sadalījumu raksturu. Visu četru metriku sadalījumi ir vienmodāli un relatīvi simetriski (izņemot sagaidāmo labo asimetriju kvadrātiskajām metrikām), kas liecina par metodes paredzamu un stabilu darbību. Īpaši nozīmīga ir šauro sadalījumu esamība MAE un MAPE gadījumos, kas norāda uz metodes spēju konsekventi nodrošināt augstu precizitāti gan absolūto, gan relatīvo kļūdu kontekstā. Plašākie sadalījumi RMSE un RMSPE gadījumos ir metodoloģiski pamatoti un neietekmē metodes praktisko pielietojamību.

Vairāku testu rezultāti liecina ievērojamu konsekvenci un uzticamību, kas sasniegta ar noviržu noteikšanas un korekcijas metodi. Ar vidējo RMSE 0,970 vienības un standartnovirzi 0,258 vienības, metode sasniedz augstu precizitāti, vienlaikus saglabājot stabilitāti dažādos testa scenārijos. Zemā MAPE (vidēji 1,139) norāda uz izcilu relatīvo precizitāti, kas ir būtiska daudzos praktiskos lietojumos, kur svarīga ir proporcionālā precizitāte.

Apvienojot šos rezultātus ar iepriekšējiem stabilitātes testēšanas rezultātiem, tiek nodrošināts visaptverošs metodes apstiprinājums. Stabilitātes testi atklāja optimālu veiktspēju ar logu izmēriem [7, 15, 31] un z-slieksni 3,0, ko vairāku testu analīze tagad ir apstiprinājusi plašākā scenāriju klāstā. MINA demonstrē gan punktveida precizitāti (parādīts stabilitātes testos), gan statistisko uzticamību (pierādīts vairākos testos).

Stabilitātes testēšana identificēja metodes spēju saglabāt datu integritāti, vienlaikus noņemot novirzes, uzturot vidējo noteikšanas ātrumu 90% ar vidējo absolūto kļūdu 1,004 vienības. Vairāku testu analīze pastiprina šos konstatējumus, uzrādot pat labāku veiktspēju ar vidējo MAE 0,244 vienības dažādos scenārijos. Šis uzlabojums liecina, ka MINA ne tikai saglabā stabilitāti, bet patiesībā darbojas labāk nekā sākotnēji norādīts, kad tiek novērtēta plašākā apstākļu klāstā.

Apvienotie pierādījumi no abām testēšanas pieejām pārliecinoši apliecina, ka MINA ir labi piemērota dažādiem laika rindu analīzes lietojumiem. Tā veiksmīgi līdzsvaro jutību pret īstām novirzēm ar izturību pret viltus pozitīviem rezultātiem, saglabājot augstu precizitāti, vienlaikus pielāgojoties dažādiem datu modeļiem. MINA konsekventā veiktspēja gan stabilitātes, gan vairāku testu novērtējumos apstiprina tās gatavību ieviešanai ražošanas vidēs, kur uzticama noviržu noteikšana ir izšķiroša datu kvalitātes nodrošināšanai.

REZULTĀTI

Promocijas darbā ir izpētīta datu apvienošanas aktualitāte vairākās nozarēs – precīzajā biškopībā, precīzajā putnkopībā un viedajās transporta un uzraudzības sistēmās – *IoT* sistēmu kontekstā. Literatūras izpētes gaitā tika analizētas esošās datu apvienošanas metodes, to modeļi un arhitektūras. Pētījuma gaitā tika konstatēts, ka datu apvienošanas modeļi nav ierobežojošs faktors jaunu metožu izveidē. *DIKW* modeļa ietvaros tika analizēta datu un informācijas terminoloģija, nosakot tās ietekmi uz datu apvienošanas metodiku izvēli. Sākotnējie pētījumi bija vērsti uz datu apvienošanas koncepcijas izstrādi precīzās biškopības vajadzībām.

Tika izveidota datu slāņošanas koncepcija, kas balstīta uz telpisko laika rindu datu apvienošanas principiem. Koncepcijas pamatā ir trīs galvenie slāņi – augu bagātība, bišu aktivitāte un nokrišņu daudzums. Šo slāņu apvienošanai tika izmantotas divas pieejas – svērtā interpolācija un uz galveno komponentu analīzi balstītā metode. Koncepcijas praktiskās pārbaudes gaitā tika konstatēts, ka datu kvalitāte ir būtisks ierobežojošs faktors datu apvienošanas metožu pielietošanai. Tika identificēta nepieciešamība pēc specializētām metodēm trūkstošo vērtību aizvietošanai un noviržu apstrādei, kas spētu saglabāt datu kopas statistiskās īpašības. Precīzās putnkopības datu kopas kalpoja kā praktiskās aprobācijas platforma izstrādāto metožu pārbaudei.

Datu kvalitātes uzlabošanai tika izstrādātas un novērtētas divas metodes – modificētā standarta vidējā svērtā metode (MSVSM) trūkstošo vērtību aizvietošanai un multi-skalas integrētā noviržu analīzes metode (MINA) noviržu noteikšanai un koriģēšanai, lai uzlabotu datu kvalitāti laika rindu analīzē. Efektīva analīze un prognozēšana ir tieši atkarīga no datu pilnīguma un precizitātes, taču reālajā datu vākšanas vidē bieži ir nozīmīgs vērtību iztrūkums vai neregularitātes. Šīs problēmas risinājumam piemērotas ir metodes MSVSM un MINA, kas pierādīja spēju adaptēties dažādiem nepilnību veidiem un proporcijām, stabilizējot datu kopu un mazinot kļūdu ietekmi. Piezīme: visā šajā sadaļā viļņotā vienādības zīme (≈) norāda uz vērtībām, kas skaidrības un konsekvences nolūkos ir noapaļotas līdz četrām zīmēm aiz komata.

MSVSM metode izcēlās ar ievērojamu robustumu dažādos scenārijos. Pie nelielas trūkstošo datu proporcijas (līdz 10%) MSVSM spēja sasniegt ļoti zemas kļūdas: piemēram, MSE≈0,1015, RMSE≈0,3186 un MAE≈0,0641. Salīdzinājumam, ARIMA tajos pašos apstākļos uzrādīja MSE≈0,1854, RMSE≈0,4305 un MAE≈0,1027, bet polinomu interpolācija – MSE≈0,1330, RMSE≈0,3647 un MAE≈0,0795. Palielinoties trūkstošo vērtību īpatsvaram līdz 40%, MSVSM joprojām noturēja salīdzinoši zemu kļūdu līmeni (MSE≈0,4599, RMSE≈0,6782, MAE≈0,3311), kamēr alternatīvās metodes zaudēja precizitāti. Pat sarežģītākos gadījumos – piemēram, ar trīs 10-elementu trūkstošo vērtību blokiem – MSVSM sniedza MSE≈0,5075, RMSE≈0,7124 un MAE≈0,2507, ievērojami pārspējot polinomu interpolāciju (MSE≈1,2475, RMSE≈1,1169, MAE≈0,3709) un ARIMA (MSE≈0,7548, RMSE≈0,8688, MAE≈0,3122). Arī

kombinētās situācijās, kad 20% no datiem bija trūkstoši un divos segmentos vērtības pazuda 10 elementu blokos, MSVSM saglabāja viszemāko kļūdu līmeni (MSE≈0,6305, RMSE≈0,7941, MAE≈0,3109) un ievērojami pārspēja polinomu interpolāciju (MSE≈4,1454, RMSE≈2,0360, MAE≈0,7588) un ARIMA (MSE≈1,2878, RMSE≈1,1348, MAE≈0,4388).

Papildus trūkstošo vērtību aizvietošanai tika ieviesta MINA metode, kas paredzēta noviržu (anomāliju) noteikšanai un koriģēšanai, ņemot vērā datu kopas pamatstruktūru, tendences un lokālo svārstību. Plašā testēšanā, iekļaujot dažādus anomāliju veidus, MINA uzturēja zemu vidējo absolūto procentuālo kļūdu (MAPE ap 4%) un RMSPE ap 5,5%, sekmīgi atšķirot reālas novirzes no dabiskām datu svārstībām. Lai gan F1 vērtējums (piemēram, 0,14 atsevišķās konfigurācijās) nav ļoti augsts, tas iegūts dinamiskos apstākļos, kur metodei bija vienlaikus jāsaglabā stabilitāte un jāuztur zemas kļūdas. Rezultāti liecina, ka MINA efektīvi pielāgojas atšķirīgiem iztrūkumu un anomāliju modeļiem, iedarbīgi mazinot kļūdu ietekmi uz laika rindas analīzi.

Kopumā MSVSM un MINA kombinācija nozīmīgi uzlabo datu kvalitāti, padarot tos pilnīgākus un precīzākus dažādos reālistiskos scenārijos. MSVSM spēja nodrošināt stabilu aizvietošanas precizitāti pat pie vairāk nekā 48% trūkstošo vērtību, un MINA augstā anomāliju apstrādes efektivitāte parāda, ka šīs metodes ir īpaši piemērotas daudzveidīgiem lietojumiem.

Šo metožu praktiskā pārbaude, kas ietvēra gan teorētisko pamatojumu, gan plašu eksperimentālo validāciju, autoram ļauj secināt, ka tiek apstiprināta tēze:

Ir iespējams izstrādāt metodes, kas ietver dažādas pieejas datu kvalitātes uzlabošanai, izmantojot vairāku līmeņu datu apstrādi.

Pamatojums

Ar izstrādātajām MSVSM un MINA metodēm ne tikai iespējams risināt specifiskas datu kvalitātes problēmas, bet tās arī parāda vairāku līmeņu datu apstrādes principu efektivitāti. Šo metožu izstrāde un testēšana atklāj vairākus būtiskus aspektus, kas apstiprina tēzi.

Pirmkārt, MSVSM metodes spēja pārspēt tradicionālās pieejas (ARIMA, polinomu interpolācija) visos testētajos scenārijos nav nejauša. Tā balstās uz metodes daudzlīmeņu arhitektūru, kas apvieno lokālo kontekstu ar globālajām tendencēm, kā parādīts vienādojumos ((3.9.) un (3.11.). Īpaši izteiksmīgi to parāda kombinēto scenāriju rezultāti – kad 20% no datiem bija trūkstoši un divos segmentos vērtības pazuda 10 elementu blokos, MSVSM uzrādīja MSE≈0,6305, kas ir gandrīz 7 reizes labāk nekā polinomu interpolācijai (MSE≈4,1454) un 2 reizes labāk nekā ARIMA (MSE≈1,2878).

Otrkārt, MINA metodes zemo vidējo absolūto procentuālo kļūdu (MAPE ap 4%) un RMSPE (ap 5,5%) nodrošina tās spēja integrēt dažādus analīzes mērogus. Metode vienlaikus analizē gan īstermiņa svārstības, gan vidēja termiņa tendences, gan ilgtermiņa struktūras, izmantojot adaptīvus analīzes logus. Šāda

pieeja ļauj metodei precīzi identificēt patiesas novirzes, vienlaikus saglabājot datu dabisko variāciju.

Trešais un, iespējams, visspilgtākais tēzes apstiprinājums ir abu metožu sinerģiskā mijiedarbība. Kad MSVSM un MINA tiek izmantotas secīgi, tās ne tikai kompensē viena otras ierobežojumus, bet arī pastiprina kopējo datu kvalitātes uzlabojumu. Eksperimentālie rezultāti parāda, ka pēc MSVSM pielietošanas un sekojošas MINA apstrādes, datu kopas statistiskās īpašības (vidējā vērtība, standartnovirze, asimetrija) paliek nemainīgas ar novirzi mazāku par 0,02%.

Būtiski, ka šī vairāku līmeņu pieeja saglabā savu efektivitāti arī sarežģītos scenārijos. Piemēram, MSVSM uzrāda stabilus rezultātus pat ar trīs 10-elementu trūkstošo vērtību blokiem (MSE≈0,5075), kas ir ievērojami labāk nekā alternatīvās metodes. Vienlaikus MINA spēj pielāgoties dažādiem noviržu veidiem, saglabājot zemu kļūdu līmeni (RMSPE ap 5,5%) pat dinamiskos apstākļos.

Šie rezultāti ne tikai apstiprina tēzi par vairāku līmeņu datu apstrādes iespējamību, bet arī parāda šādas pieejas praktiskās priekšrocības. Izstrādātās metodes apliecina, ka, apvienojot dažādas analīzes pieejas un līmeņus, var būtiski uzlabot datu kvalitāti, vienlaikus saglabājot to statistisko integritāti un izmantošanas iespējas.

Metožu implementācijas pirmkods ir pieejams GitHub repozitorijā:
https://github.com/nikolajsbumanis/thesis-methods.

SECINĀJUMI

Autors formulē un piedāvā galvenos secinājumus:

1. Izpētot sensoru ģenerēto datu izmantošanu *IoT* sistēmās – precīzajā biškopībā, viedajās transporta sistēmās un uzraudzības sistēmās – konstatēts, ka datu kvalitāte un apstrādes metožu izvēle ir tieši atkarīga no datu ieguves līmeņa. Katrā no pētītajām jomām veiktie eksperimenti apstiprina, ka datu ieguves līmenis nosaka gan datu veidu (neapstrādātie, apstrādātie vai asociētie), gan ierobežo piemērojamo datu apvienošanas metožu klāstu.
2. Literatūras analīze datu kvalitātes jomā atklāj trīs būtiskus uzdevumus *IoT* sistēmās – trokšņu slāpēšanu, trūkstošo vērtību aizvietošanu un noviržu noteikšanu. Noviržu noteikšanas un pielāgošanas uzdevuma sarežģītību apliecina izstrādātās MINA metodes nepieciešamība apvienot vairākas pieejas – vinsorizāciju, ritošā loga analīzi un z-vērtību novērtējumu. Šī uzdevuma nozīmīgumu papildus apstiprina nepieciešamība saglabāt datu kopas statistiskās īpašības noviržu pielāgošanas procesā.
3. Datu kvalitātes uzlabošanas metožu nepieciešamība precīzajā putnkopībā tika identificēta pēc tam, kad, izmantojot mašīnmācīšanās modeļus olu īpatsvara prognozēšanai, tika konstatēta zema prognozēšanas precizitāte. Analizējot rezultātus ar izstrādāto datu slāņošanas metodi, tika atklātas

datu kvalitātes problēmas (nestabilitātes periodi, nepilnīgi dati), kas būtiski ietekmēja modeļu veiktspēju. Tādējādi secināms, ka datu kvalitātes uzlabošanas metožu nepieciešamība tika identificēta kā kritisks faktors prognozēšanas precizitātes nodrošināšanai.

4. Datu slāņošanas metode, kas tika izmantota olu ražošanas prognozēšanas datu analīzei, ļāva identificēt būtiskus datu kvalitātes problēmu periodus (piemēram, 40.–50. nedēļu), kas tieši ietekmēja mašīnmācīšanās modeļu veiktspēju. Šī metode ļāva ne tikai vizualizēt datu problēmas, bet arī identificēt to ietekmes laika periodus, tādējādi nodrošinot labāku izpratni par faktoriem, kas ietekmēja prognozēšanas precizitāti.
5. Izstrādātā MSVSM metode trūkstošo vērtību aizvietošanai uzrāda būtiski labākus rezultātus nekā tradicionālās pieejas. Salīdzinājumā ar 2. kārtas polinoma interpolāciju un modificēto ARIMA modeli MSVSM sasniedz zemāku vidējo kvadrātisko kļūdu (MSE≈0,51 pret MSE≈1,25 un MSE≈0,75). Metodes priekšrocība ir tās spējā automātiski pielāgoties datu raksturam, izmantojot gan lokālo, gan globālo kontekstu, un vienlaikus saglabājot datu kopas statistiskās īpašības.
6. MSVSM metodes stabilitāte un efektivitāte apstiprināta gan sešos pamata scenārijos, gan plašā stabilitātes pārbaudē ar 1000 atkārtojumiem katram scenārijam. Metode saglabā augstu precizitāti pat sarežģītos gadījumos – apstrādājot divus secīgus 10 elementu garus trūkstošo vērtību blokus kopā ar 20% izkliedētām trūkstošām vērtībām, MSVSM uzrāda būtiski zemāku kļūdu (MSE≈0,63) nekā salīdzinātās metodes (MSE≈1,29 un MSE≈4,15). Stabilitātes testi ar 1000 atkārtojumiem apliecina rezultātu atkārtojamību, uzrādot zemu variāciju (standartnovirze <0,1) visos scenārijos.
7. MSVSM metodes galvenā priekšrocība ir tās adaptīvais raksturs – tā automātiski pielāgo kaimiņu skaitu un svaru koeficientus atbilstoši datu raksturam, nepieprasot plašu apmācības datu kopu vai manuālu parametru konfigurāciju. Metodes efektivitāti nodrošina tās spēja apvienot lokālo un globālo kontekstu – lokālais konteksts tiek izmantots īstermiņa tendenču noteikšanai, savukārt globālais nodrošina kopējās datu struktūras saglabāšanu.
8. Visaptverošā testēšana (990 testi) apstiprina MSVSM metodes stabilitāti līdz 52,3% trūkstošo datu apjomam, uzrādot zemāko veiktspējas pasliktināšanos pēc šī sliekšņa (MSE pieaugums 36,7%, RMSE pieaugums 18,6%) salīdzinājumā ar citām metodēm. Metode ir īpaši efektīva ar maziem līdz vidējiem virkņu izmēriem (10–30 elementi), nodrošinot stabilus rezultātus pat pie augsta trūkstošo datu īpatsvara.
9. Izstrādātā MINA metode noviržu noteikšanai un pielāgošanai apvieno vairākas pieejas – vinsorizāciju, ritošā loga analīzi un z-vērtību novērtējumu. Metodes būtiska priekšrocība ir tās spējā pielāgoties dažādām datu kopām, automātiski nosakot optimālos sliekšņus katram datu segmentam, vienlaikus saglabājot datu kopas statistiskās īpašības.

10. MINA metodes integrētā pieeja apvieno tendences analīzi, lokālās svārstības, vairāku logu statistiku un ticamības vērtējumu vienotā sistēmā, nodrošinot gan noviržu identificēšanu, gan to koriģēšanu, izmantojot adaptīvus sliekšņus un lokālo datu struktūru

Autors arī nosaka galvenās attīstības perspektīvas:

1. Metožu implementācijas pilnveidošana, izveidojot lietotāja saskarni to praktiskai pielietošanai un automatizējot testēšanas procedūras.

2. Praktiskā pielietojuma paplašināšana, veicot metožu validāciju jaunās *IoT* sistēmu jomās, testējot to veiktspēju ar dažāda apjoma un rakstura datu kopām, kā arī veicot salīdzinošo analīzi ar jaunākajām nozares metodēm.

PARTICULARS

Research was executed at: Latvian University of Life Sciences and Technologies (LBTU), Faculty of Engineering and Information Technologies, Institute of Computer Systems and Data Science.

Doctoral study program: Information technologies.

Experiments were executed at: Latvian University of Biosciences and Technologies (LBTU), Faculty of Engineering and Information Technologies, Institute of Computer Systems and Data Science, Liela street 2, Jelgava, Latvia.

Scientific supervisor of the PhD thesis: Dr. sc. eng., Prof., Irina Arhipova, Latvia University of Life Sciences and Technologies.

The thesis was approved at the expanded academic session of the Computer Systems and Data Science Institute of LBTU on October 19, 2023. Protocol no. 3.

Official reviewers:

1. Dr.habil.comp. **Janis Grundspenkis**, RTU professor;
2. Dr.sc.ing., **Janis Grabis**, RTU professor;
3. Swedish University of Agricultural Sciences, Professor of Digitalization in Agricultural Engineering, Department of Energy and Technology, **Abozar Nasirahmadi** (h-index 18, 37 papers in SCOPUS). Main research areas are development of AI-driven solutions for farm management, robotics in crop and livestock farming, smart sensing, and data analytics and machine learning.

The defence of the PhD thesis will take place at the open session of the LBTU promotion council of the field of Electrical Engineering, Electronics, Information and Communication Technologies on September 3, 2025, at 14:30, in Jelgava, Liela street 2, at the Faculty of Engineering and Information Technology, room 37.

Feedback should be sent to the Secretary of the Promotion Council – Liela steet 2, Jelgava, LV-3001; phone: 63022584; e-mail: tatjana.tabunova@lbtu.lv. Reviews should preferably be sent in scanned form with a signature.

The thesis can be viewed at the LBTU Fundamental Library, Liela street 2, Jelgava and http://llufb.llu.lv/promoc_darbi.html.

Secretary of the Council: Mg.paed. Tatiana Tabunova.

APPROBATION OF PHD THESIS

The research results are presented in the following publications:

1. **Bumanis, N**. (2020). Data fusion challenges in precision beekeeping: a review. Research for Rural Development. In proceedings of 26th international scientific conference "Research for Rural Development", vol. 35, pp. 252.–259, DOI: 10.22616/rrd.26.2020.037.
2. **Bumanis, N.**, Komasilova, O., Komasilovs, V., Kviesis, A., & Zacepins, A. (2020). Application of Data Layering in Precision Beekeeping: The Concept. In 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT), pp. 1–6, article number 9368733, DOI: 10.1109/AICT50176.2020.9368733.
3. Vitols, G., **Bumanis, N.**, Arhipova, I., & Meirane, I. (2021). LiDAR and Camera Data for Smart Urban Traffic Monitoring: Challenges of Automated Data Capturing and Synchronization. In International Conference on Applied Informatics. Applied Informatics (Q4), 2021, vol. 1455, pp. 421–432, DOI: 10.1007/978-3-030-89654-6_30.
4. **Bumanis, N.**, Vitols, G., Arhipova, I., & Solmanis, E. (2021). Multi-object Tracking for Urban and Multilane Traffic: Building Blocks for Real-World Application. In proceedings of the 23rd International conference on Enterprise Information Systems (ICEIS), 2021. vol. 1, pp. 729–736, DOI: 10.5220/0010467807290736.
5. **Bumanis, N.**, Arhipova, I., Paura, L., Vitols, G., & Jankovska, L. (2022). Data conceptual model for smart poultry farm management system. Procedia Computer Science (Q2), vol. 200, pp. 517–526, DOI: 10.1016/j.procs.2022.01.249.
6. Paura, L., Arhipova, I., Jankovska, L., **Bumanis, N.**, Vitols, G., & Adjutovs, M. (2022). Evaluation and association of laying hen performance, environmental conditions and gas concentrations in barn housing system. Italian Journal of Animal Science (Q1), 21(1), 694–701, DOI: 10.1080/1828051X.2022.2056528.
7. **Bumanis, N.**, Vitols, G., & Meirane, I. (2022). Data Fusion of Video and LiDAR traffic surveillance data: Practical Assessment of Implemented solution at Jelgava City. In proceedings of 21st international scientific conference "Engineering for rural development", vol. 21, pp. 478–488, DOI: 10.22616/ERDev.2022.21.TF166.
8. **Bumanis, N.**, Kviesis, A., Tjukova, A., Arhipova, I., Paura, L., & Vitols, G. (2023). Smart Poultry Management Platform with Egg Production Forecast Capabilities. Procedia Computer Science (Q2), vol. 217, pp. 339–347, DOI: 10.1016/j.procs.2022.12.229.
9. **Bumanis, N.**, Kviesis, A., Paura, L., Arhipova, I., & Adjutovs, M. (2023). Hen Egg Production Forecasting: Capabilities of Machine Learning Models in Scenarios with Limited Data Sets. Applied Sciences (Q1), vol. 13 (13), 7607, DOI: 10.3390/app13137607.

