Latvijas Biozinātņu un tehnoloģiju universitāte
Inženierzinātņu un informācijas tehnoloģiju fakultāte
Datoru sistēmu un datu zinātnes institūts

Mg. sc. ing. **Nikolajs Būmanis**

promocijas darbs

VAIRĀKU AVOTU DATU APVIENOŠANA UN ANALĪZE INTERAKTĪVAI APSTRĀDEI UN VIZUALIZĀCIJAI

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

zinātnes doktora grāda
zinātnes doktors (Ph. D.) inženierzinātnēs un tehnoloģijās
iegūšanai

Promocijas darba vadītāja
Prof. *Dr. sc. ing.* Irina Arhipova _____

Promocijas darba autors _____

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.

Zinātniskā 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.

ANOTĀCIJA

Nikolaja Būmaņa promocijas darbs "**Vairāku avotu datu apvienošana un analīze interaktīvai apstrādei un vizualizācijai**" tika izstrādāts Latvijas Biozinātņu un tehnoloģiju universitātē (LBTU), Inženierzinātņu un informācijas tehnoloģiju fakultātē, Datoru sistēmu un datu zinātnes institūtā, laika posmā no 2019. gada septembra līdz 2025. gada februārim.

Darba apjoms ir 140 lappuses, kas ietver 58 attēlus, 15 tabulas un 3 pielikumus. Darbā veiktas atsauces uz 318 literatūras avotiem.

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;
2. izpētīt pastāvošās datu apvienošanas pieejas, tai skaitā noteikt to klasifikācijas un darbības principus;
3. izstrādāt datu apvienošanas metodi vairāku avotu datu savstarpējās saistības vizualizācijai, balstoties uz svarīgāko periodu noteikšanu;
4. pārbaudīt izstrādāto metodi precīzās putnkopības problemātikas jomā;
5. izstrādāt datu kvalitātes (pēc pilnīguma un precizitātes kritērijiem) uzlabošanas metodi trūkstošo vērtību aizvietošanai un noviržu pielāgošanai;
6. veikt izstrādāto metožu novērtēšanu.

Promocijas darbs strukturēts 6 nodaļās.

Pirmajā nodaļā tiek iedziļināts lietu interneta (*IoT*) pasaulē, īpašu uzmanību pievēršot tam, kā sensori vāc datus un cik kvalitatīva ir šī informācija. Šeit tiek apskatīti arī veidi, kā uzlabot šo datu kvalitāti.

Otrajā nodaļā veikts pārskats par dažādām datu apvienošanas metodēm un to klasifikāciju, piemēram, sensoru datu apvienošanu un filtrēšanas algoritmiem. Ir noteikti darbības principi un piemērošanas jomas dažādām pieejām.

Trešajā nodaļā analizēta datu apvienošanas un datu kvalitātes jautājumu problemātika reālā vidē. Ir apskatītas trīs jomas – precīzā biškopība, precīzā putnkopība, kā arī viedās transporta un uzraudzības sistēmas.

Ceturtajā nodaļā ir aprakstīta izstrādātā jaunā konceptuālā metode vairāku avotu datu apvienošanai un vizualizēšanai, kas parāda saistības starp datiem svarīgākajos laikposmos.

Piektajā nodaļā aprakstīts precīzās putnkopības produktivitātes prognozēšanas uzdevums un izstrādātas metodes pārbaude šī uzdevuma rezultātiem.

Sestajā nodaļā aprakstītas divas izstrādātās datu kvalitātes uzlabošanas metodes: trūkstošo vērtību aizvietošanas metode un noviržu pielāgošanas metode. Novērtēta abu metožu efektivitāte.

ANNOTATION

Doctoral thesis "**Multi-source data fusion and analysis for interactive processing and visualization**" was developed by **Nikolajs Bumanis** at the Latvian University of Life Sciences and Technologies (LBTU), Faculty of Engineering and Information Technologies, Institute of Computer Systems and Data Science, in the period from September 2019 to February 2025.

The volume of the work is 140 pages, which includes 58 figures, 15 tables and 3 appendices. References to 318 literary sources are made in the work.

The **objective** of the doctoral thesis is to develop methodological solutions for data integration and quality improvement in order to enhance the efficiency and accuracy of data analysis, providing deeper and more well-founded insights in the field of research.

To achieve the goal, the following **tasks** were defined:

1. analyze data quality characteristics and improvement techniques for sensor-generated data according to their completeness and accuracy criteria;
2. analyze the existing approaches to combining data, including determining their classification and operating principles;
3. develop a method of combining data for visualizing the interrelationship of data from several sources, based on the determination of the most important periods;
4. test the developed method in the field of precision poultry farming;
5. develop a method of improving data quality (according to completeness and accuracy criteria) for replacing missing values and adjusting deviations;
6. evaluate the developed methods,

The thesis is structured in 6 chapters.

In the **first** chapter, the use of sensor-generated data within the framework of the Internet of Things is discussed, focusing on data completeness and accuracy metrics.

In the **second** chapter, an overview of various data fusion methods and their classification is provided, including sensor data fusion and filtering algorithms. The operational principles and application areas of different approaches were determined.

In the **third** chapter, the issues of data fusion and data quality in real-world environments are examined. Three areas were discussed – precision beekeeping, precision poultry farming, and smart transportation and surveillance systems.

In the **fourth** chapter, a developed concept is described – a new method for multi-source data fusion and visualization, which reveals the relationships between data during the most important periods.

In the **fifth** chapter, the productivity prediction in precision poultry farming is described, together with testing of the developed method for evaluating the results.

In the **sixth** chapter, two developed methods for improving data quality are described: a method for replacing missing values and a method for adjusting deviations. The effectiveness of both methods was evaluated.

SATURS

ATTĒLU SARAKSTS

TABULU SARAKSTS

DARBĀ LIETOTIE SIMBOLI UN SAĪSINĀJUMI

Termins vai termina saīsinājums	Skaidrojums
AMP	*Arbitrary Missing Pattern*. *Nejauši trūkstošo datu modelis.* Modelis, kurā dati trūkst bez noteikta secīguma vai modeļa, t. i., trūkstošo datu izkārtojums ir nejaušs.
Association rule mining	*Asociācijas likumu atklāšana.* Datu ieguves tehnika, kas atrod saistības starp datu punktiem lielās datu kopās, piemēram, preču grozu analīzē.
Bluetooth	*Bezvadu tehnoloģija*, kas nodrošina īsviļņu datu pārraidi starp ierīcēm nelielā attālumā. Izmanto ierīču savienošanai, piemēram, austiņu, tastatūru vai citu perifēriju.
Cloud Computing	*Mākoņdatošana.* Tehnoloģija, kas ļauj lietotājiem internetā piekļūt skaitļošanas resursiem un tos izmantot, piemēram, serverus, uzglabāšanas vietas, datubāzes un lietojumprogrammas.
DAI-DAO	*Data in – Data out. Dati iekšā – dati ārā*. Datu avoti tiek apvienoti bez iepriekšējas apstrādes.
DAI-DEO	*Data in – Decision out. Dati iekšā – lēmumi ārā*. Dati tiek tieši apstrādāti, lai iegūtu gala lēmumus.
DAI-FEO	*Data in – Feature out*. *Dati iekšā – pazīmes ārā*. Dati tiek apstrādāti, lai iegūtu pazīmes pirms apvienošanas.
Data layer	*Datu slānis datu apvienošanas kontekstā,* kur katra laika rinda tiek attēlota kā atsevišķs slānis, ir informācijas slānis, kas satur specifiskus datus vai mērījumus noteiktā laika intervālā. Katrs slānis atspoguļo konkrētus sensora datus vai laika rindas datus, ļaujot efektīvi apvienot un salīdzināt dažādu avotu informāciju.
DB	*Datubāze.* Savstarpēji saistītu informacionālu objektu tematisks kopums, kas ar speciālas pārvaldības sistēmas starpniecību izveidots un organizēts tā, lai nodrošinātu tajā ievadītās informācijas izguvi, veiktu tās atlasi un kārtošanu.
Decision	*Lēmums.* Iznākums, ko pieņem, pamatojoties uz datu analīzi, modeli vai algoritmu. Tas var būt automatizēts lēmums, piemēram, algoritma noteikts risinājums, vai cilvēka veidots lēmums, balstoties uz tehnoloģiju sniegto informāciju.
DEI-DAO	*Decision in – Data out*. *Lēmumi iekšā – dati ārā*. Lēmumu rezultāti tiek pārveidoti par datiem.
DEI-DEO	*Decision in – Decision out*. *Lēmumi iekšā – lēmumi ārā*. Apvieno vairākus atsevišķi veiktus lēmumus.
DEI-FEO	*Decision in – Feature out*. *Lēmumi iekšā – pazīmes ārā*. Lēmumi tiek pārveidoti par pazīmēm nākamajam apstrādes posmam.
DIKW	*Data, Information, Knowledge, Wisdom. Dati, informācija, zināšanas, gudrība.* Hierarhiskais modelis, kas parāda, kā dati tiek pārvērsti informācijā, informācija zināšanās un zināšanas gudrībā, uzlabojot lēmumu pieņemšanu.

Termins vai termina saīsinājums	Skaidrojums
DNN	*Deep Neural Network. Dziļais neironu tīkls*. Neironu tīkla struktūra ar vairākiem slāņiem, kas tiek izmantota sarežģītu datu modelēšanai un izpētei.
DST	*Dempster-Shafer theory. Dempstera-Šafera teorija*. Matemātiska metode, kas ļauj apvienot dažādus pierādījumus un secinājumus nenoteiktos apstākļos.
Feature	*Pazīme.* Pazīme ir noteikta datu īpašība vai atribūts, ko izmanto datu apstrādē vai analīzē. Mašīnmācīšanās un datu analīzes kontekstā pazīme var būt atsevišķa mainīgā vērtība vai īpašība, kas palīdz raksturot datus un veikt klasifikāciju vai prognozēšanu.
FEI-DAO	*Feature in – Data out. Pazīmes iekšā – dati ārā*. Pazīmes tiek atgrieztas sākotnējā datu formātā.
FEI-DEO	*Feature in – Decision out. Pazīmes iekšā – lēmumi ārā*. Pazīmes tiek apstrādātas, lai iegūtu gala lēmumus.
FEI-FEO	*Feature in – Feature out. Pazīmes iekšā – pazīmes ārā.* Apstrādātas pazīmes tiek tieši apvienotas.
FTP	*File Transfer Protocol. Standartizēts protokols*, kas tiek izmantots failu pārraidīšanai starp serveriem un klientiem internetā.
Fuzzy Clustering	*Faziklasterizācija*. Datu analīzes metode, kurā objekti tiek sadalīti vairākās grupās (klasteros) ar neskaidrām robežām. Tas nozīmē, ka viens objekts var piederēt vairākiem klasteriem ar dažādu piederības pakāpi, nevis stingri tikai vienam klasterim.
GAN	*Generative adversarial network. Ģeneratīvie sacensību tīkli.* Dziļās mācīšanās modelis, kurā divi neironu tīkli (ģenerators un diskriminators) savstarpēji sacenšas, lai ģenerētu reālistiskus datus, piemēram, attēlus vai video.
GDPR	*General Data Protection Regulation.* Eiropas Savienības regula, kas aizsargā personas datus un stiprina privātuma tiesības, nosakot stingras prasības datu apstrādei un glabāšanai.
IoT	*Internet of Things. Lietu internets.* Tehnoloģiju tīkls, kurā ierīces ir savienotas un spēj savstarpēji apmainīties ar datiem, uzlabojot automatizāciju un kontroli.
Isolation Forest	*Isolation Forest. Izolācijas meža algoritms.* Neuzraudzītas mācīšanās metode, ko izmanto galvenokārt anomāliju (vai ārpusnieku) atklāšanai.
ISTM	*Inverse Space Transformation Method. Inkrementālais telpas-laika modelis.* Modelis, kas iteratīvi atjaunina un prognozē datus, ņemot vērā gan telpiskās, gan laika dimensijas, piemēram, reģionālās un laika secības.
JDL	*Joint Directors of Laboratories. JDL datu apvienošanas modeli*, kas klasificē datu apstrādes līmeņus dažādiem sensoru datu avotiem militārajās un citās pielietojuma jomās.

Termins vai termina saīsinājums	Skaidrojums
K-NN	*K-Nearest Neighbors algorithm. K-tuvāko kaimiņu algoritms.* Mašīnmācīšanās algoritms, kas klasificē vai prognozē vērtības, pamatojoties uz tuvāko kaimiņu (punktu) datiem noteiktā daudzumā.
LiDAR	*Light detection and ranging.* Gaismas avota atklāšana un attāluma noteikšana līdz tam.
LOF	*Local Outlier Factor. Lokālā izlecēja koeficients*. Metode, kas nosaka datu punkta novirzi, salīdzinot tā novietojumu ar tuvāko kaimiņu punktu novietojumu.
LSTM	*Long short-term memory. Ilgtermiņa īstermiņa atmiņa*. Dziļās mācīšanās neironu tīklu arhitektūra, kas spēj saglabāt un apstrādāt secīgus datus, kuri ir būtiski laika rindu prognozēšanai un dabiskās valodas apstrādei.
LTE/4G	*Long-term evolution.* Ceturtās paaudzes bezvadu sakaru tehnoloģija, kas nodrošina ātrāku un stabilāku datu pārraidi mobilajos tīklos.
MAR	*Missing at Random. Nejauši trūkstošie dati.* Datu trūkuma mehānisms, kurā iztrūkstošo vērtību trūkums ir saistīts ar novērotajām vērtībām, bet ne ar neredzamām vērtībām.
MCAR	*Missing Completely at Random. Pilnīgi nejauši trūkstošie dati.* Datu trūkuma mehānisms, kurā datu iztrūkums nav saistīts ar novērojamo vai neredzamo mainīgo vērtībām.
MHT	*Multiple Hypothesis Tracking. Vairāku hipotēžu izsekošana.* Objektu izsekošanas metode, kas uztur vairākas iespējamās hipotēzes par objektu trajektorijām, līdz tiek iegūti pietiekami daudz datu.
MINA	Multi-skalas integrēto noviržu analīzes metode. Promocijas darbā izstrādātā metode noviržu noteikšanai un pielāgošanai.
MMP	*Monotone Missing Pattern. Monotonu trūkstošo datu modelis.* Modelis, kurā trūkstošie dati ir sakārtoti secīgā veidā, un vēlākās vērtības ir atkarīgas no iepriekšējo vērtību pieejamības.
MSVSM	Modificētā standarta vidējā svērtā metode. Promocijas darbā izstrādātā metode trūkstošo vērtību aizvietošanai.
NFC	*Near Frequency Communication. Tuvā lauka komunikācija.* Bezvadu sakaru metode, kas nodrošina datu pārraidi un savienojumu izveidi starp ierīcēm nelielā attālumā (parasti mazāk par 10 cm), bieži izmantota maksājumiem un ierīču autentifikācijai.
NMAR	*Not Missing at Random. Nav nejauši trūkstošie dati.* Datu trūkuma mehānisms, kurā iztrūkstošie dati ir atkarīgi no neredzamām vērtībām, nevis novērotajām vērtībām.
PCA	*Principal Component Analysis. Galvenā komponenta analīze.* Statistikas metode, kas tiek izmantota lielu datu kopu dimensiju samazināšanai, pārvēršot mainīgos jaunos komponentos, saglabājot pēc iespējas vairāk informācijas no sākotnējiem datiem.

Termins vai termina saīsinājums	Skaidrojums
PCHIP	*Piecewise Cubic Hermite Interpolation Polynomial*. Ermita interpolācijas polinoms.
PDA	*Probabilistic Data Association. Varbūtiskā datu asociācija*. Izsekošanas metode, kas katram novērojumam piešķir varbūtību, lai noteiktu, kuram objektam šis novērojums pieder.
PDAF	*Probabilistic Data Association Filter. Varbūtiskās datu asociācijas filtrs.* Objektu izsekošanas metode, kas balstās uz varbūtībām, lai pareizi piesaistītu neskaidrus vai trokšņainus mērījumus konkrētiem objektiem, novēršot kļūdainu mērījumu piešķiršanu vairākiem objektiem.
PLS-DA	*Partial Least Squares Discriminant Analysis. Daļējā mazāko kvadrātu diskriminantas analīze*. Statistikas metode, kas savieno pazīmju atlasi ar diskriminantu analīzi, lai klasificētu mainīgos lielumus ar korelētām datu kopām.
PMF	*Probabilistic Matrix Factorization. Varbūtības matricas faktorizācija*. Mašīnmācīšanās metode, kas izmanto latentās mainīgās, lai modelētu lietotāju un materiālu mijiedarbību, piemēram, rekomendāciju sistēmās.
Point cloud	*Punktu mākonis*. Datu kopums, kas sastāv no daudziem telpiskiem punktiem, kuri attēlo objektus vai vidi trīsdimensiju telpā.
Puzzle tile maps	*Flīžu kartes*. Flīžu kartes ir kartēšanas metode, kurā ģeogrāfiskās kartes tiek sadalītas atsevišķās, regulārās flīzēs (parasti kvadrātiņos), lai atvieglotu karšu datu ielādi un attēlošanu tiešsaistē vai mobilajās lietotnēs. Šī pieeja ļauj ātri vizualizēt lielus ģeogrāfisko datu kopumus, pieprasot tikai vajadzīgās flīzes.
R-CNN	*Region-based Convolutional Neural Networks. Reģionu bāzēts konvolūcijas neironu tīkls*. Neironu tīkla metode, kas lokalizē un klasificē objektus attēlos, identificējot attēla reģionus, kuros atrodas objekti, un pēc tam tos klasificējot.
Smart City	*Viedpilsēta*. Pilsēta, kas izmanto digitālās tehnoloģijas un *IoT*, lai efektīvāk pārvaldītu resursus, infrastruktūru un pakalpojumus, uzlabojot iedzīvotāju dzīves kvalitāti un ilgtspēju.
VPN	*Virtual private network. Virtuālais privātais tīkls*. Tehnoloģija, kas šifrē interneta savienojumu un izveido privātu un drošu kanālu starp lietotāju un tīklu, aizsargājot datus no nesankcionētas piekļuves.
Wi-Fi	*Bezvadu tehnoloģija*, kas ļauj ierīcēm, piemēram, datoriem, viedtālruņiem un citām ierīcēm, izveidot savienojumu ar internetu vai tīklu bezvadu režīmā.
Winsorization	*Vinsorizācija.* Statistiska metode, kurā ekstremālās novērojumu vērtības tiek aizstātas ar robežvērtībām, lai samazinātu ārkārtējo vērtību ietekmi uz analīzi.

Termins vai termina saīsinājums	Skaidrojums
XML	*Extensible Markup Language. Paplašināma iezīmēšanas valoda*. Marķēšanas valoda, ko izmanto, lai strukturētu datus, apmainītos ar datiem un uzglabātu datus, bieži lietota tīmekļa lietojumprogrammās datu aprakstīšanai un transportēšanai.
Zig Bee	*Bezvadu komunikācijas protokols*, kas paredzēts mazas jaudas ierīcēm un sensoru tīkliem, nodrošinot zemu datu pārraides ātrumu un energoefektīvu darbību lielos tīklos.

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 jauninājums 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 (*Principal Component Analysis, 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.

1. SENSORU ĢENERĒTIE DATI

Lietu internets jeb angliski *IoT* (*Internet of Things*) ir nozīmīga mūsdienu digitālās vides daļa, kas nodrošina datu vākšanu un analīzi rūpniecībā un pilsētvides attīstībā. Sensoru tīkli un ierīču sistēmas sniedz ieskatu vides procesos, kas ļauj labāk izprast un kontrolēt dažādas sistēmas. Datu vākšana ar *IoT* sensoriem nodrošina nepārtrauktu informācijas plūsmu, kas ir būtiska procesu uzlabošanai un lēmumu pieņemšanai. Sensoru savāktie dati sākotnēji ir neapstrādāti signāli, bet pēc to apstrādes tie kļūst par lietderīgu informāciju, ko var izmantot tālākai analīzei. Sistēmu darbības kvalitāte ir atkarīga no ievaddatu precizitātes. Tas ietver sensoru tehnisko stāvokli, to kalibrāciju un datu pārraides drošību. *IoT* tehnoloģijas nodrošina papildu iespējas procesu automatizācijā un analīzē.

1.1. Sensoru ģenerēto datu izmantošana lietu internēta ietvaros

IoT, kas apvieno sensorus, aparatūru un programmatūru, ir kļuvis par svarīgu informācijas sistēmu elementu (Nižetić et al., 2020). Tas ļauj iegūt datus gan no tiešajiem sensoriem, gan no saistītajām sistēmām. *Nižetić* u. c. (Nižetić et al., 2020) pētījumā norāda uz *IoT* tehnoloģiju attīstības saistību ar digitalizācijas procesiem un elektronisko ierīču savienojamību. Pētījumā uzsvērta nepieciešamība pēc efektīvākiem pakalpojumiem un elastīgākiem procesiem *IoT* tehnoloģiju ieviešanā. Rezultātā *IoT* potenciāls ļauj tam kļūt par vienu no būtiskākajiem sasniegumiem mūsdienu tehnoloģiju attīstībā (Alam et al., 2016).

Visprogresīvākās *IoT* lietojumu jomas ir saistītas ar Industriju 4.0 (*Industry 4.0*) (Osterrieder et al., 2020) un viedās pilsētas koncepciju (Eremia et al., 2017), bet nevar izslēgt arī transporta (Porru et al., 2020) un lauksaimniecības (Villa-Henriksen et al., 2020) jomas. Industrija 4.0 nosaka to (Osterrieder et al., 2020), ka ražotnē jau iezīmējas četri atšķirīgi darbības slāņi: fiziskais, datu, mākoņdatošanas (*Cloud Computing*), ka arī izlūkošanas un kontroles slānis. Sistēmas fiziskajā slānī darbojas mašīnas un iekārtas, kas veic faktiskās darbības. Datu slānis nodrošina informācijas plūsmu starp mašīnu sensoriem un mākoņa serveri, kontrolējot datu apjomu, veidu un pārraides ātrumu.

Pēc tam dati tiek uzglabāti, parasti īslaicīgi, mākonī, kur tos var apstrādāt sarežģītas analīzes nolūkos. Uzraudzības un kontroles slānī tiek nodrošināta viedā rūpnīcas uzraudzība, un šo kontrolējošo programmatūru var pielāgot cilvēka iejaukšanās gadījumā. Līdzīgas koncepcijas var piemērot arī citās iepriekšminētajās nozarēs (Villa-Henriksen et al., 2020).

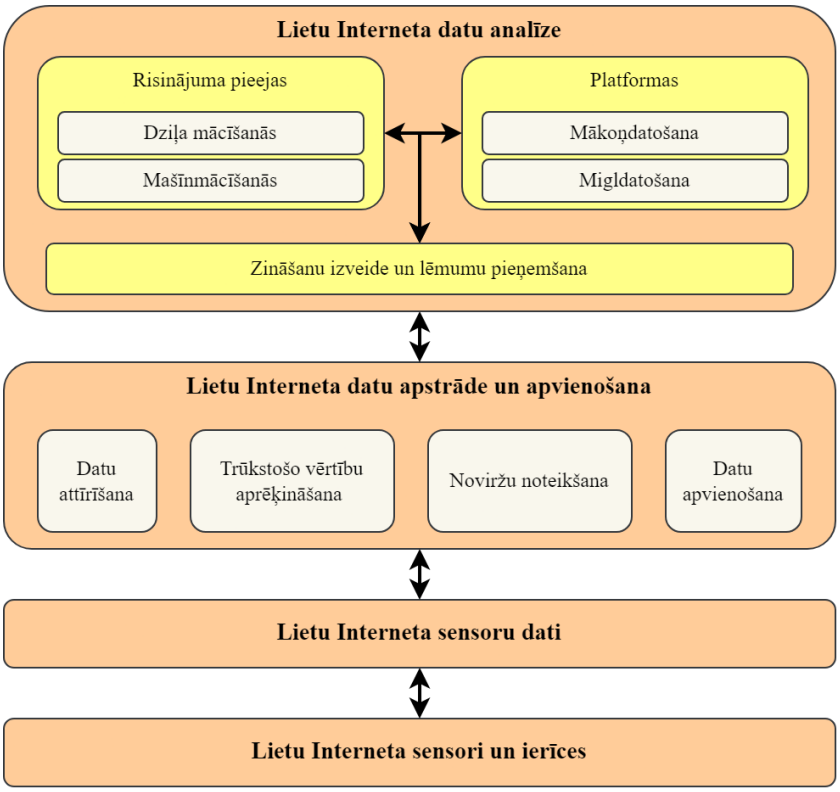
Bezvadu tīklu attīstība un jauno tehnoloģiju, īpaši mākoņu platformu, izmantošana veicina *IoT* tirgus izaugsmi visā pasaulē (Krishnamurthi et al., 2020). Līdz ar to strauji pieaug pieprasījums pēc viedierīcēm un to sniegtajiem pakalpojumiem. Lielākā daļa *IoT* sensoru datu tiek izmantota reāllaika apstrādei rūpnieciskos lietojumos, veselības aprūpē un zinātniskajos pētniecības projektos. Piemēram, veselības aprūpes jomā ķermeņa sensori (Jayakumar et al., 2021) pacientu stāvokļa uzraudzībai rada apjomīgus datus, kas prasa rūpīgu apstrādi, lai nodrošinātu precīzu datu analīzi, zināšanu attīstību un efektīvu lēmumu pieņemšanu (Mavrogiorgou et al., 2019). Respektīvi, *IoT* galvenais mērķis ir integrēt fizisko un virtuālo pasauli, izmantojot internetu kā galveno saziņas un informācijas apmaiņas kanālu (Patel et al., 2016). *IoT* sensoru tīkli galvenokārt darbojas trīs virzienos (Govinda & Saravanaguru, 2016; Sanyal & Zhang, 2018): tie uztver svarīgu informāciju no apkārtējās vides, seko līdzi sistēmas iekšējiem procesiem un pārveido savāktos datus noderīgā informācijā, kas palīdz pieņemt lēmumus.

Analizējot *IoT* sistēmu īpašības, var izdalīt sešas būtiskākās iezīmes (Patel et al., 2016). Pirmkārt, *IoT* sistēmām piemīt universāla savienojamība – tās var integrēt jebkuru ierīci globālajā informācijas un sakaru tīklā. Šī savienojamība nodrošina ne tikai tīkla piekļuvi, bet arī spēju gan izmantot, gan radīt datus. *El-Mawla* un *Badawy* (El-Mawla & Badawy, 2023) pētījumā apraksta *IoT* sistēmu struktūru. Tā ietver dažādas aparatūras platformas un tīklus ar

savstarpējo mijiedarbību. Sistēmām raksturīgi vairāki ierobežojumi, tostarp privātuma aizsardzība un nepieciešamība pēc saskaņotības starp fiziskajām un virtuālajām ierīcēm. Sistēmu dinamiskums ietver ierīču stāvokļu maiņu (aktivizēšanos un deaktivizēšanos) un konteksta izmaiņas (atrašanās vietu, ātrumu).

Parasti *IoT* tiek aplūkots viedās pilsētas kontekstā kā rīks, kas automatizē pilsētas procesus (Lau et al., 2019). Tā kā viedās pilsētas rada milzīgus datu apjomus, nepieciešamība pēc datu apvienošanas un analīzes nepārtraukti pieaug (Khan et al., 2021). Līdz ar to *IoT* ir svarīga sastāvdaļa dažādu starpnozaru uzraudzības un pārvaldības sistēmu izstrādē (M. Zhang et al., 2021). *IoT* ieviešana dažādās nozarēs ir veicinājusi to pārvēršanos par viedajām nozarēm, piemēram, precīzo biškopību (Zacepins et al., 2015) un precīzo putnkopību (Astill et al., 2020).

IoT sistēmu arhitektūras elastīgums ļauj izmantot dažādas sensoru un programmatūras kombinācijas. Pētnieki identificē četru slāņu arhitektūras modeli (sk. 1.1. att.)(Krishnamurthi et al., 2020) kas fokusējas uz datu apstrādes posmiem, tomēr neaptver visus *IoT* ieviešanas aspektus, piemēram, uzraudzību, savienojuma vārtejas un saskarnes, gala lietojumprogrammatūru un specifiskās lietojuma jomas (Patel et al., 2016).



1.1. att. **Tipveida arhitektūra *IoT* sistēmām** (Krishnamurthi et al., 2020).

Kviesis un Zacepins (Kviesis & Zacepins, 2015) savā pētījumā aplūko sensoru sistēmu īpatnības. Autori īpašu uzmanību pievērš sensoru tehniskajiem ierobežojumiem – ierobežotai skaitļošanas jaudai un atmiņas resursiem. Pētījumā atklāts interesants aspekts: šie šķietamie trūkumi noteiktos lietojumos var kļūt par sistēmas priekšrocībām. Sensoru savstarpējai saziņai izmanto plaši pazīstamus protokolus – *Wi-Fi, Bluetooth, ZigBee* un *LTE/4G* tehnoloģijas. Datu apstrādes slānis risina četrus būtiskus uzdevumus: trokšņa samazināšanu (G. Han et al., 2022), noviržu noteikšanu (Gaddam et al., 2020), trūkstošo datu aizvietošanu (Yuehua Liu et al., 2020) un datu apkopošanu (Sanyal & Zhang, 2018). Savukārt datu apvienošanas slānis integrē informāciju no dažādām neviendabīgām sensoru ierīcēm (Teh et al., 2020), sagatavojot to tālākai analīzei. *Anawar* u. c. (Anawar et al., 2018) pētījumā analizē viedās funkcionalitātes nozīmi *IoT* sistēmās. Autori identificē vairākus būtiskus aspektus sistēmu optimizācijā, īpaši uzsverot resursu pārvaldību un sistēmu drošību. Pētījuma rezultāti norāda uz analītisko risinājumu lomu sistēmu veiktspējas uzlabošanā.

Pētnieki norāda uz sarežģījumiem, ko rada lielais datu apjoms un to daudzveidība (Krishnamurthi et al., 2020). Īpaši aktuāli ir lielo datu apstrādes jautājumi (Ding et al., 2019), kas ietver trīs galvenās problēmas (Cai & Zhu, 2015; Merino et al., 2016; Z. Yan et al., 2018): grūtības atšķirt viltus datus, apvienošanas precizitātes problēmas saistībā ar datu atkarību un korelāciju, kā arī lielu saziņas slogu un nopietnu aizkavēšanos centralizētajos datu apvienošanas modeļos. Būtiskas problēmas rada arī privātuma aizsardzība (X. Su et al., 2019) un datu aktualitāte (Okafor et al., 2020). *IoT* sensoru dati izceļas ar piecām raksturīgām iezīmēm (Patel et al., 2016), kas būtiski ietekmē gan sistēmu izstrādi, gan to praktisko pielietojumu (G. Su & Kensek, 2021; M. Zhang et al., 2021). Tehniskie ierobežojumi ietver ierobežotu skaitļošanas jaudu, akumulatora darbības laiku un atmiņas apjomu. Šie ierobežojumi padara sistēmas vieglāk ievainojamas pret kļūmēm un uzbrukumiem. Reāllaika apstrādes prasības nosaka, ka neapstrādātie dati ātri jāpārveido noderīgā informācijā. Mērogojamības aspekts ir īpaši svarīgs, jo sensoru skaits tīklā strauji pieaug. Papildu sarežģījumus rada datu daudzveidība – tie var būt gan vienkārši (Būla, binārie), gan sarežģīti (nepārtraukti dati, skaitliskās vērtības). Sistēmu neviendabīgums izpaužas kā dažādi datu avoti, sākot no strukturētām kopām līdz sociālo tīklu plūsmām.