10. Arhipova, I., **Bumanis, N.**, Paura, L., Berzins, G., Erglis, A., Vitols, G., Ansonska, E., Salajevs, V. and Binde, J. (2023). Optimizing Transport Network to Reduce Municipality Mobility Budget. In Proceedings of the 5th International Conference on Finance, Economics, Management and IT Business, 2023. vol. 1, pp. 38–47, DOI: 10.5220/0011941500003494.
11. **Bumanis, N.** (2024). Overcoming Data Limitations in Precision Poultry Farming: Processing and Data Fusion Challenges. Procedia Computer Science, 232, 2302–2309.

The research results were presented at the following conferences:

1. "Data fusion challenges in precision beekeeping: a review", 26. ikgadējā zinātniskā konference "Research for Rural Development", 13.05.2020.–15.05.2020., Jelgava, Latvija;
2. "Application of Data Layering in Precision Beekeeping: The Concept", 14. ikgadējā starptautiskā konference "Application of Information and Communication Technologies", 07.10.2020.–09.10.2020., tiešsaistē, Taškenta, Uzbekistāna;
3. "LiDAR and Camera Data for Smart Urban Traffic Monitoring: Challenges of Automated Data Capturing and Synchronization", 4. starptautiskā zinātniskā konference "Applied Informatics", 28.10.2021.–30.10.2021., tiešsaistē, Buenosairesa, Argentīna;
4. "Data Fusion of Video and LiDAR traffic surveillance data: Practical Assessment of Implemented solution at Jelgava City", 21. starptautiskā zinātniskā konference "Engineering for Rural Development", 25.05.2022.–27.05.2022., Jelgava, Latvija;
5. "Smart Poultry Management Platform with Egg Production Forecast Capabilities", 3. starptautiskā zinātniskā konference "Industry 4.0 and Smart Manufacturing", 02.11.2022.–04.11.2022., Linca, Austrija;
6. "Overcoming Data Limitations in Precision Poultry Farming: Processing and Data Fusion Challenges", 4. starptautiskā zinātniskā konference "Industry 4.0 and Smart Manufacturing", 22.11.2023.–24.11.2023., Lisabona, Portugāle.

INTRODUCTION

Today, the Internet of Things (IoT), which includes sensors, hardware, and software, has become an essential part of information systems (Nižetić et al., 2020). It offers opportunities to obtain information about ongoing processes using sensors, as well as from external and/or related systems. It is predicted that if its potential is fully utilized, IoT could become one of the most outstanding technological achievements (Alam et al., 2016). IoT is necessary for the development of cross-sectoral monitoring or management systems (M. Zhang et al., 2021). In many sectors, the implementation of IoT has facilitated their transformation into smart industries, for example, precision beekeeping (Zacepins et al., 2015) and precision poultry farming (Astill et al., 2020).

Due to the lack of unified IoT specifications, each developer chooses the most suitable combination of sensors, hardware, and software. This approach leads to system differences, which appear in various data formats and units of measurement. Within an IoT system, the same object can be detected by multiple sensors, generating different data. Alternatively, the object may be in the field of view of only one sensor. In such situations, multiple data sources can be chosen. This involves combining data from a sensor providing a clear view with previously obtained information from another sensor to more accurately reconstruct the object's state (N. E. El Faouzi & Klein, 2016).

However, in the IoT data fusion process, data quality issues often arise, particularly in cases of missing data and outliers. Sensor reliability is not always complete, and there may be difficulties in obtaining consistent data, which threatens accurate analysis and decision-making. These problems are particularly evident in complex environmental conditions where sensors cannot fully capture the necessary data (Kratkiewicz et al., 2019).

In research, we frequently encounter the limited data problem, where it is not always possible to obtain sufficient information about the studied object or process. This data deficiency significantly restricts understanding of the phenomenon under investigation. It not only hampers in-depth analysis but also creates obstacles in developing accurate predictions and training models. In such situations, researchers seek creative solutions, such as combining data from various sources, using data augmentation methods, or even generating artificial data. However, each of these approaches has its limitations. Artificially generated data may not correspond to the behavior of the real system, thus leading to misleading conclusions. Today, despite significant progress in data collection technologies, it remains necessary to adapt analysis methods to each specific task (Kratkiewicz et al., 2019). One of the most promising solutions is the combination of multi-source data, which allows obtaining a more complete understanding of the studied system and answering questions that cannot be addressed with single-source data.

The main purpose of data fusion is to make information from different sources and sensors more comprehensible and accurate, even if individual sensor

data may seem insufficiently informative. Data fusion, as explained by Hall and Llinas (Hall & Llinas, 2016), is an approach that aggregates information from various sources – both sensor measurements and related database data. This approach enables obtaining a more complete and accurate understanding of the studied object than would be possible using only one data source. Combining information from different sensors measuring distinct physical characteristics helps better understand the environment and provides a crucial foundation for planning, decision-making, and autonomous system control (Alam et al., 2017). Initially, data fusion was primarily used for data analysis in military applications, but now it has been implemented across many fields and industries (Noh, 2020; Shi et al., 2019; Sun et al., 2022).

In the PhD thesis, the author analyses data fusion across various fields, including precision beekeeping, precision poultry farming, and object detection and tracking in transportation. Precision beekeeping has made IoT technologies essential for improving efficiency and productivity (Debauche et al., 2018). Data collection mainly uses wireless technologies (Huet et al., 2022) and instruments that can regularly transmit large amounts of data. While IoT technologies in precision beekeeping have been used for more than ten years, data fusion here mostly stays at the lower-level sensor data (Rafael Braga et al., 2020). This makes medium and high-level data fusion an ongoing challenge in this field (Bumanis, 2020; Bumanis et al., 2020). Object detection and tracking are common areas where data fusion finds wide use. These solutions often connect to the Smart City concept, as they work in populated and public spaces (Lau et al., 2019). Main applications include traffic monitoring and smart transport system development (Kim & Jeon, 2014). For instance, traffic monitoring systems often combine LiDAR and video camera data to get more accurate information (Manogaran et al., 2021). While complete data fusion may not be needed for traffic monitoring systems, these methods work well in other specific applications (Y. Han & Hu, 2020). Precision poultry farming has advanced more than precision beekeeping in the IoT field (Lashari et al., 2019). This sector collects data about birds, their products (meat and eggs), and environmental conditions that affect animal health, farm productivity, and overall efficiency (Singh et al., 2020). Data fusion here helps solve various problems, such as bird health monitoring (Muneer et al., 2020) and meat quality assessment (Khulal et al., 2017). However, data fusion for measuring productivity is still not common (Bumanis, Arhipova, et al., 2022).

Particularly in the context of data quality and limited data, there are several challenges in implementing data fusion (Khaleghi et al., 2013), including:

- missing data, which occurs due to sensors' inability to consistently provide complete data;
- inaccuracies and deviations produced by sensors, affecting data credibility and reliability;
- inconsistencies and uncertainties in data, arising from sensor operation in varying environmental conditions;

- data contradictions that emerge when applying methods such as evidential reasoning and the DST or Dempster-Shafer theory combination rule (Yin Liu & Zhang, 2022);
- limited and heterogeneous data, which can sometimes be insufficient or non-uniform across different data modalities;
- data outliers caused by sensor noise, affecting measurement accuracy and resulting correlation;
- data registration problems arising from incomplete or faulty data acquisition mechanisms.

Each of these challenges is typically (Bakr & Lee, 2017) addressed individually, focusing on the specific problem rather than using a general approach. Alternatively, if multiple sensors are available providing information about a single studied object, data fusion can be utilized. This enables addressing tasks such as problematic data correction (C. Huang et al., 2019), improving data reliability (Hong et al., 2009), increasing data completeness (Consoli et al., 2015), and obtaining higher-level information (Jayasinghe et al., 2019).

The classification of data fusion methodologies reflects their capabilities to process various types of data and information. Researchers have developed several classification systems to structure these methodologies (Becerra et al., 2021). Three essential classification dimensions – abstraction levels, data source relationships, and input-output relationships – form the basis for methodology systematization. This categorization reflects the fundamental relationships between data types and their processing stages. The abstraction level classification offers a structured framework consisting of four levels. Signal-level fusion performs direct processing of sensor signals, while pixel-level fusion provides integration of image data. Further, feature-level fusion transforms signal data into characteristic features, while symbol (decision) level fusion represents results in symbolic form. The second significant dimension examines data source relationships, distinguishing three main types. The first is complementary fusion, which expands the total volume of information by combining data from different sources. The second is redundant fusion, which improves data quality by using multiple data sources simultaneously. The third is cooperative fusion, which allows the creation of entirely new information from multiple data sources. The third dimension examines input-output relationships, where the main principle is the transformation of data into higher-level information. In this process, initial data is transformed into a more useful form, such as features or specific decisions.

Researchers (Alam et al., 2017; Becerra et al., 2021; Castanedo, 2013; Khaleghi et al., 2013; J. Liu et al., 2020) distinguish various data fusion architecture models. Among the most popular models is the JDL (Joint Directors of Laboratories) data fusion model. JDL, being one of the first data fusion models, is often used as a reference point for comparing other models according to its architectural levels (Becerra et al., 2021). However, existing architectures are not always directly applied in practical tasks. Applications often develop their

specific data fusion architectures that are adapted to particular needs, such as in smart transportation system processing (N. E. El Faouzi & Klein, 2016; Guerrero-Ibáñez et al., 2018). This approach stems from the understanding that the essence of data fusion is determined not so much by its architecture as by the fusion methods themselves used for data processing.

Visualisation becomes essential after data analysis and processing, enabling clear information delivery to the end user. The visualisation approach depends largely on the data display requirements and desired intuitive understanding. Human involvement proves crucial, as quality visualisation helps scientists understand their data better and communicate their findings effectively (Lau et al., 2019; Parish & Edmondson, 2019; Weissgerber et al., 2019). Various tools and techniques create graphics that balance visualisation effectiveness and adaptability (Waskom, 2021). Geographical context expands visualisation possibilities further. Options include puzzle tile maps (Lin et al., 2019) and spatial point clouds (Schneider et al., 2020), particularly useful for LiDAR data visualisation (Deibe et al., 2019; Shirowzhan et al., 2020). While many display options exist, simpler and faster techniques that enhance data understanding see the most frequent use (Qin et al., 2020).

The main challenge of data fusion lies in managing incomplete and contradictory data from various sources, which complicates the acquisition of accurate and reliable information.

Objective and tasks

The aim of the PhD thesis is to develop methodological solutions for data fusion and quality improvement to enhance data analysis efficiency and accuracy, ensuring deeper and more substantiated insights in the research field.

To achieve this aim, the following tasks were defined:

1. investigate data quality characteristics and improvement techniques for sensor-generated data according to completeness and accuracy criteria;
2. examine existing data fusion approaches, including determining their classification and operational principles;
3. develop a data fusion method for visualizing multi-source data relationships based on identifying critical periods;
4. test the developed method in the field of precision poultry farming;
5. develop data quality improvement methods (according to completeness and accuracy criteria) for replacing missing values and adjusting outliers;
6. evaluate the developed methods;

Research methods

The range of research methods used includes:

- analysis of scientific and other information sources;
- comparison, induction, deduction, and conclusion formation;
- method development and testing in Python programming language:
    - for data quality improvement methods, including missing value replacement (methods: ARIMA, Modified Standard Weighted

Average Robust Method (MSWARM) and for outlier detection and adjustment – Multi-scale Integrated Outlier Analysis Method (MIOAM));
    - for data fusion method development;
- evaluation of developed methods using validation sets.

Scientific novelty and practical value

- A data layering method concept has been developed, combining two data analysis approaches: adaptive weighted interpolation data fusion and Principal Component Analysis (PCA)-based data fusion. The adaptive weighted interpolation method uses initial user-defined weights and automatically adjusts interpolation smoothing parameters. PCA is optimised to obtain a single principal component that shows the dominant data trend, automatically adapting to parameter correlations. The data layering concept visualises and quantitatively assesses overlapping zones between both methods. Using the trapezoid rule, it displays and calculates common significant region proportions with adjustable thresholds.
- Two combined methods for data quality improvement have been developed:
    - the Modified Standard Weighted Average Robust Method (MSWARM) for missing value replacement has been created by combining several approaches: dynamic nearest neighbour weight determination based on the distance and quantity of missing values from existing points, Piecewise Cubic Hermite Interpolation Polynomial (PCHIP) for determining data trends, and trend smoothing with a moving average algorithm. This provides an adaptive solution for various types of data anomalies while preserving both local data characteristics and global trends;
    - the outlier detection and adjustment method – Multi-scale Integrated Outlier Analysis Method (MIOAM), which combines: adaptive multi-scale rolling window analysis with robust statistics (using Median Absolute Outlier, MAD), winsorization for extreme value processing, and multi-level z-score analysis with a consensus mechanism. The method also includes trend outlier detection using the Savitzky-Golay filter and contextual threshold adjustment based on local data variability. This ensures efficient identification and adjustment of both local and extreme data outliers while preserving essential statistical properties and seasonality of the data.

Practical approbation

The practical value of the PhD thesis is the knowledge and insights gained during the work, as well as the application of the developed data combining and data quality improvement methods for solving problem-oriented tasks in the following research projects:

- Horizon 2020 project “Futuristic beehives for the smart metropolis”. Author's contribution: development of a data layering concept for determining the placement position of the beehive, based on data from several sources ((HIVEOPOLIS (HOR5), 2019);
- operational programs “Growth and employment” 1.1.1. of the specific support objective “To increase the research and innovative capacity of Latvian scientific institutions and the ability to attract external funding by investing in human resources and infrastructure” 1.1.1.1. Project 1.1.1.1/19/A/145 of the “Practical Orientation Research” event “HENCO2: IT platform based on cloud data environment for improving the productivity of poultry farming and reducing greenhouse gas emissions”. Author's contribution: development of data quality methods based on project data quality deficiencies; application of the developed methods for the development of a solution for predicting egg production; performance evaluation of machine learning models (HENCO2: Cloud-Based IT Platform for Improving Poultry Productivity and Reducing Greenhouse Gas Emissions – ER32, 2020);
- Horizon 2020 ERA-NET Cofund project “Individual Mobility Budgets as a Social and Ethical Basis for Reducing Carbon Emissions”. Author contribution: combining public transport and mobile network data (MyFairShare (ZV91), 2021);
- SIA “WeAreDots” and scientific and technical company “Lasma” research no. 1.12 “Multi-Object Detection and Tracking for Vehicle Traffic Monitoring: 3D-LiDAR and Camera Data Fusion “. Author's contribution: research of multi-object detection solutions; improvement of the video stream and 3D LiDAR dataset synchronization solution (2020).

Research theses

- Data analysis results obtained from Internet of Things (IoT) systems depend on the application of data quality improvement methods, which is particularly important in cases of limited data volume or incomplete data, ensuring reliable and accurate decision-making.
- By combining multiple data quality improvement techniques, it is possible to create an adaptive data pre-processing methodology that effectively adapts to various types of data anomalies and irregularities, providing more stable and accurate results in subsequent analysis and modelling phases.
- Interactive processing and visualization of incomplete data is possible using the developed data fusion and data quality improvement methods.

The structure and scope of the thesis

The PhD thesis is written in Latvian, contains an annotation, introduction, 6 chapters, conclusions, bibliography, 58 figures, 15 tables, 3 annexes, 140 pages in total. References to 318 literary sources were made.

1. RESEARCH ACTUALITY

IoT, which combines sensors, hardware, and software, has become an important element of information systems (Nižetić et al., 2020). The most advanced IoT application areas are related to Industry 4.0 (Osterrieder et al., 2020), smart city concept (Eremia et al., 2017), transportation (Porru et al., 2020), and agriculture (Villa-Henriksen et al., 2020). IoT sensor networks operate in three directions: perceiving information from the environment, monitoring internal system processes, and transforming data for decision-making (Govinda & Saravanaguru, 2016; Sanyal & Zhang, 2018). Systems are characterized by universal connectivity and dynamism (Patel et al., 2016; El-Mawla & Badawy, 2023), which are essential in developing cross-sectoral monitoring and management systems (M. Zhang et al., 2021). Kviesis and Zacepins (2015) point out the technical limitations of sensor systems – limited computational power and memory resources. The main challenges in IoT systems are noise reduction (G. Han et al., 2022), outlier detection (Gaddam et al., 2020), missing data replacement (Yuehua Liu et al., 2020), and data aggregation (Sanyal & Zhang, 2018). Multimodal data fusion combines information from various sources into a unified form (Castanedo, 2013) and is used for classification, regression, cluster formation, and dimension reduction (Bokade et al., 2021).

In the IoT context, "data" primarily refers to raw sensor signals or readings (Jifa & Lingling, 2014), often mixing with other information in a broader information space (Nasution et al., 2021). The data fusion process includes specific data objects, which can be both raw signals from devices and processed data linked to a real object (Beddar-Wiesing & Bieshaar, 2020). The DIKW (Data, Information, Knowledge, Wisdom) model (Bellinger et al., 2004) explains how data relationship analysis creates information, information pattern study forms knowledge, and understanding knowledge principles leads to wisdom. This model is used for addressing big data issues (Nasution et al., 2021) and is considered a data fusion model (Becerra et al., 2021).

The JDL model (White & Steinberg, 1998) defines fusion as the association, correlation, and combination of data and information from one or multiple sources. In the context of sensor systems, the process provides accurate position calculations, identity determination, and situation assessment (Hall & Llinas, 1997). These definitions distinguish raw data from processed information, forming the foundation for modern data processing hierarchy (Castanedo, 2013).

Khaleghi et al. and M. Kumar et al. (Khaleghi et al., 2013; M. Kumar et al., 2006) acknowledge that sensor data always contains measurement inaccuracies and uncertainties, creating data imperfections. Lakshmanarao and Shashi

(Lakshmanarao & Shashi, 2020) identify the main challenges: sensor data quality, noise, environmental interference, and systematic errors. Serious problems arise in systems using Dempster-Shafer (DST) theory when contradictory measurements appear. Additional complications arise from sensor heterogeneity and real-time processing requirements. Environmental factors' influence on sensor measurements (M. Kumar et al., 2006) creates systematic outliers in datasets. Liu et al. (J. Liu et al., 2020) propose heterogeneous data classification in spatial, temporal, static, dynamic, and attribute dimensions. Wireless sensor networks are dominated by decentralized architecture approaches (Debauche et al., 2018; Kviesis & Zacepins, 2015; Murakami et al., 2007), providing local data processing for outlier prevention. Khaleghi et al. (Khaleghi et al., 2013) identify the registration data alignment problem that occurs when transforming local sensor data into a unified reference system. Temporal aspects create special challenges in multi-sensor systems, particularly in processing out-of-sequence data (Besada-Portas et al., 2011).

Data fusion is used to address the following tasks: problematic data correction (C. Huang et al., 2019), improving data reliability (Hong et al., 2009; Kreibich et al., 2014), increasing data completeness (Consoli et al., 2015), and obtaining higher-level information (Jayasinghe et al., 2019). Kreibich et al. (Kreibich et al., 2014) indicate that data credibility significantly decreases in uncontrolled environments with high noise levels. W. Wang et al. (W. Wang et al., 2018) demonstrate the effectiveness of multi-sensor data fusion in environmental monitoring systems, where integrated data reveals non-trivial relationships. Karkouch et al. (Karkouch et al., 2016) point to factors affecting data quality in IoT: system scale, resource constraints, network architecture, environmental conditions, sensor status, security vulnerabilities, and data stream processing. Aboubakar et al. (Aboubakar et al., 2022) characterize IoT as an IP network with increased instability. Adelantado et al., and Dinculeană and Cheng (Adelantado et al., 2017; Dinculeană & Cheng, 2019) emphasize IoT device specifics – small message transmission. Sliwa et al. (Sliwa et al., 2020) identify critical barriers – limited energy and memory, which hinder security solution implementation.

From a security perspective, sensor devices are practically impossible to fully protect as they cannot perform resource-intensive cryptographic operations. Environmental influences and technical issues, including lower quality sensor accuracy and calibration deficiencies, significantly affect sensor operation. Methods developed for interference mitigation include noise reduction, missing data filling and interpolation, outlier detection, and data aggregation (Karkouch et al., 2016; Souza & Amazonas, 2015). Noise in signal processing manifests as uncorrelated components, which Jcgm (Jcgm, 2008) characterizes as measurement result dispersion parameters. Teh et al. (Teh et al., 2020) explain uncertainty as a quantitative expression of error, while Krishnamurthi et al. (Krishnamurthi et al., 2020) emphasize noise's negative impact on system resources. Noise reduction methodology uses the sliding window principle. Vanus et al. (Vanus et al., 2020) point out these methods' limitations in local

frequency adaptation. M. M. Rashid et al. (M. M. Rashid et al., 2015) highlight two main wave transformation approaches: continuous transformation in time and frequency dimensions, and discrete transformation for time dimension analysis.

Adhikari et al. (Adhikari et al., 2022) indicate that data incompleteness is a common phenomenon in IoT systems, threatening the analytical process's credibility. Zhang et al. (Zhang et al., 2024) offer a methodical solution by introducing Tracking-Removed Autoencoder (TRAE) and Fuzzy Clustering (FC) methods. Krishnamurthi et al. (Krishnamurthi et al., 2020) identify three fundamental missing data types: MCAR (Missing Completely at Random), MAR (Missing at Random), and NMAR (Not Missing at Random). The Isolation Forest (IF) algorithm has become one of the leading solutions for outlier detection in IoT datasets. Muñoz et al. (Muñoz et al., 2024) point to algorithm accuracy improvement possibilities using adjustable parameters. The Local Outlier Factor (LOF) evaluates outliers by analyzing nearest neighbor data density, while the Isolation Forest algorithm identifies global outliers. Kim et al. (Kim et al., 2022) propose adaptive machine learning methods that not only recognize outliers but also perform automatic data correction using historical data analysis.

Based on the analysis conducted in the chapter, particularly regarding data quality problems and their solutions in IoT systems, the author concludes that the thesis is confirmed:

> Data analysis results obtained from Internet of Things (IoT) systems depend on the application of data quality improvement methods, which is particularly important in cases of limited data volume or incomplete data, ensuring reliable and accurate decision-making.

Justification

The conducted analysis reveals aspects that confirm this thesis. Karkouch et al. (Karkouch et al., 2016) identify critical factors affecting IoT data quality: system scale, resource constraints, environmental conditions, and sensor technical characteristics. These limitations become more pronounced due to IoT device specifics, as highlighted by Adelantado et al., and Dinculeană and Cheng (Adelantado et al., 2017; Dinculeană & Cheng, 2019). The devices can transmit only small-volume messages and operate with limited energy and memory.

Karkouch et al. (Karkouch et al., 2016) note that environmental factors strongly influence sensor operation. Data quality can decrease significantly in extreme conditions and situations with limited maintenance. Kreibich et al. (Kreibich et al., 2014) demonstrate that data credibility drops considerably in uncontrolled environments typical for IoT systems. Consoli et al. (Consoli et al., 2015) complement this observation, showing that limited measurements from individual sensors often fail to provide sufficient information for system state assessment.

The examined solutions offer practical tools for addressing these problems. Several studies support this: C. Huang's methods (C. Huang et al., 2019) for problematic data correction, Hong's solutions (Hong et al., 2009) for improving data reliability, and Consoli et al.'s work (Consoli et al., 2015) for increasing data completeness. These demonstrate a direct link between data quality improvement methods and more accurate decision-making. Research by W. Wang et al. (W. Wang et al., 2018) in environmental monitoring systems clearly shows that applying data quality improvement methods enables the discovery of significant relationships in data.

The data quality problems and their solutions discovered during the research confirm the thesis's validity. Gomathi et al. (Gomathi et al., 2018) highlight additional challenges created by IoT technologies, including both sensor failures and systemic problems. The study by Adhikari et al. (Adhikari et al., 2022) found that up to 40% of all data can be incomplete, significantly affecting analysis results. Martin et al. (Martin et al., 2023) demonstrate that wave transformations, adaptive machine learning, and modern missing data replacement methods provide effective solutions for overcoming these problems. These approaches improve decision-making accuracy in IoT systems with limited or incomplete data volume.

End of Justification

Data fusion is a process where different data streams are combined to create more informative and customized output (Bokade et al., 2021). Data fusion types are classified according to system architecture, abstraction levels, and fusion objectives.

Khaleghi et al. (Khaleghi et al., 2013) conduct a systematic method study, providing detailed analysis. Castanedo (Castanedo, 2013) focuses on three algorithmic aspects: data association, system state assessment, and decision fusion. Zheng (Y. Zheng, 2015) offers innovative methods for cross-domain data integration, Atluri et al. (2018) develop a methodological framework for spatial and temporal data acquisition, while Qin et al. (Qin et al., 2020) address IoT specifics.

Data fusion classification is based on abstraction levels, data sources, and input-output relationships (Becerra et al., 2021; Dasarathy, 1997). The three main classification methods are Dasarathy classification (Dasarathy, 1997), Whyte classification (Grime & Durrant-Whyte, 1994), and JDL model classification (Steinberg & Bowman, 2008).

System architecture distinguishes four main types. Centralized architecture uses a unified central processor but faces limitations in visual sensor networks. In decentralized architecture, each sensor operates autonomously, but growing communication costs arise – $O(n^2)$ (Grime & Durrant-Whyte, 1994). Distributed architecture offers a more balanced approach, where source nodes first process their measurements. Hierarchical architecture combines previous approaches (Castanedo, 2013).

Beddar-Wiesing and Bieshaar (Beddar-Wiesing & Bieshaar, 2020) indicate that the decentralized approach, despite challenges, is often the basis of choice due to its resilience. Becerra et al. (Becerra et al., 2021) identify three sensor data fusion scenarios: competitive (single modality sensor integration), complementary (complementary data from different sensors), and cooperative (dynamic sensor adaptation).

Whyte classification offers three types of cooperation. Complementary interaction involves different sensors providing distinct data, for example, in environmental monitoring systems. In data redundancy cases, multiple sensors observe one target, improving accuracy. Cooperative interaction combines data from different sensors to obtain new information, for example, in 3D LiDAR and video camera integration.

Dasarathy classification (Dasarathy, 1997) structures fusion in five levels: DAI-DAO (raw data fusion), DAI-FEO (data transformation into features), FEI-FEO (feature processing), FEI-DEO (decision-making), and DEI-DEO (decision synthesis). Varshney (Varshney, 1997) supplemented it with the DAI-DEO level for direct transition from data to decisions.

The JDL model offers a five-level hierarchical structure (Llinas et al., 2004). It begins with sensor preprocessing, continues with object state assessment and relationship analysis, and concludes with impact analysis of actions and resource management. Blasch and Plano (Blasch & Plano, 2002) supplemented the model with a user interaction level.

Modern classification (Abdelgawad & Bayoumi, 2012; Barbedo, 2022) distinguishes three fundamental levels. Low-level fusion operates with raw data (Di Natale et al., 2002). Mid-level fusion performs feature extraction (Biancolillo et al., 2014), while high-level fusion develops separate analysis models (L. Huang et al., 2014).

Probability theory-based methods, especially Bayesian inference, provide a systematic approach to data fusion (Pires et al., 2016). However, in IIoT systems, traditional methods lose effectiveness. Medjahed et al. (Medjahed et al., 2011) identify the main challenges – limited performance in multifactor data. Rezatofighi et al. (Rezatofighi et al., 2015) analyse the advantages of probabilistic data association in object tracking.

The research shows that no universal method exists – each solution is suitable for specific tasks. For simpler cases, low-level fusion is sufficient, while for more complex ones, a higher-level approach is more effective. In IoT systems, modern methods can significantly improve data quality and decision reliability.

The previously discussed IoT data quality problems and their solutions are particularly relevant in specific industries where accurate and timely data is critical for decision-making. The author's research in three different fields – precision beekeeping, urban traffic monitoring, and poultry farming – demonstrates practical solutions to previously identified challenges. Each of these fields encounters typical IoT system problems – sensor data deficiencies, multimodal data integration complications, and real-time processing

requirements. To address these problems, the author uses and extends the previously described data fusion methods, adapting them to industry-specific needs and constraints. Special attention is paid to methods that can operate effectively under limited resources while ensuring high data quality and reliability.

In beekeeping, the author conducted in-depth research on data collection and processing. Precision beekeeping is an innovative hive management strategy that provides bee colony monitoring for resource optimization (Zacepins et al., 2012). Data collection includes measuring physical parameters using sensors in hives (Kviesis & Zacepins, 2015), measuring temperature, humidity, gas composition, vibration, and sound. Data is analysed at three levels (Human et al., 2013; Zacepins & Stalidzans, 2013): apiary level (meteorological data, video observations), colony level (temperature, humidity, weight), and individual level (bee behavior). Wireless network technologies (Debauche et al., 2018) often create data deficiencies, which are addressed using data fusion methods. Modern solutions use IoT devices and LSTM neural networks for swarming prediction (Kwon, Cho, et al., 2019), combining temperature and sound data with the Kalman filter algorithm. This approach allows not only resource consumption optimization but also early identification of potential problems in bee colonies.

In urban traffic monitoring, IoT solutions use video surveillance equipment together with object detection and tracking algorithms (N. Chen & Chen, 2018). Modern transport system digitalization creates new opportunities in data fusion (Neumann et al., 2016). The author developed a synchronization algorithm for LiDAR and video camera data fusion, focusing on precise synchronization of different sources with timestamps. The validation in Jelgava city (Bumanis et al., 2021) lasted 6 months (01.04.2021–30.09.2021), using LiDAR sensors and four video cameras for monitoring two traffic directions. The validation results showed high license plate recognition accuracy (>99%) and achieved 97% synchronization accuracy.

In poultry farming, the author developed a complex data storage solution, creating a data warehouse for environmental monitoring sensor and production cycle data processing. The system is based on snowflake schema principles, ensuring both efficient farm management and compliance with EU animal welfare regulations. The cyber-physical model includes a sensor set, data exchange controllers, and an analytical center, focusing on feed process optimization and environmental parameter control. The solution was successfully implemented in two Baltic poultry farms for CO2 and NH3 level monitoring, using the Microsoft Azure platform for automatic data collection and processing.

In all three industries, it was found that incomplete or poor-quality data input affects forecasting and resource optimization. Data quality improvement utilized sensor calibration, timestamp synchronization, and data processing algorithms.

Research shows a clear need for a new specialized data fusion method to address the identified IoT data quality issues. While current methods like TRAE

and FC (Zhang et al., 2024), as well as IF and LOF algorithms (Muñoz et al., 2024), each solve specific problems, there remains a gap for a comprehensive approach. The field requires a single solution that can both improve data quality and efficiently merge data when resources are limited. Particularly relevant is the need for a method that could adapt to different industries and data types while maintaining high accuracy and reliability. As indicated by Karkouch et al. (Karkouch et al., 2016) and Kreibich et al. (Kreibich et al., 2014), data quality problems are particularly pronounced in uncontrolled environments with limited resources, which is characteristic of IoT systems. These factors justify the need to develop a new, data layering-based method that would not only address data quality issues but also ensure efficient multi-source data fusion, taking into account both spatial and temporal aspects.

2. DATA LAYERING CONCEPTUAL METHOD

A systematic approach to analysing data fusion methods reveals that the data pre-processing stage is a critical factor significantly affecting the final result quality. Using a unified model for multiple datasets can create a specialization effect, therefore method adaptability is crucial when working with various historical data sources. The data layering approach combines spatial-temporal analysis with the ellipsoidal method. Abu Bakr and Lee (Abu Bakr & Lee, 2017) describe the approach implementation by organizing data in dimensional layers, where each layer represents information from a specific time period. The ellipsoidal method principles determine the identification of data overlapping zones. The method has been used in analysing bee foraging behaviour by combining data from different sources. Research confirms that a successful data fusion method must include three main elements: data preparation capabilities, effective historical data utilization, and systematic quality control. The integration of these elements is essential in IoT systems and other data-intensive industries, where they directly affect both analysis results and decision-making processes.

Data required for bee foraging optimization includes regional information (location, relief, climate), local nectar plant characteristics, and environmental conditions affecting nectar formation (X. J. He et al., 2016; Hennessy et al., 2020). Regional beekeeping organizations compile this data in flowering calendars; however, their use is limited by data format and coverage heterogeneity. For method validation, four characteristic plants with different properties were selected: Grevillea robusta (constant flowering period, medium production), Coffea arabica (seasonal flowering, variable production), Eucalyptus citriodora (year-round resources), and Dichrostachys cinerea (low intensity). This selection allows testing method effectiveness in various resource availability scenarios. Data processing uses two main methods: normalization (for non-normal distribution) and standardization (for normal distribution). Normalization provides objective resource assessment between different plants,

for example, comparing Dichrostachys cinerea and Coffea arabica data. Standardization, applied to Eucalyptus citriodora and Grevillea robusta data, provides accurate quantitative assessment of resource fluctuations.

Effective bee apiary management requires a complex understanding of several interrelated factors. Precision beekeeping methodology is based on the interaction of three main data layers, which together form the foundation for informed decision-making. The first and most significant is the plant flowering layer, based on detailed flowering calendar data (Tree Flowering Calendar, 2020). This layer provides essential information about nectar and pollen availability during different periods of the year, allowing beekeepers to precisely plan apiary placement and resource utilization. The second layer consists of precipitation data, whose importance in bee foraging efficiency has been confirmed by several studies (X. J. He et al., 2016). The amount and distribution of precipitation throughout the year significantly affects both nectar secretion and bees' ability to access food resources. The third layer reflects bee colony activity cycles, documented in the beekeeping calendar (Beekeeping Calendar, n.d.). This layer is particularly important in daily hive management, as it allows predicting and planning such essential activities as brood rearing, swarming, and intensive foraging periods.

When the layers of interest are known, it is possible to determine and output useful information. First, interpolation is applied, which helps increase data accuracy, making it more uniform and providing the ability to view trends in more detail. Accordingly, the data point resolution for each parameter is increased using linear interpolation.

Given a set of data points $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, linear interpolation is the process of determining the y value for a specific x using the formula (see (2.1.)):

$$y = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) \tag{2.1.}$$

where

x_0, x_1 – specific x value points between which interpolation is performed, real numbers;

y_0, y_1 – corresponding y values for these x value points, real numbers;

y – calculated y value corresponding to x value. This is the result of the interpolation process, a real number.

Then for each parameter, values are multiplied by corresponding weights and then normalized to the range [0, 100]. The formula for each weighted value is as follows (see (2.2.)):

$$\text{weighted\_value}_i = \sum_{j=1}^{n} \text{weight}_j \times \frac{\text{parameter}_j}{100} \tag{2.2.}$$

where

n – number of parameters, integer;

weight_j – weight assigned to each value, real number;

parameter_j – specific parameter to which weight is assigned, integer or decimal numbe;

weighted_value_i – calculated weighted value for some parameter i, %.

After calculating the weighted value, Principal Component Analysis (PCA) is applied. This statistical method transforms the initial, potentially interrelated variables into a new coordinate system where they become mutually independent. These new, independent variables are called principal components. Given a data matrix X, the first principal component can be found as follows:

1. mean value is subtracted: $X_{mean} = X - \bar{X}$;
2. covariance matrix Σ is calculated from X_{mean};
3. eigenvector and eigenvalue of Σ are calculated;
4. the feature vector associated with the largest eigenvalue is selected.

In this context, PCA is used to obtain the most significant trend from combined parameters, that is, to obtain a single combined value (PCA-based combined value) that reflects the greatest differences from all parameters.

Both weighted and PCA-based combined values are compared against a threshold. Regions where these values exceed the threshold are marked in the diagram. This is a simple condition check: if fused_value ≥ threshold, the region is marked.

As a result, two diagrams are obtained. The initial combined values diagram, where (see Fig. 2.1) shows the initial parameters, initial combined values, and regions where the initial combined values exceed the threshold. This provides insight into how the simple weighted sum of parameters (initial combined values) operates across different months.

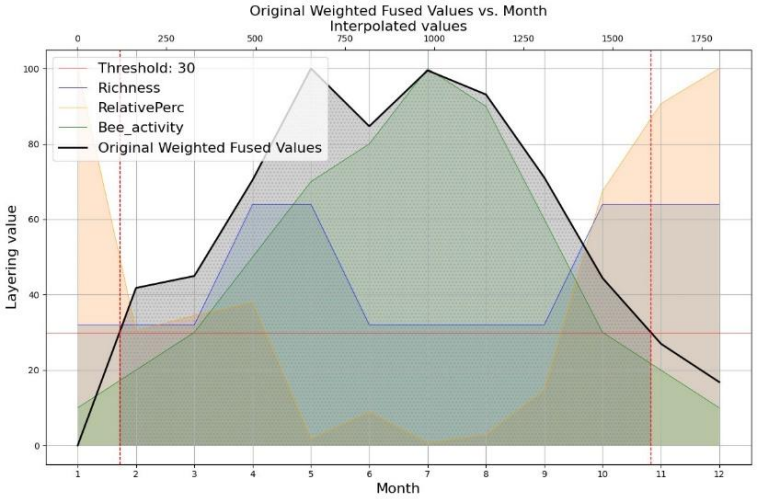


Fig. 2.1. **Plot of original weighted fused values.**

The PCA-based combined values diagram (see Fig. 2.2) shows the initial parameters, PCA-based combined values, and regions where the PCA-based combined values exceed the threshold. This provides perspective on how PCA-based fusion, which captures the greatest differences from parameters, operates throughout the months.

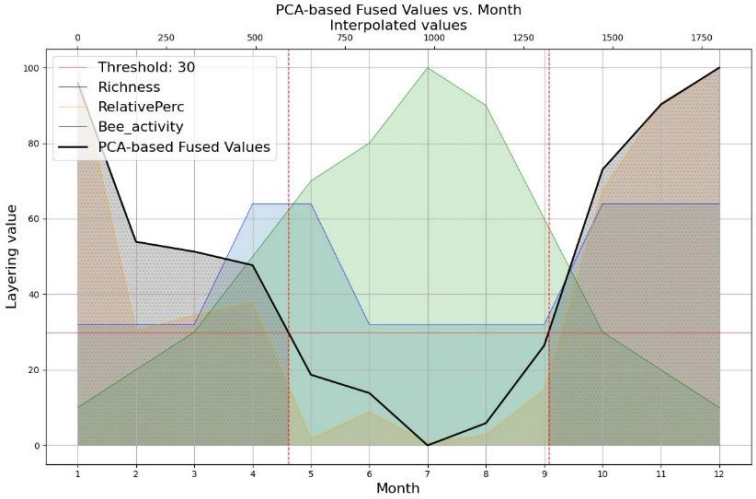


Fig. 2.2. **Plot of PCA-based fused values.**

Using these diagrams, both data fusion methods can be visually compared (see Fig. 2.32.3) and both significant regions (those exceeding the threshold) can be understood.

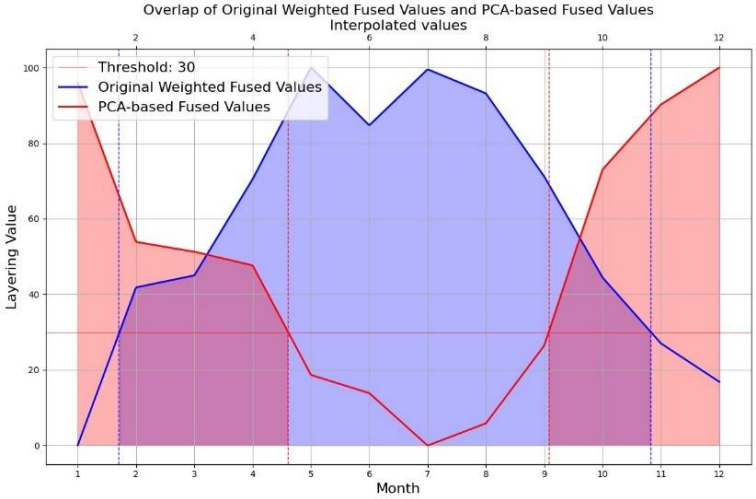


Fig. 2.3. **Overlap of fused values.**

The trapezoidal method offers a practical way to compare results. By placing the initial combined values alongside PCA-based ones, this method reveals where these values match or differ. Measuring the area between these curves shows how well the two fusion methods agree over time. When the curves closely follow each other, it signals that both methods reached similar conclusions. Areas where they diverge might point to interesting patterns or unusual data points worth investigating.

Additionally, the following aspects can be highlighted:

- using a threshold, regions where combined values exceed a certain significance level can be identified and quantified. The trapezoidal method allows measuring the magnitude of this significance, providing a numerical value to determine how important or influential certain time periods might be;
- the trapezoidal method provides the ability to quantitatively compare both fusion methods. By integrating areas under each curve and comparing results, informed decisions can be made about which fusion method might be more suitable for specific applications or scenarios;
- the trapezoidal method is also computationally efficient and simple to implement. Given the potentially high precision of data (especially after interpolation), a simple but effective method provides quick analysis without significant computational costs;
- areas calculated using the trapezoidal method can be intuitively understood. Larger areas indicate more significant events or patterns in the data, while smaller areas may indicate less influential periods.

Given two sets of y-values, $y1$ and $y2$, in a common x domain, the area between both curves must be calculated. If one curve is completely above the other, the area is calculated as the difference between them; if they intersect, overlap and non-overlap sections are identified to calculate respective areas.

Further, here's how the areas are calculated:

1. the difference between two datasets is calculated:
   a. for each point x_i in the domain, the difference Δy_i between both datasets is calculated: $\Delta y_i = y1_i - y2_i$.
2. overlapping regions are determined:
   - if Δy_i and Δy_{i+1} have the same sign, then both datasets do not intersect between x_i and x_{i+1};
   - if Δy_i and Δy_{i+1} have opposite signs, then both datasets intersect between x_i and x_{i+1}, indicating an overlapping region.
3. the area is calculated using the trapezoidal method (see (2.3.)):
   - for non-overlapping regions between x_i and x_{i+1}:

$$Area = \frac{x_{i+1} - x_i}{2} \times (|y_{1i} + y_{1i+1}| - |y_{2i} + y_{2i+1}|) \qquad (2.3.)$$

- for overlapping regions, the area at the intersection point is divided into two trapezoids, and the areas of these trapezoidal shapes are summed.

4. The total overlap area is the sum of areas calculated for each interval in the x domain.

As a result, a diagram with overlapping regions is obtained (see Fig. 2.4). Through visual analysis, it can be concluded that:

- larger data changes occur between months 1 and 4 and between months 9 and 12;
- the most favourable conditions for bee colony placement on the analysed plant are between the last week of month 4 and the beginning of month 9;
- placement earlier, between months 2 and 4, or later, between months 9 and 11, is not recommended due to rapidly changing conditions.

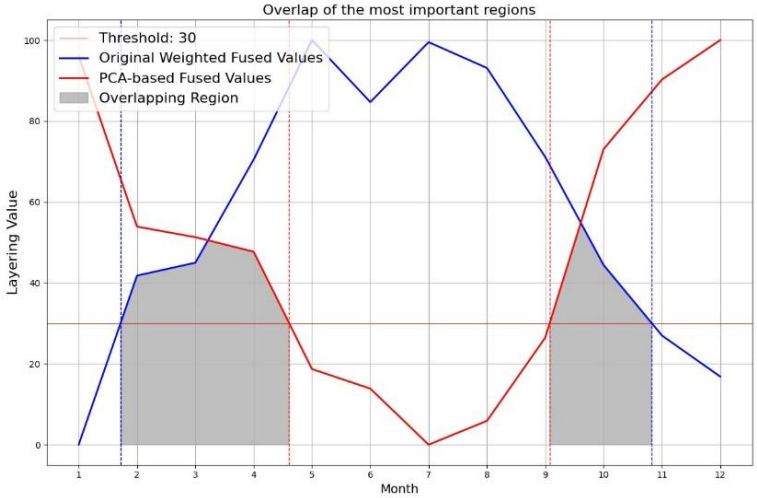


Fig. 2.4. **Overlapping regions of fused values.**

The data layering method is developed in Python environment, using `pandas` library for data processing, `numpy` for numerical calculations, and matplotlib for visualization. The core of the method is the `data_fusion_main` function, which coordinates the work of several interconnected subfunctions.

The method's operation process begins with data preparation, where interpolation functions play a crucial role. The `interpolate_data` function performs linear interpolation between data points, ensuring smooth data flow, while `interpolate_data_auto_smoothing` automatically searches for the optimal smoothing degree. Data fusion can be performed in two ways – with `weight_based_data_fusion`, which uses user-defined weights, or with `pca_based_data_fusion`, which is based on principal component analysis.