Minēto IoT sensoru datu problēmu risināšanai pētnieki izmanto datu apvienošanas metodes. Šīs metodes palīdz attīrīt un apvienot informāciju no dažādiem avotiem, iegūstot precīzākus un pilnīgākus datus. Īpaši svarīga ir datu modalitāte – veids, kādā sensori uztver un ieraksta informāciju par novēroto sistēmu (Turk, 2014). Mūsdienās, kad sensoru izmaksas ir ievērojami samazinājušās, speciālisti strādā ar daudzveidīgiem datu avotiem, aptverot arvien plašāku modalitāšu klāstu (Bokade et al., 2021). Multimodālā datu apvienošana apvieno informāciju no dažādiem avotiem un kanāliem vienotā, vieglāk izmantojamā formā (Castanedo, 2013). Šī pieeja ietver gan vienādu sensoru izmantošanu dažādās vietās, gan atšķirīgu sensoru datu apvienošanu. Multimodālo datu apvienošanu plaši izmanto vairākiem mērķiem. Klasifikācijas un regresijas gadījumos tiek analizēti marķēti dati, lai atrastu sakarības starp mainīgajiem. Klasteru veidošana grupē līdzīgos datus bez iepriekšējas marķēšanas. Savukārt dimensiju samazināšana palīdz atlasīt svarīgāko informāciju no lielām datu kopām, saglabājot precīzu saikni starp ievades un izvades datiem (Bokade et al., 2021). Datu apvienošana notiek trīs līmeņos: neapstrādātu datu, pazīmju un lēmumu līmenī (Gao et al., 2020). Vislabākos rezultātus sniedz sinhronizētu vai strukturāli līdzīgu datu apvienošana. Tomēr mūsdienu mašīnmācīšanās metodes spēj veiksmīgi apvienot arī ļoti atšķirīgus datu veidus. Praksē, ņemot vērā datu daudzveidību, vispirms tiek izdalītas būtiskās pazīmes no katras modalitātes. Pazīmju līmeņa apvienošana strādā ar vienkāršotiem datiem, kas ir piemērotāki tālākai apstrādei. Lēmumu līmeņa apvienošana vispirms analizē katru modalitāti atsevišķi un tad apvieno iegūtos secinājumus kopējā rezultātā.

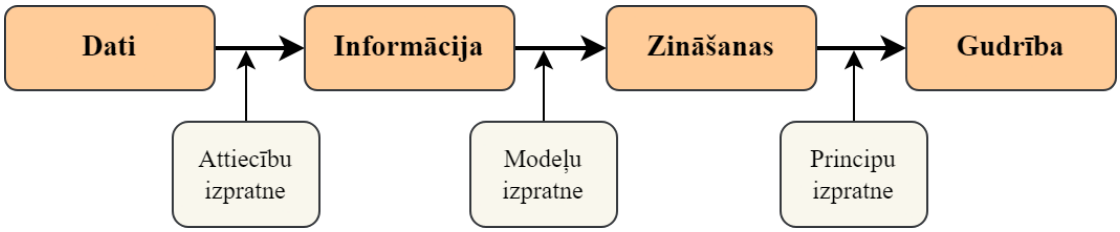
Iepriekš aprakstītās sarežģītības un raksturlielumi rada būtiskas atšķirības starp *IoT* sistēmām, īpaši datu tipu un mērvienību daudzveidības ziņā. Svarīgi atzīmēt, ka termins "dati" *IoT* kontekstā primāri attiecas uz neapstrādātiem sensoru signāliem vai rādījumiem, kā arī vienkāršām skaitļu virknēm, nevis uz apstrādātu un strukturētu informāciju ar konkrētu objektu piesaisti (Jifa & Lingling, 2014). Praksē bieži sastopamas situācijas, kad dati nav izolēti, bet atrodas plašākā informācijas telpā, sajaucoties ar citu informāciju (Nasution et al., 2021). Tādējādi terminam "dati" ir vairāki uztveršanas veidi, bet nav viennozīmīgas definīcijas, kas strikti ierobežotu tā pielietojumu (Beddar-Wiesing & Bieshaar, 2020).

1.2. Datu un informācijas terminu saistība

Datu apvienošanas process ietver noteiktus datu objektus, kas var būt pētāmi vai veidot kāda pētāma objekta daļu. Parasti ar "datiem" saprot neapstrādātus signālus vai rādījumus no ierīces, piemēram, sensora, kā arī apstrādātus datus, kas piesaistīti reālam objektam. Datu

apvienošanas metožu klasifikācija var atšķirties atkarībā no izmantotās datu apvienošanas terminoloģijas.

Izplatīts modelis šīs problēmas risināšanai ir datu, informācijas, zināšanu, gudrības modelis (sk. 1.2. att.) jeb *DIKW* (*Data, Information, Knowledge, Wisdom*), kas skaidro saistības starp šiem terminiem (Bellinger et al., 2004). Svarīgi ir saprast datu apvienošanas terminoloģiju. Datu un informācijas konceptuālai izpratnei īpaši nozīmīga ir *Ackoff* piedāvātā kategorizācija (Targowski, 2005), kas veido strukturētu pāreju no datiem uz augstākiem izpratnes līmeņiem. Šajā pieejā dati tiek definēti kā objektu vai notikumu attēlojums, kas caur apstrādes procesu kļūst par informāciju, sniedzot atbildes uz fundamentāliem jautājumiem par notikumu būtību. Tālāka informācijas transformācija rada zināšanas, kas nodrošina praktiskas vadlīnijas un instrukcijas. Dziļāka izpratne veidojas, analizējot informācijas savstarpējās sakarības un modeļus, kas ļauj izprast cēloņsakarības. Augstākajā līmenī gudrība apvieno visus iepriekšējos elementus efektīvai izmantošanai. *Bellinger* u. c. šo koncepciju papildina ar būtisku novērojumu, ka izpratne ir nepieciešams priekšnosacījums pārejai starp līmeņiem (Bellinger et al., 2004), ka izpratne ir nosacījums pārejai no zemāka līmeņa uz augstāku. Šī *DIKW* (*Data, Information, Knowledge, Wisdom*) hierarhijas modeļa principi ir kļuvuši par teorētisko pamatu mūsdienu datu apvienošanas risinājumu izstrādē.



1.2. att. ***DIKW* modeļa terminu saistības** (Allen, 2004).

DIKW hierarhijas modelis demonstrē, ka datu attiecību analīze rada informāciju, informācijas modeļu izpēte veido zināšanas, savukārt zināšanu pamatprincipu apguve noved pie gudrības. Šis modelis, kā pamats, tiek izmantots arī lielo datu jautājumu risināšanai (Nasution et al., 2021), ieviešot lielo datu raksturojošos parametrus un veidojot dažādas datu apvienošanas metodes. Tas ir efektīvs lielu datu apjomu izmantošanas veids no vairākiem avotiem (Dong et al., 2009). *DIKW* modelis tiek uzskatīts par datu apvienošanas modeli (Becerra et al., 2021).

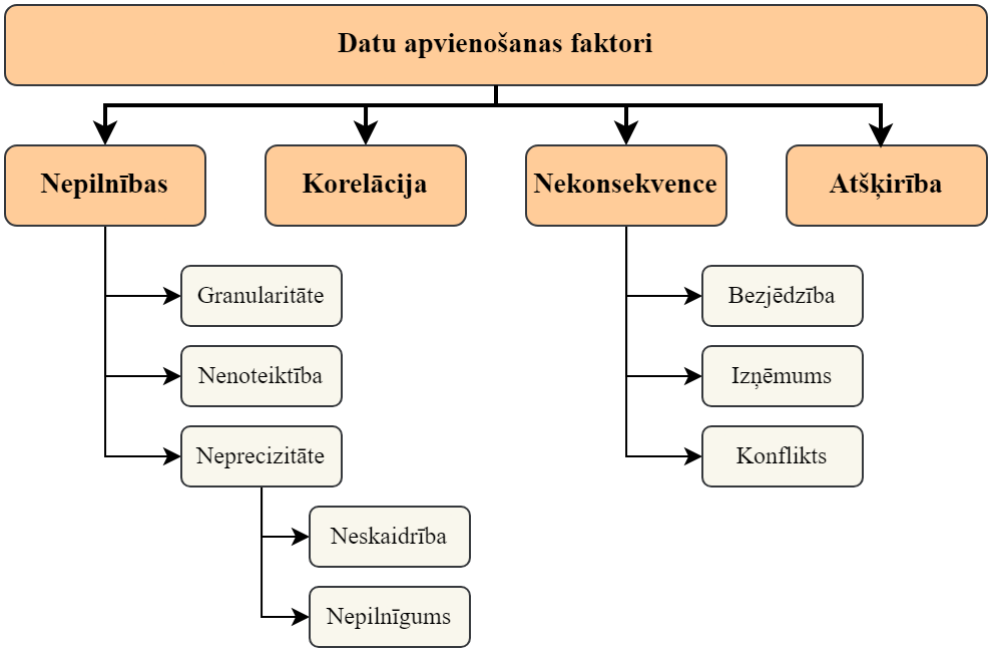
Pētījumi datu apvienošanas jomā identificē divas galvenās perspektīvas (Beddar-Wiesing & Bieshaar, 2020): universālu apvienošanas definīciju izstrādi un apvienošanu dažādos abstrakcijas līmeņos. Tomēr joprojām pastāv grūtības viennozīmīgu terminu noteikšanā, īpaši attiecībā uz atšķirībām starp datu un informācijas apvienošanu. Pamatprincipi datu apvienošanas izpratnei balstās uz trim būtiskiem aspektiem. Pirmais ir datu apvienošanas komponenti – neapstrādāti dati, kas iegūti no novērojumiem vai mērījumiem. Otrais aspekts ir datu vai informācijas izcelsme no vairākiem avotiem. Trešais skar datu attēlojuma īpatnības (Boström et al., 2007). Datu apvienošanas galvenais mērķis ir iegūtās informācijas kvalitātes uzlabošana. Šī procesa priekšrocības ietver datu līdzsvarošanu, nenoteiktības samazināšanu un uzlabotas secinājumu sniegšanas iespējas (Hall & Llinas, 2016). Tālāka attīstība šajā jomā akcentē prognozēšanas precizitātes (Steinberg & Bowman, 2008) un lēmumu pieņemšanas procesa uzlabošanu (Piechnicki et al., 2021), kā arī bagātākas un praktiskāk izmantojamas informācijas iegūšanu (Beitzel et al., 2003).

Datu apvienošanas procesa izpratnē īpaši nozīmīgs ir JDL modelis jeb *JDL* (*Joint Directors of Laboratories*) (White & Steinberg, 1998), kas definē apvienošanu kā datu un informācijas asociēšanu, korelāciju un kombinēšanu no viena vai vairākiem avotiem. Šis modelis ievieš vairāku līmeņu konceptu datu apvienošanā. Specifiskāku pieeju piedāvā sensoru sistēmu orientētā datu apvienošana. Šajā kontekstā process tiek definēts kā datu un informācijas sasaiste un apvienošana no dažādiem avotiem, lai nodrošinātu precīzus pozīcijas aprēķinus, identitātes noteikšanu un situācijas novērtējumu (Hall & Llinas, 1997). Svarīgi atzīmēt, ka šīs

definīcijas skaidri nošķir neapstrādātus datus no apstrādātas informācijas, tādējādi veidojot pamatu mūsdienu datu apstrādes hierarhijai (Castanedo, 2013).

1.3. Datu kvalitāte un raksturlielumi

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. Šīs nepilnības tiek klasificētas kā apvienošanas faktori (sk. 1.3. att.) (Khaleghi et al., 2013).



1.3. att. **Ar datiem saistītie apvienošanas faktori** (Khaleghi et al., 2013).

Datu apvienošanas jomas izpēte atklāj vairākus būtiskus šķēršļus, kurus savā pētījumā identificējuši *Lakshmanarao* un *Shashi* (Lakshmanarao & Shashi, 2020), neraugoties uz daudzveidīgo risinājumu izstrādi specifiskajām vajadzībām. Galvenais izaicinājums ir sensoru datu kvalitāte – mērījumi bieži ir neprecīzi, nenoteikti vai nepilnīgi. Lai gan statistiskās metodes un papildu informācijas avoti var uzlabot šo situāciju, problēma joprojām būtiski ietekmē rezultātu kvalitāti. Līdzīgas grūtības rada sensoru trokšņi un vides interference. Īpašu izaicinājumu rada sistemātiskās kļūdas, kas veidojas ilgstošu vai dinamisku traucējumu rezultātā. Nopietnas problēmas rodas sistēmās, kas izmanto ticamības funkcijas vai *Dempster-Shafer* (DST) teoriju, kad parādās pretrunīgi mērījumi. Nepareiza šo pretrunu apstrāde var radīt būtiskas kļūdas rezultātu interpretācijā. Izaicinājumu rada arī dažādo sensoru datu sinhronizācija vienotā atskaites sistēmā. Īpaši svarīga ir atkārtotas datu apstrādes problēma sadalītajās sistēmās. Papildu sarežģījumus rada sensoru heterogenitāte – katrs sensors uztver atšķirīgus fizikālos lielumus. Apstrādes algoritmiem jāspēj integrēt šos dažādos mērījumu veidus vienotā objekta stāvokļa novērtējumā. Būtisks ir arī datu apstrādes arhitektūras jautājums bezvadu sensoru tīklos – centralizēta vai izkliedēta apstrāde rada atšķirīgas kompromisa situācijas starp skaitļošanas resursiem un precizitāti. Mūsdienu sistēmās aizvien kritiskāka kļūst reāllaika apstrādes nepieciešamība, jo mērījumu aktualitāte strauji samazinās laikā. Šis aspekts ir īpaši svarīgs reāllaika vadības sistēmās, kur nepieciešama tūlītēja datu validācija un apstrāde.

Datu apvienošanas izaicinājumi parasti tiek risināti individuāli, nevis sistēmiski (Bakr & Lee, 2017). Vides faktoru ietekme uz sensoru mērījumiem (M. Kumar et al., 2006) rada sistemātiskas novirzes datu kopās, kuru novēršanai statistiskajā analīzē nepieciešama pirmsapstrāde (Y. Zhang et al., 2010). Īpaši aktuāla ir pretrunīgu datu problēma vairāku sensoru

sistēmās, kas mēra vienu un to pašu fizikālo lielumu. Vairāku sensoru sistēmu pētījumi (Henry et al., 2019; Kviesis & Zacepins, 2015), demonstrē to nozīmīgo lomu lēmumu pieņemšanas atbalstā, apstrādājot multimodālus datus – gan homogēnus, gan heterogēnus, ieskaitot audio, vizuālos un tekstuālos datus. *Liu* u. c. (J. Liu et al., 2020) piedāvā heterogēno datu klasifikāciju telpiskajā, temporālajā, statiskajā, dinamiskajā un atribūtu dimensijās. Šo dimensiju kombinācijas, piemēram, telpiski-temporālie dati (Atluri et al., 2018), rod praktisku pielietojumu dažādās nozarēs, tostarp biškopības meteoroloģiskajos novērojumos. Bezvadu sensoru tīklu pētījumi (Kviesis et al., 2020; Meikle & Holst, 2015; Tan et al., 2015) akcentē vides trokšņu ietekmi uz sensoru mezglu mērījumiem, uzsverot precīzas datu korelācijas nozīmīgumu. *Khaleghi* u. c. (Khaleghi et al., 2013) identificē reģistrācijas datu izlīdzināšanas problēmu, kas rodas, transformējot lokālo sensoru datus vienotā atskaites sistēmā. Vairāku mērķu izsekošanas scenārijos izdala (Sheng et al., 2018) divas būtiskas datu asociācijas formas: mērījumu-celiņu (*measurement-to-track*) un celiņu savstarpējo asociāciju (*track-to-track*). 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. Precīzās biškopības pētījumi (Chang & Bai, 2018; Ferrari et al., 2008) demonstrē šīs pieejas efektivitāti sistēmas slodzes optimizācijā. Datu kompresijas metodes var tikt izmantotas arhitektūras optimizācijai, pieļaujot kontrolētus zudumus (Zhu et al., 2005). Temporālie aspekti rada īpašus izaicinājumus vairāku sensoru sistēmās. *Besada-Portas et al.* (Besada-Portas et al., 2011) akcentē ārpus secības saņemto datu problemātiku reāllaika lietojumos. Dinamisko sistēmu gadījumā *Brooks* u. c. (Brooks et al., 2009), uzsver mērījumu vēstures nozīmi datu apvienošanā, kas, saskaņā ar *Khaleghi* u. c. (Khaleghi et al., 2013) pētījumu, ļauj novērtēt datu aktualitāti un sensoru sistēmu reaktivitāti.

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).

Datu kvalitātes problēmu risināšana ir īpaši aktuāla gadījumos, kad sensoru mērījumi uzrāda neatbilstības, nepilnības vai nekonsekvences. *Kreibich* u. c. (Kreibich et al., 2014) norāda, ka datu ticamība būtiski samazinās nekontrolētā vidē ar augstu trokšņu līmeni, tādēļ nepieciešama papildu informācijas avotu integrācija datu kvalitātes un ticamības paaugstināšanai. *Consoli* u. c. (Consoli et al., 2015) uzsver, ka ierobežota apjoma mērījumi no atsevišķiem sensoriem bieži nespēj nodrošināt pietiekamu informāciju sistēmas stāvokļa novērtēšanai. Šī iemesla dēļ vairāku sensoru datu apvienošana kļūst par būtisku instrumentu visaptverošas sistēmas izpratnes veidošanā. Tiek demonstrēta (W. Wang et al., 2018) šīs pieejas efektivitāte vides monitoringa sistēmās, kur, piemēram, ēku noslodzes novērtējums tiek iegūts, integrējot dažādu vides sensoru datus, tādējādi iegūstot netriviālas sakarības, kas nav tieši novērojamas neapstrādātos mērījumos.

1.4. Datu kvalitātes uzlabošana

IoT sensoru tīklu izpēte parāda dažādus datu kvalitātes izaicinājumus. *Karkouch* u. c. (Karkouch et al., 2016) norāda uz vairākiem faktoriem, kas ietekmē datu kvalitāti *IoT*. Tie ietver *IoT* izmēru un mērogu, resursu ierobežojumus, sakaru tīkla arhitektūru un stāvokli, vides apstākļus, sensorus, drošības ievainojamību, datu plūsmas apstrādi un īpašus gadījumus, piemēram, izkrustošanu un netīrus sensoru mezglus un vandālismu.

IoT sistēmu izpēte atklāj virkni nozīmīgu problēmu. Kā norāda *Gomathi* u. c. (Gomathi et al., 2018), *IoT* tehnoloģiju globālā izplatība rada līdz šim nepieredzētu datu avotu daudzveidību – dati vairs nenāk tikai no datorsistēmām, bet arī no ikdienišķiem priekšmetiem. *Aboubakar* u. c. (Aboubakar et al., 2022) šo situāciju trāpīgi raksturo kā IP tīklu ar paaugstinātu nestabilitāti un datu zudumu līmeni. Analizējot resursu ierobežojumus, *Adelantado* u. c., un

Dinculeană un *Cheng* (Adelantado et al., 2017; Dinculeană & Cheng, 2019) uzsver *IoT* ierīču specifiku – tās spēj pārraidīt tikai neliela apjoma ziņojumus. *Sliwa* u. c. (Sliwa et al., 2020) identificē divus kritiskos šķēršļus v ierobežoto enerģiju un atmiņu, kas būtiski kavē drošības risinājumu ieviešanu. *Nair* u. c. (Nair et al., 2016) īpaši uzsver bateriju problēmu – ierīces bieži darbojas ar kritiski zemu enerģijas līmeni. Šie ierobežojumi, kā norāda *Guan* un *Pires* (Guan et al., 2017; Pires et al., 2016), spiež pieņemt kompromisus datu vākšanas stratēģijās, tieši ietekmējot iegūto datu kvalitāti.

Drošības aspektā situācija ir īpaši sarežģīta – sensoru ierīces praktiski nav iespējams pilnvērtīgi aizsargāt pret uzbrukumiem, jo tās nespēj veikt resursietilpīgās kriptogrāfiskās operācijas. Šī ievainojamība rada nopietnus draudus datu integritātei. Vides ietekme uz sensoriem ir īpaši jūtama ekstremālos apstākļos, piemēram, meteoroloģiskajos novērojumos kalnu virsotnēs. Ierobežotā piekļuve apkopei apvienojumā ar dabas apstākļiem (sniega segas veidošanos, netīrumu uzkrāšanos) būtiski ietekmē sensoru darbību. Turklāt ierīces ir pakļautas vandālisma un pat savvaļas dzīvnieku radītiem bojājumiem. Tehniskā aspektā būtiska problēma ir zemākas kvalitātes sensoru precizitāte un kalibrācijas trūkumi. Ekstremālie laikapstākļi var fiziski bojāt gan ierīču korpusus, gan mērīšanas elementus. Īpašas grūtības rada mērījumu pārveidošana (piemēram, sprieguma rādījumu konvertēšana mitruma datos). Nopietnu problēmu rada tā sauktie "izkritušie un netīrie sensori" – ierīces, kas, lai gan bojātas, turpina sūtīt kļūdainus datus. Datu plūsmu kontekstā viedās ierīces nepārtraukti sūta informāciju uz centrālajām sistēmām. Šo datu apstrāde dažādiem mērķiem (gan zināšanu ieguvei, gan resursu taupīšanai) un dažādu apstrādes operatoru izmantošana var papildus ietekmēt sākotnējo datu kvalitāti.

Tādējādi pastāv vairākas metodes, kas ir izstrādātas, lai novērstu šos traucējumus, un dažas no tām ietver trokšņa samazināšanu, trūkstošo datu aizpildīšanu un interpolāciju, datu noviržu noteikšanu un datu apkopošanu (Karkouch et al., 2016; Souza & Amazonas, 2015), galvenokārt pievēršot uzmanību trokšņa samazināšanai un trūkstošo datu aizpildīšanai.

1.4.1. Trokšņu slāpēšana

Troksnis signālu apstrādē izpaužas kā nekorelētas sastāvdaļas, kas būtiski ietekmē un modificē sākotnējo signālu. Šī parādība cieši saistīta ar nenoteiktības jēdzienu, kas veido atsevišķu datu kvalitātes dimensiju. Kā norāda *Jcgm* (Jcgm, 2008) pētījumā, nenoteiktību var raksturot kā mērījumu rezultātu izkliedes parametru, kas atspoguļo vērtību variācijas. Šo konceptu tālāk attīsta *Teh* u. c. (Teh et al., 2020), skaidrojot nenoteiktību kā kļūdu kvantitatīvo izteiksmi. Savukārt *Krishnamurthi* u. c.(Krishnamurthi et al., 2020) uzsver trokšņa negatīvo ietekmi uz sistēmas resursiem, norādot uz nevajadzīgu apstrādes procesu un nederīgu datu izmantošanu.

Trokšņu mazināšanas metodoloģija balstās uz vairākiem pieejām, no kurām izplatītākā izmanto bīdāmā loga principu. Šī metode ļauj analizēt signāla lokālās īpašības, izmantojot dažādus statistiskos rādītājus – vidējās vērtības, filtrus un adaptīvās procedūras. Tomēr, kā norāda *Vanus* u. c. (Vanus et al., 2020), šīm metodēm piemīt būtiski ierobežojumi, īpaši attiecībā uz lokālo frekvenču pielāgošanu. Adaptīvie filtri, lai gan spējīgi risināt šo problēmu, prasa ievērojamus laika resursus un sarežģīti iegūstamus atsauces signālus.

Viļņu transformācijas metodes šajā kontekstā piedāvā efektīvāku risinājumu signālu apstrādē. To galvenā priekšrocība ir spēja saglabāt sākotnējos signāla koeficientus, vienlaikus samazinot trokšņa komponentes. *M. M. Rashid* u.c. (M. M. Rashid et al., 2015) savā pētījumā izceļ divas galvenās viļņu transformācijas pieejas: nepārtraukto transformāciju, kas analizē signālu gan laika, gan frekvences dimensijās, un diskrēto transformāciju, kas koncentrējas uz laika dimensijas analīzi. Šī metodoloģiskā daudzveidība ļauj elastīgi pielāgoties dažādām signālu apstrādes vajadzībām.

1.4.2. Trūkstošo datu aizvietošana

Eksistē būtiskā problēma datu priekšapstrādē – nepilnīgo datu aizvietošana. Lai gan mūsdienu mašīnmācīšanās algoritmi, kas analizē *IoT* datus, ir veidoti ar pieņēmumu par datu pilnīgumu, realitātē situācija ir atšķirīga. Kā norāda *Adhikari* u. c. (Adhikari et al., 2022), *IoT* sistēmās datu nepilnīgums ir izplatīta parādība, kas rada nopietnus draudus analīzes kvalitātei. Pētnieki īpaši uzsver, ka analīzes process, kas balstās uz nepilnīgiem vai trūkstošiem datiem, var novest pie maldinošiem un neuzticamiem rezultātiem, tādējādi apdraudot visa analītiskā procesa ticamību.

Trūkstošo vērtību problēma mūsdienu datu analīzē kļūst aizvien aktuālāka, īpaši strādājot ar apjomīgām un kritiskām datu kopām satiksmes vadības un klīnisko pētījumu jomās. Prakse rāda satraucošu tendenci – nereti pat 40% no visiem datiem ir nepilnīgi vai iztrūkstoši. Vienkārša šādu datu izslēgšana vai ignorēšana var radīt būtiskus kropļojumus statistiskajos rādītājos un modeļu precizitātē, kas savukārt negatīvi ietekmē lēmumu pieņemšanas procesu. *Zhang* u. c. (Zhang et al., 2024) piedāvā metodisku risinājumu, ieviešot izsekošanas noņemto autoenkoderu jeb *TRAE* (*Tracking-Removed Autoencoder*) un faziklasterizācijas jeb *FC* (*Fuzzy Clustering*) metodes. Šīs pieejas nodrošina efektīvu trūkstošo datu modelēšanu, balstoties uz datu īpašību savstarpējo korelāciju analīzi un rekonstrukciju.

Reāllaika datu plūsmu specifika *IoT* sistēmās rada īpašas prasības datu kvalitātei un apstrādei. Pētot šo problemātiku, *Martin* u. c. (Martin et al., 2023) atklāj ciešo saistību starp datu nepilnībām un to ietekmi uz sistēmu darbības savlaicīgumu, precizitāti un pilnīgumu. Īpaši izteikti šī ietekme novērojama vides monitoringa un rūpniecisko procesu vadības sistēmās, kur nepieciešama tūlītēja rīcība, pamatojoties uz ienākošajiem datiem. Pētnieku izstrādātā pieeja, kas balstīta uz mākslīgo intelektu, piedāvā sistemātisku risinājumu automātiskai datu modeļu atpazīšanai un trūkstošo vērtību prognozēšanai.

Liela mēroga *IoT* sistēmu izpētē *Kim* u. c. (Kim et al., 2022) akcentē datu daudzveidības radītās problēmas sistēmu darbībā. Viņu veiktā analīze parāda, ka datu plūsmu heterogenitāte, ko rada dažādu ierīču un sensoru tehnisko specifikāciju atšķirības, padara datu apstrādes procesu būtiski sarežģītāku. Pētnieku izstrādātās adaptīvās metodes un mašīnmācīšanās algoritmi nodrošina efektīvu pieeju trūkstošo vērtību identificēšanai, vienlaikus saglabājot datu plūsmas integritāti, kas ir kritiski svarīga precīzai lēmumu pieņemšanai.

Trūkstošo datu problēmas efektīvai risināšanai *Krishnamurthi* u. c. (Krishnamurthi et al., 2020) identificē trīs fundamentālus uzdevumus. Pirmais un būtiskākais ir trūkstošo datu cēloņu izpratne, kas var ietvert vairākus tehniskos un vides faktorus – no vājas tīkla savienojamības un bojātām sensoru sistēmām līdz vides ietekmei un sinhronizācijas traucējumiem. Šajā kontekstā trūkstošos datus klasificē trīs kategorijās: pilnīgi nejauši trūkstoši jeb *MCAR* (*Missing Completely at Random*), nejauši trūkstoši pēc noteiktiem likumiem jeb *MAR* (*Missing at Random*) un nejauši nenoteikti trūkstoši jeb *NMAR* (*Not Missing at Random*). Otrais uzdevums fokusējas uz trūkstošo datu modeļu izvērtēšanu. *Silva-Ramírez* u. c. (Silva-Ramírez et al., 2015) izceļ divas galvenās pieejas: monotoni trūkstošie modeļi jeb *MMP* (*Monotone Missing Pattern*) un nejauši trūkstošie modeļi jeb *AMP* (*Arbitrary Missing Pattern*). Šo modeļu identifikācija ir kritiska, jo tā nosaka piemērotākās aizvietošanas stratēģijas izvēli. Trešais uzdevums ietver atbilstoša trūkstošo vērtību aizvietošanas modeļa izstrādi un implementāciju, kas, balstoties uz identificētajiem modeļiem un cēloņiem, ļauj pēc iespējas precīzāk novērtēt un aizstāt trūkstošās vērtības.

Globālā līmenī ir izstrādātas daudzas metodes un modeļi, lai aizvietotu trūkstošos datus, un tie tiek pielāgoti atbilstoši trūkstošo datu veidiem. Piemēram, saraksta dzēšana (Pepinsky, 2018), kas tiek pielietota pilnīgi nejaušiem datiem. Turklāt vairums no metodēm, piemēram, maksimālās iespējamības metode, Beijesa novērtējums, nosacītā vidējā aizvietošana, ir piemērojami diviem vai trim trūkstošo datu variantiem (Pires et al., 2016). Papildus var izdalīt šādas metodes: asociāciju noteikumu ieguve (*Association rule mining*), klasteru veidošana, k-tuvākais kaimiņš jeb *K-NN* (*K-Nearest Neighbors algorithm*), vienotās vērtības dekompozīcija,

telpiskā un laika korelācija, inkrementālais telpas – laika modelis jeb *ISTM* (*Incremental Space-Time Model*), varbūtības matricas faktorizācija jeb *PMF* (*Probabilistic Matrix Factorization*).

1.4.3. Noviržu noteikšana un aizvietošana

IoT sistēmās novirzes bieži rodas datu pārraides laikā, kad ierīces, piemēram, sensori, sāk ģenerēt netipiskas vai kļūdainas vērtības. Šīs novirzes var būt saistītas ar problēmām sensora aparatūrā, signāla vājināšanos, nepareizu datu nolasīšanu vai pat ārējo faktoru izraisītām kļūdām. Piemēram, nepareizs signāls no sensora, kas pārraida datus par temperatūru, var radīt nepareizus lēmumus automatizētās sistēmās, kur precizitāte ir kritiski svarīga.

Izolācijas meža jeb *IF* (*Isolation Forest*) algoritms ir kļuvis par vienu no vadošajiem risinājumiem noviržu noteikšanai *IoT* datu kopās. Algoritma pamatā ir unikāla pieeja – datu sadalīšana lēmumu kokos, kur novirzes, pateicoties to specifiskajām īpašībām, tiek izolētas ar mazāku sadalījumu skaitu nekā normālie dati. Šī īpašība padara algoritmu īpaši efektīvu liela apjoma *IoT* sistēmās, kur dati tiek ģenerēti nepārtraukti un reāllaikā. *Muñoz* u. c. (Muñoz et al., 2024) norāda uz algoritma precizitātes uzlabošanas iespējām, izmantojot pielāgojamus parametrus – noviržu skaita aprēķinu un izolācijas koku skaitu. Šie parametri ir būtiski modeļa stabilitātes un precizitātes saglabāšanai, vienlaikus minimāli ietekmējot tā veiktspēju. Pēc noviržu identificēšanas seko to analīze, lai noteiktu, vai tās ir tehniskas kļūdas (piemēram, signāla traucējumi) vai sistēmas normālas izmaiņas (piemēram, sensora pārvietošana vai jaunas ierīces pievienošana). Šāda detalizēta izpratne ir kritiski svarīga, lai izvairītos no kļūdainām korektīvajām darbībām.