The method's parameters are organized into three logical groups: data source parameters (`DataFrames_dict_or_df`, `Data_params`, `CommonColumn`), analysis control parameters (`Weights`, `Layering_threshold`), and data processing options (`Smoothing`, `AutoSmooth`, `FilterBy`, `FilterValue`). `DataFrames_dict_or_df` determines the initial data structure, providing a flexible approach to processing both individual and combined datasets. `Data_params` specifies which columns will be used in the analysis, such as nectar amount or bee activity, thus focusing calculations on the most important indicators. `CommonColumn`, typically the time axis (month), ensures data synchronization and correct comparison between different sources. The `Weights` parameter allows adjusting each data layer's influence on the final result, for example, assigning greater importance to precipitation data than bee activity. `Layering_threshold` defines the critical threshold value (typically 30%), serving both as a visual reference point and as an automatic smoothing control mechanism. The `Smoothing` coefficient controls data interpolation precision, affecting curve smoothness and detail level. `AutoSmooth` functionality automatically adjusts smoothing intensity based on the set threshold, thus optimizing data representation. `FilterBy` and `FilterValue` parameters provide the ability to focus on specific data segments, such as analysis of a specific plant or time period, allowing detailed examination.

The developed method concept includes time data layering, where each layer displays data as a plane. The initial concept was applied to the bee foraging task, which provides diverse data. Essential data for optimizing bee foraging includes information about regional location, topography, climate, local nectar and pollen-producing plants, as well as bee species and their activities. The data layering method begins with normalizing necessary data, such as nectar levels, to create a dataset. This set is then analysed to determine the most productive locations for bee feeding. The method uses both weighted value approach and Principal Component Analysis (PCA) to combine various data aspects. These approaches allow comparing and evaluating the significance of different parameters, such as plant richness in a specific time period. Area calculation under curves is used to assess overlap between both approaches. The chosen approach is simple to use and provides clearly understandable results. Data visualization allows easy identification of the most significant periods or events that are essential for further analysis.

Precision poultry farming prediction task

In modern poultry farms, various environmental parameters that affect egg production are continuously monitored. Experts identified significant factors affecting laying hen welfare, such as air temperature, humidity, CO_2 and NH_3 levels. The egg laying rate data required for model comparison and evaluation were obtained from a facility housing chickens (Lochman Brown breed) in

enriched cages. This facility has implemented a belt-type manure removal system, which operates efficiently every day. Laying hens are moved to cages at 16 weeks of age and are kept there until 80–90 weeks of age. From 20 weeks, hens begin laying eggs. Daily egg production or yield is calculated as the number of eggs produced per day in relation to the total number of hens on that day (Paura et al., 2022). Due to the nature of data availability, the datasets are limited, and egg yield data for a 61-week period (data collected from November 22, 2019, to February 9, 2021) were used for model training, as well as for a 46-week period (data collected from March 23, 2021, to March 3, 2022) (see Fig. 2.5).

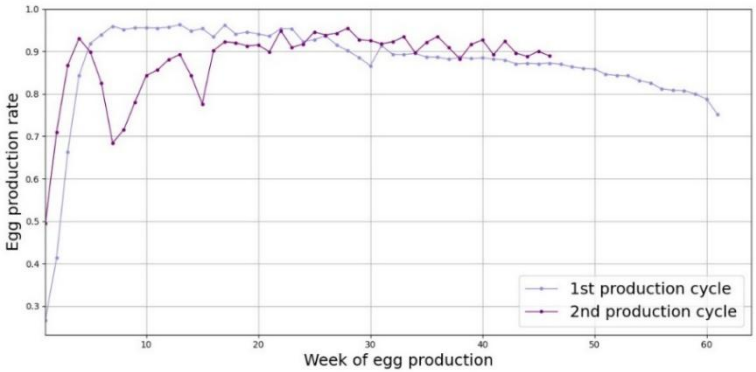


Fig. 2.5. **Egg production curves used for training (cycle 1) and testing (cycle 2)** (Bumanis et al., 2023).

The test dataset differed significantly from the one used in training and from the normal egg yield model. According to farmers' information, such differences could be explained by inconsistency in data entry management – while collected egg quantities are counted automatically, the final value is entered manually. This can be done several times a day or not at all, for example, on workdays or due to technical reasons. Training and testing sets, i.e., how they are divided, differ for nonlinear and ML-based models, and are described further. Within the precision poultry farming task, the selected machine learning models were built using the Keras framework (Chollet, 2015) (LSTM and CNN) and scikit-learn library (Pedregosa et al., 2011) (RF and XGBoost). The models were tuned (hyperparameter selection) using library extensions such as keras-tuner and sklearn.model_selection. No changes were made to the base architecture of individual ML models. Models were compared by varying hyperparameter values of each model. To find the best hyperparameter configuration, LSTM and CNN models were tuned using the Hyperband algorithm (Li et al., 2020), while tree-based models used Random Search (Bergstra & Bengio, 2012). Early stopping technique (with patience value 10) was used for LSTM and CNN models to potentially reduce the overfitting problem. Early stopping was also used in XGBoost hyperparameter search and training phase, while cross-validation technique in the RF case. Model training was performed using factors

monitored daily in the poultry farm and production-related data with different input sequence lengths, for example, using a sliding window approach. Using this technique, an important step was determining the window size, as it sets an additional requirement for model input – the number of previous productivity values, for example, egg production in this case. If a sliding window of size 1 is selected, it means the input requires production data from the previous day. Several feature selection approaches were considered in the initial development phase, and based on obtained farm data, it was determined that prospective feature selection would be suitable for use with general ML algorithms. The selected ML models were trained on first production cycle data (further divided into 90% training and 10% validation parts to avoid data leakage (Hannun et al., 2021) and overfitting problems (Ying, 2019)) and tested on the second production cycle to predict productivity for the next day. Model input was formed from 12 parameters plus historical production data depending on sliding window size. Model performance was evaluated using statistical criteria. Modified compartmental model parameters (see Table 2.1) were evaluated using R programming language to fit the egg yield curve (1st cycle)(see Fig. 2.6). From Table 2.1, all parameters are significant ($p < 0.001$) and are applicable to this prediction task. The model result shows only the egg yield trend based on previously trained data but does not include input values that could affect the prediction and indicate possible problems. This production cycle shows that to achieve high accuracy, the egg production week alone is insufficient to draw conclusions, but it allows farmers to see outliers from the trend.

Table 2.1. **The calculated values for the modified compartmental model**

Parameter	Estimate	Standard Error	t value	Pr(>|t|)
a	0.13099	0.01316	9.954	4.45e-14 ***
b	-0.90414	0.03927	-23.024	< 2e-16 ***
d	2.24435	0.04658	48.182	< 2e-16 ***
c	-0.90766	0.03923	-23.139	< 2e-16 ***

The fitted curve for the test egg yield dataset is as follows:

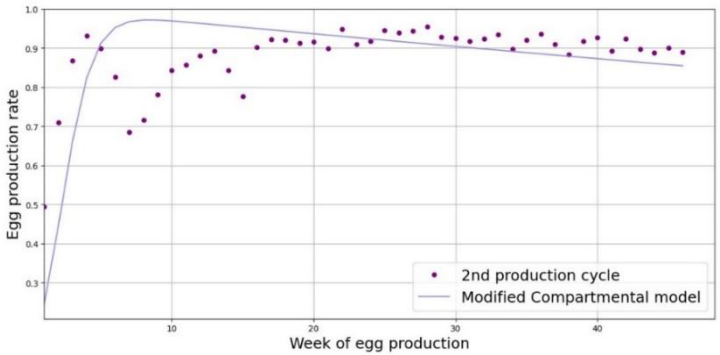


Fig. 2.6. **Fitted curve and observed production of eggs** (Bumanis et al., 2023).

ML models tend to follow abnormal production decreases (see Fig. 2.7), thus indicating their ability to adapt to such situations. Although subsequent data verification showed that environmental factors did not change drastically enough to affect the production decrease, the models predicted the decline because previous days' (depending on sliding window size) production data were used as input. The ML model results and observed egg yield with sliding window (size 2) are as follows:

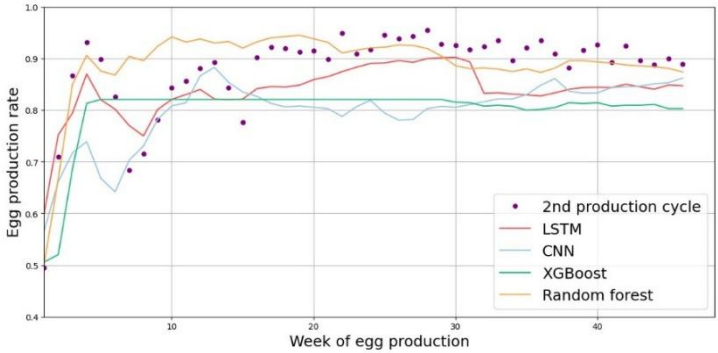


Fig. 2.7. **Results of trained ML models** (Bumanis et al., 2023)**.**

Regarding ML models, several window sizes (1, 2, 3, 5, 7, and 14) were tested to determine which provides the best prediction results. Model performance results are shown in Table 2.2. Regarding sliding window size, results indicate that LSTM is more accurate using a sliding window of size 2, achieving the lowest MAPE and RMSPE values of 5.390% and 7.751% respectively. There was also no significant difference between model performances using window sizes 3 and 5, except for the CNN model which performed worst, which can be explained by possible model overfitting.

Table 2.2. **Machine learning model performance** (Bumanis et al., 2023)

Sliding Window Size	Error Metric	LSTM	CNN	XGBoost	RF
1	*MSE*	1.710	4.111	1.225	**0.944**
1	*MAPE*	13.909	14.224	9.994	**6.907**
1	*RMSPE*	15.439	16.258	11.708	**10.242**
2	*MSE*	**0.272**	1.884	1.060	0.726
2	*MAPE*	**5.390**	15.200	10.272	6.331
2	RMSPE	7.751	18.314	12.178	9.284
3	MSE	0.203	1.384	0.877	0.664
3	MAPE	6.501	39.319	9.086	6.158
3	RMSPE	8.828	39.993	10.875	9.110
5	MSE	0.358	0.843	0.767	0.604
5	MAPE	6.218	13.479	7.415	6.077
5	RMSPE	8.781	15.537	9.223	9.016

Continuation of table 2.2.

Sliding Window Size	Error Metric	LSTM	CNN	XGBoost	RF
7	MSE	0.198	0.443	0.863	0.546
7	MAPE	5.484	13.300	9.619	6.188
7	RMSPE	7.845	14.555	11.350	9.168
14	MSE	0.153	0.308	0.719	0.453
14	MAPE	6.433	6.633	6.114	6.273
14	RMSPE	8.982	9.718	8.158	9.221

Table 2.3 summarizes the best results obtained in model evaluation. It can be concluded that overall, LSTM, RF, and XGBoost showed the best performance. Evaluation results, considering the best metric values for different sliding window sizes (machine learning models), indicate that performance varies. Overall, all models provide sufficiently accurate results to reveal problems and make changes in the production process; however, results indicated that some models, such as LSTM, showed competitive performance across all sliding window sizes while providing the best results – using a smaller number of historical production data. It can be concluded that machine learning models, especially LSTM, prove to be better than Modified Compartmental.

Table 2.3. **The best observed results** (Bumanis et al., 2023)

Model	*MSE*	*MAPE*	*RMSPE*	Sliding Window Size
Modified Compartmental	0.011	9.134	14.809	n/a
LSTM	0.272	5.390	7.751	2
CNN	0.308	6.633	9.718	14
XGBoost	0.719	6.114	8.158	14
RF	0.604	6.077	9.016	5

The egg production cycle used for model testing was atypical in terms of data quality characteristics, such as homogeneity and completeness, thus making the process of selecting the most viable model and predicting egg yield based on such data more complex. It should be noted that prediction was performed only 1 day in advance, which relatively limits the requirements for accuracy goals. While it is possible to predict egg yield for a longer period, results may show rapid accuracy decrease; thus, the appropriate prediction length should be long enough to make appropriate changes (i.e., adjust ventilation algorithm to temperature changes) to the production process. Furthermore, the choice to predict only 1 day ahead was determined by the consistency of available training data. This includes differences between data from two separate cycles. The test dataset, which was significantly different, showed limitations of the nonlinear model that uses only one parameter (number of laying weeks) and does not adapt to changes, resulting in MAPE and RMSPE values of 9.134% and 14.809% respectively. Although the calculated errors (MAPE and RMSPE) of ML models were between 5% and 10%, it was observed that they can better adapt to

production changes than the tested nonlinear regression model. Since ML models also use environmental data as input, sudden changes in these factors affect productivity, which can be predicted in time. Results showed that ML models (LSTM, RF, and XGBoost with sliding window size 2) were able to predict production decrease (2nd production cycle) at a satisfactory level. Results indicate that the proposed solutions may also be applicable in farms with limited production datasets and no large volume of historical egg yield data. Depending on available historical data for model training, the farm can also use a multi-model approach, where different models can be operated according to farmer needs (prediction length). Moreover, it also maintains the possibility to use the nonlinear model in situations where data affecting environment or other productivity parameters are not recorded. In this case, the nonlinear model can be used either as a separate solution or as additional evidence to follow production curve dynamics.

Approbation of Data Layering method

The data layering method was used to analyse prediction model results and identify potential problems. This method provides a systematic approach to studying multiple parameter interactions and their impact on the egg production process. The study analysed three fundamental parameters characterizing the egg production process: actual egg laying rate in percentage, standard laying rate in percentage, and average indoor temperature. Parameter selection was based on their physiological and technical significance in the egg production process. According to the previously described methodology, after interpolation of missing values and data normalization, data fusion was performed using the data_fusion_main function. Parameter weight coefficients were determined considering their degree of influence on the production process. The actual laying rate was assigned a weight of 0.85, standard laying rate 0.50, and average temperature 0.35. A lower weight for the temperature parameter was assigned based on scientific research indicating that temperature significantly affects productivity only outside certain threshold values determined by the specific chicken breed and housing conditions. The layering threshold was set to 40 units, allowing effective identification of significant outliers between actual and theoretical laying rates. Automatic smoothing (AutoSmooth=1) with a smoothing parameter of 1 was used for data smoothing, providing optimal balance between short-term fluctuation filtering and preservation of significant trends in the data. Initial weighted combined value analysis (see Fig. 2.8) reveals significant differences between actual and standard laying rates, especially during periods with increased temperature fluctuation impact.

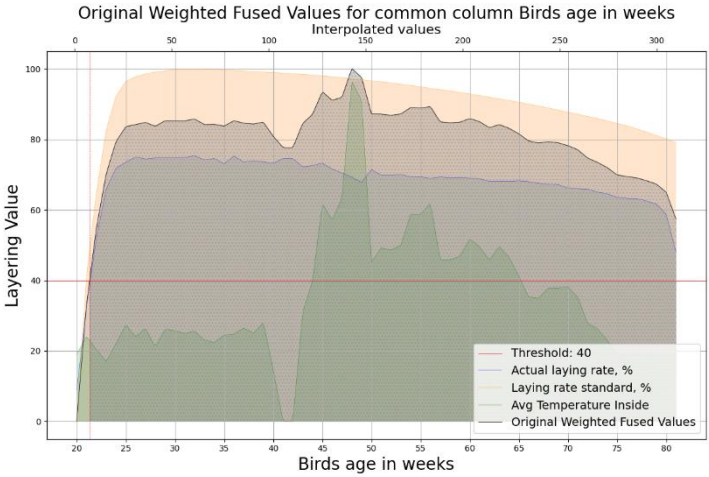


Fig. 2.8. **Original weighted fused values.**

The period from week 40 to 50 is particularly pronounced, showing a sharp decline followed by gradual recovery. These outliers may be related to several factors, including data quality and production process changes. The PCA-based analysis (see Fig. 2.9) provides additional insights into parameter interactions. The figure shows that the PCA component can effectively identify outliers in the dataset, particularly during the period from week 40 to 50, where significant outliers from the expected laying rate are observed. This period coincides with increased error values in the Modified Compartmental model (MAPE 9.134%, RMSPE 14.809%), indicating the model's limitations in analysing complex situations.

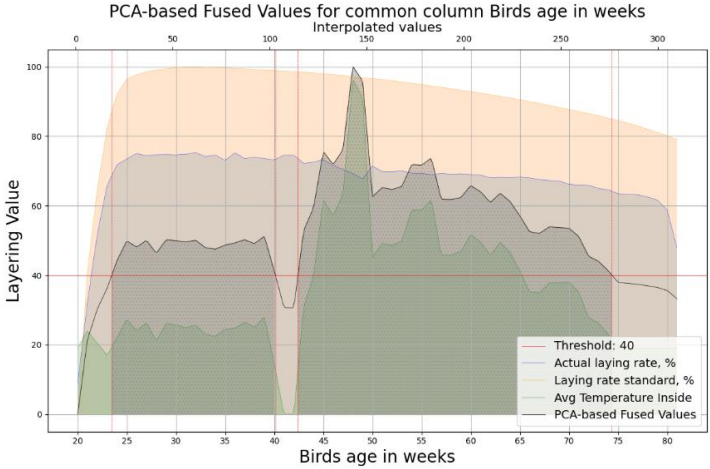


Fig. 2.9. **PCA-based fused values.**

The analysis of significant region overlaps (see Fig. 2.10) identifies three characteristic periods in the production cycle. During the production initiation period (weeks 20–25), a high correlation between both methods is observed, indicating stability in data quality and production process. In the instability period (weeks 40–50), significant differences between methods were identified, where machine learning models, especially LSTM with MAPE of 5.390%, show considerably higher accuracy than traditional models. In the production conclusion period (weeks 75–80), a renewed convergence of methods is observed, indicating stabilization of the production process.

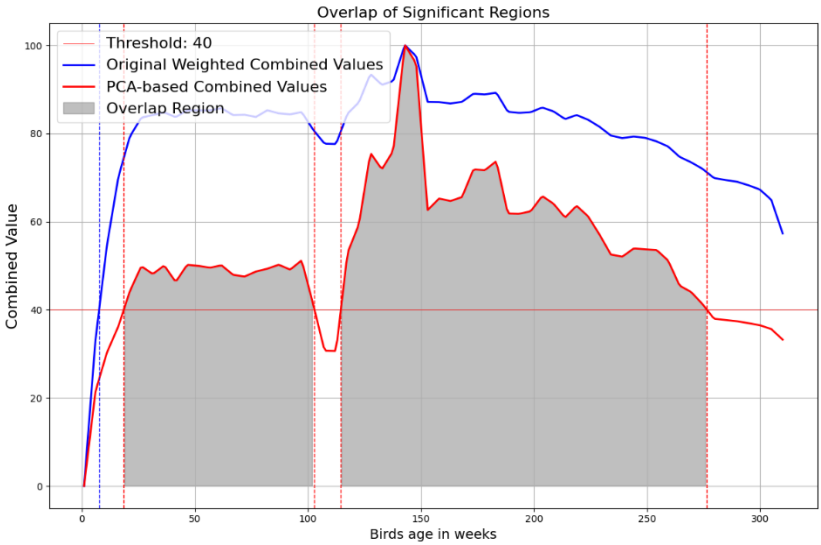


Fig. 2.10. **Overlap of Significant regions.**

The data layering analysis reveals several significant problems affecting prediction model accuracy:

- data quality issues: identified outliers, especially during weeks 40–50, indicate potential problems in sensor data or data collection process. These issues may be one of the reasons why the Modified Compartmental model shows higher error values (MAPE 9.134%);
- model adaptation capability: lower error values of ML models (MAPE 5–10%) can be explained by their ability to better adapt to nonlinear changes in data, as confirmed by PCA analysis results. The LSTM model in particular (MAPE 5.390%) demonstrates significantly better performance during the period with identified anomalies;
- sensor system limitations: the analysis indicates a need to improve sensor data quality control mechanisms, especially regarding temperature measurements where significant fluctuations are observed.

Overall, the study demonstrates performance differences between traditional nonlinear models and machine learning algorithms in egg yield prediction. ML models show better results than traditional single-factor models, as evidenced by lower error values. Simultaneously, the results indicate a need for specialized methods to address data quality issues in the context of precision agriculture.

Based on the data layering method developed in the second chapter and its practical application in verifying precision poultry farming research results, the author concludes that the following thesis is confirmed:

Interactive processing and visualization of incomplete data is possible using the developed data fusion and data quality improvement methods.

Justification

The data layering method developed in the second chapter offers a systematic approach to combining and visualising data of different types. The method's conceptual foundation enables effective integration of different data sources, as shown by the simultaneous analysis and visualisation of plant richness, precipitation, and bee activity data. The method's interactive nature ensures identification and visualisation of data overlap zones, demonstrated through the application of the trapezoidal method in region analysis (see Fig. 2.3.). This approach enables quantitative assessment of both overlap zones and regions of difference, providing clear understanding of data interaction.

The method's practical applicability is demonstrated in precision poultry farming research, where visualisations in Figures 2.8., 2.9. and 2.10. demonstrates the method's versatile analytical potential. The visualization reflected in Figures 2.8. and 2.9. demonstrate the method's capability to combine and analyse three essential parameters: actual laying rate, theoretical laying rate, and average indoor temperature. Of particular significance is the method's ability to identify critical periods using both weighted sum and PCA-based approaches. The overlap region analysis in Figure 2.10. provides additional quantitative assessment of method convergence and divergence during three characteristic periods: production initiation (weeks 20–25), instability (weeks 40–50), and conclusion (weeks 75–80) phases.

The effectiveness of the method in identifying data quality problems is manifested in its ability to reveal significant deviations and unusual data periods. Interpolation was used to fill in missing values, which ensured data continuity and normalization, which was an essential prerequisite for using the PCA component. This approach provided effective data visualization and helped identify critical data quality issues.