Lokālais ārpusnieku faktors jeb *LOF* (*Local Outlier Factor*) un izolācijas meža algoritmi tiek izmantoti noviržu identificēšanai *IoT* datu kopās. *LOF* novērtē, vai konkrēts datu punkts ir novirze, analizējot tā tuvāko kaimiņu datu blīvumu, un ir īpaši noderīgs, ja novirzes ir saistītas ar vietējiem datiem, piemēram, sensora darbību konkrētā reģionā. Savukārt izolācijas meža algoritms identificē globālas novirzes, analizējot visu datu kopumu un atrodot tos datus, kas būtiski atšķiras no parastajām vērtībām. Abi algoritmi palīdz noteikt novirzes dažādos līmeņos un nodrošina datu tīrību pirms turpmākās analīzes. Pēc noviržu noteikšanas dati tiek apstrādāti, izmantojot imputācijas metodes, piemēram, interpolāciju vai *K-NN* algoritmu, lai aizstātu nepareizos vai trūkstošos datus, tādējādi atjaunojot datu plūsmas integritāti (Martin et al., 2023).

Liela mēroga IoT sistēmās, kur datu plūsmas ir gan apjomīgas, gan daudzveidīgas, noviržu radītie draudi kļūst īpaši nozīmīgi. Neidentificētas un neapstrādātas novirzes var ne tikai pasliktināt datu kvalitāti, bet arī būtiski traucēt automatizēto procesu darbību. *Kim* u. c. (Kim et al., 2022) šīs problēmas risināšanai piedāvā inovatīvu pieeju – adaptīvas mašīnmācīšanās metodes. Šīs metodes darbojas divējādi – tās ne tikai atpazīst novirzes, bet arī veic automātisku datu korekciju, aizstājot kļūdainās vērtības ar aprēķinātām normālām vērtībām, kas balstītas uz vēsturisko datu analīzi. Šāda pieeja ļauj novērst potenciālās problēmas, pirms tās paspēj negatīvi ietekmēt *IoT* sistēmu darbību. Vēsturisko datu izmantošana korekciju veikšanai nodrošina, ka aizstātās vērtības saglabā datu kopas dabisko struktūru un tendences, tādējādi uzturot augstu datu kvalitāti pat negaidītu noviržu gadījumos.

1.5. Sensoru datu nodaļas kopsavilkums

IoT sistēmu izpēte kvalitatīvu datu nodrošināšanai atklāj vairākus būtiskus izaicinājumus. Sistēmu darbību ietekmē gan tehniskie faktori – tīkla uzbūve un sensoru ierobežojumi, gan ārējie apstākļi – vides ietekme un drošības riski. Īpaši izaicinoši ir ierīču tehniskie ierobežojumi – ierobežotā skaitļošanas jauda, atmiņas apjoms un akumulatora darbības laiks, kas nozīmīgi iespaido to, kā varam vākt un apstrādāt datus.

Lai risinātu šīs problēmas, ir izstrādātas trīs galvenās datu apvienošanas pieejas – centralizētā, decentralizētā un sadalītā. Centralizētā pieeja apvieno visus datus vienā vietā, bet prasa daudz resursu. Decentralizētā pieeja nozīmē sadalītu datu apstrādi, samazinot slodzi un padarot sistēmu elastīgāku. Sadalītā pieeja veiksmīgi apvieno abu iepriekšminēto pieeju priekšrocības. Efektīvai datu apstrādei izmanto trīs līmeņus – sākotnējo sensoru datu apstrādi, to pārveidošanu un gala lēmumu pieņemšanu.

Datu kvalitātes uzlabošanai izmanto trīs galvenās metodes. Trokšņu mazināšanai izmanto viļņu transformācijas, kas saglabā svarīgāko informāciju, bet samazina traucējumus. Trūkstošos datus atjauno ar *TRAE* un *FC* metodēm, kas izmanto datu savstarpējās sakarības. Novirzes atrod ar izolācijas meža un *LOF* algoritmiem, kas spēj pamanīt gan lielās, gan mazās neprecizitātes datos.

Multimodālā pieeja ļauj apvienot informāciju no dažādiem avotiem – gan līdzīgiem sensoriem dažādās vietās, gan pilnīgi atšķirīgiem datu avotiem. Šo pieeju izmantošanas mērķi – datu grupēšana, sakarību meklēšana un lielu datu kopu vienkāršošana. Datus var apvienot trīs līmeņos – sākotnējo datu, to raksturlielumu vai gala rezultātu līmenī. Veiksmīgākā ir līdzīgu datu apvienošana, bet mūsdienu tehnoloģijas spēj labi apvienot arī ļoti atšķirīgus datu veidus.

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.

2. DATU APVIENOŠANAS KLASIFIKĀCIJA

Datu apvienošana ir sarežģīts process, kurā dažādas datu plūsmas tiek apvienotas, lai radītu izvadi, kas ir informatīvāka un pielāgota gala lietotājam vai sistēmai (Bokade et al., 2021). Šajā saistībā datu apvienošanas veidi tiek klasificēti, ņemot vērā apvienošanas sistēmu arhitektūru, abstrakcijas līmeņus un datu apvienošanas mērķus.

Ir vairāki pētījumi (Atluri et al., 2018; Castanedo, 2013; Khaleghi et al., 2013; Y. Zheng, 2015), kuru mērķis ir analizēt datu apvienošanu no dažādiem skatu punktiem. Piemēram, tiek sniegts ieskats (Lau et al., 2019) uz datu apvienošanu viedās pilsētas mērogā, ka arī piedāvāta sava datu apvienošanas klasifikācijas kategoriju koncepcija. *Khaleghi* u. c. (Khaleghi et al., 2013) veic sistemātisku datu apvienošanas metožu izpēti, sniedzot detalizētu analīzi par to konceptuālajiem pamatiem, praktiskajām priekšrocībām un galvenajiem izaicinājumiem. *Castanedo* (Castanedo, 2013) piedāvā visaptverošu pārskatu par datu apvienošanas arhitektūrām un metodēm, īpaši koncentrējoties uz trim būtiskiem algoritmiskiem aspektiem: datu asociāciju, sistēmas stāvokļa novērtējumu un lēmumu apvienošanas procesiem. *Zheng* (Y. Zheng, 2015) izstrādā sistemātisku pieeju dažādu datu avotu apvienošanai, piedāvājot inovatīvas metodes starpdomēnu datu integrācijai un analīzei. *Atluri* u. c. (2018) izstrādā detalizētu metodisko ietvaru telpisko un laika datu ieguves jomai, sniedzot gan teorētisko pamatojumu, gan praktiskos risinājumus. *Qin* u. c. (Qin et al., 2020) sniedz sistemātisku pārskatu par datu apvienošanas algoritmiem *IoT* domēnos, īpaši pievēršoties datu iegūšanas specifiskajām īpašībām. Būtiski atzīmēt, ka visi minētie pētījumu autori nonāk pie diviem svarīgiem secinājumiem: a) pārsvarā tiek identificētas līdzīgas datu apvienošanas arhitektūras un metodes, un b) joprojām trūkst vienotas vispārīgas klasifikācijas, turklāt esošās arhitektūras un metodes ir nepieciešams papildināt vai pielāgot atbilstoši konkrētiem

Parasti datu apvienošanu klasificē pēc abstrakcijas līmeņiem, datu avotu attiecībām un ieejas-izejas attiecībām. Šī klasifikācija ļauj analizēt datu veidu attiecības un apstrādes līmeņus (Becerra et al., 2021; Dasarathy, 1997). Zināmās trīs galvenās datu apvienošanas klasifikācijas metodes ietver *Dasarathy* klasifikāciju (Dasarathy, 1997), *Whyte* klasifikāciju (Grime & Durrant-Whyte, 1994) un *JDL* modeļa klasifikāciju (Steinberg & Bowman, 2008). Bieži tiek pieminēta arī arhitektūrā balstīta klasifikācija (Castanedo, 2013).

2.1. Klasifikācija pēc arhitektūras principiem

Datu apvienošanas sistēmu izstrādes procesā viens no galvenajiem jautājumiem ir, kur precīzi jeb kurā sistēmas daļā notiks datu apvienošana.

Analizējot datu apvienošanas arhitektūras principus, var identificēt četrus galvenos veidus, katram no tiem piemīt savas raksturīgās iezīmes un pielietojuma scenāriji.

Centralizētā arhitektūra izmanto vienotu centrālo procesoru, kas saņem un apstrādā visu sensoru mērījumus, izmantojot pakešu metodi. Lai gan šī pieeja teorētiski var nodrošināt optimālu rezultātu gadījumos ar minimālu aizturi, praksē tā bieži saskaras ar būtiskiem ierobežojumiem. Īpaši nozīmīgas problēmas rodas vizuālo sensoru tīklos, kur nepieciešama liela datu caurlaidība un notiek intensīva datu pārraide.

Decentralizētajā arhitektūrā katrs sensors darbojas kā autonoma apstrādes vienība, veicot lokālo datu apvienošanu ar informāciju no citiem sensoriem. Tomēr, kā norāda *Grime* un *Durrant-Whyte* (Grime & Durrant-Whyte, 1994) raksturīgas augošas komunikācijas izmaksas – $O(n^2)$ katrā saziņas posmā, kur n ir sensoru skaits. Tas rada nopietnus mērogojamības izaicinājumus, īpaši situācijās ar intensīvu sensoru savstarpējo komunikāciju.

Sadalītā arhitektūra piedāvā līdzsvarotāku pieeju – katrs avota mezgls vispirms patstāvīgi apstrādā savus mērījumus un tikai pēc tam nosūta informāciju centrālajam apvienošanas mezglam. Šī pieeja izceļas ar savu elastību, piedāvājot dažādas konfigurācijas iespējas – no vienkārša viena apvienošanas mezgla līdz sarežģītām vairāku starpposma mezglu sistēmām.

Hierarhiskā arhitektūra apvieno iepriekšējo pieeju elementus, veidojot daudzlīmeņu struktūru ar datu apvienošanas procesiem dažādos hierarhijas līmeņos. *Castanedo* (Castanedo, 2013) uzsver, ka šāda arhitektūra nodrošina elastīgu un pielāgojamu risinājumu dažādu sistēmu vajadzībām.

Detalizēti pētot dažādās arhitektūras, atklājas vairākas būtiskas atšķirības to praktiskajā lietojumā. *Castanedo* (Castanedo, 2013) īpaši uzsver decentralizēto sistēmu ieviešanas izaicinājumus, kas galvenokārt saistīti ar datortīklu noslodzi un skaitļošanas jaudas prasībām. Interesanti, ka *Beddar-Wiesing* un *Bieshaar* (Beddar-Wiesing & Bieshaar, 2020) savā pētījumā norāda – neskatoties uz šiem izaicinājumiem, tieši decentralizētā pieeja bieži ir izvēles pamatā, pateicoties tās augstajai noturībai un uzticamībai. Datu apstrādes sistēmu izveide prasa rūpīgu pieeju arhitektūras izvēlē. Katrai sistēmai nepieciešams individuāls risinājums, ņemot vērā tās tehniskās iespējas un darbības specifiku. Būtiska loma ir arī resursiem – gan skaitļošanas jaudai, gan tīkla kapacitātei.

Arhitektūru salīdzinājums atklāj vairākas nozīmīgas atšķirības to darbības principos. Sadalītās sistēmas raksturo secīga datu apstrāde, kur katra komponente veic noteiktu uzdevumu kopējā datu analīzes procesā. Šāda pieeja ļauj efektīvi sadalīt resursus starp sistēmas mezgliem. Decentralizētā pieeja piedāvā atšķirīgu skatījumu uz datu apstrādi. Tās pamatā ir autonomu mezglu tīkls, kur katrs elements spēj veikt pilnvērtīgu datu analīzi. Šī arhitektūra izmanto specifiskus matemātiskos modeļus un informatīvos mērījumus, kas nodrošina efektīvu datu apstrādi lokālā līmenī. Īpaša uzmanība tiek pievērsta datu plūsmas organizācijai. Sadalītās sistēmas koncentrējas uz pakāpenisku datu apstrādi, izmantojot vairākus starpposmus. Savukārt decentralizētā pieeja akcentē tiešu informācijas apmaiņu starp mezgliem, kas ļauj ātrāk reaģēt uz izmaiņām sistēmā.

Sensoru datu apvienošanas metodoloģijā iezīmējas vairāki būtiski virzieni. *Becerra* u. c. (Becerra et al., 2021) piedāvā trīs galvenos sensoru datu apvienošanas scenārijus, kas balstīti uz *Whyte* arhitektūras principiem.

Pirmais scenārijs – konkurētspējīgā sensoru apvienošana – paredz vienas modalitātes sensoru datu integrāciju vai sensoru datu transformāciju uz vienotu bāzes līmeni. Šī pieeja ir īpaši efektīva trokšņu un nenoteiktības mazināšanai datu analīzē. Piemēram, objektu izsekošanas sistēmās vairāku kameru attēlu vienlaicīga apstrāde ļauj būtiski uzlabot prognožu precizitāti, vienlaikus samazinot trokšņu ietekmi uz rezultātiem. Otrais scenārijs ietver papildinošo sensoru apvienošanu, kur dažādi sensori sniedz atšķirīgus, bet savstarpēji papildinošus datus par vienu un to pašu notikumu. Šāda pieeja ļauj veidot pilnīgāku notikuma raksturojumu. Labs piemērs ir vairāku kameru sistēma, kas nodrošina paplašinātu notikuma redzamību un detalizētāku informāciju, salīdzinot ar vienas kameras ierobežoto skatu. Trešais scenārijs – kooperatīvā sensoru apvienošana – paredz sensoru konfigurācijas dinamisko pielāgošanu, balstoties uz citu sensoru sniegtajiem datiem. Lai gan šī pieeja rada zināmu laika aizturi un prasa ekspertu iesaisti, tā nodrošina precīzākus rezultātus. Piemēram, video novērošanas sistēmās kameru leņķu automātiska korekcija, reaģējot uz novērojamā objekta kustību.

Datu apvienošanas metodoloģiju klasifikācijā pastāv vairākas pieejas. *Castanedo* (Castanedo, 2013) izvirza trīs galvenās kategorijas: datu asociāciju, stāvokļa novērtējumu un lēmumu apvienošanu. Šis iedalījums balstās uz vairākiem kritērijiem, ieskaitot datu avotu savstarpējās sakarības, datu tipus un to organizāciju. Alternatīvu skatījumu piedāvā *Luo* u. c. (R. C. Luo et al., 2002) un *Zheng* (Y. Zheng, 2015), klasificējot datu apvienošanu pēc pazīmju līmeņa, apstrādes stadijām un semantiskās nozīmes. Interesanti, ka *Zheng* piedāvātā stadiju balstītā pieeja sasaucas ar *Castanedo* stāvokļa novērtējuma konceptu – abas metodes paredz secīgu neapstrādātu datu apstrādi atbilstoši datu kopu skaitam. Īpaši nozīmīgi, ka, palielinoties datu kopu vai telpisko slāņu skaitam, pieaug rezultātu kvalitāte, tomēr vienlaikus rodas arī jaunas integrācijas problēmas, kas prasa papildu risinājumus.

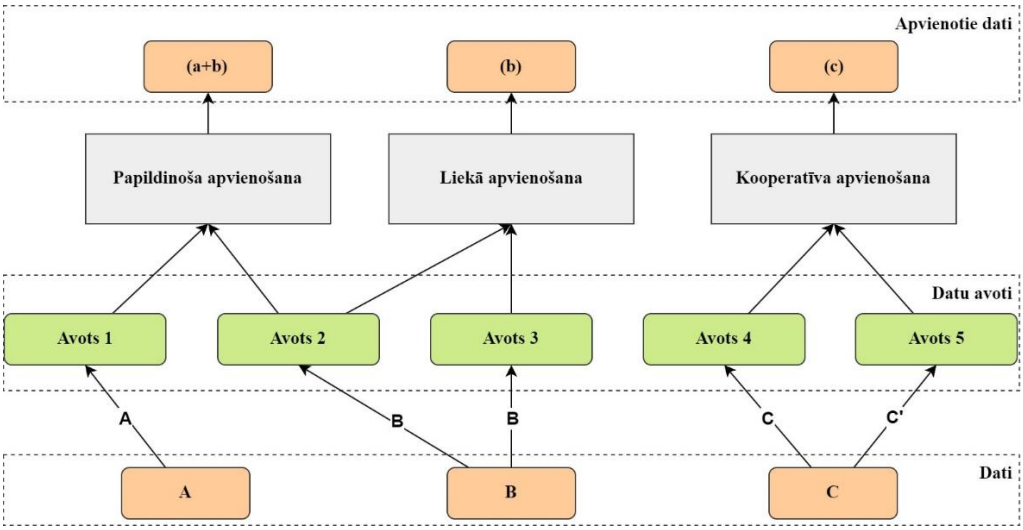
2.2. Klasifikācija pēc attiecībām starp datu avotiem

Whyte izstrādātā klasifikācija (Grime & Durrant-Whyte, 1994) piedāvā strukturētu pieeju datu avotu mijiedarbības analīzei, izdalot trīs galvenos sadarbības veidus sensoru sistēmās (sk. 2.1. att.).

Papildinošā mijiedarbība raksturo gadījumus, kad dažādi sensori sniedz atšķirīgus, bet savstarpēji papildinošus datus par novērojamo objektu vai vidi. Piemēram, vides monitoringa sistēmā temperatūras, mitruma un gaisa kvalitātes sensori kopā sniedz visaptverošu priekšstatu par telpas mikroklimatu, kur katra sensora dati papildina kopējo ainu.

Datu dublēšanās gadījumā vairāki sensori vienlaikus vēro vienu un to pašu mērķa objektu vai zonu. Šāda pieeja, lai gan šķietami neefektīva, būtiski uzlabo sistēmas kopējo precizitāti un uzticamību. Piemēram, vairāku *3D LiDAR* sensoru izmantošana vienā zonā ļauj precīzāk rekonstruēt objektu ģeometriju un kompensēt atsevišķu sensoru nepilnības.

Kooperatīvā mijiedarbība paredz dažādu sensoru datu integrāciju kvalitatīvi jaunas informācijas iegūšanai. Piemēram, *3D LiDAR* un augstas izšķirtspējas video kameru datu apvienošana ļauj ne tikai noteikt objektu formu un izmērus, bet arī to tekstūru un semantisko nozīmi, kas nav iespējams ar katru sensoru atsevišķi.



2.1. att. **Klasifikācija pēc *Whyte*** (Grime & Durrant-Whyte, 1994).

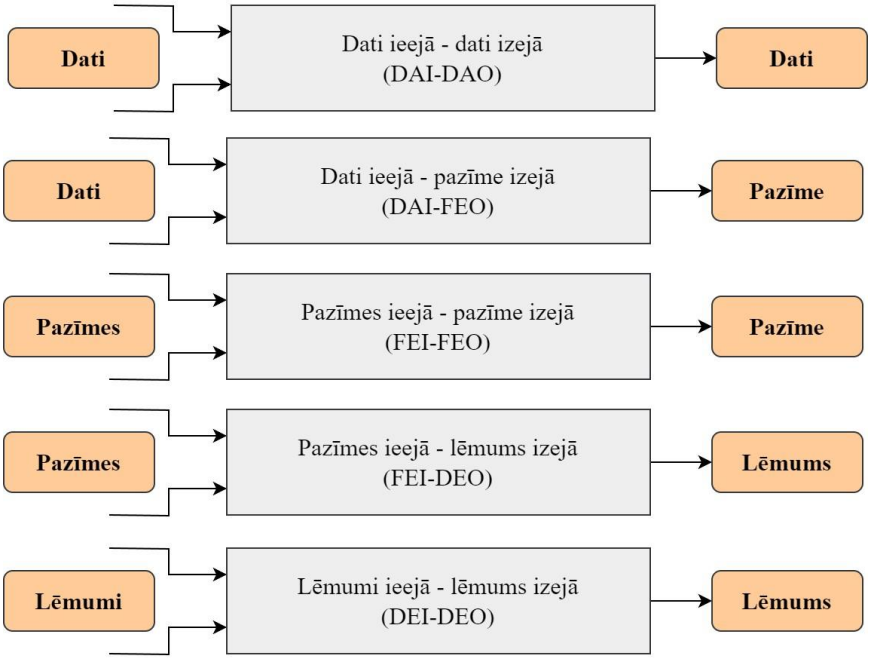
Whyte implementēja šos principus decentralizētā sistēmā, kur katrs sensoru mezgls darbojas autonomi, vienlaikus sadarbojoties ar citiem mezgliem. Izmantojot Kalmana filtrus datu apstrādē, sistēma nodrošina augstu precizitāti bez centralizētas vadības struktūras, kas padara to elastīgu un pielāgojamu dažādiem lietojumiem.

2.3. Klasifikācija pēc *Dasarathy*

Dasarathy (Dasarathy, 1997) izstrādātā klasifikācija ir kļuvusi par vienu no nozīmīgākajām datu apvienošanas jomā, cieši sasaucoties ar *DWIK* principiem. Viņa piedāvātā pieeja strukturē datu apvienošanu piecos hierarhiskos līmeņos (sk. 2.2. att.).

Pirmais līmenis (DAI-DAO) pārstāv vienkāršāko datu apvienošanas formu, kur neapstrādāti sensoru dati tiek tieši apvienoti, izmantojot signālu un attēlu apstrādes metodes. Šī pieeja nodrošina ātru un tiešu datu integrāciju, saglabājot augstu precizitāti, jo dati tiek apstrādāti to sākotnējā formā. Otrais līmenis (DAI-FEO) transformē neapstrādātus datus pazīmju kopā. Šis process ietver datu pārveidi raksturīgās pazīmēs, kas reprezentē būtiskās vides īpašības. Šāda pieeja ļauj efektīvāk apstrādāt un analizēt sarežģītus datus, izdalot no tiem nozīmīgākos raksturlielumus. Trešais līmenis (FEI-FEO) darbojas tikai ar pazīmēm, gan ieejā, gan izejā. Šis process ir īpaši nozīmīgs mūsdienu dziļās mašīnmācīšanās kontekstā. Ceturtais

līmenis (FEI-DEO) pārveido pazīmju kopu lēmumos. Šis līmenis ir raksturīgs klasifikācijas sistēmām, kas, balstoties uz ievades pazīmēm, pieņem konkrētus lēmumus par novērotajiem objektiem vai procesiem. Piektais līmenis (DEI-DEO) darbojas tikai ar lēmumiem, apvienojot vairākus atsevišķus lēmumus vienā galīgajā lēmumā. Šis process neizmanto sākotnējos datus vai pazīmes, bet koncentrējas uz augsta līmeņa lēmumu sintēzi.



2.2. att. **Klasifikācija pēc *Dasarathy*** (Dasarathy, 1997).

Dasarathy klasifikācijas tālākā attīstība iezīmējas ar vairākiem būtiskiem papildinājumiem. *Varshney* (Varshney, 1997) ieviesa DAI-DEO līmeni, kas nodrošina tiešu pāreju no neapstrādātiem datiem uz lēmumu pieņemšanu. Šis process ir īpaši nozīmīgs veidņu atpazīšanas uzdevumos, kur nepieciešama tieša klasifikācija no sākotnējiem datiem.

Beddar-Wiesing un *Bieshaar* (Beddar-Wiesing & Bieshaar, 2020) pētījums identificēja trīs papildu datu apvienošanas līmeņus. FEI-DAO līmenis realizē uz modeļiem balstītu pazīmju transformāciju atpakaļ datos. Šis process ir būtisks Gausa apvienošanas un paraugu ģenerēšanas uzdevumos, kur no pazīmēm nepieciešams rekonstruēt sākotnējos datus. DEI-DAO līmenis nodrošina datu ģenerēšanu no apvienotiem lēmumiem, izmantojot modeļu balstītu pieeju. Šis process ir īpaši noderīgs ekspertu zināšanu apvienošanā un tālākā datu ģenerēšanā. DEI-FEO līmenis transformē apvienotos lēmumus pazīmēs, nodrošinot iespēju iegūt jaunas pazīmes no augsta līmeņa lēmumiem. Šis process ir nozīmīgs situācijās, kur nepieciešams izveidot pazīmju reprezentācijas no pieņemtajiem lēmumiem.

Šie papildinājumi būtiski paplašina sākotnējo *Dasarathy* klasifikāciju, nodrošinot pilnīgāku datu apvienošanas procesu aprakstu un ļaujot aptvert plašāku lietojumu spektru. *Dasarathy* klasifikācijas nozīmīgākais devums ir sistematizēta pieeja datu apstrādes abstrakcijas līmeņu definēšanai. *Luo* u. c. (R. C. Luo et al., 2002) tālāk attīstīja šo konceptu, identificējot četrus galvenos abstrakcijas līmeņus datu apvienošanas procesā.

Zemā līmeņa apvienošana strādā ar neapstrādātiem sensoru datiem. Tas ir līdzīgi kā vairāku IoT sensoru datu plūsmu apvienošana – piemēram, kad vairāki temperatūras sensori vienā telpā sniedz precīzāku mērījumu, jo tiek samazināta atsevišķu sensoru kļūdu ietekme. Vidējā līmeņa apvienošana transformē sākotnējos datus raksturīgās pazīmēs. Šis process līdzinās attēlu apstrādei datorredzē, kur no pikseļu matricas tiek izdalītas augstāka līmeņa pazīmes – malas, formas vai tekstūras, kas tālāk tiek izmantotas objektu atpazīšanai. Augstā līmeņa apvienošana izmanto Beijesa metodes lēmumu pieņemšanai. Tas ir līdzīgi kā anomāliju detektēšanas sistēmās, kur vairāki klasifikatori kopīgi lemj par potenciāliem drošības incidentiem tīklā, balstoties uz iepriekš apgūtiem modeļiem. Vairāku līmeņu apvienošana

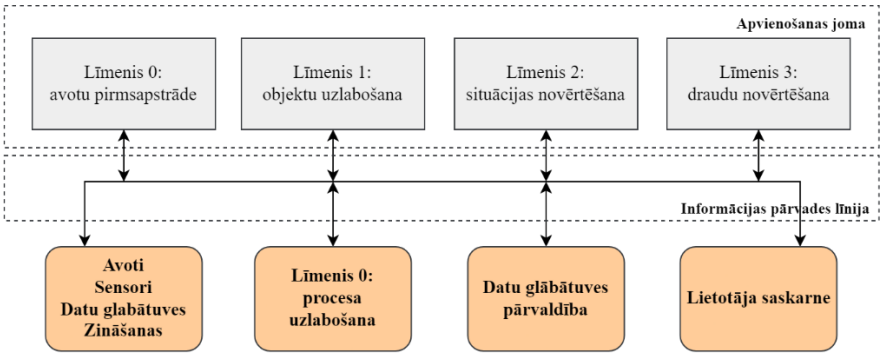
integrē informāciju no visiem iepriekšējiem līmeņiem. Šī pieeja tiek plaši izmantota autonomajās sistēmās, kur, piemēram, robotam jāapvieno gan tiešie sensoru dati, gan atpazītie objekti, gan augsta līmeņa navigācijas lēmumi.

Šī hierarhiskā pieeja ne tikai palīdz sistematizēt datu apstrādi, bet arī nodrošina teorētisku pamatu modernu sensoru tīklu un mākslīgā intelekta sistēmu izstrādei, ļaujot pielāgot apstrādes metodes konkrētā uzdevuma vajadzībām.

2.4. Klasifikācija pēc datu apvienošanas procesa

Analizējot datu apvienošanas procesu klasifikācijas, īpaši jāizceļ Apvienotās direktoru laboratorijas (*JDL*) pieeja, kas joprojām tiek plaši citēta mūsdienu pētījumos (Castanedo, 2013). Kā redzams (sk. 2.3. att.) *JDL* modelis sākotnēji piedāvāja piecu līmeņu hierarhisku struktūru, kur katram līmenim ir specifiska funkcija datu apstrādes ciklā.

JDL modeļa struktūra, ko detalizēti apraksta *Llinas* u. c. (Llinas et al., 2004), sākas ar 0. līmeni, kas veic sensoru mērījumu priekšapstrādi signālu/pikseļu līmenī. Pirmais līmenis fokusējas uz objektu stāvokļu novērtēšanu un prognozēšanu, balstoties uz tiešiem novērojumiem. Otrais līmenis novērtē un prognozē objektu stāvokļus, pamatojoties uz secināto attiecību starp objektiem. Trešais līmenis analizē plānoto vai prognozēto darbību ietekmi uz situāciju, savukārt ceturtais līmenis nodrošina adaptīvu datu iegūšanu un apstrādi resursu pārvaldībai.

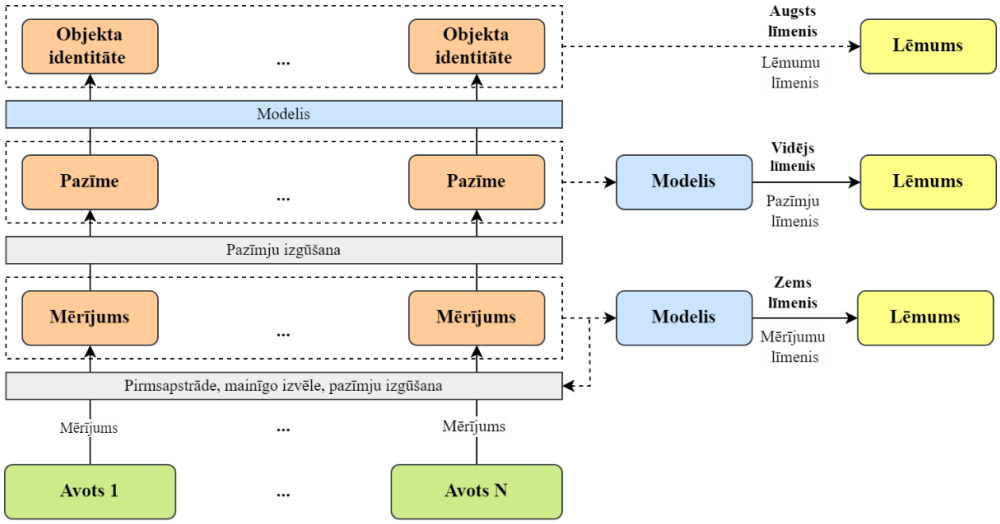


2.3. att. **Klasifikācija pēc *JDL*** (Llinas et al., 2004).

Modeļa galvenie ierobežojumi, ko identificē *Khaleghi* u. c. (Khaleghi et al., 2013), saistās ar tā militāro izcelsmi. Tomēr laika gaitā tas ir ticis papildināts – piemēram, *Blasch* un *Plano* (Blasch & Plano, 2002), ieviesa papildu līmeni lietotāja mijiedarbībai, skaidri nodalot cilvēka un mašīnas lomas sistēmā. Mūsdienu kontekstā *JDL* modeļa evolūcija atspoguļo būtisku paradigmas maiņu – no sākotnējās ieceres aizstāt cilvēka lēmumus ar automatizētiem procesiem līdz mūsdienu pieejai, kur cilvēka loma datu apvienošanas procesā kļūst aizvien nozīmīgāka. *Llinas* u. c. (Llinas et al., 2004) norāda uz nepieciešamību pievērst īpašu uzmanību kvalitātes kontroles mehānismiem, paralēlās apstrādes iespējām, izkliedētajai datu apstrādei un datu uzticamības novērtējumam, kas joprojām ir aktuāli modernajās sistēmās.

2.5. Klasifikācija pēc abstrakcijas līmeņiem

Ļoti bieži tiek pieminēta (Abdelgawad & Bayoumi, 2012; Barbedo, 2022; Borràs et al., 2015) klasifikācija pēc abstrakcijas līmeņiem, kur izdala trīs līmeņus (sk. 2.4. att.).



2.4. att. **Klasifikācija pēc abstrakcijas līmeņiem** (Borràs et al., 2015).

Datu apvienošanas metodoloģijas var tikt strukturētas trīs fundamentālos abstrakcijas līmeņos, katrs no tiem piedāvājot atšķirīgu pieeju informācijas integrācijai un apstrādei.

Zemā līmeņa datu apvienošana operē ar neapstrādātiem sensoru datiem, organizējot tos vienotā matricā. Šajā pieejā katra matrica satur pilnu informāciju par visiem analizētajiem paraugiem (rindās) un mērītajiem signāliem (kolonnās). *Di Natale* u. c. (Di Natale et al., 2002) paplašināja šo konceptu, ieviešot signālu "ārējā produkta" metodi. Šī metode paver jaunas iespējas datu analīzē, ļaujot izmantot gan specializētas daudzu līmeņu metodes, gan klasiskās analīzes pieejas.

Vidējā līmeņa jeb elementu apvienošana, kā to definē *Biancolillo* u. c. (Biancolillo et al., 2014), vispirms veic pazīmju ekstrakciju no katra datu avota atsevišķi. Šajā līmenī būtiska ir spēja atrast līdzsvaru starp datu redukciju un informatīvo komponenšu saglabāšanu. Pētnieki īpašu uzmanību pievērš optimālu metožu kombināciju noteikšanai, kas nodrošina gan precīzu datu reprezentāciju, gan efektīvu tālāko analīzi.

L. Huang u. c. (L. Huang et al., 2014) savā darbā aplūko augstā līmeņa datu apvienošanu, kas balstās uz atsevišķu analīzes modeļu izveidi katram datu avotam. Šī pieeja ļauj precīzāk pielāgot analīzes metodes katra avota specifikai. Pētnieki uzsver, ka šāda stratēģija palīdz norobežot problemātisko datu ietekmi uz kopējo analīzes kvalitāti, vienlaikus saglabājot būtiskās sakarības starp dažādiem datu avotiem.

Datu apvienošanas līmeņi atšķiras pēc to īpašībām un pielietojuma. Zemā līmeņa pieeja strādā ar neapstrādātiem signāliem, kas rada lielāku datu apjomu. Vidējā līmeņa apstrāde samazina datu dimensijas, izmantojot pazīmju ekstrakciju. Augstā līmeņa pieeja balstās uz atsevišķu modeļu izveidi katram datu avotam. Stratēģijas izvēli ietekmē vairāki faktori: nepieciešamā precizitāte, pieejamie skaitļošanas resursi, rezultātu interpretācijas prasības un sistēmas noturība. Katra pieeja piedāvā atšķirīgas iespējas konkrētu uzdevumu risināšanai.

2.6. Datu apvienošanas metožu klasifikācija

Datu apvienošanas metožu klasifikācijā vērojamas dažādas pieejas. Vairāki pētnieki (Dalla Mura et al., 2015; Ding et al., 2019; Pires et al., 2016) pētījumos iezīmē trīs līmeņu struktūru. Pirmajā līmenī notiek tiešs darbs ar neapstrādātiem datiem, kur galvenā uzmanība tiek pievērsta datu elementu integrācijai. Otrajā līmenī process koncentrējas uz raksturlielumu ekstrakciju, kas ļauj identificēt un apvienot būtiskākās datu pazīmes. Trešais līmenis paredz atsevišķu analīzes modeļu izveidi katram datu avotam.

Ouhami u. c. (Ouhami et al., 2021) piedāvā atšķirīgu skatījumu, klasificējot metodes pēc to teorētiskā pamatojuma. Viņu pieeja izdala trīs galvenos virzienus: varbūtību teorijā balstītās metodes, pierādījumu teorijas metodes un zināšanu bāzēs balstītās metodes. Šāds dalījums ļauj

labāk izprast katras metodes teorētiskos pamatus. Šāds iedalījums sniedz papildu perspektīvu metožu izvēlē atkarībā no pieejamo datu un uzdevuma rakstura. Apskatot, piemēram, uz varbūtību balstītas datu apvienošanas metodes, var uzskaitīt:

- Beijesa secinājumus (*Bayesian inference*) (Pans*IoT* et al., 2007);
- Kalmana filtru (Y. Yin et al., 2023);
- stāvokļa telpas modeļus (*state-space models*) (Becker & Neumann, 2022),
- Markova modeļus (*Markov models*) (X. Chang et al., 2019);
- uzskatu izplatīšanas metodes (*belief propagation*) (Gaglione et al., 2018);
- maksimālās varbūtības metodes (*maximum likelihood methods*) (Sur & Candès, 2019);
- iespēju teoriju (*possibility theory*) (Dubois & Prade, 2000);
- pierādījumu spriešanu (*evidential reasoning*) (Xiao, 2020);
- pierādījumu teoriju (*evidence theory*) (Song & Deng, 2019);
- mazāko kvadrātu novērtēšanas metodes (*least square-based estimation methods*) (R. Yang & Ying, 2023).

Kā piemēru no šīm metodēm autors detalizētāk apskata Beijesa secinājumu, kas ir matemātiska metode, ko izmanto datu apvienošanā, lai apvienotu informāciju no dažādiem avotiem, pieņemtu apzinātus lēmumus vai novērtētu nezināmas vērtības. Tas ir sistemātisks veids, kā atjaunināt uzskatus, pamatojoties uz jauniem datiem, un rīkoties ar nenoteiktību.

Beijesa metodoloģija balstās uz sistemātisku pieeju varbūtību atjaunināšanai, ņemot vērā jaunus pierādījumus. Šī pieeja sākas ar sākotnējo varbūtības sadalījumu $P(\theta)$, kas reprezentē mūsu sākotnējos pieņēmumus par pētāmo parametru θ. Šis sadalījums kalpo par atskaites punktu turpmākai analīzei. Datu apstrādes procesā katra informācijas avota ieguldījums tiek izteikts caur iespējamības funkciju $P(Data \mid \theta)$, Šī funkcija kvantitatīvi novērtē, cik labi novērotie dati atbilst dažādām parametra θ vērtībām. Pēc jaunu datu iegūšanas notiek uzskatu atjaunināšana, izmantojot Beijesa teorēmu (sk. (2.1.)), kas matemātiski apraksta, kā apvienot sākotnējo informāciju ar jaunajiem datiem.

Beijesa teorēmas pamatvienādojums:

$$P(\theta|\text{Data}) = \frac{P(Data|\theta) \cdot P(\theta)}{P(Data)} \tag{2.1.}$$

kur

P(θ | Data) – aizmugurējā varbūtība – parametra θ stāvoklis pēc datu apstrādes, %;
P(Data | θ) – iespējamības funkcija – datu atbilstības vērtība pie konkrētām θ vērtībām, %;
P(θ) – iepriekšējā varbūtība – sākotnējais parametra θ stāvoklis, %;
P(Data) – normalizācijas konstante – kopējā datu varbūtība, %.

Datu apvienošanā ir vairāki datu avoti ar to iespējamības funkcijām un dažreiz atšķirīgiem sākotnējiem uzskatiem. Lai apvienotu šos informācijas avotus, tiek izmantots Beijesa secinājums:

- katram datu avotam metode aprēķina aizmugurējo varbūtības sadalījumu θ, izmantojot Beijesa teorēmu;
- pēc tam šie atsevišķie aizmugurējie sadalījumi tiek apvienoti. To var izdarīt, izmantojot svērto vidējo noteikšanu vai citas apvienošanas metodes, nosakot svarus, pamatojoties uz katra avota uzticamību;
- rezultāts ir apvienots aizmugurējais sadalījums, kas atspoguļo atjaunināto uzskatu par θ, ņemot vērā visus datu avotus.

Beijesa secinājumi sniedz spēcīgu instrumentu nenoteiktības kvantitatīvai novērtēšanai, kur aizmugurējais (*posterior*) sadalījums nodrošina ne tikai parametra θ punkta novērtējumu, bet arī tā ticamības intervālu. Šī īpašība ir īpaši vērtīga, kad strādājam ar datu avotiem, kuriem ir atšķirīgi uzticamības līmeņi.

Pires u. c. (Pires et al., 2016) veiktais pētījums atklāj vairākas fundamentālas priekšrocības datu sapludināšanas metodēm. Pētnieki demonstrē šo metožu spēju nodrošināt sistemātisku modeļu validāciju un veikt klasifikāciju bez tiešas cilvēka iejaukšanās. Izvēlētie algoritmi, darbojoties kontrolētos apstākļos, efektīvi novērtē mainīgo stāvokļus, vienlaikus samazinot kļūdas apvienotajos novērtējumos. Īpaši nozīmīga ir sistēmas spēja pielāgoties pieaugošam datu apjomam, saglabājot nemainīgu bāzes struktūru. Metodes nodrošina apvienotas kovariācijas matrices izveidi, kas precīzāk atspoguļo paredzamās novirzes aizmugurējā sadalījumā.

Tomēr praktiskajā pielietojumā atklājas vairāki būtiski ierobežojumi. Viens no centrālajiem izaicinājumiem ir precīzu blīvuma funkciju iegūšana un prioritāšu varbūtību definēšana datu sapludināšanas metodēs, īpaši Beijesa secinājumos un maksimālās varbūtības pieejās. Šo problēmu pastiprina fakts, ka reālās sistēmās bieži trūkst precīzu sākotnējo varbūtību, sevišķi darbā ar jauniem vai ierobežotiem datu avotiem.

Tradicionālo datu apstrādes metožu ierobežojumi kļūst īpaši redzami iIoT sistēmu kontekstā. Kalmana filtri un stāvokļu telpas modeļi, kas lieliski darbojas vienkāršās lineārās sistēmās, zaudē efektivitāti sarežģītākos scenārijos. Viedā transporta sistēmās, piemēram, vienlaicīgi tiek apstrādāti dati par transportlīdzekļu kustību, satiksmes intensitāti un vides apstākļiem – situācija, kas būtiski pārsniedz klasisko metožu iespējas.

Nenoteiktības pārvaldības jautājums izvirza papildu izaicinājumus. Lai gan iespēju teorija un pierādījumu teorija piedāvā teorētiskus risinājumus, to praktiskā implementācija saskaras ar nopietnām grūtībām. Mūsdienu pētnieki meklē risinājumus, apvienojot tradicionālās pieejas ar jaunākajiem sasniegumiem dziļajā mašīnmācīšanā, izmantojot tādas inovatīvas metodes kā dziļās mācīšanās Kalmana filtri vai pastiprinātās mašīnmācīšanās algoritmi.

Īpašu uzmanību pelna Beijesa secinājumu metožu implementācijas izaicinājumi. Neprecīzi definētas sākotnējās varbūtības var radīt sistemātiskas nobīdes parametru novērtējumos, īpaši situācijās ar ierobežotu novērojumu skaitu. Decentralizētās sistēmās šī problēma kļūst vēl aktuālāka, jo kļūdas var kumulatīvi ietekmēt visu sistēmu. Līdzīgas problēmas novērojamas arī Markova modeļos un uzskatu izplatīšanas metodēs.

Visbeidzot, liela mēroga *IoT* sistēmās parādās fundamentāli skaitļošanas ierobežojumi. Matricu inversijas operāciju sarežģītība pieaug kubiskā proporcijā, bet atmiņas prasības – kvadrātiski līdz ar sistēmas izmēru. Kovariācijas matricu precīza novērtēšana kļūst arvien sarežģītāka, jo sensoru mērījumu precizitāte mainās atkarībā no vides apstākļiem, datu ievākšana nav vienmērīga, un dažādu sensoru mērījumiem ir atšķirīgas skalas. Pētnieku centieni izstrādāt adaptīvas novērtēšanas pieejas turpinās, taču to praktiskā pielietojamība joprojām ir ierobežota sistēmās ar augstu datu mainīgumu.

Medjahed u. c. (Medjahed et al., 2011) pētījums identificē vairākus būtiskus varbūtību apvienošanas algoritmu trūkumus: grūtības iegūt precīzas blīvuma funkcijas un definēt prioritāšu varbūtības, ierobežotu veiktspēju darbā ar daudzfaktoru datiem, kā arī nepietiekamu spēju apstrādāt nenoteiktību. *Pires* u. c. (Pires et al., 2016) analīze papildus norāda uz metodoloģiskiem ierobežojumiem – nepieciešamību pēc sākotnējām varbūtību zināšanām, kas bieži nav pieejamas reālās situācijās, klasifikācijas rezultātu atkarību no sākuma stāvokļa un ierobežoto pielietojamību liela mēroga sistēmās. Īpaši kritiski ir jautājumi par nepieciešamajām sākotnējām zināšanām par nenoteiktību kovariācijas matricēm, kas saistītas ar sistēmas modeli un tā mērījumiem.

Varbūtību balstītās metodes, neskatoties uz to ierobežojumiem, joprojām veido datu apvienošanas pamatu. Varbūtiskā datu asociācija (*PDA*) un tās paplašinātā versija – apvienotā varbūtiskā datu asociācija (*JPDA*) – ir plaši izmantotas objektu izsekošanā. *Rezatofighi* u. c. (Rezatofighi et al., 2015) norāda uz *JPDA* priekšrocībām vairāku objektu izsekošanā, tomēr *Alam* u. c. (Alam et al., 2016) atklāj gan *PDA*, gan *JPDA* neobjektivitātes problēmas. *PDA* algoritmam piemīt izteiktāka neobjektivitāte, savukārt *JPDA* demonstrē specifiskas novirzes kontrolētā vidē – gan noraidīšanas, gan apvienošanas kontekstā. Vairāku hipotēžu izsekošanas (MHT) algoritms uzrāda augstāku efektivitāti nekā PDA, ko *Kennedy* (Kennedy, 2008) saista

ar uzlaboto iespējamības novērtējumu. Tomēr MHT piemīt būtisks ierobežojums – eksponenciāli pieaugošas atmiņas prasības līdz ar apstrādājamo video kadru skaita palielināšanos. *Barbedo* (Barbedo, 2022) uzsver, ka datu apvienošanas kontekstā izmantotās regresijas metodes pamatā ir lineāras, mēģinot pielāgot funkciju novērotajiem datiem turpmāko vērtību prognozēšanai. Šāda pieeja ir efektīva tikai gadījumos, kad datu avoti ir homogēni vai nodrošina augstu savstarpējo saderību.

2.7. Klasifikācijas nodaļas kopsavilkums

Šajā nodaļā tiek apskatīta datu apvienošanas klasifikācija un sniegts plašs pārskats par dažādiem apvienošanas veidiem un kritērijiem. Datu apvienošana tiek definēta kā process, kurā no dažādiem avotiem iegūtie dati tiek apvienoti, lai radītu vērtīgāku un noderīgāku rezultātu galalietotājam vai sistēmai. Tiek izskatītas vairākas pieejas, piemēram, *Dasarathy*, *Whyte* un *JDL* modeļi, kas balstās uz arhitektūras principiem, datu avotu attiecībām un apstrādes līmeņiem.

Datu apvienošanas sistēmu arhitektūras būtiski atšķiras pēc to pieejas datu apstrādei. Centralizētajā arhitektūrā visi sensoru savāktie dati nonāk vienā apstrādes centrā, kas nodrošina pilnīgu kontroli pār datu apstrādi un rezultātu kvalitāti. Tomēr šāds risinājums rada būtisku slodzi tīkla infrastruktūrai – jo vairāk sensoru sistēmā, jo lielāks nepieciešamais datu pārraides joslas platums. Decentralizētā pieeja ļauj katram sensora mezglam veikt sākotnējo datu apstrādi uz vietas. Lai gan šāds risinājums ievērojami uzlabo sistēmas mērogojamību, tas rada papildu izaicinājumus – pieaug kopējās komunikācijas izmaksas starp mezgliem un sarežģītāka kļūst datu interpretācijas saskaņošana dažādos sistēmas līmeņos. Sadalītās apvienošanas modelis apvieno abu iepriekšējo pieeju elementus. Dati vispirms tiek apstrādāti katrā sensora mezglā, un tikai pēc tam tiek nosūtīti uz centrālo apvienošanas mezglu galīgajai apstrādei. Šāda pieeja samazina kopējo tīkla noslodzi un optimizē sistēmas resursu izmantošanu, vienlaikus saglabājot centralizētas kontroles iespējas. Hierarhiskā arhitektūra apvieno dažādu līmeņu apvienošanas pieejas, nodrošinot lielāku elastību un precizitāti. Svarīgi ir arī dažādi datu apstrādes līmeņi. Zemā līmeņa apvienošanā neapstrādātie dati tiek apstrādāti tieši pēc to iegūšanas no sensoriem, nodrošinot sākotnējo datu precizitāti. Vidējā un augstā līmeņa apvienošana pārvērš datus pazīmēs vai lēmumos, kas palīdz labāk interpretēt informāciju. Datu apvienošana tiek klasificēta arī pēc datu avotu attiecībām. Papildinošā apvienošana apvieno dažādu sensoru sniegtos datus, lai iegūtu pilnīgāku ainas attēlojumu, savukārt liekā apvienošana apstrādā līdzīgu informāciju no dažādiem avotiem, lai uzlabotu ticamību. Kooperatīvā apvienošana ļauj sensoriem sadarboties, lai iegūtu precīzākus datus. Nodaļā ir apskatītas arī tādas metodes kā Beijesa secinājumi, Kalmana filtri un Markova modeļi, kas sevišķi noder, lai apstrādātu nepilnīgus vai trokšņainus datus, īpaši viedajā transportā un precīzās lauksaimniecības jomā, kur efektīva datu apvienošana uzlabo lēmumu pieņemšanu.

Pētījuma gaitā kļuva skaidrs, ka universāla datu apvienošanas metode nepastāv. Katrs risinājums – centralizēts, decentralizēts vai sadalīts – ir piemērots specifiskiem uzdevumiem un situācijām. Būtiska loma ir arī datu abstrakcijas līmeņiem, kas nosaka apstrādes detalizāciju un rezultātu specifiku. Vienkāršākos gadījumos, kur ir precīzi dati, pietiek ar zemā līmeņa apvienošanu, savukārt sarežģītāku datu gadījumā efektīvāka ir augstāka līmeņa pieeja. Prakse rāda, ka metožu izvēle ir jāpielāgo konkrētajai situācijai, ņemot vērā gan datu struktūru, gan apstrādes vajadzības. Īpaši svarīgi tas ir *IoT* sistēmās, kur bieži ir nepilnīgi vai neprecīzi dati. Mūsdienīgās datu apvienošanas metodes spēj ievērojami uzlabot šādu datu kvalitāti, tādējādi paaugstinot uz tiem balstīto lēmumu ticamību. Nodaļā aplūkotie piemēri apliecina datu apvienošanas un kvalitātes uzlabošanas metožu nozīmi *IoT* sistēmu kontekstā. Elastīga arhitektūra, pārdomāta datu apstrāde dažādos abstrakcijas līmeņos un efektīva noviržu identificēšana ļauj pielāgot šīs metodes dažādām vajadzībām, nodrošinot uzticamus rezultātus.

3. DATU APVIENOŠANAS AKTUALITĀTES IDENTIFICĒŠANA

Nodaļas mērķis ir pamatot risināmo problemātiku, ar kuras palīdzību tiek izstrādāti nepieciešamie risinājumi datu apstrādes un apvienošanas jomā, kā arī demonstrēt daļu no promocijas darba rezultātiem.

Autors ir veicis padziļinātu izpēti par datu vākšanas un apstrādes procesiem, kas ir kritiski nozīmīgi globālu problēmu risināšanā, tostarp bišu saimju uzraudzībā. Īpaši izcelta ir datu apvienošanas nozīme, kas līdz šim biškopības nozarē nav plaši pielietota, taču var palīdzēt uzlabot bišu veselības uzraudzību un prognozēt būtiskus bišu saimes stāvokļus, piemēram, spietošanu un saimes sabrukšanas sindromu. Autors analizē dažādus datu līmeņus (dravas, bišu saimju un individuālā līmenī) un piedāvā vairākus risinājumus, kas balstīti uz sensoru datiem un to apstrādi, lai optimizētu resursu patēriņu un bišu produktivitāti. Analīzē īpaša uzmanība pievērsta datu apvienošanas un prognozēšanas metodēm, kas balstītas uz modernām tehnoloģijām, piemēram, *LSTM* neironu tīkliem un *IoT* ierīcēm.

Viedās transporta sistēmās autors ir ieguldījis darbu sinhronizācijas algoritma izveidē, lai apvienotu *LiDAR* un videokameru datus pilsētas satiksmes novērošanas sistēmā. Viens no galvenajiem uzdevumiem bija izstrādāt sinhronizācijas risinājumu, kas balstās uz laika zīmogu izmantošanu, lai precīzi apvienotu dažādus datu avotus – transportlīdzekļu skenējumus no *LiDAR* un attēlus no videokamerām. Sinhronizācijas algoritms ir izstrādāts, ņemot vērā sarežģītību, ko rada transportlīdzekļu kustība, laikapstākļi un aparatūras konfigurācija, lai samazinātu kļūdas ātruma noteikšanā un transportlīdzekļu klasifikācijā. Tika izpētītas dažādu sensoru un videokameru sinhronizācijas iespējas, ņemot vērā reālos pilsētas apstākļus, piemēram, apgaismojuma maiņu, lietu un citus ārējos faktorus. Tika izmantoti dažādi paņēmieni, lai apstrādātu un sinhronizētu datus, piemēram, kontrasta izmaiņas videokameru kadros un *LiDAR* punktu mākoņu datus. Izveidotais risinājums tika validēts Jelgavas pilsētā, un rezultāti pierādīja sinhronizācijas algoritma spēju nodrošināt augstu precizitāti (līdz 97% sinhronizācijas ātrumu), kas ir īpaši svarīgi, lai precīzi identificētu transportlīdzekļus un novērotu satiksmes pārkāpumus. Šis darbs apliecina autora ieguldījumu precīzas datu apstrādes un apvienošanas tehnoloģijas izveidē, kas spēj risināt pilsētas satiksmes novērošanas izaicinājumus, balstoties uz daudzavotu datiem reāllaika apstākļos.

Putnkopības nozarē autors ir izstrādājis datu glabāšanas risinājumu kā galveno elementu precīzās putnkopības pārvaldībā. Darba gaitā tika izveidota datu noliktava, kas apstrādā datus no dažādiem avotiem – vides monitoringa sensoriem un ražošanas ciklu ierakstiem. Risinājuma galvenais uzdevums bija nodrošināt efektīvu datu apstrādi un glabāšanu, kas atbalsta gan saimniecības pārvaldību, gan atbilstību ES regulām par dzīvnieku labturību un ražošanas standartiem. Tā kā sensori rada ievērojamu datu apjomu, bija nepieciešams centralizēts risinājums informācijas apkopošanai no dažādām vietām. Sistēmai bija jānodrošina piekļuve gan reāllaika, gan vēsturiskajiem datiem, kas ir būtiski ikdienas lēmumu pieņemšanā un ražošanas optimizācijā. Izstrādātais risinājums tika ieviests divās Baltijas putnu fermās, kur tika uzstādītas sensoru sistēmas CO2 un NH3 līmeņa uzraudzībai. Par tehnisko pamatu kalpoja *Microsoft Azure* mākoņdatošanas platforma, kas nodrošināja automātisku datu vākšanu un apstrādi.

Kopumā autora ieguldītais darbs ne tikai identificē un risina būtiskos jautājumus katrā no šīm nozarēm, bet arī demonstrē, kā piedāvātie risinājumi var tikt veiksmīgi ieviesti praksē, tā apliecinot promocijas darba nozīmīgumu un praktisko vērtību.

3.1. Precīzā biškopība

IT risinājumu ieviešana starpnozaru projektos, kas balstās uz viedajām tehnoloģijām, prasa īpašu uzmanību pievērst datu iegūšanas un apstrādes procesiem, kas ir būtiski konkrētu globālu problēmu risināšanai. Piemēram, precīzā biškopība, kas ir inovatīva dravas pārvaldības

stratēģija, nodrošina atsevišķu bišu saimju monitoringu ar mērķi optimizēt resursu patēriņu un uzlabot bišu produktivitāti (Zacepins et al., 2012). Šī stratēģija sastāv no trim galvenajiem posmiem: datu vākšanas, apstrādes un rezultātu analīzes. Datu vākšana ietver informācijas iegūšanu par dažādiem fiziskiem mainīgajiem, kas tieši ietekmē bišu saimes (Meikle & Holst, 2015), piemēram, temperatūru, mitrumu, gāzu sastāvu, vibrāciju un skaņu. Šie dati tiek iegūti ar sensoru palīdzību, kas ir integrēti stropos un pieslēgti centrālajai apstrādes sistēmai (Kviesis & Zacepins, 2015).

Bišu saimju datu apstrādes posms galvenokārt balstās uz statistisko analīzi (Henry et al., 2019), kuras mērķis ir identificēt tādus svarīgus periodus kā perošana un spietošana, kā arī noteikt to sākuma laikus. Savukārt datu izvades posms nodrošina apstrādāto datu vizualizāciju vai tabulas formu, lai biškopji varētu pieņemt pamatotus lēmumus. Vairākos pētījumos (Ferrari et al., 2008; Kviesis & Zacepins, 2015; Zacepins et al., 2015) ir analizēti datu veidi un pieejas šo datu vākšanai, un dažos no tiem piedāvātas arī klasifikācijas sistēmas datu kolekcijas procesam (Human et al., 2013), uzskatot to par neatkarīgu posmu. Tomēr viens no šo pētījumu ierobežojumiem ir fokuss uz konkrētiem fiziskiem mainīgajiem, piemēram, temperatūru vai svaru, nevis uz plašākiem globāliem jautājumiem.

Pašlaik biškopības nozarē datu iegūšana bieži tiek veikta ar bezvadu tīkla tehnoloģijām (Debauche et al., 2018; Henry et al., 2019). Tomēr nestabilā savienojuma apstākļos tas var radīt datu nepilnības vai neatbilstības. Lai šīs problēmas risinātu, pirms datu apstrādes bieži tiek izmantotas datu apvienošanas metodes. Lai gan datu apvienošana biškopības nozarē vēl nav plaši atzīta kā labā prakse, tā var būt īpaši noderīga tādu problēmu risināšanā kā bišu saimju veselības monitorings un saimes sabrukšanas sindroma agrīna atklāšana.

Datu apvienošanas metodes izvēle ir cieši saistīta ar sākotnējiem ievades datiem, jo neapstrādātu datu kvalitāte būtiski ietekmē gala rezultātu. Sensori vai datu ievade paši par sevi nepadara datu apvienošanu obligātu, taču biškopībā, kā globālā mērķa kontekstā, bieži tiek izmantoti vairāku sensoru sistēmu sniegtie multimodālie dati. Praktiskie pielietojumi ne vienmēr izvirza tik stingras prasības datu apvienošanai kā teorētiskie modeļi, taču datu tipi un to īpašības ir svarīgi faktori. Precīzajā biškopībā neapstrādāto datu veids ir atkarīgs no datu avota, un pašlaik nozarē ir pieejamas dažādas datu avotu variācijas. Tiek identificēti trīs galvenie datu līmeņi: dravas līmenis, bišu saimes līmenis un atsevišķo bišu (individuālais) līmenis (Human et al., 2013; Zacepins & Stalidzans, 2013).

Dravas līmeņa dati aptver meteoroloģiskos datus un video novērojumus. Galvenie meteoroloģiskie parametri, kas tiek izmantoti dravas pārvaldībā, ir vēja ātrums un nokrišņi. Lai iegūtu šos parametrus, dravas pārvaldības programmatūras bieži izmanto trešo pušu meteoroloģiskās stacijas (Rafael Braga et al., 2020). Telpiskie novērojumi ļauj identificēt lauku un kultūraugu veidus (Atluri et al., 2018; Calatayud-Vernich et al., 2019), kas ir bišu barības avoti. Dravas līmeņa datu avoti ietver platleņķa videokameras, vietējās meteoroloģiskās stacijas, publiski pieejamās meteoroloģiskās stacijas un satelītattēlu pakalpojumus.

Saimes līmeņa dati ietver temperatūras, mitruma, svara, skaņas, vibrācijas un video datus. Populārākie parametri šajā līmenī ir temperatūra, mitrums un svars (Meikle & Holst, 2015; Stalidzans & Berzonis, 2013), bet biežākie monitoringa mērķi ir spietošana un bišu saimes bojāeja (Ferrari et al., 2008; Kridi et al., 2014). Šie parametri tiek izmantoti, lai noteiktu stropu stāvokļus, piemēram, bezperu stāvokli, intensīvu peru audzēšanu, spietošanas periodus (pirms un pēc spietošanas), stropa pārkaršanu un bišu saimes nāvi. Skaņas un video dati tiek izmantoti arī, lai noteiktu gaisa un trokšņa piesārņojuma līmeni. Piemēram, skaņas un vibrāciju dati var palīdzēt noteikt stropu stāvokļus, kad nav bišu mātes, ir maz peru vai strops ir pārapdzīvots, kā arī novērot bišu saimes bojāeju (Bencsik et al., 2015). Saimes līmeņa datu avoti ir temperatūras un mitruma sensori, svara mērītāji, skaņas uztvērēji, kā arī mono un daudzspektru videokameras.

Individuālā līmeņa uzraudzība ir vērsta uz atsevišķu bišu novērošanu, piemēram, bišu ienākšanas/iznākšanas no stropa monitoringu (Cunha et al., 2020), inficēto bišu skaita noteikšanu (Bjerge et al., 2019) un bišu aktivitātes līmeņa novērtēšanu (Ngo et al., 2019). Šie

novērojumi galvenokārt tiek veikti, izmantojot mono un daudzspektru videokameras. Atkarībā no datu līmeņa un parametriem pētnieki un dravas pārvaldības sistēmu izstrādātāji izveido arhitektūras risinājumus, kas spēj optimizēt šos procesus (Debauche et al., 2018; Kridi et al., 2014; Murakami et al., 2007; Zacepins et al., 2015). Šie risinājumi ietver: 1) minimālu sensoru izmantošanu, lai samazinātu enerģijas patēriņu; 2) sensoru darba efektivitātes optimizēšanu, izmantojot ieslēgšanas-izslēgšanas ciklus; 3) tīmekļa vai mākoņa bāzes datu krātuves izmantošanu; un 4) mērogojamību, kas nodrošina turpmākus sistēmas jauninājumus. Datu apvienošanas pieejas precīzajā biškopībā nosaka prasības datu kopām, kas tiek izmantotas lietojumprogrammās ar plašāku pielietojumu, piemēram, bišu saimju telpiskās pozicionēšanas un laika apstākļu īstermiņa prognozēšanas uzdevumos.

Telpiskā pozicionēšana attiecas uz bišu dravu optimālo izvietojumu, balstoties uz pieejamām dravas robežām. Efektivitātes mērījums ir medus daudzums, ko bites saražo noteiktā laika posmā. Lai veiktu šo uzdevumu, datu apstrādei nepieciešams izmantot vairākas datu kopas un neapstrādātus datu slāņus datu apvienošanai. Šie slāņi ietver dravas atrašanās vietu, lielumu un robežas, tuvumā esošo lauku veidus un izmērus, pieejamo veģetāciju (tostarp sezonālo ziedēšanu), kā arī pesticīdu lietošanu vai citu kaitīgo ķimikāliju klātbūtni šajās teritorijās. Papildus tam jāņem vērā arī tuvumā esošie un bišu barošanās apvidu šķērsojošie ceļi, kā arī Zemes reljefs. Tāpat svarīgi ir dati par bišu skaitu, kas noteiktā laika posmā ienāk un iziet no dravas, kā arī bišu masas izmaiņas šajā periodā.

Īstermiņa laika apstākļu prognozēšana attiecas uz vēja un nokrišņu prognozēšanu tuvāko divu stundu laikā, lai pārvaldītu bišu aktivitātes un stropu apstākļus. Piemēram, automātiska bišu stropa skrejas aizvēršana vai iekšējās temperatūras regulēšana ir svarīga stropa pārvaldības daļa. *Henessy* u. c. (Hennessy et al., 2020) pētījumā tika pierādīta vēja tiešā un netiešā ietekme uz darba bišu barošanos. Tiešā ietekme izpaužas kā barības savākšanas ātruma samazināšanās, palielinoties vēja ātrumam, savukārt netiešā ietekme rada vilcināšanos bišu pacelšanā no ziediem pēc nektāra savākšanas. Vējš arī spēj pārnest kaitīgās vielas no tuvējiem laukiem uz bišu barības vietām (Gamboa et al., 2020). Nokrišņi tieši ietekmē bišu barošanās ātrumu un to dzīves ilgumu, jo stiprs lietus var bojāt bišu spārnus. *He* u. c. pierādīja (X. J. He et al., 2016), ka bites intensīvāk strādā pirms spēcīgiem nokrišņiem. Lai veiktu šāda veida īstermiņa prognozēšanu, nepieciešamas datu kopas par gaisa mitrumu noteiktā laika periodā, stropa iekšējo un ārējo temperatūru, vēja ātrumu, vēja virzienu un aktuālo laika prognozi.