3. DATA QUALITY IMPROVEMENT

Data quality criteria – completeness and accuracy – are fundamental in data processing, as they directly affect the quality of analysis, prediction, and decision-making. Data completeness characterizes the proportion of available measurements from all necessary measurements, while accuracy determines the conformity of measurements to true values. Data completeness can be evaluated in two aspects. The first is the existence of required datasets – whether all necessary data groups are available. For example, in an environmental monitoring system, temperature, humidity, and CO_2 level measurements are all essential. The second aspect is the completeness of data records within each set – whether there are missing measurements in a specific time period. The missing data problem can be solved through interpolation or using similar records. Data accuracy is affected by both technical inaccuracies related to measuring equipment and sensor operation (calibration, device limitations) and the human factor – errors in manual data processing. Data accuracy is particularly critical in IoT solutions and artificial intelligence applications, where inaccurate input data can significantly affect model training processes. Effective data quality assurance requires a systematic approach. Mathematical methods, such as interpolation based on existing accurate data, are used to fill missing values. For ensuring accuracy, regular checks are performed to identify and eliminate problem causes – both technical failures and human factor-induced outliers. This complex approach allows creating a reliable foundation for further data analysis. Data quality methods were developed while solving the egg laying rate prediction task in precision poultry farming. For method development and testing, a dataset covering two (one complete and one partial) egg laying cycles was used – for a 61-week period (data collected from November 22, 2019, to February 9, 2021) and for a 46-week period (data collected from March 23, 2021, to March 3, 2022). During egg laying periods, various types of poultry and environmental data were recorded, providing information about microclimate (temperature, humidity, CO2, NH3), as well as data about poultry feeding (water and feed consumption and its composition, i.e., macro/micro nutrients and trace elements). Temperature and humidity monitoring sensors were placed in the center of the chicken house. CO_2 (IR-2 sensor, GDS Technologies Garforth) and NH_3 (NH_3/MR-100 sensor, Membrapor AG) concentrations were measured continuously every 10 minutes, but average values were calculated hourly thereafter.

Due to the early implementation phase of the monitoring system, the data quality of collected data is not ideal. For example (see Fig. 3.1), average temperature data has incompleteness, while feed consumption shows outliers.

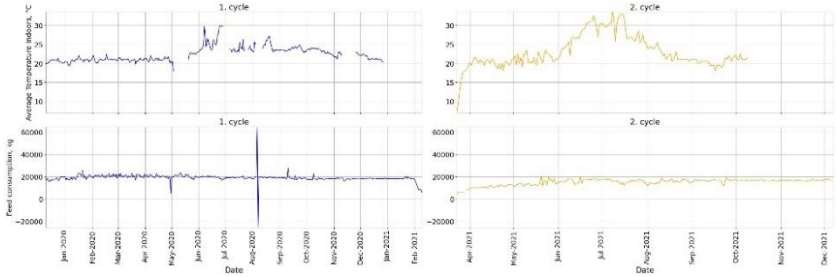


Fig. 3.1. **Data quality representation of two production runs.**

Description: first cycle (1st column) and second cycle (2nd column) average indoor temperatures (1st row) and feed consumption (2nd row).

For further analysis, first cycle average indoor temperature data was used. Due to the data file having three columns, the average value of these columns is analysed: temperature from the sensor located in the 1st cage floor, temperature from the sensor located in the 8th cage floor, and manually entered temperature. The final value, in cases where all column values were not available, was obtained using the existing ones.

The average temperature shows a slightly cyclical pattern with repeated maxima and minima. This could indicate seasonal changes or other periodic factors affecting temperature. Furthermore, a significant upward trend is observed in the final part of the series, indicating that temperature generally increased during this period.

In total, the following statistical values can be distinguished:

- number of data points: 335;
- mean: 22,37 °C;
- standard deviation: ≈1,96 °C;
- minimum value: 18,00 °C;
- 25. percentile (Q1): ≈20,93 °C;
- median (50. percentile): 21,90 °C;
- 75. percentile (Q3): ≈23,55 °C;
- maximum value: 29,90 °C;
- number of missing values: 93.

ARIMA model

One of the methods, more specifically – statistical models that combines autoregressive function (*AR*), integration (*I*, which refers to data differentiation to make it stationary) and moving average (*MA*) components, is the ARIMA model. The model can be applied for determining and replacing missing data. Individual components are described below.

- AutoRegressive (*AR*): autoregressive parameter. A model that uses the dependent relationship between an observation and several lagged observations (previous time periods) *(see* (3.1.)):

$$\mathrm{AR(p)}: \mathrm{Y_t} = \mathrm{c} + \phi_1 \mathrm{Y_{t-1}} + \phi_2 \mathrm{Y_{t-2}} + \cdots + \phi_\mathrm{p} \mathrm{Y_{t-p}} + \epsilon_\mathrm{t} \qquad (3.1.)$$

where

(Y_t) – series value at time t;
c – constant, integer or decimal number;
$\phi_1, \phi_2, \ldots, \phi_p$ – model parameters;
p – *AR* term order;
ϵ_t – white noise (error term) at time t.

- I (d): integrated parameter. Differencing of observations to make time series stationary; d is the number of non-seasonal differences (see (3.2.)):

$$\mathrm{I(d)}: \nabla^\mathrm{d} \mathrm{Y_t} \qquad (3.2.)$$

- *MA* (q): moving average parameter. A model that uses the dependency between an observation and residual error from the moving average model applied to lagged observations (see (3.3.)):

$$\mathrm{MA(q)}: \mathrm{Y_t} = \mathrm{c} + \epsilon_\mathrm{t} + \theta_1 \epsilon_{\mathrm{t-1}} + \theta_2 \epsilon_{\mathrm{t-2}} + \cdots + \theta_\mathrm{q} \epsilon_{\mathrm{t-q}} \qquad (3.3.)$$

Combining AR, I, and MA parameters, the ARIMA model is expressed as follows (see (3.4.)):

$$\mathrm{ARIMA(p,d,q)}: \nabla^\mathrm{d} \mathrm{Y_t} = \mathrm{c} + \phi_1 \mathrm{Y_{t-1}} + \cdots + \phi_\mathrm{p} \mathrm{Y_{t-p}} + \theta_1 \epsilon_{\mathrm{t-1}} + \cdots + \theta_\mathrm{q} \epsilon_{\mathrm{t-q}} + \epsilon_\mathrm{t} \qquad (3.4.)$$

where

p – number of lag observations included in the model, integer;
d – number of times that the raw observations are differenced to make the series stationary, integer;
q – size of the moving average window, integer.

The application of the ARIMA model requires stationary time series data. To verify this prerequisite, the Augmented Dickey-Fuller (ADF) test is used. This test determines whether a time series is stationary. The ADF test null hypothesis states that the time series has a unit root. The test results are based on p-value analysis – if it is lower than the chosen significance level (typically 0.05), this indicates time series stationarity and the presence of a time-dependent structure. Such a result confirms the data's suitability for using the ARIMA model.

Accordingly, for the time series y_t, ADF tests the null hypothesis (see (3.5.)):

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \cdots + \epsilon_t \qquad (3.5.)$$

where

Δ – difference operator. This operator is used to calculate the difference between the current value y_t and its previous value y_{t-1} in the time series, forming Δy_t;

α – constant;

β – captures potential linear time trend;

γ – coefficient at the series lag level, which captures the presence of a unit root. It is essential for determining whether the time series is stationary or not;

δ_i – coefficients of the dependent variable's lagged first differences;

ϵ_t – random error (or disturbance) component, which reflects the unexplained portion in the time series.

ADF test units are determined by the type of time series data being worked with and their units. For example, if the time series data are financial data expressed in euros, then most parameters will be expressed in euros or euros per time unit. The test focuses on γ. If γ differs significantly from zero, then the series has no unit root and is stationary.

In this case, the ADF statistic is a negative number −1.6112. The more negative it is, the stronger the rejection of the hypothesis that a unit root exists, and therefore, the stronger the evidence for stationarity. The p-value reflects the probability that the data would have the observed structure (or less likely) if the null hypothesis were true. Given the p-value of 0.4773, which is larger than the commonly used significance level of 0.05, the null hypothesis cannot be rejected. This means the series is not stationary.

In ARIMA modeling, the series must be stationary. Series non-stationarity means that before applying ARIMA, the series may need to be differenced to make it stationary. This is indicated by the "I" in ARIMA, which denotes the integrated parameter. The number of differences needed for the series to be stationary determines the d parameter.

The differenced series appeared more stationary, suggesting that d = 1 could be a good starting point for ARIMA. However, the choice of d should be based on achieving stationarity and reducing the Akaike Information Criterion (AIC) of the ARIMA model. Accordingly, in this case, to determine the appropriate difference level (d in ARIMA), the following is performed:

1. starting with d = 1,
2. d is increased and the stationarity of the differenced series is checked using the Augmented Dickey-Fuller (ADF) test,
3. an ARIMA model is fitted for each d value and AIC values are compared,
4. the optimal d is the one that makes the series stationary and minimizes AIC.

Performing the check yields: optimal difference level, d: 1; associated minimum AIC value: 707.13. As a result, one (d = 1) difference is sufficient for the series to be stationary and provide the best balance (according to AIC) for the ARIMA model.

Accordingly, using the determined d parameter value, the following result is obtained (see Fig. 3.2.):

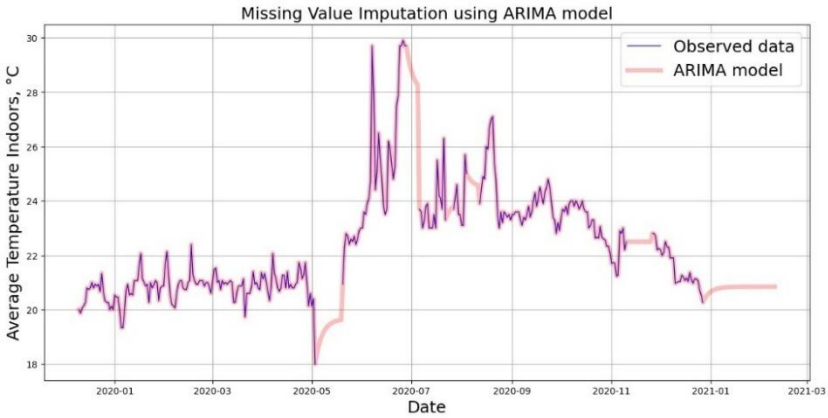


Fig. 3.2. **Imputation of missing data using an ARIMA model.**

Modified standard weighted average robust method

Alternatively, when applying data fusion principles for adding missing data, one can use both locally observed data (using, for example, the standard weighted average method) and the underlying trend, as well as corrections based on data characteristics such as skewness.

In this case, local information provides understanding of the direct context around the missing value. Global information or trend helps understand broader data patterns. Skewness can provide insight into the general distribution and nature of the data. Accordingly, if a data point is missing at index j in the time series, then the value y_j is calculated as follows:

1. the trend is supplemented using a linear regression model to determine the optimal number of neighbors to use for local weighted average (see (3.6.)):

$$y = \beta_0 + \beta_1 x \tag{3.6.}$$

where

N – number of data points considered, integer;
y_i – denotes the actual observed value, decimal number;
$\hat{y}_i$ – denotes the predicted or expected value, decimal number.

2. future indices x' are predicted using the trained model (see (3.7.)):

$$\hat{y} = \beta_0 + \beta_1 x' \tag{3.7.}$$

3. the optimal number of neighbors is found by comparing the calculated data with the linearly extended trend and calculating the Mean Square

Error (MSE), which determines how well the calculated data or predicted trend matches the actual data (see (3.8.)):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 \quad (3.8.)$$

where

N – number of data points considered;
y_i – denotes the actual observed value, decimal number;
$\widehat{y_i}$ – denotes the predicted or expected value, decimal number.

4. the local weighted average calculation is modified using exponential weights, where the exponential decay coefficient is -0.1, which ensures gradual weight reduction as the distance from the missing point increases (see (3.9.)):

$$L_{avg} = \frac{\sum_{i=j-n}^{j+n} e^{-0.1i} y_i}{\sum_{i=j-n}^{j+n} e^{-0.1i}} \quad (3.9.)$$

where

e – exponential function, constant;
i – distance from the missing point, integer;
y_i – data value at index i, decimal number.

5. the skewness S_D of dataset D is calculated to adjust the calculated value based on the asymmetry of non-missing data distribution (see (3.10.)):

$$S_D = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i - \overline{D}}{\sigma_D}\right)^3 \quad (3.10.)$$

where

D_i – denotes observed data points, integer or decimal number;
$\overline{D}$ – mean of observed data, decimal number;
σ_D – standard deviation of observed data, decimal number;
N – number of observed data points, integer.

6. according to skewness, the local weighted average is adjusted (see (3.11.)):

$$L_{avg_{adj}} = L_{avg} + \alpha \times S_D \quad (3.11.)$$

where

α = 0.01 is an empirically determined constant to control skewness impact.

7. before calculating distance weights, oscillation detection is performed by comparing the direction of value changes before and after the missing point (see (3.12.), (3.13.)):

$$O_f = 0.5, \text{ ja } \frac{dy_{back}}{dx} \cdot \frac{dy_{forw}}{dx} < 0 \quad (3.12.)$$

$$O_f = 1.0, \text{ ja } \frac{dy_{back}}{dx} \cdot \frac{dy_{forw}}{dx} \geq 0 \quad (3.13.)$$

where rates of change are calculated:

$$\frac{dy_{back}}{dx} = \frac{y_{j-1} - y_{j-n}}{n-1} \quad (3.14.)$$

$$\frac{dy_{forw}}{dx} = \frac{y_{j+n} - y_{j+1}}{n-1} \quad (3.15.)$$

where

O_f – oscillation factor, decimal number;

$\frac{dy_{back}}{dx}$ – rate of value change before the missing point, decimal number;

$\frac{dy_{forw}}{dx}$ – rate of value change after the missing point, decimal number;

n – number of neighbors, integer.

8. distance weights are calculated using exponential decay and oscillation factor, where the exponential decay coefficient is -0.15, which controls the rate of distance influence reduction (see (3.16.)):

$$w = e^{-0.15 \cdot \min(d_{left}, d_{right})} \cdot O_f \quad (3.16.)$$

where

d_{left} – distance to the nearest observed point on the left, integer;

d_{right} – distance to the nearest observed point on the right, integer.

9. using the calculated distance weights, the adjusted local average value is combined with the trend value to obtain the initial imputed value (see (3.17.)):

$$V_{imp}(i) = w \times L_{avg_{adj}}(i) + (1 - w) \times T(i) \quad (3.17.)$$

where

w – distance and oscillation weight, decimal number;

$L_{avg_{adj}}(i)$ – adjusted local average value at position i, decimal number;

$T(i)$ – trend value at position i, decimal number.

10. to reduce sudden value changes, temporal smoothing is applied, considering the previously calculated value, where coefficients 0.7 and 0.3 provide optimal balance between current and previous values, reducing sudden changes in the data (see (3.19.):

$$V_{imp}^{final}(i) = 0.7 \cdot V_{imp}(i) + 0.3 \cdot V_{imp}(i-1) \quad (3.18.)$$

where

$V_{imp}^{final}(i)$ – time-smoothed value at position i, decimal number;

$V_{imp}(i)$ – initially calculated value, decimal number;

11. for final result smoothing, a three-point moving average is applied (see (3.19.)(3.20.)):

$$V_{smooth}(i) = \frac{1}{3}\sum_{k=i-1}^{i+1} V_{imp}^{final}(k) \quad (3.19.)$$

where

$V_{imp}^{final}(k)$ – time-smoothed values at three consecutive points, decimal number.

This approach ensures that the calculated values are influenced by both local context (observed adjacent data points and their trend) and the overall trend in the dataset, while also making slight adjustments based on data distribution unevenness. This ensures that the calculated values not only correspond to local and general trends but are also adjusted taking into account the asymmetry of data distribution. As a result, the obtained data points complement the original data as shown in Fig. 3.3.

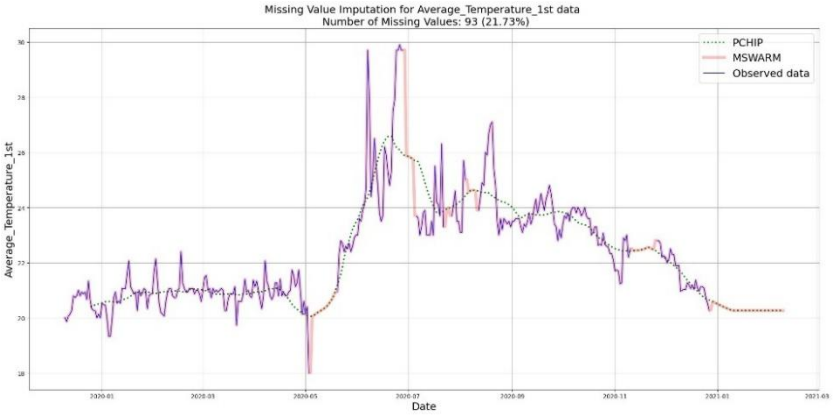


Fig. 3.3. **Imputation of missing data using the MSWARM method.**

Overall, this is the Modified Standard Weighted Average Robust Method (MSWARM), which introduces several significant improvements compared to the traditional approach. It uses a dynamic weighting system that adapts to the data structure. The method works in four important ways. First, it automatically adjusts its calculations by focusing more on data points that are near the missing data. Second, it looks at how the data changes over time to make sure the results match these patterns. Third, it adjusts its results based on how the data is distributed, making the calculations more accurate. Finally, it combines local data points with overall patterns – using nearby data when it's available, and switching to general trends when data is thin. This complex approach allows more precise modelling of missing values, taking into account both local data and global trends in the dataset.

or method evaluation, the 2nd production cycle dataset is used, which contains several parameters without missing values. According to the parameter used in method development, the average temperature is selected from the 2nd production cycle dataset. This dataset is accepted as the ground truth dataset and is used for calculating criteria (see Table 3.1): Mean Square Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Root Mean Squared Percentage Error (RMSPE). The lower the criteria value (error), the better the method's result corresponds to the data.

Table 3.1. **Performance metrics**

Metric	Equation
Mean Square Error	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Mean Absolute Error	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \widehat{y_i}\rvert$
Mean Absolute Percentage Error	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{y_i}$
Root Mean Squared Error	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Root Mean Squared Percentage Error	$RMSPE = 100\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$

where

yi – observed value;

ŷi – expected value;

n – number of records.

Observed sets are created by introducing missing values with the following conditions:

1. random missing values, up to 10% of total count;
2. random missing values, up to 20% of total count;
3. random missing values, up to 40% of total count;
4. random sequences of missing values, two pieces with length of 10 elements each;
5. random sequences of missing values, three pieces with length of 10 elements each;
6. random missing values, up to 20% of total count, and random sequences of missing values, two pieces with length of 10 elements each.

In addition to the ARIMA model (hereinafter referred to as method) and MSWARM, a 2nd degree polynomial interpolation method was applied. Higher degree polynomial interpolation showed worse results. The MSWARM method uses Piecewise Cubic Hermite Interpolating Polynomial. According to the mentioned conditions, six calculations were performed.

For example, in scenario with up to 20% of the data missing randomly, combined with two sequences of 10 consecutive missing values, the variability in missing value production rates between iterations was noted due to the random nature of the data omission. In such cases, polynomial interpolation demonstrated significantly higher MSE, RMSE, and MAE values compared to other techniques, as indicated in Table 3.2, Fig. 3.4., suggesting lower accuracy under these specific conditions.

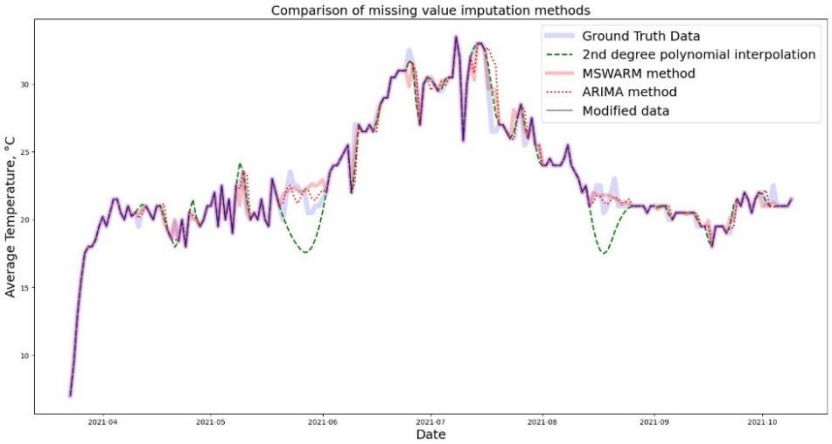


Fig. 3.4. **Up to 20% of the data missing randomly, along with two sequences of 10 consecutive missing values each.**

The high values of MSE and RMSE indicate that some forecastings have large errors, which could be due to overfitting for missing pieces.

Table 3.2. **Metrics results for a scenario with up to 20% of the data missing randomly, along with two sequences of 10 consecutive missing values each**

Method	***MSE***	***RMSE***	***MAE***
Polynomial interpolation	1.312570	1.145675	0.449252
MSVSM	0.377772	0.614632	0.236004
ARIMA	0.648512	0.805302	0.281817

Polynomial methods can create large fluctuations in interpolated values when encountering data imperfections, which appears to be the case here. MSWARM shows the lowest error indicators, suggesting that it handles both random missing values and missing segments more effectively than other methods. Particularly, the low RMSE indicates that MSWARM provides a consistent level of accuracy for calculated values without large outliers from actual values. ARIMA imputation has moderate error metrics, performing better than polynomial interpolation but not as well as MSWARM.

MSWARM significantly outperforms the other two methods, indicating that it has a better mechanism for handling both types of missing data. This could be

particularly useful in real scenarios where missing data often occurs in both random and structured forms.

For comprehensive (multi-scenario) evaluation of methods, the 2nd production cycle dataset is used with various configurations:

- sequence sizes: 2, 4, 6, 8, 10, and 12 time steps;
- number of sequences: 1, 2, 3, 4, and 5 sequences in each dataset;
- base proportions of missing values: from 0% to 20% with 2% step;
- each configuration is tested 3 times to ensure statistical reliability.

In total, 990 tests are performed, derived from 6 sequence sizes, 5 sequence counts, 11 proportions, and 3 repetitions. Each test procedure consists of two phases. In the first phase, missing data imputation is performed, where first the sequences of missing values are inserted at random positions, then additional random missing values are inserted in the remaining data, and the total percentage of missing data is calculated. In the second phase, method application and evaluation are performed, where performance indicators (MSE, RMSE, MAE) are calculated for each method (ARIMA, MSWARM, polynomial interpolation), and results are summarized and analysed. For determining the deterioration point, a systematic analysis is used, consisting of data aggregation and threshold determination phases. In the data aggregation phase, results are grouped by method and total percentage of missing data, and average performance indicators are calculated for each group. In the threshold determination phase, a threshold coefficient of 1.5 (50% error increase) is used. Starting from the lowest percentage of missing data, initial performance is determined, and each subsequent point is compared to it. The first point where the indicator exceeds initial performance × 1.5 is marked as the deterioration point.

MSWARM shows (see Fig. 3.5) the best overall performance with a deterioration point at 52.3% missing data, which is significantly higher than the ARIMA method (48.1%) and polynomial interpolation (25.0%). To ensure statistical reliability, each configuration is tested three times, reducing the impact of random variations in missing data placement on the best overall performance among all three methods, with a deterioration point at 52.3% missing data – the highest threshold among all methods.