Datu apvienošana precīzajā biškopībā vēl nav plaši attīstīta un izmantota (Bumanis, 2020). Tomēr pastāv iniciatīvas, kas cenšas integrēt datu apvienošanas risinājumus gan viedā stropa arhitektūrā, gan specifisku bišu stāvokļu prognozēšanā. Piemēram, pētījumā (Kwon, Cho, et al., 2019; Kwon, Kim, et al., 2019) *IoT* ierīce tika izmantota, lai uzraudzītu bišu vides apstākļus un prognozētu spietošanas laiku. Datus autori ieguva no svara, temperatūras, mitruma un skaņas sensoriem, un temperatūras un skaņas datus apvienoja, izmantojot Kalmana filtra algoritmu un ieguves apvienošanas algoritmu, kas balstās uz *LSTM* neironu tīklu, lai prognozētu spietošanas laiku.

3.2. Viedās transporta un uzraudzības sistēmas

Viens no *IoT* pielietojumiem ir pilsētas satiksmes novērošana, izmantojot dažādu videonovērošanas aparatūru kopā ar objektu noteikšanas un izsekošanas algoritmiem (N. Chen & Chen, 2018; Putra & Warnars, 2019). Datu apvienošanai bieži atrod vietu objektu detektēšanas un izsekošanas risinājumos. Viedā autostāvvieta pašlaik ir arī viena no visvairāk attīstītajām *IoT* jomām, ja ņem vērā transporta nozari kopumā. Šajā ziņā tiek veikti dažādi pētījumi, kuru galvenais mērķis ir nodrošināt jaunāko pieejamo stāvvietu statusu, kontrolēt un uzraudzīt dažādu noderīgu stāvvietu informāciju reāllaikā (Luque-Vega et al., 2020).

Objektu noteikšanas un izsekošanas uzdevumi ir izplatīti datorredzes jomā (W. Luo et al., 2021), un kamēr joprojām tiek piedāvāti jauni algoritmi un modeļi (Vu et al., 2022; Xia et al.,

2022), praktiskie pielietojumi, politikas un mērķi ir galvenie ierobežojumi reālajā dzīvē. Piemēram, videonovērošana pašlaik ir galvenā pilsētas mēroga drošības risinājuma un satiksmes novērošanas stratēģija (Agustina & Clavell, 2011). Par vienu no populārākiem sektoriem var nosaukt satiksmes novērošanu un viedā transporta sistēmas (N.-E. El Faouzi et al., 2011; Favalli et al., 1996; Kim & Jeon, 2014). Pēdējos gados pieaug interese par reālajām pielietojumiem saistībā ar gājēju un transportlīdzekļu noteikšanu un izsekošanu. Satiksmes uzraudzības tēma ietilpst arī šajā tendencē, jo bieži tiek piedāvātas dažādas jaunas metodes, pieejas un pielietojumi (Adamová & Boroš, 2021; V. Kumar et al., 2022; Patil et al., 2022; Pawar & Attar, 2022), kur pašreizējā tendence lēnām pāriet no optimālas transportlīdzekļu noteikšanas metodes uz satiksmes anomāliju, negadījumu un satiksmes negadījumu identificēšanu. ceļu satiksmes noteikumu pārkāpumu atklāšana un prognozēšana (Minnikhanov et al., 2020; Y. Zhao et al., 2021).

Mūsdienu transporta sistēmu digitalizācija rada jaunas iespējas un izaicinājumus. Transportlīdzekļos iebūvētās modernās sakaru un skaitļošanas ierīces, kā arī sensoru tehnoloģijas, veido plašu datu avotu tīklu. Šī tehniskā evolūcija rada jaunu izaicinājumu – kā efektīvi apvienot un analizēt datus no dažādiem avotiem, lai uzlabotu transporta sistēmu darbību (Neumann et al., 2016). Inteliģentās transporta sistēmas (*ITS*) izmanto šos datus dažādiem mērķiem. *Cui* u. c. (Cui et al., 2018) pētījums demonstrē, kā sensoru dati var uzlabot satiksmes drošību, īpaši joslu noteikšanā un sadursmju novēršanā. *Radak* u. c. (Radak et al., 2015)) savukārt koncentrējas uz transporta sistēmu efektivitātes paaugstināšanu, analizējot vilcienu kustību un incidentu atklāšanu. *Ahmed* un *Abdel-Aty* (Ahmed & Abdel-Aty, 2013)) pētījums vērsts uz ceļojuma informācijas precīzāku novērtēšanu, kamēr *Liou* u. c. (Liou et al., 2014) piedāvā integrētu pieeju autovadītāju palīg-sistēmu un satiksmes vadības uzlabošanai. Sensoru datu integrācija transporta sistēmās parasti ietver ierobežotu modalitāšu skaitu. Galvenie datu veidi ir video plūsmas, statiskie attēli un viendimensijas sensoru mērījumi – ātrums, satiksmes intensitāte un blīvums. *Meuter* u. c. (Meuter et al., 2011) norāda, ka lielākā daļa lietojumu balstās uz lēmumu līmeņa datu apvienošanu. *T. Liu* u. c. (T. Liu et al., 2021) pētījums apstiprina, ka videokameras un *LiDAR* sensori joprojām ir visbiežāk izmantotās tehnoloģijas transporta uzraudzības sistēmās.

Novērošanas sistēmas veiktspēju nosaka vairāki faktori, t. i., iekšējie faktori – aparatūra un programmatūra, kā arī ārējie faktori, piemēram, apgaismojums, laikapstākļi un šķēršļi. Aparatūru mēdz izvēlēties programmatūras izstrādes sākumposmā atbilstoši ārējo faktoru analīzei (Finogeev et al., 2019; Fu et al., 2022), bet, lai gan tas ir plaši pazīstams (Bahnsen & Moeslund, 2019; Pramanik et al., 2021), ka ārējie faktori ietekmē objektu noteikšanai un izsekošanai nepieciešamo video kvalitāti, lielākā daļa risinājumu tiek izstrādāti, neiekapsulējot visus iespējamos scenārijus (Pramanik et al., 2021). Ārējo faktoru ietekmes pakāpe ir atkarīga no faktora rakstura – nokrišņiem un sniegputenim var būt lielāka ietekme uz veiktspēju, kas ir īslaicīga, savukārt saulrietam var būt objektu noteikšanu traucējošs efekts, kas ilgst aptuveni stundu (Bahnsen & Moeslund, 2019). Turklāt monitoru ierīču izvietojums nosaka potenciāli sasniedzamo mērķu apjomu. Piemēram, *He et al.* (Y. He et al., 2017) ierosināja piecas dažādas videokameru izvietošanas stratēģijas, bet koncentrējās uz maksimālu pārklājuma palielināšanu metropoles krustojumos.

Datus no *LiDAR* un videokamerām var apvienot, mazinot *LiDAR* vai videokameru atsevišķas izmantošanas negatīvās puses (Y. He et al., 2017). Ir dažādi apvienošanas algoritmi, piemēram, Dempstera-Šefera pierādījumu teorija, Beijesa secinājumi, Montekarlo metodes un Kalmana filtrs (tostarp paplašinātais Kalmana filtrs), ko parasti izmanto ar satiksmi saistītiem scenārijiem (N. E. El Faouzi & Klein, 2016). Šie algoritmi ir izstrādāti, ņemot vērā datu apvienošanas problēmas, piemēram, datu nepilnības, pretrunīgus datus, novirzes un viltus ierakstus (N. E. El Faouzi & Klein, 2017), un to ieviešana galvenokārt koncentrējas uz satiksmes prognozēšanu un ceļojuma laika novērtēšanu. Lai gan satiksmes uzraudzībai var nebūt nepieciešama pilnīga šādu algoritmu ieviešana, var izmantot datu zinātnes principus, savukārt pielietojumam reālajā pasaulē joprojām ir nepieciešama pielāgota izstrādes pieeja (Y.

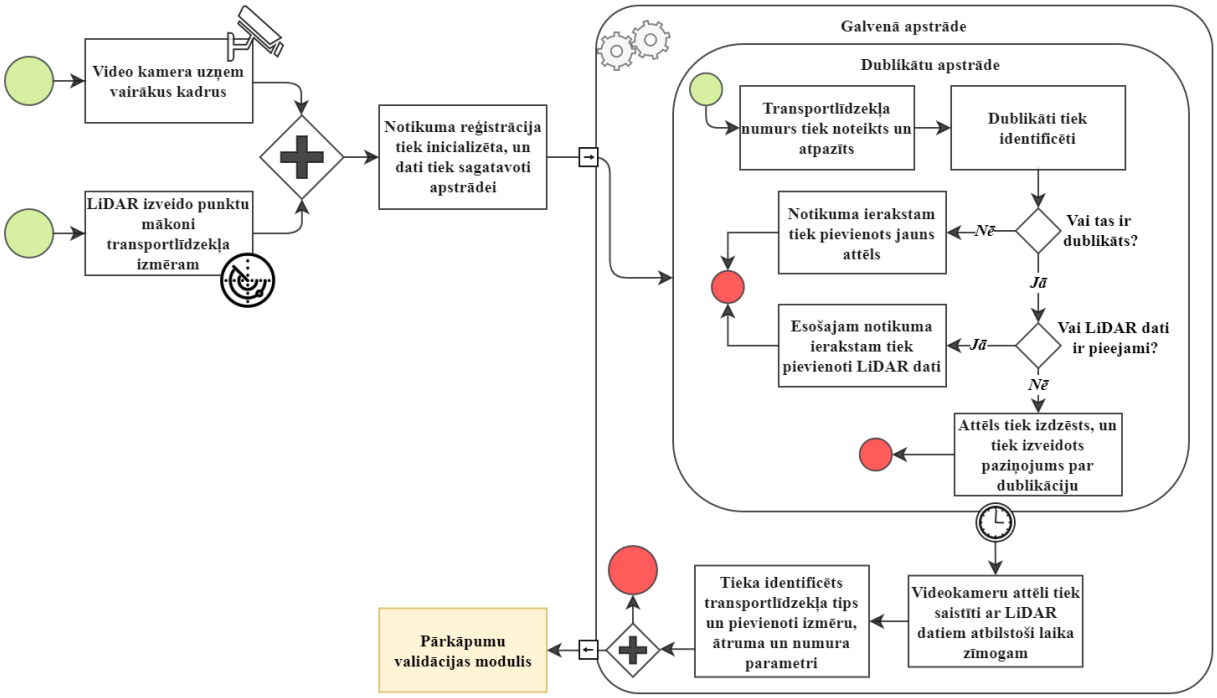
Han & Hu, 2020; Kenk & Hassaballah, 2020). Datu apvienošana starp *LiDAR* un videokamerām ieviešana satiksmes uzraudzības risinājumos bieži tiek veikta (Manogaran et al., 2021) bez pienācīgas ārējo faktoru sākotnējās izpētes, un ir tendence koncentrēties vai nu uz vienu konkrētu kļūdu izraisošu faktoru atsevišķi, vai uz konkrētu uzdevumu. Tipiski *LiDAR* un videokameru datu apvienošanas scenāriji ietver gājēju noteikšanu un autonomu transportlīdzekļu braukšanu. Piemēram, *Wu et al.* (Wu et al., 2018) ierosināja izmantot *LiDAR 3D* punktu mākoņdatus, lai uzlabotu atklātā objekta formas kvalitāti, un *Zhao et al.* (G. Zhao et al., 2014) apvienoja *LiDAR* un video datus, lai radītu vides situācijas autonomiem transportlīdzekļiem. Lai gan šo priekšlikumu mērķis ir uzlabot risinājumu, tie neattiecas uz vispārējo sistēmu un visiem iespējamiem reālas lietošanas gadījumiem.

Tehniski šie risinājumi attiecas uz viedās pilsētas risinājumiem, jo bieži tiek pielietoti tieši apdzīvotās un publiskās vietās (Lau et al., 2019). Apskatot Jelgavas pilsētu (Latvija) kā video testa vidi, tika izstrādāts (Bumanis et al., 2021) un testēts (Bumanis, Vitols, et al., 2022) *LiDAR* un videokameru datu sinhronizācijas risinājums, kas balstās uz sākotnējo datu apvienošanu, izmantojot laika zīmogus par kopējo īpašību, un vēlāk, izmantojot *LiDAR* mākoņdatus videokameru datu pareizības noteikšanai. Risinājuma mērķi iekļauj satiksmes intensitātes fiksēšanu, nelegālās stāvēšanas apzināšanu, transportlīdzekļa gabarītu mērīšanu u. c. Laika zīmogi ir viena no visizplatītākajām metodēm, kuru pielieto sinhronizēšanai starp sensoriem un kamerām ar dažādiem iestatījumiem, ka arī, lai sinhronizētu valkājamo ierīču datus (Y. C. Zhang et al., 2020), sensoru datus robotikā (Tschopp et al., 2020) un autonomas braukšanas automašīnu sensoru datus (Rangesh et al., 2017). Laika sinhronizācijas precizitāte šajos gadījumos ir svarīga, jo var būt tādi ierobežojumi kā nepareizi konfigurētas ierīces, operētājsistēmu aizkavēšanās, signāla piegādes kavējumi utt. Šādos gadījumos var izmantot IEEE 1588 precizitātes laika protokols (*Precision Time Protocol)*(Rangesh et al., 2017).

Risinājuma izstrādei tika veikta dažādu sensoru un kameru sinhronizācijas iespēju izpēte, akcentējoties uz *LiDAR* un videokameru datu izmantošanu. Piemēram, smago transportlīdzekļu autonomās sistēmas bezceļa vidē izmanto kameras, *LiDAR* un radara sensorus, kas nodrošina adekvātu datu sinhronizāciju (Yeong et al., 2020), vai pilsētas vidē, izmantojot hibrīda *LiDAR* kameru datus, kurus apstrādā, lietojot reģionu konvolūcijas neironu tīklus (*Region-based Convolutional Neural Networks, R-CNN)* tīklus un to variantus (piemēram, *Faster R-CNN*), lai iegūtu uzlabotu objektu turpmākai sinhronizācijai (K. Banerjee et al., 2018). Lielākā daļa autonomās braukšanas risinājumu ietver vairāku sensoru sinhronizāciju, parasti kameras, savukārt *LiDAR* un radari – ar dziļu neironu tīklu pievienošanu. Tās specifiskā iestatīšana un tipiskie risinājumi ietver iekšējo un ārējo parametru konfigurēšanu, kā arī īpašu algoritmu izmantošanu gājiena punktu mākoņiem, ko uztver sensori. Situācijā, kad nav būtiski veikt sinhronizāciju reāllaikā, var izmantot citas metodes (Ravindran et al., 2021).

Uzstādīšanas laikā katra kamera tika iestatīta ar noteiktu skata lauku, lai tās darbība būtu ierobežota uz divām joslām. Tika pieņemts, ka vairākas kameras ar pārklājošu skata lauku filmēs dažādos spektros – infrasarkanajā un redzamajā gaismā. Turklāt, palielinot kameru skaitu, varētu tikt atrisināta oklūzijas problēma, kas rodas, ja lielāka izmēra transportlīdzeklis, it īpaši vidēji līdz lieli kravas transportlīdzekļi, aizsedz aiz muguras braucošos transportlīdzekļus. Tika konstatēts, ka līdzīga pieeja ir efektīva vairāku kameru iestatījumos gājēju izsekošanai (Y. Yan et al., 2021), kur tika pārbaudīti divi un trīs pārklājoši kameru redzeslauku iestatījumi un pierādīts, ka tie ir labāki par monokulāro sistēmu. Otrkārt, *LiDAR* ir iestatīts, lai uzraudzītu visas ceļa līnijas, izmantojot tā skata lauka priekšrocības. Sinhronizācija ir nepieciešama starp *LiDAR XML* datiem un vairākiem videokameru nodrošinātajiem attēliem.

Prototipēšanas un izstrādes laikā datu sinhronizācijai tika izmantots notikuma princips, kur notikums tiek definēts kā datubāzes ieraksts ar noteiktu datumu un laiku, saistīto attēlu identifikācijas numuriem un *XML* datiem. Lai reģistrētu notikumu, ir jāveic vismaz viena datu tveršanas un sinhronizācijas iterācija (sk. 3.1. att.).



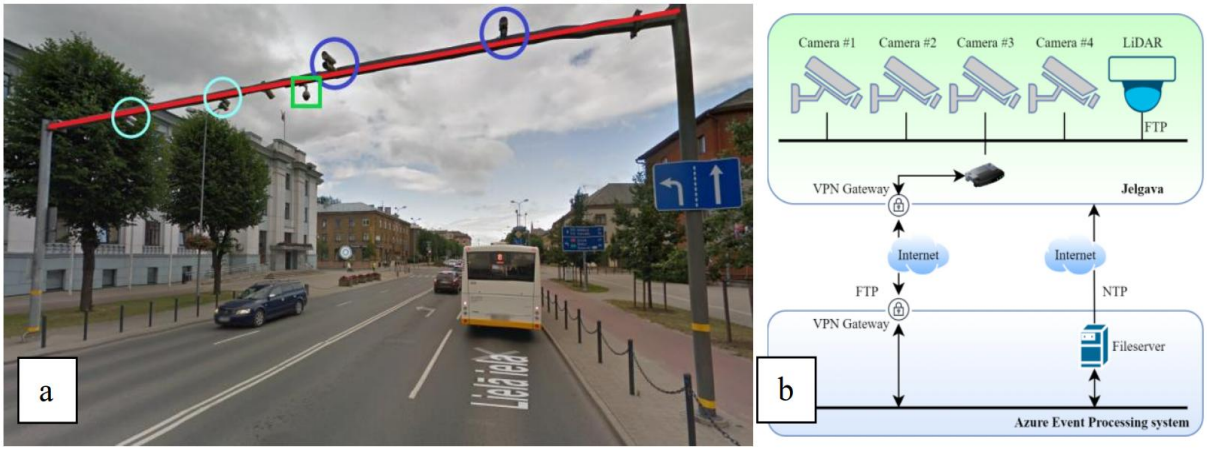
3.1. att. ***LiDAR* un videokameru datu apstrādes process** (Vitols et al., 2021)**.**

Reģistrācija tiek inicializēta tiklīdz vismaz viena videokamera spēj nodrošināt kādus datus. *LiDAR* dati vienmēr tiek pievienoti esošam notikumu ierakstam; taču pats *LiDAR* sensors vienmēr sniedz datus. Konkrēti, katru transportlīdzekli pirms videokameras noteikšanas jau ir fiksējis *LiDAR* sensors, un par to ir atbilstoši dati. Viena no problēmām ir tā, ka dati nav saistīti ar transportlīdzekli. *LiDAR XML* satur datus par visiem transportlīdzekļiem, tāpēc apstrādes mērķis ir anotēt un sinhronizēt (t. i., apvienot) *LiDAR* sensora savāktos datus ar videokameru savāktajiem. Pirmkārt, katra videokamera uzņem kadru, kad tiek konstatētas jebkādas uz kontrastu balstītas izmaiņas, t. i., ir iespējama priekšplāna segmentācija, izmantojot kontrastu. Līdzīga pieeja tika izstrādāta (Fuentes & Velastin, 2001) piedāvātajā metodē, kur tika norādīts, ka šāda pieeja var samazināt apstrādes jaudas vajadzības, jo ir nepieciešams tikai viens fona attēls un nav nepieciešama papildu informācija par pikseļiem. Videokamera fiksē kadrus, kamēr attēlā notiek būtiskas izmaiņas. Rezultātā katram transportlīdzeklim tiek uzņemti vairāki kadri, īpaši joslu maiņas laikā, kad rodas izteiktākas kontrasta izmaiņas. *LiDAR* sensors vienlaicīgi veido punktu mākoņus visiem redzamajiem transportlīdzekļiem. Šie dati tiek sakārtoti pēc to laikspiedoliem un saglabāti *XML* formātā tālākai apstrādei. Atšķirībā no videokameras, *LiDAR* neveido jaunus *XML* ierakstus, kad transportlīdzeklis maina joslas, jo visi transportlīdzekļi jau ir iekļauti sākotnējā punktu mākonī. Kad sistēma konstatē, ka dati ir gatavi eksportēšanai, tiek aktivizēta notikumu reģistrācija un sākas datu sagatavošana apstrādei.

Galvenā apstrāde sākas ar reģistrēto notikumu dublikātu noteikšanu. Pirmkārt, tikko reģistrēta notikuma ieraksta attēls tiek izmantots, lai atpazītu transportlīdzekļa numura zīmi. To var veikt ar atsevišķu numura zīmes atpazīšanas procesu, izmantojot atpazīšanas algoritmu. Otrkārt, cits atpazīšanas algoritms nosaka transportlīdzekļu kustības virzienu. Lai noteiktu, vai notikums ir dublēšanās, tiek izmantoti šādi parametri: transportlīdzekļa kustības virziens, eksportētā attēla datums un laiks (jābūt konfigurējamam, lai optimizētu veiktspēju/precizitāti), transportlīdzekļa numurs. Ja ir kāds notikums, kas atbilst šiem parametriem, proti, ar šo notikumu jau ir saistīts attēls, tiek veikta *LiDAR* datu pieejamības pārbaude. Pieejamie *LiDAR* dati tiek pievienoti esošam ierakstam. Vairāku dublikātu apstrādes rezultāts ir viens notikums ar video un *LiDAR* datiem, kas tiek sinhronizēti atbilstoši laika zīmogiem un iepriekš minētajiem parametriem. Šie dati tiek izmantoti, lai iegūtu šādu informāciju par transportlīdzekli: tā tips, izmēri, ātrums un numura zīme. Pēc apstrādes dati tiek pievienoti esošam notikumam. Pilna informācija par transportlīdzekli tiek pārsūtīta uz pārkāpumu apstrādes moduli. Šis modulis veic validāciju atbilstoši ceļu satiksmes noteikumiem. Ja tiek

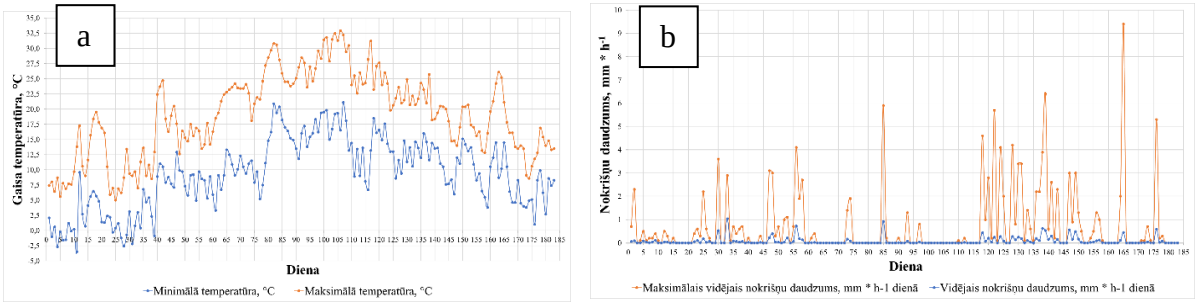
konstatēts pārkāpums, attiecīgs paziņojums tiek nosūtīts vietējai pašvaldības policijas nodaļai. Šī informācija tiek pievienota arī notikumu datiem.

Satiksmes uzraudzības risinājuma validācija tika veikta netālu no Lielās un Pasta ielas krustojuma Jelgavas pilsētā, Latvijā, (56° 39.1115' N, 23° 43.3802' A) (sk. 3.2. att., a) kopumā 6 mēnešus (183 dienas), no 01.04.2021. līdz 30.09.2021. (tas neietver sākotnējo aparatūras instalēšanu un testēšanu, kas notika 2021. gada martā). Divi satiksmes virzieni tika uzraudzīti ar videokamerām un *LiDAR* sensora iestatīšanu, lai tie darbotos 24 stundas diennaktī, 7 dienas nedēļā (tas ietver periodus, kad dati nebija pieejami savienojuma vai strāvas problēmu dēļ).



3.2. att. **Eksperimentālais uzstādījums** (Bumanis, Vitols, et al., 2022)**.**
Apraksts: (a), kur sarkanā līnija – konstrukcijas atbalsta josla; zili apļi – videokameras satiksmes dalībniekiem, kas iebrauc pilsētā un pagriežas pa kreisi; gaiši zili apļi – videokameras satiksmes dalībniekiem, kas izbrauc no pilsētas; zaļais kvadrāts – *LiDAR* sensors; aprīkojuma topoloģijas vizuālais attēlojums (b).

Laikapstākļi eksperimenta periodam tika iegūti, izmantojot SIA "Latvijas Vides, ģeoloģijas un meteoroloģijas centrs" nodrošināto interneta pakalpojumu, kas pieejams https://www.meteo.lv/. Vidējā temperatūra bija 9,3 °C (sk. 3.3. att., a). Maksimālais novērotais nokrišņu daudzums lietus veidā reģistrēts 12.09.2021. ar vērtību 9,4 mm * h^{-1} laika posmā no 17.00 līdz 18.00.

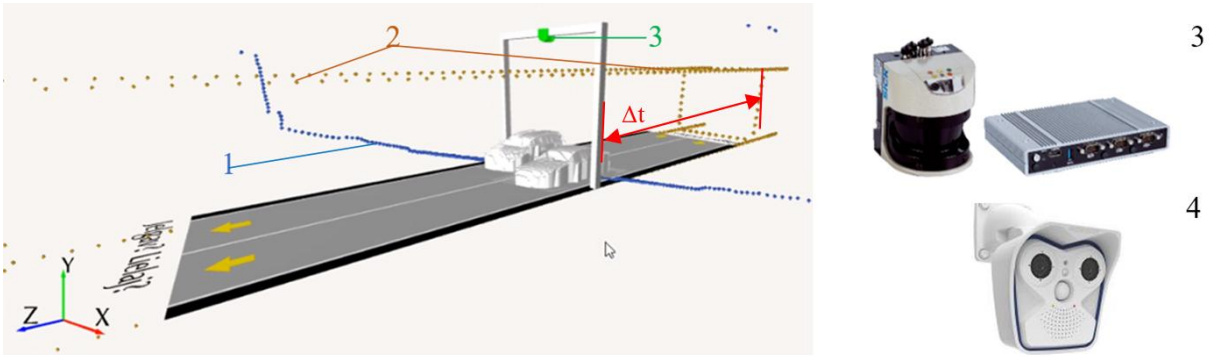


3.3. att. **Laika apstākļi eksperimenta periodā** (Bumanis, Vitols, et al., 2022)**.**
Apraksts: a – minimālā un maksimālā temperatūra, °C dienā; b – maksimālais un vidējais nokrišņu daudzums, mm diennaktī.

Atbilstoši aprobācijas mērķiem un pastāvīgi pilnveidojot konfigurāciju un kalibrēšanu, ilgtermiņa statistiskie dati netika vākti. Īstermiņa un testēšanas dati ietvēra šādu informāciju: transportlīdzekļa ātrums (km·h^{-1}), transportlīdzekļa izmēri – platums, augstums un garums (m), transportlīdzekļa kategorija pēc klasifikācijas (automašīna, automašīna ar piekabi, furgons, pikaps, furgons ar piekabi, kravas automašīna), braukšanas josla (0–4), braukšanas virziens (uz priekšu, pagrieziens pa kreisi, maina joslu). Papildus tika veikta numura zīmes atpazīšana, izmantojot algoritmu, būvētu uz divām metodēm, – *Single Shot Multibox Detector* transportlīdzekļa noteikšanai, numura zīmes noteikšanai un apgriešanai; un Siāmas neironu tīkla numura zīmes identifikācijai (t. i., atsevišķu simbolu noteikšanai, simbolu izvadīšanai teksta virknes veidā un šīs virknes salīdzināšanai ar iepriekš noteiktajiem parametriem).

Transportlīdzekļa numura zīme tika izmantota ceļu satiksmes noteikumu pārkāpumu konstatēšanai.

Ņemot vērā Vispārīgo datu aizsardzības regulu (*General Data Protection Regulation, GDPR)* un autovadītāju privātumu, datu vākšanu atbalstīja juridiska vienošanās ar vietējām iestādēm. Vidējā datu caurlaidspēja bija aptuveni 20 GB dienā, ko nodrošināja *LiDAR* sensorā sistēma (*TIC501* kontrolieris ar *LMS511 2D* sensoru, kas aizsargāts ar laikapstākļu aizsargpārsegu 2063050, visu nodrošina *Sick AG*), kura novietota virs vidējās līnijas, un četras videokameras (četras M16B korpusi, katra ar *Mx -O-SMA-S-6L237* nakts objektīvs un *Mx-O-SMA-S-6D23* dienas objektīvs, visi no *Motobix* un infrasarkanā apgaismojuma M-sērijas 860nm ar *Emitlight*) (sk. 3.4. att.).



3.4. att. **Eksperimenta novērošanas interešu zona** (Bumanis, Vitols, et al., 2022).
Apraksts: 1 – *LiDAR* sensora mērījumu ņemšanas zona; 2 – video kameras redzes lauks; 3 – LiDAR sensors; 4 – videokameras; Δ*t* – aizkavē starp videokameras noteikšanu un *LiDAR* noteikšanu.

Praktiskā eksperimenta ietvaros tika izstrādāta un pārbaudīta hibrīda sensoru sistēma transportlīdzekļu uzraudzībai. Sistēmas pamatā ir divu sensoru veidu – videokameru un *LiDAR* – datu integrācija. Videokameras nodrošināja transportlīdzekļu identifikāciju, savukārt *LiDAR* sniedza precīzus mērījumus par to fiziskajiem parametriem. Normālos braukšanas apstākļos no transportlīdzekļa noteikšanas ar kameru līdz *LiDAR* noteikšanai paiet aptuveni 200 ms līdz 1200 ms (sk. 3.4. att., Δ*t*).

Datu apstrādes process tika organizēts divās paralēlās plūsmās. Videokameru dati tika apstrādāti, izmantojot gan infrasarkanā, gan redzamā spektra attēlus, kas tika glabāti *Microsoft Azure BLOB* krātuvē. *LiDAR* mērījumi tika apkopoti *XML* formātā un uzglabāti *FTP* serverī. Sistēmas drošība tika nodrošināta ar Site-to-site *VPN IKEv2* protokolu.

Datu integrācijas metodoloģija (Vitols et al., 2021) balstās uz pazīmju līmeņa datu apvienošanu. Sistēma katru transportlīdzekļa atpazīšanu apstrādā kā notikumu, kas apvieno vairākus datu elementus: augstākās kvalitātes attēlu ar transportlīdzekļa tipu un numuru, 3D punktu mākoni ar precīziem izmēriem, kustības virzienu un ātrumu, kā arī notikuma metadatus. Transportlīdzekļu korelācija starp abiem sensoru avotiem tiek nodrošināta, izmantojot koordinātu sistēmu, objekta formu un laika zīmogus.

Eksperimentālā validācija tika vērsta uz trim praktiskiem lietojumiem:

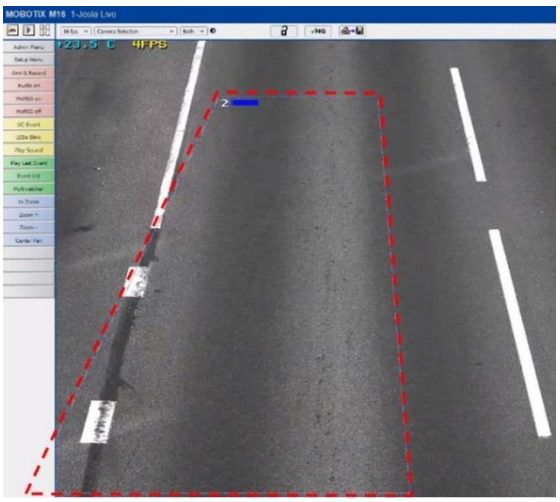
- stāvēšanas pārkāpumu konstatēšana;
- transportlīdzekļu kustības monitorings pēc numura zīmes;
- kravas transporta plūsmas kontrole pilsētā.

Validācijas rezultāti atklāja vairākas būtiskas tendences. Numura zīmju atpazīšanas precizitāte sasniedza augstu līmeni (virs 99%), taču datu sinhronizācijas efektivitāte svārstījās plašā diapazonā (67,52-99,8%). Tika identificēti vairāki ārējie faktori, kas ietekmē sistēmas veiktspēju, tostarp laikapstākļi, diennakts periods un satiksmes intensitāte. *LiDAR* mērījumu analīze atklāja specifiskas problēmas transportlīdzekļu klasifikācijā, īpaši sarežģītām transportlīdzekļu konfigurācijām. Šie novērojumi norāda uz nepieciešamību pēc uzlabotiem klasifikācijas algoritmiem un precīzākas sensoru kalibrācijas.

Sistēmas darbību būtiski ietekmē vides apstākļi, īpaši nokrišņi un apgaismojuma izmaiņas. Videokameru darbības pamatā ir kontrasta izmaiņu analīze, kas var radīt viltus pozitīvus notikumus nelabvēlīgos laikapstākļos. Lietus un sniega gadījumā katra lāse vai pārsla var tikt interpretēta kā potenciāls transportlīdzeklis. Līdzīgas problēmas novērotas arī putnu aktivitātes dēļ sensoru darbības zonā. Šo problēmu risināšanai tika izstrādāta vairāklīmeņu pieeja. Jelgavas eksperimentā tika veikta kameru parametru kalibrēšana, pakāpeniski samazinot kontrasta izmaiņu slieksni no 20% līdz 5%. Rezultātā viltus pozitīvo notikumu skaits stipra lietus laikā samazinājās no 42% (244 no 582) līdz 16% (74 no 461), bet mērena lietus laikā līdz 3% (8 no 279). Diennakts cikla ietekme izpaužas īpaši saullēkta un saulrieta periodos (sk. 3.5. att.). Šajos laikos neizmantojamo attēlu īpatsvars svārstījās no 6% līdz 34% stundas laikā, atkarībā no satiksmes intensitātes. Problēmas rada gan tiešā saules gaisma, kas atstarojas no numura zīmēm, gan slapja vai apledojuša ceļa seguma radītie atspīdumi nakts laikā. Interesanti, ka pilnmēness fāzē ar skaidrām debesīm infrasarkanā gaisma nodrošināja īpaši kvalitatīvus numura zīmju attēlus ar precīzām kontūrām.

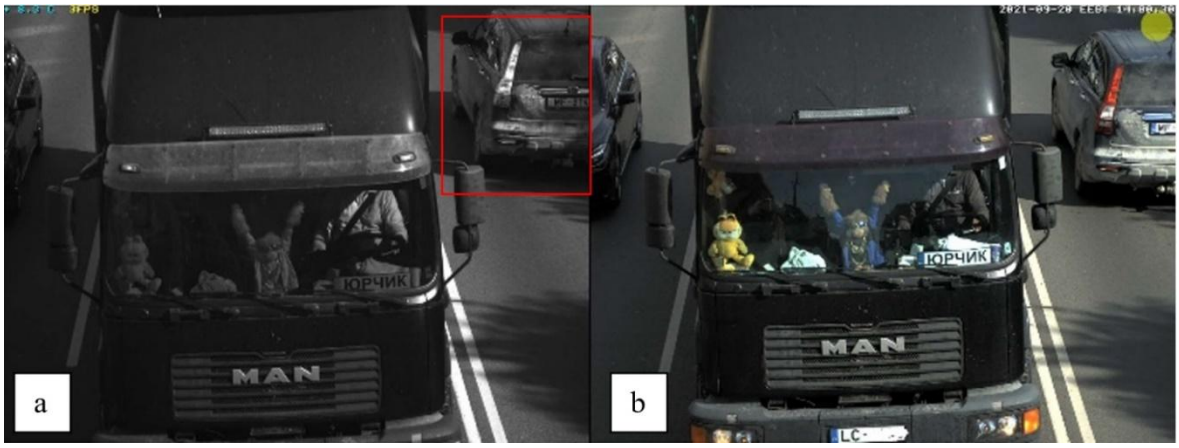
3.5. att. **Gaismas atstarojums** (Bumanis, Vitols, et al., 2022)**.**

Sistēmas optimizācija ietvēra vairākus tehniskos risinājumus. Apgaismojuma problēmu novēršanai tika ieviesta paralēlā attēlu uzņemšana gan redzamajā, gan infrasarkanajā spektrā. Šī pieeja ļāva izvēlēties augstākās kvalitātes attēlu datu sinhronizācijai, vienlaikus izmantojot abus attēlus numura zīmes atpazīšanai. Datu plūsmas optimizācija kļuva par būtisku izaicinājumu, jo dienā tika pārsūtīti aptuveni 20 GB datu. Satiksmes intensitāte tieši ietekmēja datu apjomu – sastrēgumu stundās viens transportlīdzeklis varēja radīt pat 120 attēlus, bet mierīgākos periodos mazāk nekā 5 attēlus. Lai optimizētu datu plūsmu, tika veiktas vairākas korekcijas. Sākotnēji kadru skaits tika samazināts no 10 uz 6, vēlāk uz 4 kadriem sekundē. Tika veikta arī kameras darbības zonas samazināšana no 20% līdz 5% (sk. 3.6. att.). Papildus tam tika pielāgots kontrasta slieksnis attēlu uzņemšanai.



3.6. att. **Kameras darbības diapazons** (Bumanis, Vitols, et al., 2022)**.**

Koordinātu sistēmas konfigurācija ceļu joslu uzraudzībai atklāja papildu izaicinājumus. Četru joslu divvirzienu ceļā pretējo virzienu transportlīdzekļu oklūzija būtiski palielināja apstrādājamo datu apjomu. Numura zīmju atpazīšanas algoritms veidoja objektus ap visām konstatētajām numura zīmēm neatkarīgi no transportlīdzekļa orientācijas, radot tā saukto datu sadursmes troksni. Šis troksnis apgrūtināja ceļu satiksmes noteikumu pārkāpumu konstatēšanu (sk. 3.7. att.).



3.7. att. **Datu sadursmes "troksnis"** (Bumanis, Vitols, et al., 2022)**.**

Apraksts: a) kravas automašīnas numurs nav redzams, bet pretējā virziena automašīnas numurs ir (sarkanais kvadrāts); b) ir redzamas abas numura zīmes.

Sistēmas optimizācijas procesā katrai joslai tika piešķirts atsevišķs koordinātu plaknes segments. Šis risinājums kopā ar samazināto kameras darbības diapazonu būtiski uzlaboja sistēmas precizitāti – vairāku transportlīdzekļu attēlu skaits četru dienu periodā samazinājās no 17% līdz 4%.

Laika sinhronizācijas problēmu risināšanai tika ieviests tīkla laika protokola (NTP) serveris. Sākotnējā pieeja, izmantojot konvolūcijas neironu tīklu vidējās laika aizkaves (30-60 s) noteikšanai, nenodrošināja pietiekamu precizitāti. Kalibrēšana Jelgavā tika veikta, izmantojot paralēlu manuālu filmēšanu ar mobilo tālruni kā references avotu.

Veiktie uzlabojumi ļāva sasniegt 97% vidējo datu sinhronizācijas ātrumu. Īpaši nozīmīgs progress tika panākts ātruma mērījumu precizitātē – kļūda samazinājās no 8% pirmajās trīs dienās līdz mazāk nekā 0,5% vēlāk. Četru dienu periodā no 60 381 konstatētā transportlīdzekļa tikai 69 gadījumos (0,114%) tika konstatētas ātruma mērījumu kļūdas, savukārt ātruma pārsniegšana tika fiksēta 2850 gadījumos (4,72%). Eksperiments atklāja vairākus būtiskus aspektus satiksmes uzraudzības sistēmu izveidē:

1. sensoru izvietojums ir kritiski svarīgs (H. T. Rashid & Ali, 2021). Jelgavas gadījumā četru videokameru vienmērīgs izvietojums ar pārklājošiem redzeslaukiem nodrošināja efektīvu pārkāpumu fiksēšanu. Tīkla caurlaidspēja jāpielāgo konkrētajam uzdevumam – ne vienmēr nepieciešams augsts kadru ātrums;
2. reālās vides konfigurācija prasa vai nu statisku vai dinamisku parametru pielāgošanu (Jiang et al., 2019). Satiksmes intensitātes analīze un dinamiska pielāgošanās ir būtiska sistēmas komponente, ko var optimizēt ar neironu tīklu palīdzību (Khazukov et al., 2020);
3. sistēmas dublēšana ar papildu sensoriem, piemēram, infrasarkano sensoru vai asfaltā iestrādātu induktoru, var uzlabot precizitāti, īpaši nakts apstākļos (Gulati & Srinivasan, 2019). Automātiska kalibrēšana var būtiski atvieglot sistēmas uzturēšanu(Hu & Ma, 2020).

Eksperimentu rezultāti parādīja, ka sistēmas veiktspēju visvairāk ietekmēja aparatūras iestatījumi, apgaismojums un nokrišņi. Stāvēšanas pārkāpumu uzraudzībai pietiek ar ierobežotiem resursiem, bet reāllaika satiksmes uzraudzībai ar numuru atpazīšanu nepieciešama kompleksa optimizācija un mašīnmācīšanās risinājumi. Lai gan katram faktoram pastāv atsevišķi risinājumi, to efektīva kombinācija prasa rūpīgu pielāgošanu konkrētajam uzdevumam un videi.

3.3. Precīzā putnkopība

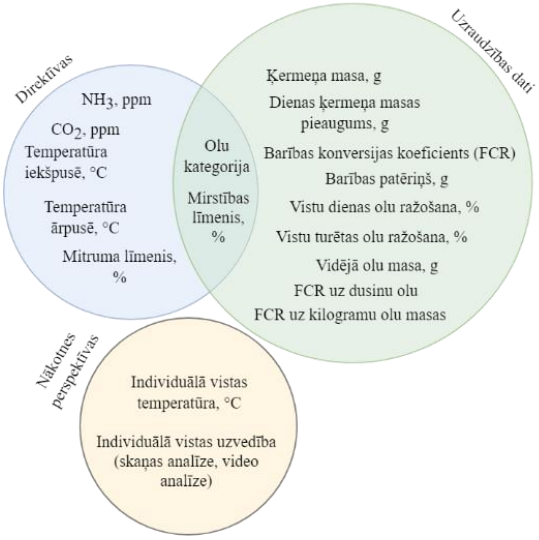
Industrija 4.0 iekļauj lietu interneta ieviešanu ražošanas sektoros, kur tas līdz šim nebija pieejams, pārveidojot šos sektorus tā sauktajos "precīzajos" variantos (Fresco & Ferrari, 2018; Mirkouei, 2020). Tas var attiekties uz precīzo lauksaimniecību un agronomiju, precīzo lopkopību (An et al., 2017). Kamēr precīzās putnkopības termins, kas definēts *IoT* risinājuma ieviešanā mājputnu nozarē, kļūst arvien plašāk izmantots (Astill et al., 2020), pareizās datu apstrādes aspekti padziļinās (Debauche et al., 2021). Precīzā putnkopība cieši seko Industrijas 3.5 (C. Y. Wang et al., 2021) un 4.0 ietvariem (Pitesky et al., 2020). Pašreizējās tendences precīzajā putnkopībā ir vērstas uz parametru uzraudzību (izmantojot dažādus sensorus), kas ietekmē olu/gaļas ražošanu, dzīvnieku labklājību, uzvedību un izaugsmi (Ahmad, 2011; Orakwue et al., 2022; Revanth et al., 2021). Šo novērojumu mērķis ir palielināt kopējo ražošanas daudzumu un kvalitāti. *IoT* saskaņā ar Industriju 4.0 nodrošina veidus, kā iegūt informāciju par lauksaimniecības produkciju; tomēr, jo vairāk datu, jo grūtāk ir noteikt pareizo veidu, kā šos datus apstrādāt (Kärner, 2017). Pirmkārt, nav tiesību aktu, kas noteiktu, kā tieši dati ir jāapstrādā. Tirgū piedāvātie risinājumi ir īpaši vērsti uz datu vākšanu, nevis to apstrādi un analīzi. Otrkārt, datu īpašumtiesības, īpaši apkopotie dati, kļūst par aktuālu tēmu (Raina & Palaniswami, 2021; Wiseman et al., 2019) – vai saimniecību īpašniekiem ir jāglabā dati privāti, vai arī ar tiem jādalās, lai veicinātu viedās mājputnu nozares attīstību kopumā? Apkopotos datus var izmantot, lai izstrādātu precīzas lēmumu atbalsta sistēmas un analīzes sistēmas, kas neaprobežojas tikai ar vienu konkrētu saimniecību.

Putnkopība ir viena no jomām, kurā Industrijas 4.0 principu pielietošana kļūst par nepieciešamību. Tas ir īpaši svarīgi, kontrolējot siltumnīcefekta gāzes (SEG), kuru līmenis ir jāpārvalda saskaņā ar ES klimata politiku (European Commission, 2003) un *KIoT*o protokola līgumu (Böhringer, 2003). Vairāki SEG ietekmējošie faktori ir uztura sastāvs un barības konversijas attiecība, kūtsmēslu ražošana un kūtsmēslu apsaimniekošana; tie visi tieši vai netieši ietekmē vides parametrus (Wang-Li et al., 2020) – temperatūru, mitrumu un gaisa sastāva gāzes: amonjaks (NH_3), oglekļa dioksīds (CO_2), metāns (CH_4) un slāpekļa oksīds (N_2O). Ir dažādas Eiropas Savienības (ES) direktīvas, kuru mērķis ir ierobežot, regulēt un/vai pārvaldīt dējējvistu turēšanas un audzēšanas aspektus, tostarp prasības attiecībā uz novietņu veidiem (The Council of the European Union & C E C (Commission of the European Communities), 1999), iekštelpu vidi (Council of the European Union, 2007), putnu fermu

darbinieku labklājību (EU Commission Directive 2000/39/EC, 2000). Turklāt, saskaņā ar dažādiem Eiropas Savienības putnkopības noteikumiem un direktīvām (Council of the European Union, 2007; EU Commission Directive 2000/39/EC, 2000; The Council of the European Union & C E C (Commission of the European Communities), 1999), siltumnīcefekta gāzu uzturēšanai vajadzētu būt vienam no galvenajiem uzdevumiem, lai sasniegtu minēto mērķi. Ir arī stingri ES noteikumi attiecībā uz olu tirdzniecības standartiem (Commission Regulation, 2008). Problēma ir šo direktīvu uztverē un piemērošanā, jo īpaši ņemot vērā to, ka šīs direktīvas ir rekomendējošas. Lielākā daļa saimniecību īpašnieku ignorēs direktīvas, ja vien tās nepieprasa kādu juridisku sodu, kas ir pietiekami liels, lai ietekmētu uzņēmējdarbību. Piemēram, saskaņā ar minimālajām prasībām, kas ir jānodrošina (Council of the European Union, 2007), galvenie parametri, kuri pastāvīgi jāuzrauga un jāpārvalda, ir NH_3 koncentrācija (nedrīkst pārsniegt 20 ppm), CO_2 koncentrācija (nedrīkst pārsniegt 3000 ppm), iekštelpu temperatūra (nedrīkst būt augstāka par 33°C,) iekštelpu mitrums (nedrīkst pārsniegt 70% vidēji 48 stundu laikā, ja āra temperatūra ir zemāka par 10°C).

Izplatīta (Debauche et al., 2020) pieeja mājputnu fermas vadībai bez sensoriem ir manuāli uzraudzīt dzīvnieku uzvedību un korelēt to ar pozīcijām attiecīgajā etogrammā. To parasti (Omomule et al., 2020) veic kā pēcfaktu analīzi, pamatojoties uz katru nedēļu savāktajiem datiem. Galvenie trūkumi ir saistīti ar problēmām, kas tiek atklātas vēlākā posmā saistībā ar vienlaikus dažādu dzīvnieku medicīnisko aprūpi un produkcijas zudumu. Ieviešot *IoT*, šīs problēmas lielākoties tiek novērstas; tomēr īstenošanas process nav definēts, un tam ir nepieciešama konkrētā gadījuma analīze un pielāgojumi. Šādas sistēmas galvenie mērķi ir 1) nodrošināt datu kvalitāti un kvantitāti turpmākai analīzei un 2) samazināt lauksaimniecības darbinieku darba slodzi, kuri citādi būtu atbildīgi par šo datu manuālu vākšanu.

No datu nepieciešamības viedokļa var definēt trīs galvenās grupas – ES direktīvu prasību nodrošināšanai nepieciešamie dati, saimniecību uzraudzības dati biznesa analīzei un optimizācijai, un papildu dati, kas var ietekmēt kopējo putnu fermas pārvaldību (sk. 3.8. att.).



3.8. att. **Apkopoto datu sinerģija** (Bumanis, Arhipova, et al., 2022).

Jebkura veiksmīga darbība ar novērojumu analīzi prasa sistemātisku datu ieguvi un apstrādi. To var atrisināt, ieviešot viedās putnkopības pārvaldības sistēmu. Ir pieejami dažādi komerciāli risinājumi (piemēram, (*Baku: Poultry IOT Solution*, 2022; *Fancom: Smart Farming*, 2022)), tomēr tie ir paredzēti vidējām un lielām saimniecībām, parasti ir dārgi un prasa palīdzību no risinājumu izstrādātāja, lai tos veiksmīgi izmantotu. Pēdējā laikā datu avotu apvienošanas jautājums tirgū tiek virzīts, piedāvājot vairākus risinājumus (*Layer Farm Manager – Poulty Layer Farm Management Software*, 2021; *Poultry ERP Software for Profitable Poultry Business / PoultryCare ERP*, 2021; *SmartBird Poultry Manager Features / SmartBird*, 2021). Šo risinājumu galvenās iezīmes attiecas uz datu laukiem, kas tika apspriesti

iepriekš. Šo risinājumu vispārējā arhitektūra ir balstīta uz iespēju ievadīt datus mobilajās vai tīmekļa lietojumprogrammās, lai vēlāk ģenerētu atskaites. Komerciāli pieejamo sistēmu ierobežojums ir stingrība, t. i., nespēja funkcionāli pielāgot un/vai paplašināt minēto sistēmu (t. i., ar reāllaika analīzi un lēmumu atbalsta sistēmām). Papildus minētajām komerciāli pieejamajām vadības sistēmām *So-In* (So-In et al., 2014) piedāvā hibrīdu pieeju, kas ietver sensoru tīklus (datu vākšanai) un mobilos tīklus kopā ar datu analīzes funkciju, lai nodrošinātu konkrētus lēmumus par saimniecības vides izmaiņām. Putnu fermu uzraudzībai (kā arī automatizācijai), ieviešot *IoT*, ir veltīti vairāki pētījumi (Handigolkar et al., 2016; Jayarajan et al., 2021). Uz *IoT* balstītu pārvaldības sistēmu mazām putnu fermām ierosināja Batuto u. c. (Batuto et al., 2020). Viņu pētījumā tika izstrādāta mobilā lietojumprogramma, lai palīdzētu lauksaimniekiem ar barības un ūdens apsaimniekošanu. *Zheng* (H. Zheng et al., 2021) piedāvā pētījumu apraksta mākonī balstītas mājputnu pārvaldības sistēmas koncepciju un ieviešanu, uzsverot lielo datu apstrādes sistēmu nepieciešamību putnkopībā. Var secināt, ka apsaimniekošanas sistēmu skaits tirgū lauksaimnieku atbalstam, īpaši Baltijas reģionā, ir salīdzinoši neliels. Tomēr tas ir ļoti svarīgi, jo trūkst pieredzes Baltijas saimniecībās (Arhipova et al., 2022).

ES direktīvu prasības galvenokārt attiecas uz vispārīgajiem vistu labturības aspektiem – nosakot maksimālo mirstības līmeni, minimālo izmēru un blīvuma prasības novietnēm un vides parametriem. Stingrākā ES regula nosaka olu kategorijas un prasības iepakojumam. Ar uzņēmējdarbību saistītie dati attiecas uz investīcijām, t. i., barības apjomu, uzturēšanas un loģistikas izdevumiem un atdevi – saražotās broileru gaļas, olu vai vistu apjomu. Uzņēmējdarbības dati ir saistīti ar saimniecību monitoringa datiem, kas jāapkopo katru dienu. Turklāt var izmantot padziļinātus analītiskos indeksus, piemēram, neto barības efektivitātes indeksu vai olu un barības cenas attiecību. Nākotnes perspektīvas attiecas uz jebkādiem papildu datiem, kas vēl nav reglamentēti nevienā ES direktīvā un nav biznesa optimizācijas prioritāte. Tomēr šie dati var ietekmēt abas iepriekšminētās datu grupas. Piemēram, novērojot atsevišķas vistas temperatūru, var prognozēt iespējamo kaitīgo stāvokli, tādējādi samazinot kopējo mirstības līmeni (Okinda et al., 2020). Ir vairāki veidi, kā ieviest lietu internetu mājputnu fermās, tomēr visbiežāk (Astill et al., 2020; Choosumrong et al., 2019) tiek izmantoti vairāki sensori citā vietā. Bieži sensori tiek uzstādīti pa pāriem, kur katrs sensors ir atbildīgs par konkrētiem datiem (t. i., viens sensors par CO_2, otrs par NH_3). Šos sensorus var vai nu grupēt pēc mērījumu veida, piemēram, temperatūras, NH_3 un CO_2, vai arī izvietot lieki, lai nodrošinātu datu apkopošanu. Gadījumos, kad vienam mērījumam tiek izmantoti vairāki sensori, var izmantot vairāku sensoru datu apvienošanu, lai nodrošinātu precizitāti un konsekvenci (Handcock et al., 2009; Naehev et al., 2019).

Jāuzsver, ka jo vairāk ir sensoru un jo dažādāki ir to veidi, it īpaši, ja vairākus sensorus var izvietot vienā ēkā dažādās vietās, jo sarežģītāka kļūst šīs informācijas uzglabāšana.

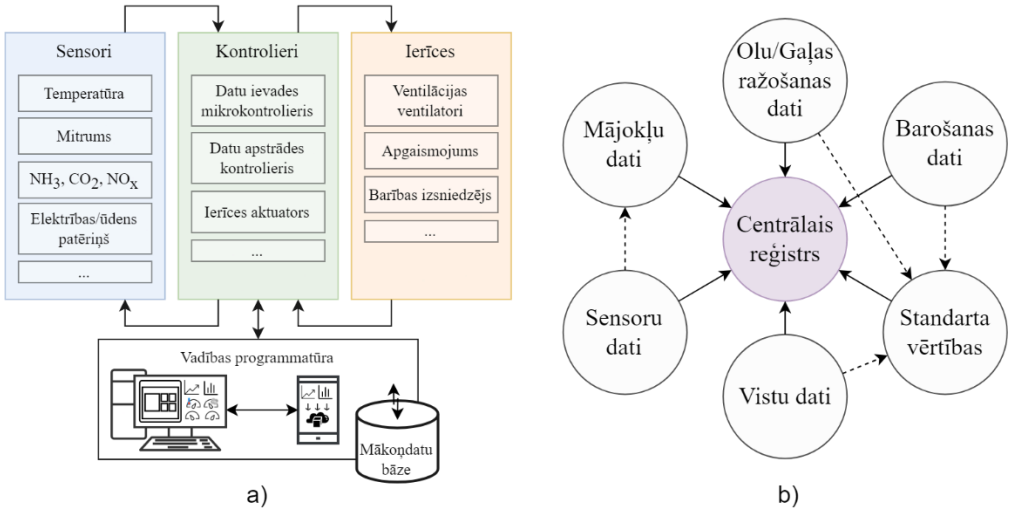
Jebkura veiksmīga darbība ar novērojumu analīzi prasa sistemātisku datu ieguvi un apstrādi. To var atrisināt, ieviešot precīzās putnkopības pārvaldības sistēmu.

Sistēmas arhitektūra, kas izveidota saskaņā ar Industrija 4.0 politikām, tiek saukta par kibernētiski-fizisko modeli (Sadiku et al., 2017). Putnu sektors nav izņēmums – ar ražošanas mēroga pieaugumu, īpaši saistībā ar kvalitātes kontroles intensitātes palielināšanos gaļas un olu ražošanā, precīzai pārvaldībai nepieciešama kibernētiski-fiziskās sistēmas ieviešana. Šajā kontekstā kibernētiski-fiziskais modelis ir trīs komponentu sistēma: sensoru kopums mērījumu ieguvei, datu apmaiņas kontrolieri, kas pilda starpnieka lomu starp sensoriem un datu centru, un datu centrs kā analītiskais centrs ar lēmumu pieņemšanas spējām. Terminu "kibernētiski-fiziskais modelis" var izmantot gandrīz jebkurā digitālajā un automātiski vadītajā ražošanas sistēmā (Okinda et al., 2020; C. Y. Wang et al., 2021).

Izveidotais precīzās putnkopības pārvaldības sistēmas dizains koncentrējas uz lēmumu pieņemšanu par piemērotāko barības procesu, vienlaikus optimizējot CO_2 un NH_3 līmeņus. Kibernētiski-fiziskais modelis tika izveidots paredzētajai pārvaldības sistēmai un ietver trīs datu grupas – datus, lai apmierinātu noteiktas prasības, uzraudzības datus un biznesa saistītos datus.

Šajā kontekstā dati tiek grupēti pēc to avotiem un mērķa. Piemēram, regulas sniedz prasības un ieteikumus minimālo vides parametru veidā, piemēram, CO_2 un NH_3, kamēr uzraudzības dati ietver visus datus, kas iegūti, izmantojot sensoru aprīkojumu.

Koncepcija ir izstrādāta, ņemot vērā vairākas neatkarīgas putnu fermas, kuras pārvalda viens procesors (sk. 3.9. att., a). Tāpēc, lai arī Industrija 4.0 mērķis ir atvieglot centralizāciju, galvenā datu apstrāde ir koncentrēta vienā centrālajā datu centrā, kas ir atbildīgs par pārvaldības programmatūras piekļuvi visām saimniecībām.



3.9. att. **Kibernētiski-fiziskā modeļa koncepcija** (Bumanis, Arhipova, et al., 2022).

Datu noliktava ir jāveido, balstoties uz pieņēmumu, ka putnu fermu var izmantot vairākiem mērķiem – olu ražošanai, broileru gaļas ražošanai un vistu audzēšanai.

Viena no bieži lietotām datu noliktavas projektēšanas shēmām ir zvaigžņu tipa shēma (Dori et al., 2008). Šī shēma ir izveidota, izmantojot vienu galveno faktu tabulu un vairākas vienas dimensijas tabulas katrai saistītai datu grupai. Lai uzlabotu datu hierarhiju, zvaigžņu shēmas tabulas var normalizēt, rezultātā iegūstot shēmu "Sniegpārsla". Abas šīs shēmas ir labas iespējas datu noliktavas projektēšanai.

Koncepta datu noliktavas struktūra izmanto sniegpārslu shēmas stilu, jo:

1. dati tiek apkopoti nelielos periodos datu ieguves un analīzes nolūkos vai katru dienu;
2. jānormalizē datubāzes struktūra, lai tā būtu pielāgojama konkrētiem putnu fermu datu avotiem.

Sniegpārslu shēmas stila galvenais trūkums ir sarežģīto vaicājumu veidošanas komplicētība. Tomēr tā nav problēma, ņemot vērā, ka dati tiek apkopoti vienu vai divas reizes dienā. Sniegpārslu shēmas priekšrocība ir tā, ka datus katrā dimensiju tabulā var atjaunināt atsevišķi, neietekmējot citas tabulas, kamēr nav bojāta atsauces integritāte (Levene & Loizou, 2003).

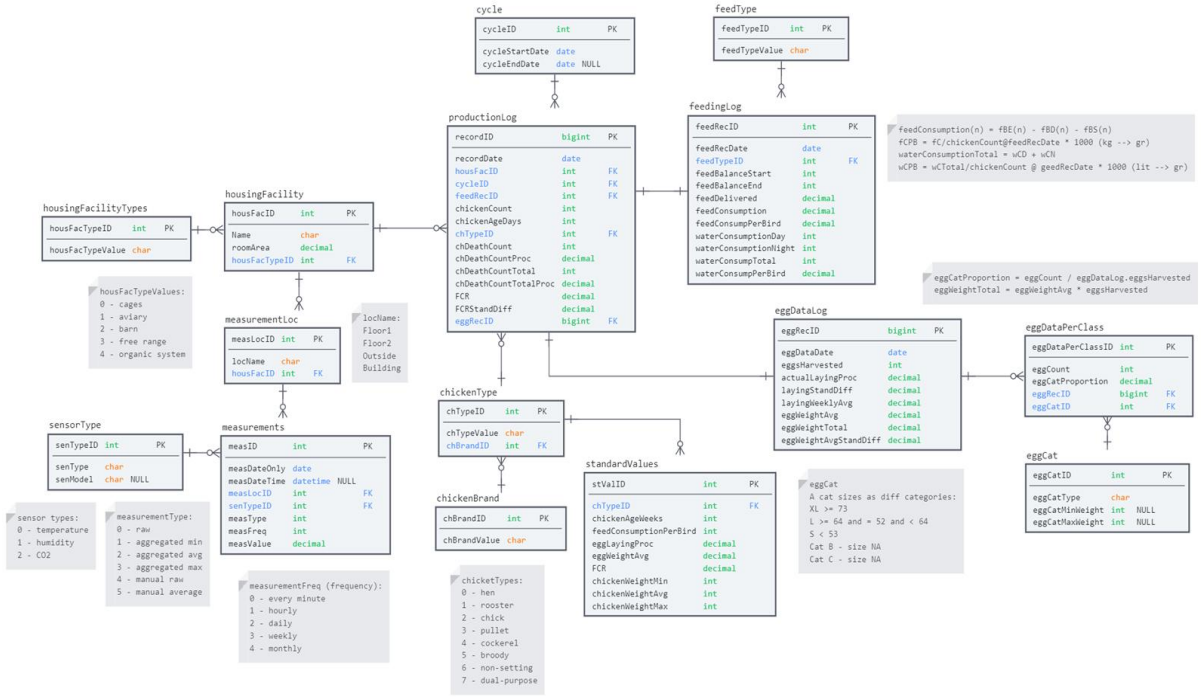
Tādējādi sniegpārslu normalizācija ņem vērā datu grupas un ir veidota, lai nodrošinātu universālu datu struktūru atšķirīgi organizētām putnu fermām.

Lai sasniegtu šos mērķus, ir jāapkopo atbilstoši dati. Sakarā ar to, ka vides mainīgie, standarta vērtība, barošanas dati un izmitināšanas dati ir statiski, dizains var izmantot iespēju izveidot universālu ietvaru, kas aptvertu aptuveni 90% līdzīgo datu bāzu (DB) arhitektūru.