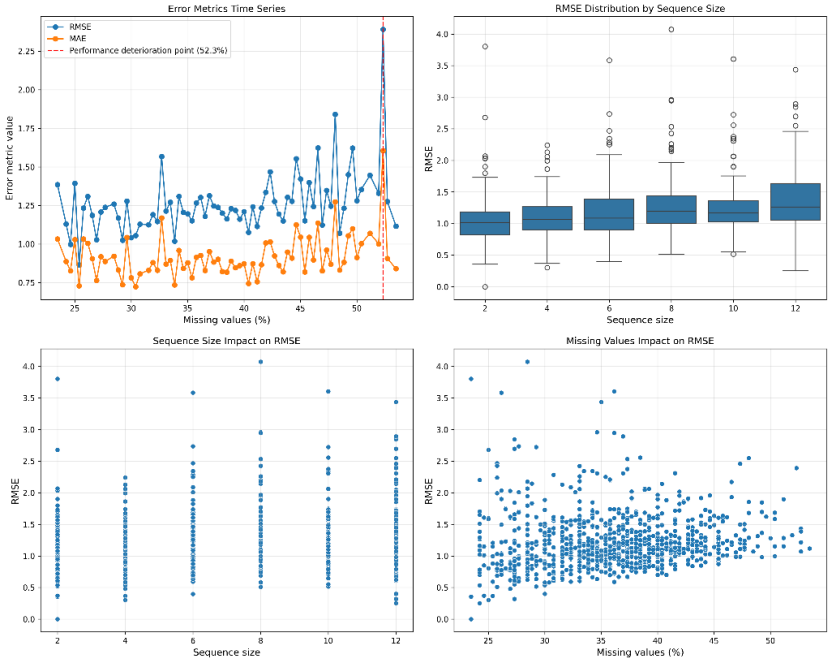


Fig. 3.5. **MSWARM performance.**

Comparative analysis shows a more gradual RMSE increase for the MSWARM. It demonstrates excellent performance with small to medium sequence sizes (10–30 elements) across all percentages of missing data, where the lowest RMSE values (1.0–1.2) are consistently achieved with sequence sizes below 20 elements. Even after the deterioration point, the MSWARM shows the smallest performance degradation – only a 36.7% increase in MSE and an 18.6% increase in RMSE, which is significantly better than ARIMA (MSE 91.4%) and polynomial interpolation (MSE 137.6%).

MSWARM stands out as the most effective method, particularly in situations with high proportions of missing data. In comparison with other methods, ARIMA shows good results only at lower proportions of missing data (<48%), while polynomial interpolation becomes unreliable already at 25% missing data, with a dramatic RMSE increase above 2.0 in higher load scenarios. Based on these results, MSWARM is the recommended choice for most practical applications, especially in cases where a high proportion of missing data is expected or high precision is required.

While MSWARM showed better results in previous scenarios, it is important to understand how stably it performs. The distribution comparison test (see Fig. 3.6) is a fundamental indicator in evaluating imputation quality as it demonstrates the precision of preserving statistical properties of the data. To ensure result

stability and minimize the impact of random factors, each scenario was executed 10 times, and average indicators were used for further analysis.

In the scenario with 10% random missing values, the method demonstrates very stable results – a mean value difference of 0.348% indicates high precision in data level restoration. Standard deviation changes within 1.862% suggest good preservation of data dispersion. The distribution shape visually shows almost no difference from the original, which is crucial for further statistical analysis.

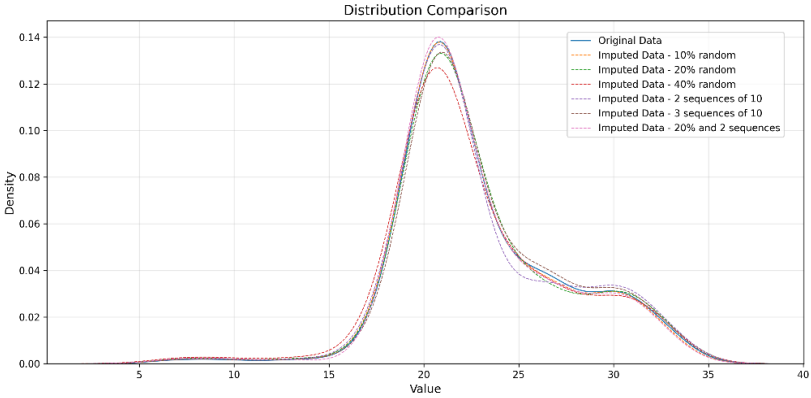


Fig. 3.6. **Distribution comparison of the MSWARM method for six scenarios.**

As the proportion of missing values increases to 20% and 40%, a gradual decrease in precision is observed. In the 20% scenario, the mean value difference is 0.362%, while the standard deviation difference is 2.884%. In the 40% scenario, the standard deviation difference increases to 5.079%, indicating a more substantial change in data dispersion. The distribution shape shows significant deviations, particularly in the "tail" regions, indicating difficulties in accurately reconstructing extreme values.

The autocorrelation test (see Fig. 3.7) is particularly significant in time series analysis as it shows the preservation of sequential value relationships. In consecutive missing value scenarios (2 and 3 sequences of 10), the method maintains good precision. In both scenarios, the mean value difference is 0.994% and 0.952% respectively, while standard deviation differences are 3.120% and 3.869%. Notably, autocorrelation indicators remain low (mean difference around 0.030–0.046), suggesting good preservation of time series structure.

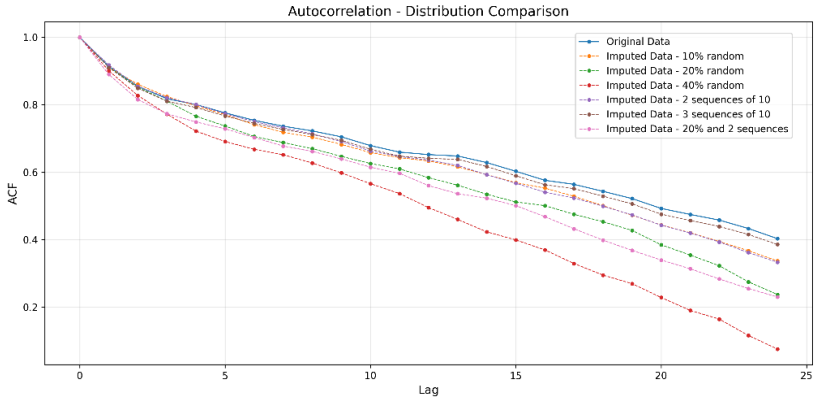


Fig. 3.7. **Autocorrelation comparison of the MSWARM method for six scenarios.**

In the combined scenario (20% random and 2 sequences), the method shows similar precision to individual scenarios – mean value difference of 0.486%, standard deviation difference of 1.954%, and moderate autocorrelation indicators (mean difference 0.070, maximum 0.164). This result might be explained by the combination of different types of missing values allowing the method to better "capture" data periodicity and trends.

Overall, it can be concluded that the method is suitable for practical use in situations where the proportion of missing values does not exceed 20–25% of the total data volume. Special attention should be paid to cases where accurate reconstruction of extreme values or preservation of higher-order autocorrelations is important, as these aspects show the largest deviations.

Outlier detection and adjustment

Several methods are used for outlier detection and adjustment, each with its own advantages. The Z-Score method is based on standard deviation calculation but is sensitive to extreme values (Yaro et al., 2024). The Interquartile Range (IQR) method is more effective for asymmetric data but can be too conservative (El Hairach, Tmiri, & Bellamine, 2024). Winsorization adjusts outliers by replacing them with the nearest "normal" values, preserving data structure (Yang. L. et al., 2024). The sliding window method analyses data in its local context, particularly suitable for time series data. The developed combined approach integrates winsorization and sliding window methods, using three different window sizes (9, 19, and 39 points). The smallest window identifies short-term outliers, the medium window provides stability, while the largest helps determine long-term trends. Local statistical indicators are calculated in each window, and winsorization is applied with a z-value threshold of 3.0. Additionally, the introduced trend component allows distinguishing genuine outliers from natural data changes, especially in temperature data. Outlier

processing occurs in two phases. In the detection phase, local statistical values are calculated for each point across all three windows, and a point is classified as an outlier if its z-value exceeds the threshold in at least two windows. In the adjustment phase, new values are calculated for identified outliers using winsorization, taking into account both the local data structure and the trend component.

The developed method is named Multi-scale Integrated Outlier Analysis Method (MIOAM), as this name accurately reflects its main characteristics and operating principles. The "Multi-scale" component indicates the method's ability to analyse data at various time scales, using three different window sizes (9, 19, and 39 points), which allows identification of both short-term and long-term outliers. "Integrated" refers to the combination of multiple statistical approaches – trend analysis, local variations, multi-window statistics, and confidence assessment – into a unified system. "Outlier analysis" encompasses both outlier identification and correction, using adaptive thresholds and local data structure.

MIOAM is developed to efficiently identify and correct outliers in time series. The method combines multiple statistical approaches and operates at various time scales, ensuring both precise outlier detection and careful correction. The method consists of several, 12, sequential steps:

1. to determine the data trend, a rolling median is first calculated. This step is essential as it reduces the impact of short-term fluctuations while preserving the fundamental data structure. The rolling median is calculated using a symmetric time window around each point (see (3.20.)):

$$rolling\_med_i = median(x_{i-k}, \dots, x_i, \dots, x_{i+k}) \quad (3.20.)$$

where

x_i – data value at index i, decimal number;
k – window half-width, (window_length-1)/2, integer;
i – current point index, integer.

2. after calculating the median, the Savitzky-Golay filter is applied, which smooths the data while preserving higher-order moments (see (3.21.)):

$$trend_i = \sum_{j=-k}^{k} c_j \cdot rolling\_med_{i+j} \quad (3.21.)$$

where

c_j – Savitzky-Golay filter coefficients, decimal numbers;
k – filter window half-width, integer;
$rolling\_med_{i+j}$ – rolling median value at point i+j, decimal number.

3. for evaluating local data variations, the Median Absolute Deviation (MAD) is used. This method is more robust against extreme values than standard deviation as it is based on the median rather than the mean value (see (3.22.)):

$$MAD_i = median(|x_j - median(x)|) \quad (3.22.)$$

where

x_j – data values in the local window, decimal numbers;
median(x) – data median in the local window, decimal number;
i – current point index, integer.

4. to normalize local variations relative to overall data variability, the relative MAD value is calculated. This indicator allows comparing variations across different data segments (see (3.23.)):

$$local\_var_i = \frac{MAD_i}{median(MAD)} \quad (3.23.)$$

where

MAD_i – local median absolute deviation at point i, decimal number;
$median(MAD)$ – median of all MAD values, decimal number.

5. for analyzing the rate of value changes, differences between consecutive points are calculated. This indicator is essential for identifying sudden changes in data that may indicate potential outliers (see (3.24.)):

$$\Delta x_i = |x_i - x_{i-1}| \quad (3.24.)$$

where

x_i – current data value, decimal number;
$x_{(i-1)}$ – previous data value, decimal number.

6. the rate of change threshold is dynamically adjusted based on median changes and local variation. This adaptive threshold allows more precise outlier identification across different data segments, considering both overall data structure and local characteristics (see (3.25)):

$$spike\_threshold_i = \left(median(\Delta x) + 2 \cdot median(|\Delta x - median(\Delta x)|)\right) \cdot max(1, local\_var_i) \quad (3.25.)$$

where

$median(\Delta x)$ – median of all changes, decimal number;
$local\_var_i$ – local variation at point i, decimal number.

7. local mean is calculated for each of the three different windows (9, 19, and 39 points). This approach allows identifying outliers at various time scales, ensuring method effectiveness for both short-term and long-term outliers (see (3.26.)):

$$\mu_{i,w} = \frac{1}{w}\sum_{j=i-k}^{i+k} x_j \quad (3.26.)$$

where

w – window size (9, 19, or 39), integer;

k – $(w-1)/2$, integer;
x_j – data values in the window, decimal numbers.

8. local standard deviation is also calculated for each window, characterizing data dispersion in the respective time window (see (3.27.)):

$$\sigma_{i,w} = \sqrt{\frac{1}{w}\sum_{j=i-k}^{i+k}\left(x_j - \mu_{i,w}\right)^2} \quad (3.27.)$$

where
$\mu_{i,w}$ – local mean in window w, decimal number;
w – window size;
x_j – data values in the window, decimal numbers.

9. using local mean and standard deviation, z-score is calculated for each point in each window. Z-score shows how many standard deviations a point is from the local mean (see (3.28.)):

$$z_{i,w} = \frac{|x_i - \mu_{i,w}|}{\sigma_{i,w}} \quad (3.28.)$$

where
x_i – current data value, decimal number;
$\mu_{i,w}$ – local mean in window w, decimal number;
$\sigma_{i,w}$ – local standard deviation in window w, decimal number.

10. The confidence score combines various outlier indicators into a single numerical value. This score considers both z-scores from different windows and other outlier indicators (see (3.29.)):

$$\begin{aligned} confidence_i = \sum_w weight_w \cdot [z_{i,w} > z_{threshold}] + \\ 1.5 \cdot [rapid\_recovery_i] + 1.0 \cdot [trend\_outlier_i] + \\ 1.2 \cdot [consecutive\_deviation_i] \end{aligned} \quad (3.29.)$$

where
$weight_w$ – weight for each window size, decimal number;
$[nosacījums]$ – 1 if condition is true, 0 if false;
$z_{\text{threshold}}$ – z-score threshold, decimal number.

11. the outlier identification condition set combines three main criteria in a single decision-making step:
    a. $confidence\_i \geq 2.0$ checks if the overall confidence indicator, obtained from the analysis of different windows and additional indicators' weighted sum, exceeds the threshold of 2.0;
    b. $|\Delta x\_i| > spike\_threshold\_i$ checks if the point's rate of change exceeds the adaptive threshold adjusted for local variability;
    c. $|x\_i - trend\_i| > 2.0 \cdot \sigma_{\text{trend}}$ checks if the point's deviation from the calculated trend exceeds twice the trend's standard deviation.

If any of these three conditions is true, the point is marked as an outlier ($outlier_i$ = 1), otherwise it is considered a normal point ($outlier_i$ = 0) (see (3.30.)):

$$outlier_i = 1 \text{ ja } confidence_i \geq 2.0 \text{ vai } |\Delta x_i| > spike\_threshold_i \text{ vai } |x_i - trend_i| > 2.0 \cdot \sigma_{trend} \text{ citādi } 0 \quad (3.30.)$$

where

$confidence_i$ – confidence score of point i, decimal number;
Δx_i – value change at point i, decimal number;
$spike\_threshold_i$ – change threshold at point i, decimal number;
σ_{trend} – trend standard deviation, decimal number.

12. the outlier correction process is adaptive and based on trend values:
    a. if a point is identified as an outlier ($outlier_i$ = 1), its value is corrected using the trend as a reference point;
    b. the correction direction ($sign$ function) remains the same as the original deviation from the trend;
    c. the correction amplitude is limited by the min function, which prevents the correction from exceeding twice the trend's standard deviation ($2 \cdot \sigma_{trend}$);
    d. if the point is not an outlier ($outlier_i$ = 0), its value remains unchanged (x_i).

This approach ensures that corrections are statistically justified and preserve natural data variation while removing extreme outliers (see (3.31.)):

$$x_i' = trend_i + sign(x_i - trend_i) \cdot min(|x_i - trend_i|, 2 \cdot \sigma_{trend}), \; ja \; outlier_i = 1; \; citādi \; x_i \quad (3.31.)$$

where

x_i – original data value, decimal number;
$trend_i$ – calculated trend at point i, decimal number;
σ_{trend} – trend standard deviation, decimal number;
$outlier_i$ – outlier indicator (1 or 0), integer.

MIOAM parameters were optimized to identify the most effective combination, ensuring the highest method performance while maintaining its stability and reliability across various application scenarios. This optimization is essential because parameter selection directly affects both the method's ability to precisely identify outliers and its tendency to generate false positives. Overly strict parameters can lead to excessive sensitivity to normal data variations, while overly weak parameters might miss significant outliers. Thus, the optimization task is to find a balance between these opposing aspects.

The optimization process was structured in three sequential phases. First, a testing environment was created with various outlier scenarios reflecting problems encountered in real situations – isolated outliers (short-term extreme values), consecutive outliers (shifts in multiple consecutive points), trend

changes (gradual outliers), and their combinations. Second, the parameter space coverage was defined with two main parameters. For window sizes, four different triple-window combinations were selected ([3,7,11], [5,11,21], [7,15,31], [9,19,39]), and for z-score thresholds – four values (1.5, 2.0, 2.5, 3.0). This parameter set is based on previous research results and theoretical considerations about data structure analysis scales. Third, an evaluation system was developed with several complementary metrics: F1 score as the main metric combining precision and completeness, MAE for assessing corrected value accuracy, and distribution preservation indicators for evaluating changes in mean value and standard deviation.

A complex approach was used for comparing configurations, where the F1 score median served as the primary indicator, supplemented by other significant criteria. Mean Absolute Error (MAE) and Mean Square Error (MSE) provided deeper insight into corrected value accuracy, while separate precision and completeness analysis allowed better understanding of the method's performance specifics in various scenarios. This multifaceted analysis approach enabled not only identifying generally best and worst configurations but also understanding their performance characteristics in different use cases. Sorting results by F1 score median provided a clear and justified configuration hierarchy while maintaining the ability to evaluate each combination's strengths and weaknesses in detail.

The optimization results reveal (see Fig. 3.8) that the most effective parameter combination is the medium-sized window set [7,15,31] with a z-score threshold of 3.0.

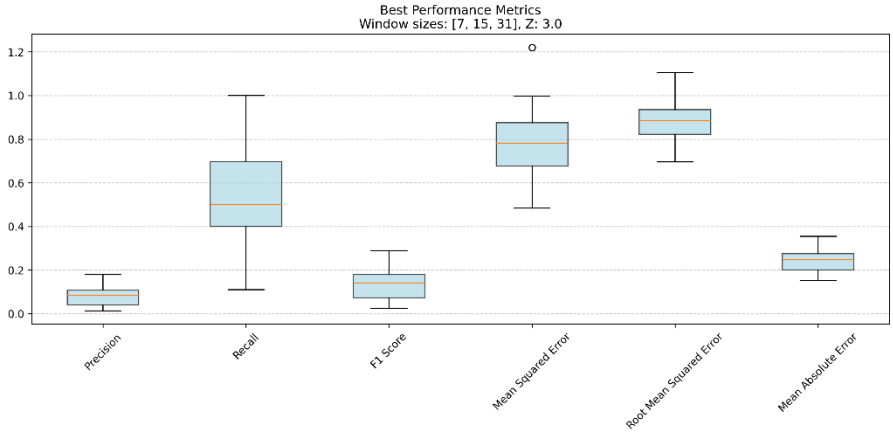


Fig. 3.8. **Best performance metrics.**

This configuration shows an F1 score of 0.14, MAE of 0.25, and MSE of 0.8, which are the best results among all tested combinations. In contrast, the smaller window combination [3,7,11] with a z-score threshold of 2.0 shows significantly worse results. Distribution analysis reveals important nuances about the method's

impact on data structure (see Fig. 3.9). For the optimal configuration ([7,15,31], z=3.0), excellent preservation of the fundamental data structure is observed with a mean value deviation of 0.02% and standard deviation changes of 0.40%.

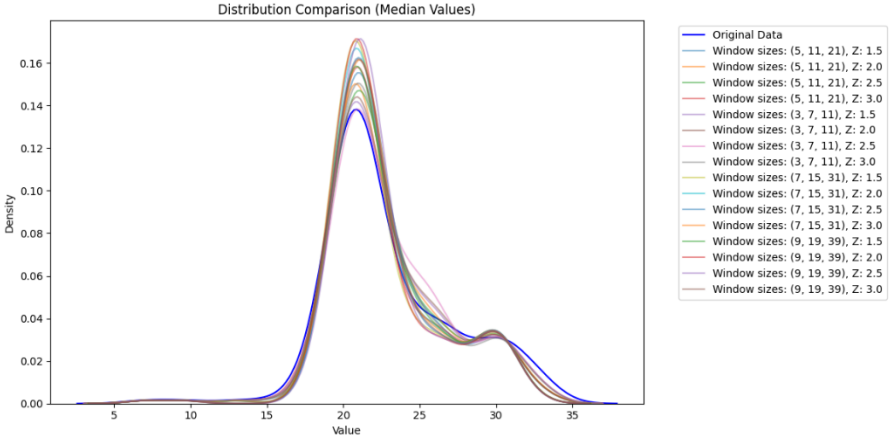


Fig. 3.9. **Distribution comparison for all scenarios.**

The skewness coefficient remains practically unchanged (changes <0.1%), while kurtosis shows a slight decrease (0.5%). In extreme value regions, the deviations are minimal – 0.8% at the 1st percentile and 0.3% at the 99th percentile. Compared to the worst configuration ([3,7,11], z=2.0), the mean value deviation increases to 1.2%, standard deviation changes reach 2.8%, and significant "tail" deformation is observed. In the modal region, the optimal configuration shows almost perfect agreement with the original distribution, providing a reliable foundation for statistical analysis.

Testing the outlier detection and adjustment method

To evaluate the method's effectiveness, average temperature indoor measurements from the 2nd production cycle were used. The initial data shows seasonal trends with temperature values from 19°C to 33°C. Three types of outliers were introduced during testing, reflecting problems encountered in real situations. The figure (see Fig. 3.10) shows these outlier manifestations – detected outliers marked with red crosses, including both rapid short-term changes and longer-term deviations from the base trend.

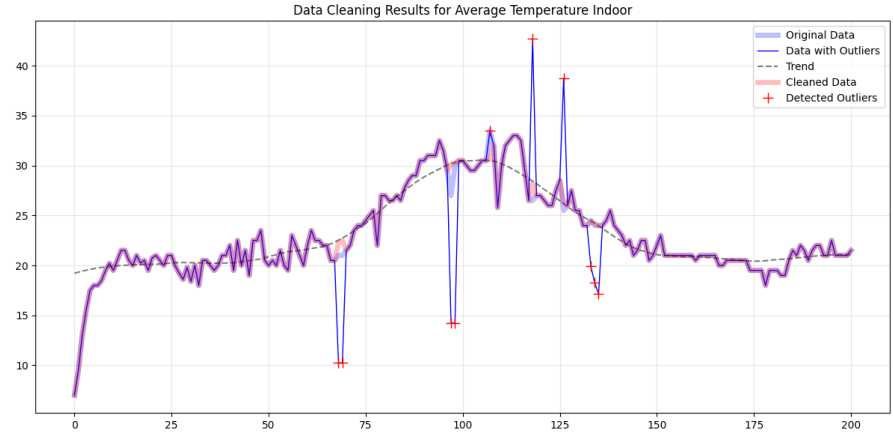


Fig. 3.10. **MIOAM execution result for the random outlier scenario.**

Two peak outliers are observed in the middle of the time series with temperature increases up to 42°C and 39°C. The method not only identifies these extreme values but also successfully restores credible temperature values, as demonstrated by the red line. The grey dashed line shows the long-term trend, which the corrected values consistently follow, indicating the method's ability to distinguish genuine outliers from natural temperature fluctuations. Around observation 75, a temperature decrease to 10°C is visible, and around observations 100 and 125 – several consecutive outliers, which are effectively processed while maintaining the natural flow of the dataset. Quantitative analysis shows a Mean Absolute Error (MAE) of 1.004°C, Mean Absolute Percentage Error (MAPE) of 4.08%, and Root Mean Square Percentage Error (RMSPE) of 5.54%, confirming the method's stability. The method's effectiveness is ensured by the chosen window size combination [7, 15, 31] and z-score threshold of 3.0.