Mājputnu pārvaldības sistēmas mērķis ir kvantitatīvi un kvalitatīvi izsekot olu vai gaļas ražošanai saistībā ar pieejamo vistu barošanu. Atkarībā no mājputnu fermas datus var vākt dažādos intervālos, parasti (W. F. Pereira et al., 2020) ik pēc 1 vai 2 stundām. Turpmāk aprakstītās datu struktūras mērķi ir 1) glabāt identificēto grupu datus, 2) nodrošināt konkrētu sensoru datu apkopojumu un glabāt tos atsevišķi, un 3) sniegt lietotājiem statistiskai analīzei un atskaitēm nepieciešamos datus. Attiecībā uz datiem, kas saistīti ar mājputniem, datu struktūrā jāņem vērā dimensiju tabulu lielums un iespējamie normalizācijas līmeņi. Lai vienkāršotu DB dizainu, tika izmantots vienots centrālais reģistrs, bet pārējās tabulas ir satelīti. Satelītu tabulas

var būt savienotas viena ar otru, var būt atkarīgas viena no otras datiem vai būt pilnīgi neatkarīgas (sk. 3.9. att., b). Mājputnu pārvaldības sistēmas galvenie dati ir vai nu olu ražošanas dati, vai gaļas ražošanas dati; tomēr šie dati ietver daudz parametru, kas veicina gala rezultātu. Apvienoto Nāciju Organizācijas Pārtikas un lauksaimniecības organizācija (*FAO*) ziņoja (MacLeod et al., 2018) par potenciālo pieprasījuma pieaugumu pēc vistas gaļas par 61% un vistas olām par 39% laikā no 2005. līdz 2030. gadam. Tieši gaļas kvalitātes jautājumam bieži tiek izmantotas datu apvienošanas metodes (Weng et al., 2020).

Visizplatītākie dati jebkurā mājputnu pārvaldības sistēmā ir vistas dati (Nyalala et al., 2021). Šie dati ietver informāciju par šķirni, veidu, pašreizējiem parametriem, piemēram, svaru, vecumu un kopējo skaitu. Termins "vista" attiecas uz putnu sugu kopumā; tas ietver vistu kā olu dējējvistu, vistu kā pirmsdējējvistu un gaili kā gaļas putnu. Ir arī statistikas un loģistikas dati, piemēram, vistu imports/eksports, kopējais mirušo skaits. Centralizētā faktu tabula DB struktūrā (sk. 3.10. att.) ir tabula [productionLog].



3.10. att. **Datu noliktavas dizains** (Bumanis, Arhipova, et al., 2022).

Šī tabula apvieno dimensiju tabulas, kurās tiek glabāti sensoru dati, barības datu un olu ražošanas dati; tā tiek izmantota arī, lai uzglabātu vistas saistītos datus. Saistītie dati tiek asociēti, izmantojot šādus laukus kā *[productionLog].housFacID*, saimniecības identifikācijas numurs, *[productionLog].cycleID*, konkrētās ražošanas cikla identifikācijas numurs un *[productionLog].recordID* kā galveno atslēgu, lai apvienotu barības un olu ražošanas datus.

Apvienotās dimensiju tabulas attiecas uz saimniecības iekārtu, ražošanas ciklu, barības žurnāla ierakstiem un olu ražošanas žurnāla ierakstiem. Turklāt visos ierakstos ir iekļauti vai nu aprēķinātie vērtējumi, piemēram, vistas nāvju skaits procentos, vai attēlo atšķirību no standarta vērtības, piemēram, barības patēriņa ātrums un vidējais vistas svars. Standarta vērtības tiek glabātas tabulā [standardValues], kurā ir standartizētas nedēļas plāna vērtības atkarībā no konkrētās vistas šķirnes, piemēram, *Lohmann Brown* (Lohmann, 2013), un tās tiek apvienotas, izmantojot noapaļotās vērtības no *[productionLog].chickenAgeDays/7* un *[standardValues].chickenAgeWeeks*. Standarta vērtības ir jāsaglabā katram parametram, kuram jābūt optimizētam. Gadījumā, ja runa ir par olu ražošanu, šajā tabulā ir jāiekļauj arī standarta vērtības olu novietošanai un vidējai olu masai.

Tabulā *[productionLog]* tiek glabātas neapstrādātas vai apkopotas ikdienas vērtības, piemēram, vistas vecums, vidējais vistas svars, pašreizējais vistu skaits un nāvju skaits. Lielākās fermās parasti ir vairāki vistas veidi, piemēram, cāļi agrīnās nedēļās un vistas vēlāk;

tomēr ir arī prakse importēt olas nesējas. Izvēlēto šķirņu skaits ir atkarīgs no fermas lieluma un pieejamības dažādām šķirnēm, jo katrai šķirnei ir savas standartizētas parametru vērtības. Tikai vienas šķirnes izmantošana ļauj precīzi uzraudzīt ražošanas datus, kas saistīti ar standarta vērtībām.

Produkcijas žurnāls ir nepieciešams, lai noteiktu piemēroto barības protokolu relatīvo kvalitāti un produktivitāti, kas rezultējas konkrētā olu vai gaļas ražošanas kvalitātē un kvantitātē. DB struktūra ietver tikai olu ražošanu; tomēr gaļas ražošanas gadījumā var tikt pievienota papildu sniegpārslu shēma, piemēram, tabula *[meatDataLog]* ar atbilstošiem laukiem un zemākā līmeņa dimensiju tabulām. Viens no galvenajiem kvalitātes un produktivitātes parametriem ir barības patēriņa ātrums vai *FCR*, kas ir barības patēriņa attiecība pret olu masu vai ražoto olu skaitu.

Olu ražošanas dati ir visi statistiskie dati, kas attiecas uz olu skaitu, olu svaru, olu sadalījumu pēc kategorijām un izmēriem. Tas ietver arī atkāpšanos no standartizētām vērtībām, kas ir attēlotas zīmola specifikācijās. Šai datu grupai pieder šādas tabulas: *[eggDataLog], [eggDataPerClass], [eggCat]*. Tabulā *[eggDataLog]* glabājas vispārīgie dati, tādi kā novākto olu skaits, olu nesamības parametrs, vidējais un kopējais novākto olu svars. Turklāt ir divas zemāka līmeņa dimensiju tabulas, kas saistītas ar olu datiem.

Tabula *[eggDataPerClass]* satur datus par dažādām savākto olu kategorijām vai klasēm. Olu ražošanai ir trīs galvenās kategorijas – kategorija A, kategorija B, kategorija C. Kategoriju A iedala vairākās apakškategorijās, kas ir definētas pēc olas izmēra. Kategorijas A apakškategorijas ir XL (ekstra lielas), L (lielas), M (vidējas) un S (mazas) olas. Dati par olu kategorijām tiek glabāti dimensiju tabulā *[eggCat],* kur attēlotas minimālās un maksimālās olu svara kategorijas A apakškategorijām.

Ēdiena veids un tā makroelementu sastāvs nosaka putnu labklājību un izmaksu/ražošanas kritisko līkni olu/gaļas ražošanā (Mota de Carvalho et al., 2021). Putniem piešķirtais ēdiens atšķiras atkarībā no to vecuma un šķirnes, tāpēc ir jāizvēlas atbilstošs veids un daudzums. Šai datu grupai pieder šādas tabulas: *[feedingLog], [feedType]*.

Ar vistu barošanu saistītie dati tiek glabāti atsevišķā dimensiju tabulā *[feedingLog],* lai atvieglotu piemēroto barošanas protokolu analīzi. Katrs ieraksts atbilst konkrētam barības veidam, proti, tā nosaukumam, formulai vai galveno uzturvielu sastāvam. Tabulā tiek uzturēts uzskaitījums par barības daudzumu katras dienas sākumā un beigās, iekļaujot arī barības daudzumu, kas piegādāts putnu fermā, un barības daudzumu, kas ir patērēts kopumā un uz vienu putnu. Šis žurnāls ietver arī ūdens patēriņa parametrus, piemēram, ūdens patēriņu dienas un nakts laikā, kopējo ūdens daudzumu, kas patērēts kopumā un uz vienu putnu.

Gudrās pārvaldības sistēma nav pilnīga bez sensoru un sensoru datu ieviešanas. Lielākā daļa sensoru ir vērsti uz vides un labturības kvalitātes optimizēšanu. Vides ievades ietver temperatūru, gaisa ātrumu, ventilācijas ātrumu, paklāju kvalitāti, mitrumu un gāzu koncentrācijas, tai skaitā oglekļa dioksīdu un amonjaku (Astill et al., 2020). Tiek pieņemts, ka sensoru dati tiek automātiski apkopoti un periodiski nosūtīti uz noliktavu, bet arī, lai gan reti, var tikt ievadīti manuāli. Šajā datu grupā ietilpst šādas tabulas: *[housingFacility], [housingFacilityTypes], [measurements], [measurementLoc], [sensorType]*.

Mājputnu pārvaldības sistēmas izmanto sensoru datus, lai uzraudzītu vides parametrus, kuriem ir tieša ietekme uz produkcijas kvalitāti un/vai kvantitāti. Sensoru dati tiek iegūti no sensoriem, kas uzstādīti pēc izmēra un novietnes veidiem atšķirīgās putnu fermās; tāpēc katram sensoram ir jāattiecas uz noteiktu sensora atrašanās vietu, savukārt atrašanās vietai ir jāattiecas uz konkrētu novietni**.** Ir Eiropas standarti attiecībā uz putnu fermu novietņu veidiem un izmēriem (EGTOP, 2012; Scientific Committee on Animal Health and Animal Welfare, 2000; Van Horne & Achterbosch, 2008) atkarībā no putnu veida. Novietņu dati attiecas uz objektu, kas tiek izmantots kā galvenā ēka putnu turēšanai. Konkrētā novietņu atrašanās vieta nosaka prasības temperatūras un ventilācijas pārvaldībai, jo optimālais iekštelpu temperatūras, mitruma un O_2 bagātības diapazons ir šaurs (*PoultryWorld – Temperature and Relative Humidity Critical in Climate Control*, n. d.).

Informācija par objektu tiek glabāta tabulā *[housingFacility]*, kur *[housingFacility]* "Name" ir objekta nosaukums un "[housingFacility].roomArea" ir objekta platība kvadrātmetros. Normalizācijas nolūkos objektu veidi tiek saglabāti atsevišķi tabulā *[housingFacilityTypes]*. Tiek ņemti vērā šādi objektu veidi: būri, voljērs, kūts, brīvā turēšana un bioloģiskā sistēma. Tehniski šajās tabulās jāiekļauj arī privāta informācija, piemēram, objekta īpašnieka vārds, uzvārds; taču šo datu nav, jo līgumā ir norādīta saistība starp saimniecības īpašnieku un konkrēto objektu. Datu noliktavas struktūrā tiek ņemti vērā arī dažādi sensoru datu avoti, frekvence un ievades veidi. Parasti šādi dati ir neapstrādāti mērījumu dati no temperatūras, mitruma un CO_2 sensoriem. Alternatīvi, mājputnu fermas personāls var ievadīt mērījumus manuāli, piemēram, visas telpas vidējo temperatūru noteiktā brīdī vai āra temperatūru, kas ņemta no stacionāra termometra. Lai saglabātu apkopotos datus ātrai piekļuvei, tiek definēti papildu mērījumu parametri. Sensoru dati tiek saglabāti tabulā [mērījumi], kur katrs ieraksts atbilst vienam mērījumam [mērījums].*measValue*. Katrs ieraksts attiecas uz noteiktu sensora atrašanās vietu, kas tiek saglabāts kā *[measurementsLoc].locName* un konkrēts sensora veids, saglabāts kā *[sensorType].senType un [sensorType].senModel*.

Lai risinātu jautājumu par datu automātisku un/vai manuālu ievadīšanu dažādās frekvencēs, ir iekļauti divi datuma lauki – *[measurements].measDateOnly* un *[measurements].measDateTime*. Ja datu noliktavas ieviešana ļauj divām no šīm vērtībām izmantot tikai vienu veidu, viens no laukiem ir jānoņem. Mērīšanas veids, piem., neapstrādāta, apkopota vai manuāla ievade, tiek saglabāts *tabulā [measType].measType,* savukārt biežums, piemēram, stundā, dienā vai nedēļā, tiek saglabāts tabulā *[measFreq].measFreq*.

Datu noliktavas arhitektūra tika ieviesta esošajā Baltijas putnu fermas infrastruktūrā, kas sastāv no divu veidu putnu novietnēm – bagātinātiem būriem un kūts sistēmas. Pabeigtais ieviestais risinājums izmanto *Microsoft Azure Cloud* kā galveno procesoru. Pilnīga ieviešana (divās Baltijas fermās) ietver 6 pāru sensoru, analizatora un mikrokontrolieru sistēmas uzstādīšanu, kur uzstādīšanas periods – no 2020. gada 13. oktobra līdz 2021. gada 21. janvārim. Piemēram, pirmā mājputnu ferma sastāv no vairākiem šķūņiem, kas sadalīti būru, brīvos un sprostos, olu dēšanas sistēmās. Vistu kūtī A (sk. 3.11. att., a), kur katrā stāvā izvietoti divi sensoru pāri (1 pāris = 1 CO_2 + 1 NH_3 sensors), putni var brīvi pārvietoties, savukārt vistu kūts B (sk. 3.11. att., b), kur atrodas viens sensoru pāris, pieder pie otrā tipa, t. i., sprosti.

(a) bagātinātie būri (b) kūts sistēma

3.11. att.

3.11. att. **CO_2 un NH_3 sensoru uzstādīšana** (Bumanis, Arhipova, et al., 2022).
Apraksts: (a) kūts sistēmas mājputnu novietnēs un (b) bagātinātos būros.

NH_3 sensori (*Membrapor* NH_3 / *MR-100*) atrodas 2,5 metru augstumā katrā stāvā, vistu kūtī B 5 metru augstumā. CO_2 (*GDS* CO_2 sensors *IR-2*) sensori atrodas 0,4 metru augstumā katrā vistu kūtī. Sensori ir savienoti ar centrālo datu uzkrāšanas analizatoru (*Combi 64, GDS Technologies, Garforth, UK*), kas ir savienots ar sadales skapi, kurš secīgi ir iestrādāts ar *Z-logger* vārtejas ierīci vai *MQTT* brokeri. *MQTT* brokeris pārvērš komunikatora signālu (RTU — īsa attāluma sakaru protokols) par *Azure* saprotamo *MQTT* tīkla protokolu. *MQTT* savienojums ir aizsargāts ar lietotājvārdu un paroli. *MQTT* brokeris apstrādā tikai ienākošos un izejošos pieprasījumus, ko veicis autorizēts lietotājs, lai izveidotu šifrētu savienojumu. Pēc datu apstrādes *MQTT* brokerī emisijas mērījumi ir pieejami jaunajā sistēmas sadaļā – Sensoru mērījumu forma, kas eksportējama *.csv* un *.xlsx* formātā. Risinājums izstrādāts uz centralizētas mākoņdatošanas datu apstrādes sistēmas. Galvenās praktiskā pielietojuma priekšrocības ir automātiskajā datu vākšanā un automātiskajā datu apstrādē. Turklāt sistēma nodrošina iespēju manuāli ievadīt mērījumus un importēt vēsturiskos datus no *Excel* izklājlapām.

Pašlaik notiek diskusijas (Rowe et al., 2019) par to, kā palielināt *IoT* iesaisti precīzās putnkopības nozarē. Daži autori (Pramir Maharjan & Yi Liang, 2020) piedāvā ieviest individuālā monitoringa risinājumus, kas jau aktīvi tiek izmantoti citās nozarēs, piemēram, precīzajā lopkopībā un precīzajā cūkkopībā. Šādi risinājumi ir, piemēram, *RFID* sensori, iekšējie sensori temperatūras uzraudzībai, uzvedības uzraudzībai un kontrolei. Ir vairāki pētnieki, kas strādā pie skaņas (Fontana et al., 2016; J. Huang et al., 2021) un video (D. F. Pereira et al., 2020) analīzes mājputnu nozarē. Nākotnes tehnoloģijām arī ir tendence attīstīties uz prognozēšanu (Chizzotti et al., 2022; Zaefarian et al., 2021), nevis uz reaktīvu reakciju. Tas galvenokārt ietver veselības stāvokli un elpošanas problēmas.

Minētā kiberfiziskā modeļa īstenošana nodrošināja pietiekami daudz datu, lai analizētu ietekmes faktorus uz vistas olu ražošanu. Pētot olu īpatsvara prognozēšanas risinājuma izpētes un izstrādes fāzes, lielāka uzmanība tika pievērsta olu īpatsvara novērtēšanai un prognozēšanai. Olu ražošana pati par sevi prasa dziļu izpratni par procesiem, kas iesaistīti tādos aspektos kā pārtikas formulas pārvaldība (Morales-Suárez et al., 2021; Sakomura et al., 2019) un dzīvnieku labklājības stāvokļa pārvaldība (Buller et al., 2020; Murillo et al., 2020).

Olu īpatsvara dinamika parasti tiek izteikta grafiskā veidā kā līkne. Saskaņā ar (A. Safari-Aliqiarloo, F. Faghih-Mohammadi, M. Zare, A. Seidavi, V. Laudadio, M. Selvaggi, V. Tufarelli, et al., 2017; Görgülü & Akilli, 2018; Grossman et al., 2000) olu ražošanā izmantotā līkne ir nelineāra un attēlo olu daudzumu konkrētā laika posmā. Kopumā tiek uzskatīts par atbilstošu piemērot nelineārus modeļus, lai gan analizētu vēsturiskos olu ražošanas datus, gan prognozētu nākotnes tendences. Tomēr, balstoties uz literatūras analīzi (Ahmad, 2011; Felipe et al., 2014; Narinc et al., 2014; Omomule et al., 2020), tika secināts, ka gudrās putnkopības pārvaldības sistēmas izveidē jāiekļauj vairāki funkcionālie interpretētāji, pamatojoties uz dažāda veida modeļiem.

Būtībā šie pētījumi noved pie nepieciešamības novērtēt vispāratzīto nelineāro un jauno mašīnmācīšanās (ML) modeļu noderīgumu, lai 1) noteiktu zināšanu potenciālo līmeni, kas iegūts to piemērošanas rezultātā, un 2) iekļautu attiecīgos modeļus izstrādātajā gudrās putnkopības pārvaldības sistēmas ietvarā.

Šis ietvars ietver: ražošanas saistīto datu uzraudzību, risinājumu datu apstrādei un analīzei, kas ietver pielāgojamu datubāzes dizainu datu glabāšanai, modeļu kopumu datu analīzei un lēmumu atbalsta mehānismu ražošanas procesu optimizēšanai. Kā minēts iepriekš, kiberfiziskā modeļa koncepcija ir balstīta uz datu apvienošanas ideju, datu apvienošanas rezultāts jāizmanto vēsturisko datu analīzei un šādu ražošanas īstermiņa un ilgtermiņa tendenču prognozēšanai. Tomēr tika konstatēts (Bumanis, Arhipova, et al., 2022), ka pašreizējie datu apvienošanas risinājumi tirgū nevar tikt ieviesti ietvarā tādi, kādi tie ir, un kā rezultātā jāpieņem daudzlīmeņu datu apvienošanas pieeja. Sistēmas arhitektūra savukārt nosaka prasības šādu līmeņu skaitam un veidiem. Tāpēc gudrās putnkopības pārvaldības sistēmai jāizmanto datu struktūra, kas 1) ir pietiekami sarežģīta, lai ietvertu visas iespējamās datu vajadzības, un 2) ir pielāgošanās līmenis, kas ir ērts gan vadībai, gan saimniecības īpašniekiem.

Vairāku modeļu kolekcija ir paredzēta, lai nodrošinātu piemērotu izvēli – vai nu algoritmisko, ko sistēma pieņem, kad dati tiek ielādēti, vai manuālo – sistēmas operatora pieņemtu, lai risinātu gadījumus, kad apmācības un testēšanas kopu parametru skaits atšķiras. Tas arī daļēji novērš problēmas, ko izraisa nepilnīgi dati – ja viens modelis nesniedz apmierinošu veiktspēju un/vai iznākumu, šo modeli var aizstāt ar citiem ieviestajiem modeļiem. Tas pieņem, ka visi modeļi tiek apmācīti vai pielāgoti tiem pašiem datiem. Lai ietvertu dažādus faktorus un to apstrādes potenciālos rezultātus, tiek iekļauti divu galveno modeļu veidi – nelineārie modeļi un mašīnmācīšanās modeļi.

3.4. Aktualitātes nodaļas kopsavilkums

Nodaļā tiek apskatītas datu apvienošanas un kvalitātes uzlabošanas metodes trīs nozarēs – biškopībā, putnkopībā un viedajās transporta sistēmās. Katrā no šīm nozarēm tika identificētas problēmas saistībā ar datu trūkumiem un nepilnībām.

Precīzās biškopības jomā tika analizēti dažādi datu līmeņi – dravas, bišu saimju un individuālais līmenis. Pētījumā tika apskatīti dažādi sensoru veidi – temperatūras, mitruma, svara, skaņas un vibrācijas sensori, kā arī videokameras. Datu apvienošanas metodes biškopības nozarē līdz šim nav plaši pielietotas. Vājš tīkla pārklājums rada datu kvalitātes problēmas, kuru risināšanai tika piedāvātas datu apvienošanas metodes.

Viedajās transporta sistēmās tika izstrādāts sinhronizācijas algoritms *LiDAR* un videokameru datu apvienošanai. Risinājums tika testēts Jelgavas pilsētas satiksmes uzraudzībā

dažādos apstākļos – mainīgā apgaismojumā un laikapstākļos. Izmantojot laika zīmogu un sensoru kalibrēšanu, tika sasniegta 97% sinhronizācijas precizitāte.

Putnkopības nozarē tika izveidots centralizēts datu glabāšanas risinājums divās Baltijas putnu fermās. *Microsoft Azure* mākoņdatošanas platforma tiek izmantota CO_2 un NH_3 līmeņa sensoru datu automātiskai vākšanai un apstrādei. Sistēma nodrošina gan reāllaika, gan vēsturisko datu analīzi, kā arī atbilstību ES regulām par dzīvnieku labturību.

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.

4. DATU SLĀŅOŠANAS KONCEPTUĀLĀ METODE

Datu apvienošanas metožu un modeļu izpēte atklāj arvien jaunas likumsakarības un principus, kas būtiski ietekmē to praktisko pielietojumu. Sistemātiska pieeja šo metožu analīzei sniedz vērtīgas atziņas turpmāko risinājumu izstrādē. Īpaši nozīmīgs ir secinājums par datu priekšapstrādes kritisko lomu – pirms apvienošanas procesa uzsākšanas nepieciešama rūpīga un metodiska datu sagatavošana, kas būtiski ietekmē gala rezultāta kvalitāti. Šī apstrāde var būt vai nu normalizācija, vai standartizācija, atkarībā no datu īpašībām. Normalizāciju izmanto gadījumos, kad datu sadalījums nav normāls, savukārt standartizāciju pielieto datiem ar normālu sadalījumu. Šī sākotnējā apstrāde ir īpaši svarīga, jo tā nodrošina, ka dati būs piemēroti tālākai apstrādei ar metodēm, kurām nepieciešamas noteikta diapazona vai mēroga vērtības.

Lielākā daļa izpētīto datu apvienošanas metožu ir orientētas uz nākotnes prognozēšanu, piemēram, ar Beijesa teoriju, kas palīdz prognozēt objekta nākotnes stāvokli, balstoties uz iepriekšējiem datiem un modeļiem. Tomēr prognozēšana ne vienmēr ir obligāta funkcionalitāte. Veidojot jaunu datu apvienošanas metodi, ir svarīgi izvērtēt, vai prognozēšana ir nepieciešama konkrētajam pielietojumam, vai arī pietiek ar datu apstrādi un apvienošanu esošās informācijas ietvaros.

Vēsturisko datu izmantošana ir metožu izstrādes komponente. Sistēmas darbojas ar dažādām datu kombinācijām – gan no individuāliem sensoriem, gan no vairākiem avotiem vienlaikus. Vienota modeļa izmantošana vairākām datu kopām var radīt specializācijas efektu, tādēļ metodes pielāgošanās spēja ir faktors darbā ar dažādiem vēsturisko datu avotiem. Datu kvalitātes nodrošināšana sākas ar sākotnējo analīzi. Šajā posmā tiek identificētas trūkstošās vērtības, anomālijas un novirzes pirms tālākas analīzes. Normalizācija un standartizācija ir datu transformācijas, kas nepieciešamas metodēm ar specifiskiem vērtību diapazoniem. 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. Jaunu datu apvienošanas metožu izstrādē nepieciešama pielāgošanās dažādiem datu veidiem un to transformācijām.

Jaunajai datu apvienošanas metodei ir jāatbilst vairākiem būtiskiem kritērijiem. Tai obligāti jāietver datu sagatavošanas iespējas – spēja veikt gan normalizāciju, gan standartizāciju atkarībā no sākotnējo datu īpatnībām. Metodei noteikti jāspēj strādāt ar vēsturiskiem datiem, tos efektīvi izmantojot un analizējot. Šī prasība ir būtiska, jo tā ļaus pielāgot metodi dažādiem datu kopu veidiem un avotiem, padarot to universālāku un praktiskāk pielietojamu. Vēl viens nozīmīgs aspekts, kas jāiekļauj jaunajā metodē, ir datu kvalitātes novērtēšana. Pirms jebkuras apvienošanas operācijas metodei jāspēj veikt rūpīgu datu kvalitātes pārbaudi, īpaši gadījumos, kad pastāv risks saskarties ar trūkstošām vērtībām vai citām nepilnībām.

Kopumā jaunajai metodei jābūt spējīgai ne tikai apstrādāt un apvienot dažādus datu avotus, bet arī nodrošināt augstu datu kvalitāti un uzticamību. Šāda metode ir īpaši nepieciešama IoT sistēmās un citās datu intensīvās nozarēs, kur tā palīdzēs uzlabot gan analīzes rezultātus, gan lēmumu pieņemšanas procesu.

4.1. Datu slāņošanas metodes koncepcijas izveide

Dati, kas nepieciešami, lai optimizētu bišu barību, ietver informāciju un zināšanas par reģionu – atrašanās vietu, reljefu, klimatu, vietējiem nektāra un ziedputekšņu augiem, interneta vai mobilā tīkla pārklājumu; dravas lielumu, bišu sugām un to darbību; laika apstākļiem, īpaši rēķinoties ar nokrišņiem un vēju. Ir arī vides apstākļi, kas ietekmē augu nektāra veidošanos – gaisa temperatūra, relatīvais gaisa mitrums, lietus, vējš, saules stabi, zibeņi un to intensitāte (X. J. He et al., 2016; Hennessy et al., 2020).

Parasti reģionālās biškopības organizācijas, piemēram, Latvijas Biškopības biedrība un Britu biškopju asociācija, ir atbildīga par pārskata sniegšanu saistībā ar vietējiem augiem, to ziedēšanas periodiem un produktivitāti. Šī informācija tiek apkopota ziedēšanas kalendāros. Tomēr ir vairākas problēmas saistībā ar ziedēšanas kalendāru sniegto datu izmantošanu: nav tāda ziedēšanas kalendāra, kas aptvertu visus augus, tāpēc jāizvēlas vietējais ziedēšanas kalendārs; kalendāri sniedz dažādus datus, t. i., ziedēšana mēnesī, ziedēšana sezonā, nektāra un ziedputekšņu daudzums abstraktā mērogā (zems, vidējs, augsts) vai reālās vērtības. Būtībā nav neviena ziedēšanas kalendāra, ko varētu izmantot bez izmaiņām. Koncepcijas demonstrēšanai tika izvēlēts precīzās biškopības projekta ietvaros izveidotais ziedēšanas kalendārs (*Tree Flowering Calendar*, 2020) (sk. 4.1. att.). Kalendārs sniedz informāciju par koku ziedēšanu Etiopijā, attēlojot koku sugas, augu dzimtu, augšanas formu, nektāra un ziedputekšņu daudzumu, izmantojot četras vērtības – nav, "~" mazi daudzumi, "o" vidēji daudzumi un "+" lieli daudzumi, un ziedēšanas statusu mēnesī, izmantojot trīs vērtības – nav, "o" ziedēšana un "+" ziedēšanas maksimuma periods.

Species	Plant family	Growth form	Amharic	Afar	Oromo	Nectar source	Pollen source	January	February	March	April	May	June	July	August	September	October	November	December	Reference
Dracaena steudneri	Agavaceae	tree	CHOWYEH, TABATOS		LANKUSO, MERKO, SHOWYE	o	o	o	o											Fichtl & Adi, 1994
Mangifera indica	Anacardiaceae	tree	MANGO		MANGO	o	o	o	o	o									o	Fichtl & Adi, 1995
Rhus glutinosa	Anacardiaceae	tree	MBS, QMMO		ADESSA, MUGAN, TATESSA	o	o	o									o	o	o	Fichtl & Adi, 1996
Schinus molle	Anacardiaceae	tree	TQUR-BERBERIE			o	+	o	o	o	o	o	o	o	o	+	+	+	+	Fichtl & Adi, 1997
Ozoroa insignis	Anacardiaceae	tree	SHELEL		GARRI	o	o	o											o	Admasu et al. 2014
Adenium obesum	Apocynaceae	tree	-			o	o	o	o	o	o	o	o	o	o	+	+	+	+	Fichtl & Adi, 1998

4.1. att. **Ziedēšanas kalendārs**, daļa (*Tree Flowering Calendar*, 2020).

Demonstrācijas nolūkos tika atlasīti četri augi ar izteiktām ziedēšanas un ražošanas iezīmēm (sk. 4.2. att.): *Grevillea Robusta, Coffea Arabica, Eucalyptus Citriodora un Dichrostachys Cinerea*.

Metodes pārbaudei tika izvēlēti četri augi ar atšķirīgām īpašībām: *Grevillea robusta, Coffea arabica, Eucalyptus citriodora* un *Dichrostachys cinerea*. Šie augi demonstrē atšķirīgus nektāra un ziedputekšņu daudzumus un ziedēšanas periodus, kas ļauj izvērtēt datu normalizācijas un standartizācijas metožu pielietojumu.

Grevillea robusta uzrāda konstantu ziedēšanas periodu ar vidēju nektāra un ziedputekšņu produkciju. Šāda veida datu normalizācija ļauj kvantitatīvi novērtēt resursu pieejamību ilgtermiņā. *Coffea arabica* demonstrē sezonālu ziedēšanas rakstu ar mainīgu nektāra un ziedputekšņu produkciju gada griezumā. Datu standartizācija šādos gadījumos nodrošina sezonālo svārstību precīzāku kvantitatīvo novērtējumu. *Eucalyptus citriodora* nodrošina nektāra un ziedputekšņu resursus visa gada garumā. Vēsturisko datu analīze šī auga gadījumā sniedz kvantitatīvu informāciju par resursu pieejamību dažādos laika periodos. *Dichrostachys cinerea* uzrāda zemāku ziedēšanas intensitāti un resursu daudzumu. Šī auga datu analīze ļauj pārbaudīt metodikas efektivitāti ierobežotu resursu apstākļos.

Datu apvienošanas procesā tiek izmantota sistemātiska pirmsapstrāde. Normalizācija un standartizācija nodrošina datu pielāgošanu analīzei neatkarīgi no to sākotnējā sadalījuma vai mēroga. Šīs metodes tiek izmantotas ziedēšanas kalendāru veidošanā, apstrādājot datus par ziedēšanas intensitāti, nektāra un ziedputekšņu daudzumu. Normalizācija tiek pielietota datiem, kas neveido normālu sadalījumu, nodrošinot kvantitatīvu salīdzinājumu starp dažādu augu ziedēšanas periodiem un resursu daudzumiem. Piemēram, *Dichrostachys cinerea* un *Coffea arabica* datu salīdzināšanā normalizācija nodrošina objektīvu resursu novērtējumu. Standartizācija tiek izmantota datiem ar normālu sadalījumu, piemēram, apstrādājot *Eucalyptus citriodora* un *Grevillea robusta* datus. Šī metode nodrošina datu svārstību kvantitatīvu novērtējumu kopējos rezultātos.

Lai izveidotu datu kopu, tika piemērota normalizācija (sk. (4.1.)), izmantojot šādus raksturlielumus:

- nektārs, *n*:

- nav: 0;
- mazs daudzums: 0,5;
- vidējs daudzums: 1;
- liels daudzums: 1.5.

- ziedputekšņi, *p*:
  - nav: 0;
  - mazs daudzums: 0,5;
  - vidējs daudzums: 1;
  - liels daudzums: 1.5.
- ziedēšana, *fl*:
  - nav: 0;
  - ziedēšana notiek: 1;
  - ziedēšanas maksimums: 2.