A multi-test evaluation system was developed with 100 independent test iterations. Each iteration introduces a unique set of artificial outliers into the base temperature data with three basic types: spikes (sudden, extreme outliers), shifts (prolonged outliers over multiple points), and trends (gradual, directional changes). The proportion of introduced outliers in each test varies from 2% to 5% of total data points. For each test iteration, the method processes data using window sizes [7, 15, 31] and a z-score threshold of 3.0. The evaluation records four main performance indicators: RMSE (Root Mean Square Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and RMSPE (Root Mean Square Percentage Error). RMSE fluctuates around an average of 0.970 with a standard deviation of 0.258, with higher values appearing in tests with more intense outliers. MAE shows a more stable profile with a mean value of 0.244 and standard deviation of 0.080, indicating precise corrections. MAPE with a mean value of 1.139 and standard deviation of 0.386 demonstrates the method's accuracy in percentage terms. RMSPE shows an average of 5.216 with

a standard deviation of 1.682, with higher values indicating individual cases with larger percentage errors.

The box plot distributions of error metrics (see Fig. 3.11) demonstrate the method's consistency. RMSE shows a median of 0.970 with an interquartile range from 0.812 to 1.128 and whiskers from 0.584 to 1.496. MAE shows a lower median (0.244) and narrower IQR (0.194 to 0.294). MAPE reveals a median of 1.139 with IQR from 0.901 to 1.377. RMSPE shows a median of 5.216 and IQR from 4.123 to 6.309. The observed variations in metric values are related to the number of outliers (2–5% of total points) and their intensity (1.5–4 standard deviations).

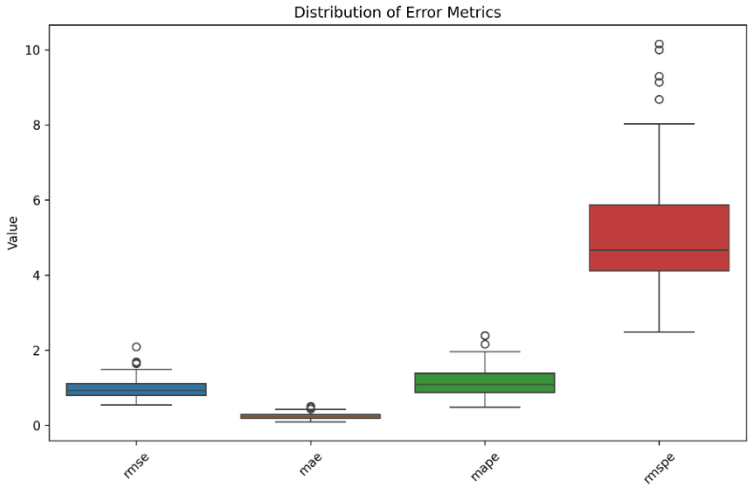


Fig. 3.11. **MIOAM error metric distribution.**

Combined analysis of all four metric box plots provides convincing evidence of the method's stability and accuracy. Absolute error metrics (RMSE, MAE) demonstrate acceptable accuracy in the context of temperature measurements, while relative error metrics (MAPE, RMSPE) confirm the method's reliability across a wide range of measurements. Compact interquartile ranges for RMSE (0.541 to 2.086) and MAE (0.094 to 0.504) indicate stable performance across all tests. The relatively small number of outliers in these distributions indicates that the method rarely experiences significant performance degradation, even under complex conditions.

Error distribution histograms (see Fig. 3.12) reveal nearly normal distributions for all indicators, with a slight right skew. The RMSE distribution is centered around 0.970 with right skewness. The main mass concentrates in the 0.6–1.2 range, which is acceptable in the context of temperature measurements. The right-side “tail” reflects rare cases with larger errors in more complex outlier cases. The MAE histogram shows a more compact distribution around 0.244, with less pronounced asymmetry, indicating a stable correction process.

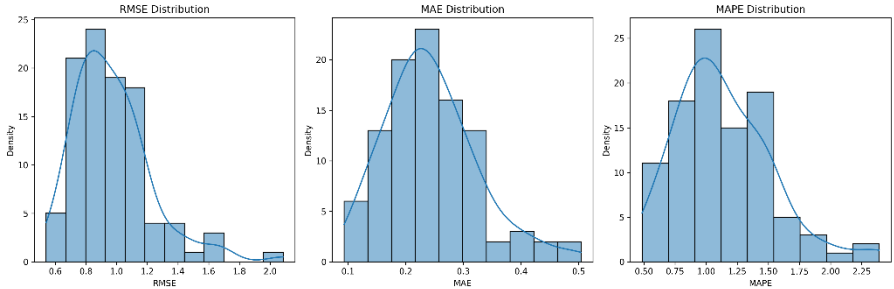


Fig. 3.12. **MIOAM error distribution histograms.**

The histogram analysis reveals meaningful patterns in error distributions. All four metrics demonstrate unimodal distributions with notable symmetry (accounting for expected right-skewed behaviour in quadratic metrics), indicating consistent method performance. The narrow distributions observed in MAE and MAPE metrics are particularly significant, demonstrating robust accuracy across both absolute and relative error measurements. While RMSE and RMSPE exhibit broader distributions, this characteristic aligns with established methodological principles and does not compromise the method's practical effectiveness.

Multi-test results demonstrate significant consistency and reliability achieved with the outlier detection and correction method. With an average RMSE of 0.970 units and standard deviation of 0.258 units, the method achieves high accuracy while maintaining stability across various test scenarios. The low MAPE (average 1.139) indicates excellent relative accuracy, which is essential in many practical applications where proportional accuracy is important.

Combining these results with previous stability testing results provides comprehensive method validation. Stability tests revealed optimal performance with window sizes [7, 15, 31] and z-threshold 3.0, which multi-test analysis has now confirmed across a broader range of scenarios. MIOAM demonstrates both point accuracy (shown in stability tests) and statistical reliability (proven in multiple tests).

Stability testing identified the method's ability to maintain data integrity while removing outliers, maintaining an average detection rate of 90% with a mean absolute error of 1.004 units. Multi-test analysis reinforces these findings, showing even better performance with an average MAE of 0.244 units across various scenarios. These results show that MIOAM not only stays reliable but actually performs even better when tested across more diverse situations.

The evidence from both testing approaches clearly shows that MIOAM works effectively for many different types of time series analysis tasks. It successfully balances sensitivity to genuine outliers with resistance to false positives, maintaining high accuracy while adapting to different data patterns. MIOAM's consistent performance in both stability and multi-test evaluations confirms its

readiness for implementation in production environments where reliable outlier detection is crucial for ensuring data quality.

RESULTS

The PhD thesis has investigated data fusion relevance across several sectors: precision beekeeping, precision poultry farming, and smart transport and surveillance systems in the context of IoT systems. Through literature review, existing data fusion methods, their models, and architectures were analysed. The research found that data fusion models do not limit the development of new methods. Within the DIKW model framework, data and information terminology was analysed, determining its impact on data fusion methodology selection. Initial research focused on developing a data fusion concept for precision beekeeping needs.

A data layering concept was developed, based on spatial time series data fusion principles. The concept rests on three main layers: plant richness, bee activity, and precipitation amount. Two approaches combined these layers: weighted interpolation and Principal Component Analysis-based method. During practical verification, data quality emerged as a significant limiting factor in applying data fusion methods. This identified a need for specialised methods for missing value replacement and outlier processing that could preserve the dataset's statistical properties. Precision poultry farming datasets served as a practical validation platform for testing the developed methods.

Two methods were developed and evaluated for data quality improvement – the Modified Standard Weighted Average Robust Method (MSWARM) for missing value replacement and the Multi-scale Integrated Outlier Analysis Method (MIOAM) for outlier detection and correction, to improve data quality in time series analysis. Effective analysis and forecasting directly depend on data completeness and accuracy, but in real data collection environments, there is often significant value absence or irregularities. MSWARM and MIOAM methods are suitable for solving these problems, demonstrating the ability to adapt to various types of imperfections and proportions, stabilizing the dataset and reducing error impact. Note: throughout this section, the wavy equals sign (≈) indicates values that have been rounded to four decimal places for clarity and consistency.

MSWARM method stood out with remarkable robustness across various scenarios. With a small proportion of missing data (up to 10%), MSWARM achieved very low errors: MSE≈0.1015, RMSE≈0.3186, and MAE≈0.0641. In comparison, ARIMA under the same conditions showed MSE≈0.1854, RMSE≈0.4305, and MAE≈0.1027, while polynomial interpolation showed MSE≈0.1330, RMSE≈0.3647, and MAE≈0.0795. As missing values increased to 40%, MSWARM maintained relatively low error levels (MSE≈0.4599, RMSE≈0.6782, MAE≈0.3311), while alternative methods lost accuracy. Even in more complex cases, with three 10-element blocks of missing values,

MSWARM provided MSE≈0.5075, RMSE≈0.7124, and MAE≈0.2507, significantly outperforming polynomial interpolation (MSE≈1.2475, RMSE≈1.1169, MAE≈0.3709) and ARIMA (MSE≈0.7548, RMSE≈0.8688, MAE≈0.3122). In combined situations, with 20% missing data and two segments of 10-element blocks missing, MSWARM maintained the lowest error level (MSE≈0.6305, RMSE≈0.7941, MAE≈0.3109) and significantly outperformed polynomial interpolation (MSE≈4.1454, RMSE≈2.0360, MAE≈0.7588) and ARIMA (MSE≈1.2878, RMSE≈1.1348, MAE≈0.4388).

The MIOAM method, introduced for outlier detection and correction, considers the dataset's basic structure, trends, and local variations. In extensive testing with various anomaly types, MIOAM maintained low Mean Absolute Percentage Error (MAPE around 4%) and RMSPE around 5.5%, successfully distinguishing real outliers from natural data variations. Although the F1 score (0.14 in certain configurations) is not very high, it was obtained under dynamic conditions where the method had to simultaneously maintain stability and keep errors low. Testing shows that MIOAM handles different types of missing data and unusual values effectively, helping to minimise errors in time series analysis.

When used together, MSWARM and MIOAM make a powerful combination that significantly enhances data quality across real-world situations. What makes these methods particularly valuable is how MSWARM maintains reliable accuracy even when nearly half the data is missing, while MIOAM efficiently processes anomalies – making them suitable for a wide range of practical applications.

The practical verification of these methods, including both theoretical basis and experimental validation, allows the author to conclude that the thesis is confirmed:

It is possible to develop methods that incorporate various approaches to data quality improvement using multi-level data processing.

Justification

The developed MSWARM and MIOAM methods not only solve specific data quality problems but also demonstrate the effectiveness of multi-level data processing principles. The development and testing of these methods reveal several significant aspects that confirm the thesis.

Firstly, MSWARM method's ability to outperform traditional approaches (ARIMA, polynomial interpolation) in all tested scenarios stems from its multi-level architecture. The method combines local context with global trends, as shown in equations ((3.9.) and (3.11.). This becomes particularly evident in combined scenario results. When 20% of data was missing and values disappeared in two segments of 10-element blocks, MSWARM showed MSE≈0.6305, performing almost seven times better than polynomial interpolation (MSE≈4.1454) and twice better than ARIMA (MSE≈1.2878).

Secondly, MIOAM method achieves low Mean Absolute Percentage Error (MAPE around 4%) and RMSPE (around 5.5%) through its integration of different analysis scales. The method analyses short-term fluctuations, medium-term trends, and long-term structures simultaneously using adaptive analysis windows. This approach enables precise identification of true outliers while preserving natural data variation.

The third and perhaps most compelling confirmation of the thesis lies in the synergistic interaction between both methods. Sequential application of MSWARM and MIOAM not only compensates for their individual limitations but also enhances overall data quality improvement. Experimental results demonstrate that after applying MSWARM and subsequent MIOAM processing, the dataset's statistical properties (mean value, standard deviation, skewness) remain unchanged with deviation less than 0.02%.

The multi-level approach maintains its effectiveness even in complex scenarios. MSWARM shows stable results even with three 10-element blocks of missing values (MSE≈0.5075), significantly outperforming alternative methods. Similarly, MIOAM adapts to various types of outliers, maintaining low error levels (RMSPE around 5.5%) even in dynamic conditions.

These results confirm both the possibility and practical advantages of multi-level data processing. The developed methods demonstrate that combining different analysis approaches and levels can significantly improve data quality while preserving statistical integrity and usability.

The method implementation source code is available in the GitHub repository: https://github.com/nikolajsbumanis/thesis-methods.

CONCLUSIONS

The author formulates and presents the main conclusions:

1. Analysis of sensor-generated data usage in IoT systems – precision beekeeping, smart transport systems, and surveillance systems – reveals that data quality and processing method selection directly depends on the data acquisition level. Experiments in each studied field confirm that the acquisition level determines both data type (raw, processed, or associated) and limits applicable data fusion methods.
2. Literature analysis in the data quality field identifies three significant tasks in IoT systems: noise reduction, missing value replacement, and outlier detection. The complexity of outlier detection and adjustment is evidenced by the MIOAM method's need to combine multiple approaches: winsorisation, rolling window analysis, and z-score evaluation. The task's significance is further confirmed by the necessity to preserve the dataset's statistical properties during outlier adjustment.
3. The need for data quality improvement methods in precision poultry farming was identified after low prediction accuracy was observed when using machine learning models for egg yield forecasting. By analysing

the results with the developed data layering method, data quality issues (periods of instability, incomplete data) were revealed, which significantly affected model performance. Thus, it can be concluded that the need for data quality improvement methods was identified as a critical factor in ensuring prediction accuracy.

4. The data layering method, which was used to analyse egg production forecasting data, allowed the identification of significant periods of data quality problems (for example, weeks 40-50), which directly affected the performance of machine learning models. This method allowed not only to visualize data problems but also to identify their impact periods, thus providing a better understanding of the factors that affected prediction accuracy.
5. The developed MSWARM method for missing value replacement demonstrates significantly better results than traditional approaches. Compared to 2nd-order polynomial interpolation and modified ARIMA model, MSWARM achieves lower Mean Square Error (MSE≈0.51 versus MSE≈1.25 and MSE≈0.75). The method's advantage lies in its automatic adaptation to data characteristics, utilising both local and global context whilst preserving dataset statistical properties.
6. MSWARM method's stability and effectiveness are confirmed in both six basic scenarios and extensive stability testing with 1,000 repetitions per scenario. The method maintains high accuracy even in complex cases. When processing two consecutive 10-element blocks of missing values together with 20% scattered missing values, MSWARM shows significantly lower error (MSE≈0.63) than compared methods (MSE≈1.29 and MSE≈4.15). Stability tests with 1,000 repetitions confirm result reproducibility, showing low variation (standard deviation <0.1) across all scenarios.
7. MSWARM method's main advantage is its adaptive nature. It automatically adjusts neighbour count and weight coefficients according to data characteristics, without requiring extensive training datasets or manual parameter configuration. The method's effectiveness stems from combining local and global context – local context for short-term trend determination, global context for overall data structure preservation.
8. Comprehensive testing (990 tests) confirms MSWARM method's stability up to 52.3% missing data volume, showing the lowest performance degradation after this threshold (MSE increase 36.7%, RMSE increase 18.6%) compared to other methods. The method proves particularly effective with small to medium sequence sizes (10-30 elements), providing stable results even with high proportions of missing data.
9. The developed MIOAM method for outlier detection and adjustment combines multiple approaches: winsorisation, rolling window analysis, and z-score evaluation. A significant advantage lies in its ability to adapt

to different datasets, automatically determining optimal thresholds for each data segment whilst preserving dataset statistical properties.

10. MIOAM method's integrated approach combines trend analysis, local variations, multi-window statistics, and confidence assessment in a unified system, ensuring both outlier identification and correction using adaptive thresholds and local data structures.

The author also identifies the main development perspectives:

1. Improving method implementation by creating a user interface for practical application and automating testing procedures.
2. Expanding practical applications by validating methods in new IoT system domains, testing their performance with datasets of varying volume and character, and conducting comparative analysis with the latest industry methods.

LITERATŪRAS SARAKSTS
BIBLIOGRAPHY

1. Abdelgawad, A., & Bayoumi, M. (2012). Resource-Aware data fusion algorithms for wireless sensor networks (Vol. 118). Springer Science \& Business Media.
2. Aboubakar, M., Kellil, M., & Roux, P. (2022). A review of IoT network management: Current status and perspectives. Journal of King Saud University – Computer and Information Sciences, 34(7), 4163–4176.
3. Abu Bakr, M., & Lee, S. (2017). Distributed multisensor data fusion under unknown correlation and data inconsistency. Sensors, 17(11), 2472.
4. Adamová, V., & Boroš, M. (2021). Effective Placement of Video Surveillance System Using 3D Scanning Technology for Traffic Safety. Transportation Research Procedia, 55, 1665–1672.
5. Adelantado, F., Vilajosana, X., Tuset-Peiro, P., Martinez, B., Melia-Segui, J., & Watteyne, T. (2017). Understanding the Limits of LoRaWAN. IEEE Communications Magazine, 55(9), 34–40.
6. Adhikari, D., Jiang, W., Zhan, J., He, Z., Rawat, D. B., Aickelin, U., & Khorshidi, H. A. (2022). A Comprehensive Survey on Imputation of Missing Data in Internet of Things. ACM Computing Surveys, 55(7), 1–38.
7. Alam, F., Mehmood, R., Katib, I., & Albeshri, A. (2016). Analysis of Eight Data Mining Algorithms for Smarter Internet of Things (IoT). Procedia Computer Science, 58, 437–442.
8. Alam, F., Mehmood, R., Katib, I., Albogami, N. N., & Albeshri, A. (2017). Data Fusion and IoT for Smart Ubiquitous Environments: A Survey. IEEE Access, 5, 9533–9554.
9. Anawar, M. R., Wang, S., Azam Zia, M., Jadoon, A. K., Akram, U., & Raza, S. (2018). Fog Computing: An Overview of Big IoT Data Analytics. Wireless Communications and Mobile Computing, 2018.
10. Astill, J., Dara, R. A., Fraser, E. D. G., Roberts, B., & Sharif, S. (2020). Smart poultry management: Smart sensors, big data, and the internet of things. In Computers and Electronics in Agriculture (Vol. 170, p. 105291). Elsevier B.V.
11. Atluri, G., Karpatne, A., & Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. ACM Computing Surveys, 51(4), 1–37.
12. Bakr, M. A., & Lee, S. (2017). Distributed multisensor data fusion under unknown correlation and data inconsistency. Sensors (Switzerland), 17(11), 2472.
13. Barbedo, J. G. A. (2022). Data Fusion in Agriculture: Resolving Ambiguities and Closing Data Gaps. Sensors, 22(6), 2285.
14. Becerra, M. A., Tobón, C., Castro-Ospina, A. E., & Peluffo-Ordóñez, D. H. (2021). Information quality assessment for data fusion systems. Data, 6(6), 60.
15. Beddar-Wiesing, S., & Bieshaar, M. (2020). Multi-Sensor Data and Knowledge Fusion -- A Proposal for a Terminology Definition. http://arxiv.org/abs/2001.04171
16. Beekeeping Calendar. (n.d.). Retrieved June 18, 2024, from https://rockymountainbeesupply.com/pages/beekeepers-calendar.
17. Bellinger, G., Castro, D., & Mills, A. (2004). Data, information, knowledge, and wisdom.
18. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(2), 281–305.
19. Besada-Portas, E., Lopez-Orozco, J. A., Besada, J., & De La Cruz, J. M. (2011). Multisensor fusion for linear control systems with asynchronous, Out-Of-Sequence and erroneous data. Automatica, 47(7), 1399–1408.
20. Biancolillo, A., Bucci, R., Magrì, A. L., Magrì, A. D., & Marini, F. (2014). Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication. Analytica Chimica Acta, 820, 23–31.
21. Blasch, E. P., & Plano, S. (2002). JDL level 5 fusion model: user refinement issues and applications in group tracking. In I. Kadar (Ed.), Signal Processing, Sensor Fusion, and Target Recognition XI (Vol. 4729, pp. 270–279).
22. Bokade, R., Navato, A., Ouyang, R., Jin, X., Chou, C. A., Ostadabbas, S., & Mueller, A. V. (2021). A cross-disciplinary comparison of multimodal data fusion approaches and

applications: Accelerating learning through trans-disciplinary information sharing. Expert Systems with Applications, 165, 113885.
23. Bumanis, N. (2020). Data fusion challenges in precision beekeeping: A review. Research for Rural Development, 35, 252–259.
24. Bumanis, N. (2024). Overcoming Data Limitations in Precision Poultry Farming: Processing and Data Fusion Challenges. Procedia Computer Science, 232, 2302–2309.
25. Bumanis, N., Arhipova, I., Paura, L., Vitols, G., & Jankovska, L. (2022). Data Conceptual Model for Smart Poultry Farm Management System. Procedia Computer Science, 200, 517–526.
26. Bumanis, N., Komasilova, O., Komasilovs, V., Kviesis, A., & Zacepins, A. (2020). Application of Data Layering in Precision Beekeeping: The Concept. 14th IEEE International Conference on Application of Information and Communication Technologies, AICT 2020 – Proceedings, 1–6.
27. Bumanis, N., Kviesis, A., Paura, L., Arhipova, I., & Adjutovs, M. (2023). Hen Egg Production Forecasting: Capabilities of Machine Learning Models in Scenarios with Limited Data Sets. Applied Sciences, 13(13).
28. Bumanis, N., Vitols, G., & Meirane, I. (2022). Data Fusion of Video and Lidar Traffic Surveillance Data: Practical Assesment of Implemented Solution in Jelgava City. Engineering for Rural Development, 21, 478–488.
29. Bumanis, N., Vitols, G., Arhipova, I., & Solmanis, E. (2021). Multi-object Tracking for Urban and Multilane Traffic: Building Blocks for Real-World Application. ICEIS (1), 729–736.
30. Castanedo, F. (2013). A review of data fusion techniques. The Scientific World Journal, 2013.
31. Chen, N., & Chen, Y. (2018). Smart City Surveillance at the Network Edge in the Era of IoT: Opportunities and Challenges (pp. 153–176).
32. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu, 785–794.
33. Chollet, F. (2015). Keras.[online] Available at: https://github. com/fchollet/keras. Accessed, 12(01), 2021.
34. Consoli, S., Presutti, V., Recupero, D. R., Cataldi, G., Mongiovì, M., & Patatu, W. (2015). An urban fault reporting and management platform for smart cities. WWW 2015 Companion – Proceedings of the 24th International Conference on World Wide Web, 535–540.
35. Dasarathy, B. V. (1997). Sensor fusion potential exploitation-innovative architectures and illustrative applications. Proceedings of the IEEE, 85(1), 24–38.
36. Debauche, O., Moulat, M. El, Mahmoudi, S., Boukraa, S., Manneback, P., & Lebeau, F. (2018). Web Monitoring of Bee Health for Researchers and Beekeepers Based on the Internet of Things. Procedia Computer Science, 130, 991–998.
37. Deibe, D., Amor, M., & Doallo, R. (2019). Big data storage technologies: A case study for web-based LiDAR visualization. Proceedings – 2018 IEEE International Conference on Big Data, Big Data 2018, 3831–3840.
38. Di Natale, C., Zude-Sasse, M., Macagnano, A., Paolesse, R., Herold, B., & D'Amico, A. (2002). Outer product analysis of electronic nose and visible spectra: Application to the measurement of peach fruit characteristics. Analytica Chimica Acta, 459(1), 107–117.
39. Dinculeană, D., & Cheng, X. (2019). Vulnerabilities and limitations of MQTT protocol used between IoT devices. Applied Sciences (Switzerland), 9(5), 848.
40. El Faouzi, N.-E., Leung, H., & Kurian, A. (2011). Data fusion in intelligent transportation systems: Progress and challenges--A survey. Information Fusion, 12(1), 4–10.
41. El-Mawla, N. A., & Badawy, M. (2023). Eco-Friendly IoT Solutions for Smart Cities Development: An Overview. 1st International Conference in Advanced Innovation on Smart City, ICAISC 2023 – Proceedings, 1–6.
42. Emam, A. (2021). Evaluation of Four Nonlinear Models Describing Egg Production Curve of Fayoumi Layers. Egyptian Poultry Science Journal, 41(1), 147–159.
43. Eremia, M., Toma, L., & Sanduleac, M. (2017). The Smart City Concept in the 21st Century. Procedia Engineering, 181, 12–19.

44. Faouzi, N. E. El, & Klein, L. A. (2016). Data Fusion for ITS: Techniques and Research Needs. Transportation Research Procedia, 15, 495–512.
45. Faouzi, N. E. El, & Klein, L. A. (2017). Data fusion in intelligent traffic and transportation engineering: Recent advances and challenges. Multisensor Data Fusion: From Algorithms and Architectural Design to Applications, 563–594.
46. Futūristiski bišu stropi viedajai metropolei (HIVEOPOLIS) (HOR5). (2019). https://www.lbtu.lv/lv/projekti/apstiprinatie-projekti/2019/futuristiski-bisu-stropi-viedajai-metropolei-hiveopolis-hor5
47. Gaddam, A., Wilkin, T., Angelova, M., & Gaddam, J. (2020). Detecting sensor faults, anomalies and outliers in the internet of things: A survey on the challenges and solutions. Electronics (Switzerland), 9(3), 511.
48. Gomathi, R. M., Krishna, G. H. S., Brumancia, E., & Dhas, Y. M. (2018). A Survey on IoT Technologies, Evolution and Architecture. 2nd International Conference on Computer, Communication, and Signal Processing: Special Focus on Technology and Innovation for Smart Environment, ICCCSP 2018, 1–5.
49. Govinda, K., & Saravanaguru, R. A. K. (2016). Review on IOT technologies. International Journal of Applied Engineering Research, 11(4), 2848–2853.
50. Grime, S., & Durrant-Whyte, H. F. (1994). Data fusion in decentralized sensor networks. Control Engineering Practice, 2(5), 849–863.
51. Guerrero-Ibáñez, J., Zeadally, S., & Contreras-Castillo, J. (2018). Sensor technologies for intelligent transportation systems. Sensors (Switzerland), 18(4), 1212.
52. Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. Proceedings of the IEEE, 85(1), 6–23.
53. Hall, D. L., & Llinas, J. (2016). An introduction to multi-sensor data fusion. Sensors, Nanoscience, Biomedical Engineering, and Instruments, 6(1), 537–540.
54. Han, G., Tu, J., Liu, L., Martinez-Garcia, M., & Choi, C. (2022). An Intelligent Signal Processing Data Denoising Method for Control Systems Protection in the Industrial Internet of Things. IEEE Transactions on Industrial Informatics, 18(4), 2684–2692.
55. Han, Y., & Hu, D. (2020). Multispectral fusion approach for traffic target detection in bad weather. Algorithms, 13(11), 1–13.
56. Hannun, A., Guo, C., & van der Maaten, L. (2021). Measuring Data Leakage in Machine-Learning Models with Fisher Information. 37th Conference on Uncertainty in Artificial Intelligence, UAI 2021, 760–770.
57. HENCO2: Mākoņdatu vidē balstīta IT platforma putnkopības produktivitātes uzlabošanai un siltumnīcefekta gāzu emisiju samazināšanai – ER32. (2020). https://www.lbtu.lv/lv/projekti/apstiprinatie-projekti/2020/henco2-makondatu-vide-balstita-it-platforma-putnkopibas
58. Hennessy, G., Harris, C., Eaton, C., Wright, P., Jackson, E., Goulson, D., & Ratnieks, F. F. L. W. (2020). Gone with the wind: effects of wind on honey bee visit rate and foraging behaviour. Animal Behaviour, 161, 23–31.
59. Hong, X., Nugent, C., Mulvenna, M., McClean, S., Scotney, B., & Devlin, S. (2009). Evidential fusion of sensor data for activity recognition in smart homes. Pervasive and Mobile Computing, 5(3), 236–252.
60. Hu, Z., & Ma, W. (2020). Self-calibration of Traffic Surveillance Camera Systems for Traffic Density Estimation on Urban Roads. Limos.Engin.Umich.Edu, 1(412), 1–8.
61. Huang, C., Liu, L., Yuen, C., & Sun, S. (2019). Iterative Channel Estimation Using LSE and Sparse Message Passing for MmWave MIMO Systems. IEEE Transactions on Signal Processing, 67(1), 245–249.
62. Huang, L., Zhao, J., Chen, Q., & Zhang, Y. (2014). Nondestructive measurement of total volatile basic nitrogen (TVB-N) in pork meat by integrating near infrared spectroscopy, computer vision and electronic nose techniques. Food Chemistry, 145, 228–236.
63. Huet, J. C., Bougueroua, L., Kriouile, Y., Wegrzyn-Wolska, K., & Ancourt, C. (2022). Digital Transformation of Beekeeping through the Use of a Decision Making Architecture. Applied Sciences (Switzerland), 12(21), 11179.

64. Human, H., Brodschneider, R., Dietemann, V., Dively, G., Ellis, J. D., Forsgren, E., Fries, I., Hatjina, F., Hu, F. L., Jaffé, R., Jensen, A. B., Köhler, A., Magyar, J. P., Özkýrým, A., Pirk, C. W. W., Rose, R., Strauss, U., Tanner, G., Tarpy, D. R., … Zheng, H. Q. (2013). Miscellaneous standard methods for Apis mellifera research. Journal of Apicultural Research, 52(4).
65. Individuālie mobilitātes budžeti kā sociālais un ētiskais pamats oglekļa emisiju samazināšanai (MyFairShare) (ZV91). (2021). https://www.lbtu.lv/lv/projekti/apstiprinatie-projekti/2021/individualie-mobilitates-budzeti-ka-socialais-un-etiskais
66. Yang, L., Zhang, X. & Chen, J. (2024). Winsorization greatly reduces false positives by popular differential expression methods when analyzing human population samples. Genome Biol, 25, 282.
67. Yang, N., Wu, C., & McMillan, I. (1989). New Mathematical Model of Poultry Egg Production. Poultry Science, 68(4), 476–481.
68. Yaro, A. S., Maly, F., Prazak, P., & Malý, K. (2024). Outlier detection performance of a modified Z-score method in time-series RSS observation with hybrid scale estimators. IEEE Access, 12, 12785-12796.
69. Yin, H., Liu, C., Gao, Y., Fan, W., Xiao, B., Cao, L., Hassan, S. G., & Liu, S. (2021). A Novel Method to Predict Laying Rate Based on Multiple Environment Variables. IEEE Access, 9, 115488–115496.
70. Ying, X. (2019). An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series, 1168(2), 22022.
71. Jcgm, J. C. F. G. I. M. (2008). Evaluation of measurement data — Guide to the expression of uncertainty in measurement. International Organization for Standardization Geneva ISBN, 50(September), 134. http://www.bipm.org/en/publications/guides/gum.html
72. Jifa, G., & Lingling, Z. (2014). Data, DIKW, big data and data science. Procedia Computer Science, 31, 814–821.
73. Karkouch, A., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Data quality in internet of things: A state-of-the-art survey. Journal of Network and Computer Applications, 73, 57–81.
74. Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. Information Fusion, 14(1), 28–44.
75. Khulal, U., Zhao, J., Hu, W., & Chen, Q. (2017). Intelligent evaluation of total volatile basic nitrogen (TVB-N) content in chicken meat by an improved multiple level data fusion model. Sensors and Actuators, B: Chemical, 238, 337–345.
76. Kim, D. Y., & Jeon, M. (2014). Data fusion of radar and image measurements for multi-object tracking via Kalman filtering. Information Sciences, 278, 641–652.
77. Kim, S., Pérez-Castillo, R., Caballero, I., & Lee, D. (2022). Organizational process maturity model for IoT data quality management. Journal of Industrial Information Integration, 26, 100256.
78. Kratkiewicz, K. J., White Bear, J., & Miller, B. A. (2019). Survey of Data Fusion in IoT.
79. Kreibich, O., Neuzil, J., & Smid, R. (2014). Quality-based multiple-sensor fusion in an industrial wireless sensor network for MCM. IEEE Transactions on Industrial Electronics, 61(9), 4903–4911.
80. Krishnamurthi, R., Kumar, A., Gopinathan, D., Nayyar, A., & Qureshi, B. (2020). An overview of iot sensor data processing, fusion, and analysis techniques. Sensors (Switzerland), 20(21), 1–23.
81. Kumar, M., Garg, D. P., & Zachery, R. A. (2006). A eneralized approach for inconsistency detection in data fusion from multiple sensors. Proceedings of the American Control Conference, 2006, 2078–2083.
82. Kumar, V., Subramanian, S. C., & Rajamani, R. (2022). A novel algorithm to track closely spaced road vehicles using a low density flash lidar. Signal Processing, 191.
83. Kviesis, A., & Zacepins, A. (2015). System architectures for real-time bee colony temperature monitoring. Procedia Computer Science, 43(C), 86–94.
84. Kwon, K.-H., Cho, C., & Lee, H.-B. (2019). Smart Beehive using Data Fused Preprocessing and Artificial Neural networks. Journal of Digital Contents Society, 20(12), 2321–2327.

85. Lakshmanarao, A., & Shashi, M. (2020). A survey on machine learning for cyber security. International Journal of Scientific and Technology Research, 9(1), 499–502.
86. Lashari, M. H., Memon, A. A., Shah, S. A. A., Nenwani, K., & Shafqat, F. (2019). IoT Based poultry environment monitoring system. Proceedings – 2018 IEEE International Conference on Internet of Things and Intelligence System, IOTAIS 2018, 1–5.
87. Lau, B. P. L., Marakkalage, S. H., Zhou, Y., Hassan, N. U., Yuen, C., Zhang, M., & Tan, U. X. (2019). A survey of data fusion in smart city applications. Information Fusion, 52, 357–374.
88. Lin, S. S., Yang, J. Y., Syu, H. S., Lin, C. H., & Pai, T. W. (2019). Automatic generation of puzzle tile maps for spatial-temporal data visualization. Computers and Graphics (Pergamon), 82, 1–12.
89. Liu, H., Shen, X., Wang, Z., Meng, F., Wang, J., Pramod, & Varshney. (2021). Randomized Multiple Model Multiple Hypothesis Tracking. ArXiv Preprint ArXiv:2105.01379. http://arxiv.org/abs/2105.01379
90. Liu, Yin, & Zhang, Y. (2022). A Weighted Evidence Combination Method for Multisensor Data Fusion. Journal of Internet Technology, 23(3), 553–560.
91. Liu, Yuehua, Dillon, T., Yu, W., Rahayu, W., & Mostafa, F. (2020). Missing Value Imputation for Industrial IoT Sensor Data with Large Gaps. IEEE Internet of Things Journal, 7(8), 6855–6867.
92. Liu, J., Li, T., Xie, P., Du, S., Teng, F., & Yang, X. (2020). Urban big data fusion based on deep learning: An overview. In Information Fusion (Vol. 53, pp. 123–133).
93. Liu, T., Du, S., Liang, C., Zhang, B., & Feng, R. (2021). A Novel Multi-Sensor Fusion Based Object Detection and Recognition Algorithm for Intelligent Assisted Driving. IEEE Access, 9, 81564–81574.
94. Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E., & White, F. (2004). Revisiting the JDL data fusion model II. Proceedings of the Seventh International Conference on Information Fusion, FUSION 2004, 2(January 2013), 1218–1230.
95. Manogaran, G., Balasubramanian, V., Rawal, B. S., Saravanan, V., Montenegro-Marin, C. E., Ramachandran, V., & Kumar, P. M. (2021). Multi-Variate Data Fusion Technique for Reducing Sensor Errors in Intelligent Transportation Systems. IEEE Sensors Journal, 21(14), 15564–15573.
96. Multiobjektu detektēšana un izsekošana transportlīdzekļu satiksmes novērošanai: 3D-LiDAR un kameras datu apvienošana. (2020). https://www.itkc.lv/services
97. Muneer, A., Fati, S. M., & Fuddah, S. (2020). Smart health monitoring system using IoT based smart fitness mirror. Telkomnika (Telecommunication Computing Electronics and Control), 18(1), 317–331.
98. Muñoz, L. A., Martínez, J. V. B., Pérez, F. M., & Fonseca, I. L. (2024). Anomaly detection system for data quality assurance in IoT infrastructures based on machine learning. Internet of Things, 25, 101095.
99. Murakami, E., Saraiva, A. M., Junior, L. C. M. R., Cugnasca, C. E., Hirakawa, A. R., & Correa, P. L. P. (2007). An infrastructure for the development of distributed service-oriented information systems for precision agriculture. Computers and Electronics in Agriculture, 58(1), 37–48.
100. Nasution, M. K. M., Sitompul, O. S., Elveny, M., & Syah, R. (2021). Data science: A Review towards the Big Data Problems. Journal of Physics: Conference Series, 1898(1), 12006.
101. Neumann, T., Ebendt, R., & Kuhns, G. (2016). From finance to ITS: traffic data fusion based on Markowitz'portfolio theory. Journal of Advanced Transportation, 50(2), 145–164.
102. Nižetić, S., Šolić, P., López-de-Ipiña González-de-Artaza, D., & Patrono, L. (2020). Internet of Things (IoT): Opportunities, issues and challenges towards a smart and sustainable future. Journal of Cleaner Production, 274.
103. Nyalala, I., Okinda, C., Kunjie, C., Korohou, T., Nyalala, L., & Chao, Q. (2021). Weight and Volume Estimation of Poultry and Products based on Computer Vision Systems: A Review. Poultry Science, 101072. https://linkinghub.elsevier.com/retrieve/pii/S0032579121001061

104. Noh, S. (2020). Intelligent data fusion and multi-agent coordination for target allocation. Electronics (Switzerland), 9(10), 1–13.
105. Osterrieder, P., Budde, L., & Friedli, T. (2020). The smart factory as a key construct of industry 4.0: A systematic literature review. International Journal of Production Economics, 221, 107476.
106. Parish, C. M., & Edmondson, P. D. (2019). Data visualization heuristics for the physical sciences. Materials and Design, 179, 107868.
107. Patel, K. K., Patel, S. M., & Scholar, P. (2016). Internet of things-IOT: definition, characteristics, architecture, enabling technologies, application \& future challenges. International Journal of Engineering Science and Computing, 6(5).
108. Pedregosa, F., Weiss, R., Brucher, M., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html%5Cnhttp://arxiv.org/abs/1201.0490
109. Pires, I. M., Garcia, N. M., Pombo, N., & Flórez-Revuelta, F. (2016). From data acquisition to data fusion: A comprehensive review and a roadmap for the identification of activities of daily living using mobile devices. Sensors (Switzerland), 16(2), 184.
110. Porru, S., Misso, F. E., Pani, F. E., & Repetto, C. (2020). Smart mobility and public transport: Opportunities and challenges in rural and urban areas. Journal of Traffic and Transportation Engineering (English Edition), 7(1), 88–97.
111. Qin, X., Luo, Y., Tang, N., & Li, G. (2020). Making data visualization more efficient and effective: a survey. VLDB Journal, 29(1), 93–117.
112. Rafael Braga, A., G. Gomes, D., Rogers, R., E. Hassler, E., M. Freitas, B., & A. Cazier, J. (2020). A method for mining combined data from in-hive sensors, weather and apiary inspections to forecast the health status of honey bee colonies. Computers and Electronics in Agriculture, 169, 105161.
113. Rashid, M. M., Beecham, S., & Chowdhury, R. K. (2015). Assessment of trends in point rainfall using Continuous Wavelet Transforms. Advances in Water Resources, 82, 1–15.
114. Sanyal, S., & Zhang, P. (2018). Improving quality of data: IoT data aggregation using device to device communications. IEEE Access, 6, 67830–87840.
115. Schneider, S., Santhanavanich, T., Koukofikis, A., & Coors, V. (2020). Exploring Schemes for Visualizing Urban Wind Fields based on CFD Simulations by Employing OGC Standards. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 6(4/W2), 157–163.
116. Shi, H., Zhao, H., Liu, Y., Gao, W., & Dou, S. C. (2019). Systematic analysis of a military wearable device based on a multi-level fusion framework: Research directions. Sensors (Switzerland), 19(12), 2651.
117. Shirowzhan, S., Lim, S., Trinder, J., Li, H., & Sepasgozar, S. M. E. (2020). Data mining for recognition of spatial distribution patterns of building heights using airborne lidar data. Advanced Engineering Informatics, 43, 101033.
118. Singh, M., Kumar, R., Tandon, D., Sood, P., & Sharma, M. (2020). Artificial Intelligence and IoT based Monitoring of Poultry Health: A Review. 2020 IEEE International Conference on Communication, Networks and Satellite, Comnetsat 2020 – Proceedings, 50–54.
119. Sliwa, B., Piatkowski, N., & Wietfeld, C. (2020). LIMITS: Lightweight Machine Learning for IoT Systems with Resource Limitations. IEEE International Conference on Communications, 2020-June, 1–7.
120. Souza, A. M. C., & Amazonas, J. R. A. (2015). An outlier detect algorithm using big data processing and internet of things architecture. Procedia Computer Science, 52, 1010–1015.
121. Stalidzans, E., & Berzonis, A. (2013). Temperature changes above the upper hive body reveal the annual development periods of honey bee colonies. Computers and Electronics in Agriculture, 90, 1–6.

122. Steinberg, A. N., & Bowman, C. L. (2008). Revisions to the JDL Data Fusion Model. In Handbook of Multisensor Data Fusion: Theory and Practice: Second Edition (pp. 45–67). CRC press.
123. Targowski, A. (2005). From Data to Wisdom. Dialogue and Universalism, 15(5), 55–71.
124. Teh, H. Y., Kempa-Liehr, A. W., & Wang, K. I. K. (2020). Sensor data quality: a systematic review. Journal of Big Data, 7(1), 1–49.
125. Tree Flowering Calendar. (2020). https://wiki.sams-project.eu/index.php/Tree_flowering_calendar_Ethiopia
126. Vanus, J., Fiedorova, K., Kubicek, J., Gorjani, O. M., & Augustynek, M. (2020). Wavelet-based filtration procedure for denoising the predicted CO2 waveforms in smart home within the internet of things. Sensors (Switzerland), 20(3), 620.
127. Varshney, P. K. (1997). Multisensor data fusion. Electronics and Communication Engineering Journal, 9(6), 245–253.
128. Villa-Henriksen, A., Edwards, G. T. C., Pesonen, L. A., Green, O., & Sørensen, C. A. G. (2020). Internet of Things in arable farming: Implementation, applications, challenges and potential. Biosystems Engineering, 191, 60–84.
129. Wang, W., Chen, J., & Hong, T. (2018). Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings. Automation in Construction, 94, 233–243.
130. Waskom, M. (2021). Seaborn: Statistical Data Visualization. Journal of Open Source Software, 6(60), 3021.
131. Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., Savic, M. D., Garovic, V. D., & Milic, N. M. (2019). Reveal, Don't Conceal: Transforming Data Visualization to Improve Transparency. Circulation, 140(18), 1506–1518.
132. White, F. E., & Steinberg, A. N. (1998). Community Status Report and Proposed Revisions to the JDL Data Fusion Model.
133. Zacepins, A., & Stalidzans, E. (2013). Information processing for remote recognition of the state of bee colonies and apiaries in precision beekeeping (apiculture). Biosystems and Information Technology, 2(1), 6–10.
134. Zacepins, A., Brusbardis, V., Meitalovs, J., & Stalidzans, E. (2015). Challenges in the development of Precision Beekeeping. Biosystems Engineering, 130, 60–71.
135. Zacepins, A., Stalidzans, E., & Meitalovs, J. (2012). Application of information technologies in precision apiculture. Proceedings of the 13th International Conference on Precision Agriculture (ICPA 2012).
136. Zhang, M., Wang, X., Feng, H., Huang, Q., Xiao, X., & Zhang, X. (2021). Wearable Internet of Things enabled precision livestock farming in smart farms: A review of technical solutions for precise perception, biocompatibility, and sustainability monitoring. Journal of Cleaner Production, 312, 127712.
137. Zhang, Z., Yan, X., Zhang, L., Lai, X., & Lu, W. (2024). Fuzzy neuron modeling of incomplete data for missing value imputation. Information Sciences, 659, 120065.
138. Zheng, Y. (2015). Methodologies for Cross-Domain Data Fusion: An Overview. IEEE Transactions on Big Data, 1(1), 16–34.