Izmantojot normalizētās vērtības, tika aprēķināta augu bagātība:

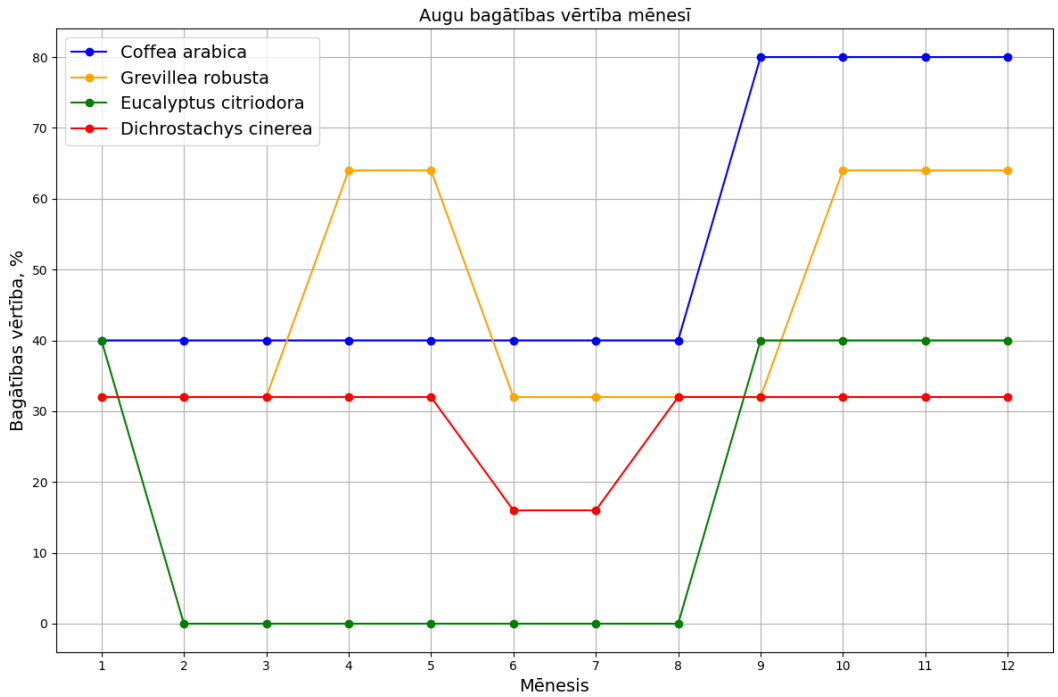
$$\mathrm{PR}_t = \left(\frac{fl_t \times (1 + n_t) \times (1 + p_t)}{MaxPR} \right) \times 100 \qquad (4.1.)$$

kur

PR – augu bagātība, %;
MaxPR – maksimāla augu bagātība, decimāldaļa;
t – laika periods, mēnesis/nedēļa/diena/stunda/minūte.

Šajā piemērā *t* vienāds ar vienu mēnesi un maksimālā augu bagātība aprēķināta, izmantojot augstākās vērtības: *n* = 1,5, *p* = 1,5 un *fl* = 2. Ziedēšanas kalendāra ietvaros Etiopijas nektāraugiem normalizācijas faktors *MaxPR* ir vienāds ar 12,5.

Atlasītie augi veido vienu datu kopu. Grafikā (sk. 4.2. att.) attēlota augu bagātības vērtības dinamika gada laikā. Kā redzams attēlā, "*Grevillea Robusta"* ir divi periodi, kad augu bagātības vērtība ir vienāda ar 65% no aprīļa līdz maijam un no oktobra līdz decembrim un ar augu bagātības vērtību 30% citos mēnešos. Salīdzinājumam, "*Coffea Arabica*" no janvāra līdz augustam ir stabila vērtība 40%, bet no septembra līdz decembrim – līdz 80%. Katra konkrēta datu ievade datu kopā pieņemta kā atsevišķs datu slānis – katram konkrētam augam.

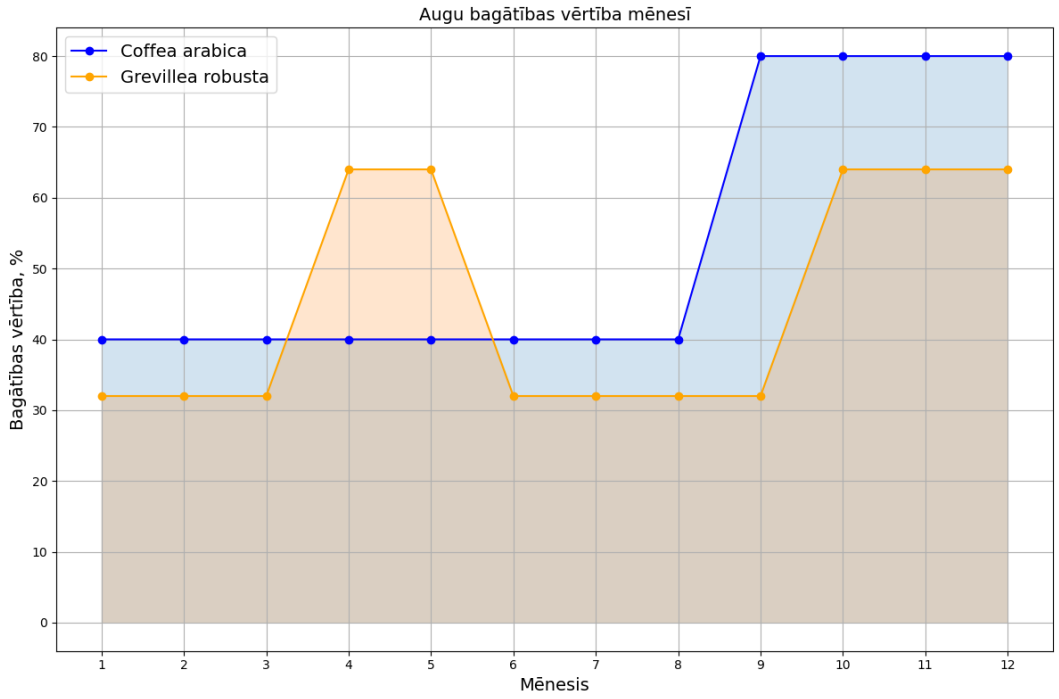


4.2. att. **Augu bagātības dati četriem atlasītajiem augiem.**

Salīdzinot *Coffea arabica* un *Grevillea* robusta, vairākos parametros iespējams izcelt katra auga stiprās un vājās puses. *Coffea arabica* ir labāks nektāra avots, jo tā nektāra daudzums ir lielāks (1,50) nekā *Grevillea robusta* (1,00), kas nodrošina vairāk enerģijas bitēm. Tomēr abi augi nodrošina vienādu ziedputekšņu daudzumu (1,00), tādējādi šajā ziņā tie ir vienādi. Ziedēšanas intensitātes ziņā *Coffea arabica* ir pārāks, jo tā ziedēšanas vērtība ir augstāka (no 5 līdz 10), salīdzinot ar *Grevillea robusta* (no 4 līdz 8), īpaši rudenī, kad *Coffea arabica* sasniedz maksimumu. Savukārt sezonalitātes ziņā *Grevillea robusta* ir labāks, jo zied visos mēnešos, nodrošinot stabilu barības avotu, kamēr *Coffea arabica* koncentrējas uz intensīvu ziedēšanu rudenī. Bagātības ziņā *Coffea arabica* ir priekšā ar vērtību līdz pat 80 rudenī, bet *Grevillea robusta* maksimums ir 64 pavasarī un rudenī, kas padara *Coffea arabica* labāku rudenī. Tomēr *Grevillea robusta* nodrošina stabilus resursus visa gada garumā, kas ir priekšrocība salīdzinājumā ar sezonālo *Coffea arabica*. Kopumā *Coffea arabica* ir labāks īstermiņā intensīvu resursu pieejamības laikā rudenī, savukārt *Grevillea robusta* piedāvā stabilu barību ilgtermiņā visa gada garumā.

Metodoloģijas izstrādei tika īpaši izraudzīti divi augi ar ļoti atšķirīgām īpašībām – *Coffea arabica* un *Grevillea* robusta. Šī izvēle nav nejauša, jo abi augi sniedz iespēju pārbaudīt metodikas efektivitāti pilnīgi pretējos apstākļos. *Coffea arabica* ir kā bagātīgs, bet īslaicīgs svētku galds bitēm – noteiktos gada periodos tas piedāvā pārsteidzoši daudz nektāra un ziedputekšņu. Šāds sezonāls raksturs ir ideāls, lai pārbaudītu, kā metodika tiek galā ar īslaicīgiem, bet ļoti intensīviem resursu periodiem. Savukārt *Grevillea robusta* ir kā uzticams ikdienas maltītes nodrošinātājs – tas piedāvā vidēju, bet stabilu nektāra un ziedputekšņu daudzumu visa gada garumā. Šis augs ļauj pārbaudīt metodikas spēju novērtēt un optimizēt pastāvīgus, ilgtermiņa resursus. Šāda pretstatu kombinācija ļauj metodikai pielāgoties dažādām biškopju vajadzībām – gan tiem, kas vēlas izmantot intensīvos sezonālos resursus, gan tiem, kas dod priekšroku stabilai, prognozējamai bišu barības bāzei visa gada garumā. Tieši šī daudzveidība padara abus augus par ideāliem kandidātiem metodikas vispusīgai pārbaudei.

Daļā no šīs datu kopas, kas attēlo divus atšķirīgākus parametrus, izmantojot apgabalu diagrammu, ir redzami (sk. 4.3. att.) divi šīs datu kopas datu slāņi, kas pārklājas.



4.3. att. **Divu atlasīto augu datu slāņi.**

Pamatojoties tikai uz šo informāciju, var pieņemt, ka laukos ar *Coffea Arabica* kopā iegūst vairāk medus; tādēļ būtu ekonomiski izdevīgi dravas atrašanās vietu izvēlēties *Coffea Arabica* lauku tuvumā. Ir iespējams noteikt apgabala proporciju katram augam:

1. tiek aprēķināts katra auga nepārklājamā reģiona laukums;
2. tiek aprēķināta kopējā platība zem katra auga līknes;
3. tiek aprēķināta nepārklājošā reģiona proporcija katra auga kopējā platībā;
4. laukums zem katras līknes tiek tuvināts, izmantojot trapecveida likumu. Apgabals, kas nepārklājas, būtībā ir atšķirība starp divām līknēm (kur viena ir lielāka par otru), kas integrēta visā domēnā.

Trapecveida metode ir metode funkcijas noteiktā integrāļa tuvināšanai. Tas darbojas, sadalot laukumu zem līknes trapecveida formās un pēc tam aprēķinot katras no tām laukumu. Šo trapecveida formu kopējais laukums tuvina funkcijas integrāli šajā intervālā.

Matemātiski ņemot vērā punktus \((x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)),integrālis tiek tuvināts šādi (sk. (4.2.)):

$$Area \approx \frac{1}{2}\sum_{i=1}^{n}(x_i - x_{i-1}) \cdot (y_i - y_{i-1}) \tag{4.2.}$$

kur

x_i un x_{i-1} – secīgas x vērtības;
y_i un y_{i-1} – ir atbilstošās funkcijas vērtības;
x_i – mēnesis, vesels skaitlis;
y_i – auga bagātības vērtība.

Attiecīgi kopējā platība zem katra auga līknes (sk. (4.3.)) ir auga bagātības vērtību integrālis pa mēnešiem:

$$\text{Total Area of Plant} \approx \frac{1}{2}\sum_{i=1}^{n}(x_i - x_{i-1})\left(y_{\text{plant},i} + y_{\text{plant},i-1}\right) \tag{4.3.}$$

Nepārklājošā reģiona laukums ir integrālis starp divu augu bagātības vērtību starpību, ja viena ir lielāka par otru (sk. (4.4.) (4.5.)).

Grevillea robusta:

$$\text{Area} \approx \tfrac{1}{2}\textstyle\sum_{i=1}^{n}(x_i - x_{i-1})(\text{difference}_i + \text{difference}_{i-1}), \text{ ja atšķirība} > 0 \tag{4.4.}$$

Coffea arabica:

$$\text{Area} \approx \tfrac{1}{2}\textstyle\sum_{i=1}^{n}(x_i - x_{i-1})(-\text{difference}_i - \text{difference}_{i-1}), \text{ ja atšķirība} > 0 \tag{4.5.}$$

Parametrs "atšķirība" atspoguļo punktu starpību starp abu augu – *Grevillea Robusta* un *Coffea Arabica* – bagātības vērtībām katrā mēnesī.

Matemātiski to aprēķina šādi (sk. 4.6.):

$$\text{difference}_i = y_{\text{Grevillea Robusta},i} - y_{\text{Coffea Arabica},i}, \tag{4.6.}$$

kur

difference_i – i^{th} mēneša atšķirība, vesels vai decimāls skaitlis;
$y_{\text{Grevillea Robusta},i}$ – *Grevillea Robusta* bagātības vērtība i^{th} mēnesī, vesels vai decimāls skaitlis;
$y_{\text{Coffea Arabica},i}$ – *Coffea Arabica* bagātības vērtība i^{th} mēnesī, vesels vai decimāls skaitlis.

Šī atšķirība nosaka, kuram augam ir lielāka bagātības vērtība konkrētajā mēnesī. Ja starpība ir pozitīva, tas norāda, ka *Grevillea Robusta* šajā mēnesī ir augstāka bagātības vērtība nekā *Coffea Arabica*; un otrādi, ja atšķirība ir negatīva, *Coffea Arabica* ir augstāka bagātības vērtība.

Platības, kas nepārklājas grafikā, nosaka šādas atšķirības:

- ja starpība ir pozitīva, nepārklājošais reģions atrodas starp *Grevillea robusta* un *Coffea Arabica* līknēm;
- ja starpība ir negatīva, nepārklājošais apgabals atrodas starp *Coffea Arabica* un *Grevillea Robusta* līknēm.

Nepārklājamā reģiona īpatsvaru katram augam aprēķina šādi (sk. (4.7.)):

$$\text{Proportion} = \left(\frac{\text{Non-overlapping Area}}{\text{Total Area of Plant}}\right) \times 100 \qquad (4.7.)$$

Šajā gadījumā, katra auga nepārklājošā reģiona (punktēta apgabala) proporcija salīdzinājumā ar tā pilno reģionu ir:

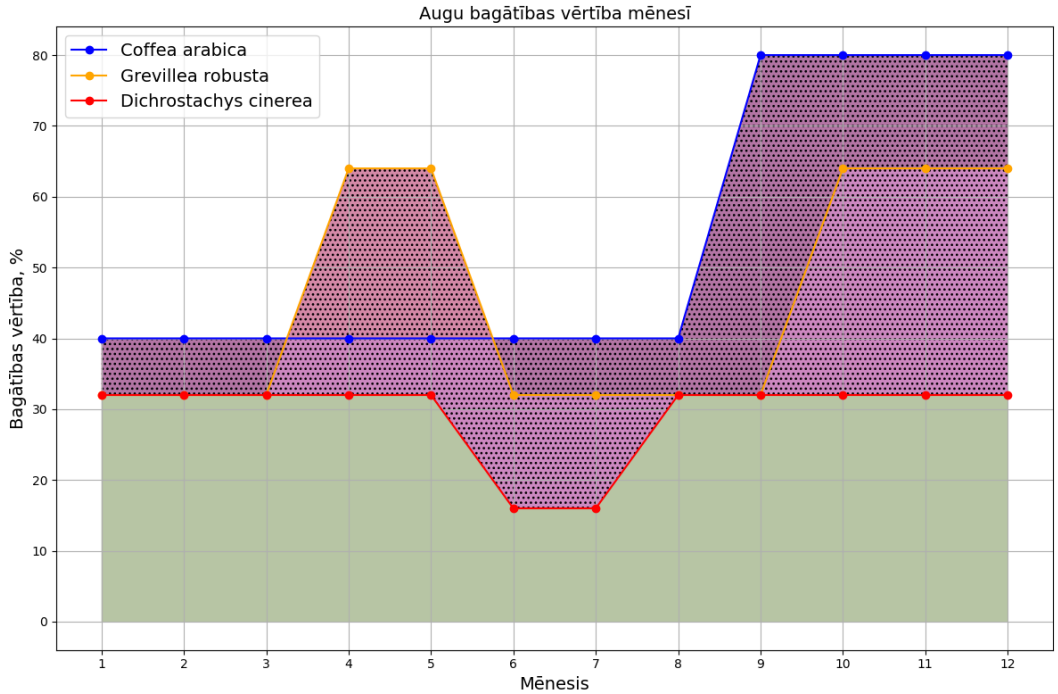
- *Grevillea Robusta*: aptuveni 9,68% no visa tās reģiona;
- *Coffea Arabica*: aptuveni 22,76% no visa tā reģiona.

Konkrēti, proporcija mēra, cik liela daļa no katra auga bagātības līknes netiek dalīta ar otru augu. Lūk, ko nosaka šī proporcija:

- ja vienam augam ir liels nepārklājošu reģionu īpatsvars, tas norāda, ka ievērojama daļa tā bagātības vērtību mēnešu laikā atšķiras no otra auga. Tas varētu nozīmēt, ka abiem augiem ir atšķirīgi bagātības maksimuma periodi vai ka vienam augam vienmēr ir augstākas bagātības vērtības nekā otram;
- proporcija palīdz izprast augu salīdzinošo uzvedību laika gaitā. Piemēram, ja vienam augam ir daudz lielāks reģionu, kas nepārklājas, īpatsvars nekā otram, tas liecina, ka tā bagātības vērtības vairāk atšķiras no otra auga;
- ekoloģiskā kontekstā, ja diviem augiem ir augsta bagātības vērtību pakāpe, kas nepārklājas, tas var norādīt, ka tiem ir atšķirīgi ziedēšanas vai ražības periodi. Tā var būt būtiska informācija, lai izprastu bioloģisko daudzveidību, apputeksnētāju darbību vai izstrādātu stādīšanas shēmas nepārtrauktai ziedēšanai dārzos.

Rezumējot, šī proporcija sniedz kvantitatīvu mērījumu tam, cik ļoti katra auga bagātības vērtības atšķiras no cita auga, sniedzot ieskatu to salīdzinošajā uzvedībā un iespējamajās ekoloģiskajās ietekmēs.

Proporciju aprēķins var tikt pielietots vairākiem objektiem, kuri tiek salīdzināti pret vienu parametru. Piemēram (sk. 4.4. att.), pievienojot trešo augu, zonas izskatās šādi:



4.4. att. **Trīs augu reģionu pārklāšanās.**

Rezultātā katra auga nepārklājošo reģionu proporcijas attiecībā pret visu reģionu ir šādas:

- *Grevillea Robusta*: aptuveni 45,16% no visa reģiona;
- *Coffea Arabica*: aptuveni 67,59% no visa tā reģiona;
- *Dichrostachys cinerea*: 0,0%. Tas norāda, ka *Dichrostachys Cinerea* bagātības vērtības attiecīgajos mēnešos pilnībā pārklājas ar citiem augiem.

Tas notiek, jo līdz ar *Dichrostachys Cinerea* iekļaušanu tagad ir papildu reģioni, kuros *Grevillea Robusta* vai *Coffea Arabica* nepārklājas ar *Dichrostachys Cinerea* . Tas palielina abu šo augu kopējo nepārklājamo platību. Veidojas kumulatīvais efekts, funkcija aprēķina reģionu, kas nepārklājas katram augam, ņemot vērā atšķirības ar citiem augiem. Koncepcija, kā aprēķināt starpību starp diviem objektiem (piemēram, augiem) noteiktam parametram (piemēram, bagātībai), paliek nemainīga. Tomēr tagad katra objekta atšķirība tiek aprēķināta ar katru citu objektu, nevis tikai ar vienu citu objektu. Tātad *Grevillea Robusta* un *Coffea Arabica* reģioni, kas nepārklājas, tagad ietver gan to atšķirības, gan atšķirības ar *Dichrostachys cinerea*. *Dichrostachys Cinerea* proporcija ir 0,0%, kas norāda, ka tās bagātības vērtības pilnībā pārklājas ar pārējiem diviem augiem. Tas nozīmē, ka pārējiem diviem augiem ir papildu reģioni, kas nepārklājas (ar *Dichrostachys Cinerea*), un tas veicina to proporcijas. Būtībā papildu auga iekļaušana analīzē var ieviest jaunus nepārklājošus reģionus, kas iepriekš netika ņemti vērā, tādējādi ietekmējot aprēķinātās proporcijas.

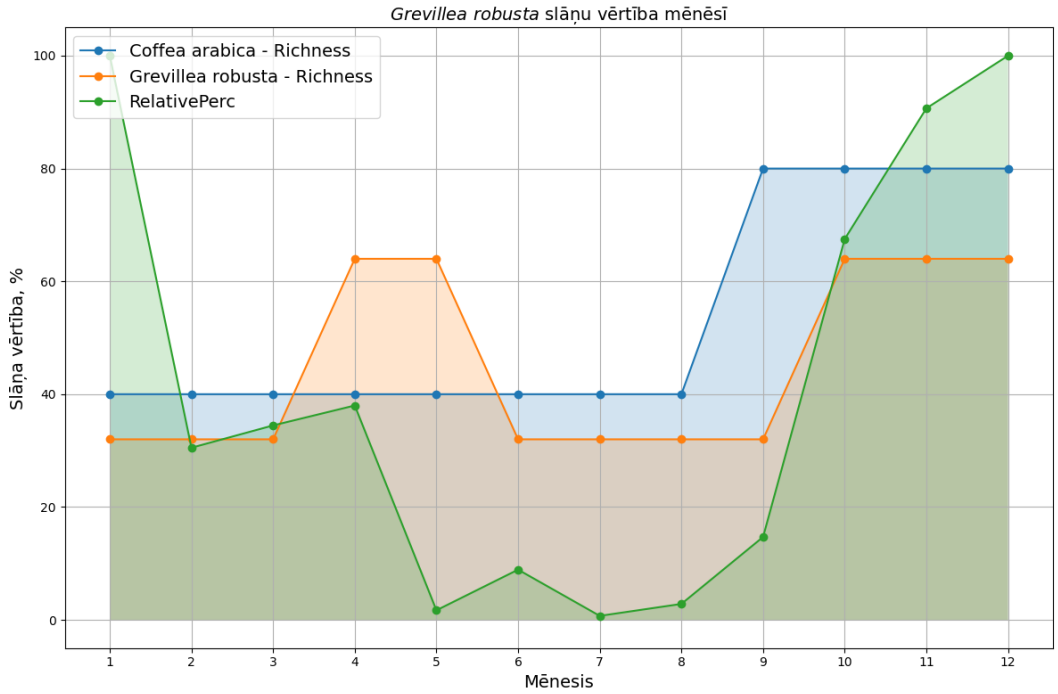
Otrais datu slānis, kas parāda slāņošanas nozīmi, ir nokrišņu slānis. Ir pierādīts (X. J. He et al., 2016), ka lietus ir viens no barības meklēšanas efektivitāti noteicošajiem faktoriem. Līdzīgi kā ziedēšanas kalendāros, arī nokrišņu dati ir reģionāli. Tabulā 4.1. attēlotie nokrišņu dati parāda mēneša nokrišņu daudzumu collās, centimetros un procentuālo vērtību (*RelativePerc*). Nokrišņu slānis ir būtisks bišu barības meklēšanas efektivitātes noteicošais faktors, jo lietus ietekmē gan augu ziedēšanu, gan nektāra izdalīšanos, kā arī bišu lidošanas iespējas. Šie dati ir parādīti parabolas veidā, kur maksimumi vērojami pirmajos un pēdējos gada mēnešos, kas atspoguļo lietus sezonas sākumu un beigas. Nokrišņu dati ir pārveidoti par relatīvajām procentu vērtībām, lai korelētu ar augu ziedēšanas datiem diapazonā no 0 līdz 100. Nokrišņu procentuālie dati palīdz iegūt skaidrāku priekšstatu par to, cik lielā mērā lietus ietekmē augu ziedēšanu un nektāra izdalīšanos katrā mēnesī. Piemēram, janvārī un decembrī nokrišņi sasniedz maksimumu – 100%, kas norāda uz intensīvu lietus sezonu, bet jūlijā tie samazinās līdz minimālajai vērtībai – 0.71%, kas ir sausā sezona.

4.1. tabula. **Nokrišņu dati**

Mēnesis	**Daudzums, collas**	**Daudzums, cm**	***RelativePerc*, %**
1	8.12	20.62	100.00
2	2.16	5.49	30.51
3	2.44	6.20	34.46
4	2.69	6.83	37.99
5	0.12	0.30	1.69
6	0.63	1.60	8.90
7	0.05	0.13	0.71
8	0.2	0.51	2.82
9	1.04	2.64	14.69
10	4.78	12.14	67.51
11	6.42	16.31	90.68
12	7.08	17.98	100.00

Nokrišņu dati parāda ievērojamas sezonālās svārstības gada laikā, ar vidējo daudzumu 3,08 collas jeb 7,95 cm mēnesī un relatīvo procentuālo daudzumu 42,88%. Maksimālais nokrišņu daudzums ir janvārī un decembrī, ar 8,12 collām jeb 20,62 cm (attiecīgi 100% relatīvais daudzums), savukārt minimālais daudzums tiek novērots jūlijā ar tikai 0,05 collām jeb 0,13 cm (relatīvais daudzums 0,71%). Standartnovirze ir 2,79 collas jeb 7,20 cm, kas norāda uz būtiskām svārstībām starp dažādiem mēnešiem. Pirmajā un pēdējā ceturksnī nokrišņu

daudzums ir ievērojami augstāks, bet sausākie mēneši ir no maija līdz augustam, kad nokrišņu relatīvais daudzums svārstās no 0,71% līdz 8,90%. Izkliede starp minimālo un maksimālo vērtību ir ievērojama – 8,07 collas jeb 20,49 cm, kas norāda uz skaidru sezonālo sadalījumu, ar lietus sezonas maksimumu ziemā un minimālu nokrišņu daudzumu vasarā. Līniju diagramma, kas attēlo mēneša nokrišņu datus, parāda izteiktu sezonālu svārstību, kurā parabolas veidā ir augstākie punkti lietus sezonā. Šajos datos ir redzama skaidra nokrišņu koncentrācija ziemas mēnešos – janvārī un decembrī nokrišņu daudzums sasniedz maksimumu ar 8,12 collām jeb 20,62 cm (100% relatīvais daudzums), kamēr vasaras mēnešos, īpaši jūlijā, nokrišņu daudzums samazinās līdz minimumam ar 0,05 collām jeb 0,13 cm (0,71% relatīvais daudzums). Pamata līmenī šis nokrišņu datu slānis var tikt balstīts uz vēsturiskajiem datiem, kas atspoguļo tipiskos sezonālos modeļus. Demonstrācijas nolūkos tika iegūti mēneša nokrišņu dati parabolai tuvā formā (sk. 4.5. att., Nokrišņu vērtība, *RelativePerc*). Augu ziedēšanas un nokrišņu datu kopu korelācijai tika izmantota konvertēšana, lai attēlotu nokrišņu datus procentu vērtībās, nevis nokrišņu daudzuma absolūtos skaitļos (citādi attēloti centimetros vai collās). No šī brīža katra datu slāņa vērtība tiek uzskatīta par vispārīgu "datu slāņa vērtību" diapazonā no 0 līdz 100.



4.5. att. **Nokrišņu datu slānis pret divu augu slāņiem.**

Nokrišņu datu pievienošana augu ziedēšanas datiem ļauj iegūt pilnīgāku ainu par to, kā klimatiskie apstākļi ietekmē augu bišu barošanas iespējas. Augiem, piemēram, *Grevillea robusta* un *Coffea arabica*, kas zied intensīvi pavasarī un rudenī, nokrišņu daudzums var būt izšķirošs faktors nektāra izdalīšanās intensitātei. *Coffea arabica* rudenī zied visintensīvāk, kad nokrišņu daudzums ir augsts – piemēram, novembrī nokrišņi ir 90,68%, kas sakrīt ar ziedēšanas maksimumu, un tas var veicināt augstāku nektāra daudzumu. Savukārt mēnešos ar zemu nokrišņu daudzumu, piemēram, jūlijā, kad *Grevillea robusta* vēl zied, lietus trūkums spēj ietekmēt nektāra un ziedputekšņu ražošanu. Pēc nokrišņu slāņa pievienošanas šai demonstrācijai kļuva grūtāk noteikt, vai *Grevillea Robusta* vai *Coffea Arabica* dod lielāku ražu. *Coffea Arabica* kopējā "bagātība" (580,0) ir augstāka nekā *Grevillea Robusta* (496,0) novērotajā periodā. Tas varētu norādīt, ka *Coffea Arabica* šajā laika posmā ir labvēlīgāka reakcija uz vides apstākļiem vai ka tai dabiski ir augstākas bagātības vērtības. Sākumā var šķist, ka nav tiešas proporcionālas attiecības starp nokrišņu daudzumu un augu bagātību abiem augiem, vismaz ne konsekventi visā periodā. Ir brīži, kad bagātība it kā palielinās līdz ar lietu, bet ir arī brīži, kad tā samazinās vai paliek nemainīga. Praksē, domājams, nokrišņu daudzums jāanalizē, koncentrējoties uz konkrētu reģionu, jo pat mazās valstīs ir atšķirīgs nokrišņu

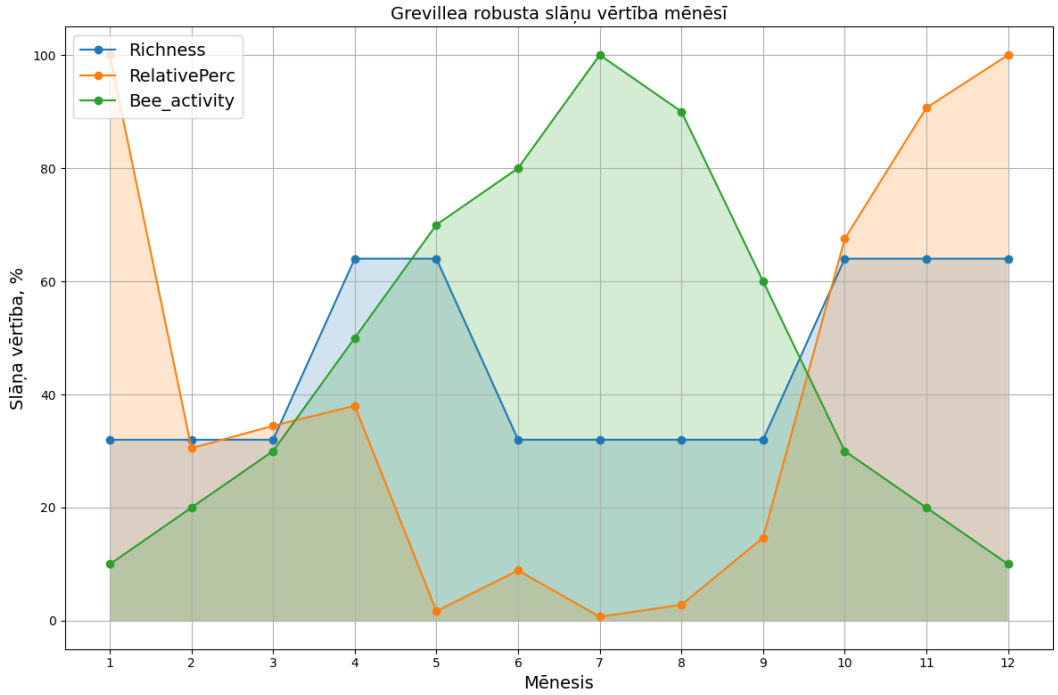
sadalījums; tāpēc datu slāņiem jābūt balstītiem uz atrašanās vietu. Turpmākai koncepta īstenošanai tiek izmantots *Grevillea Robusta* augs, jo tā dinamika ir "interesantāka".

Bišu aktivitātes datu slānis, kā arī biškopības kalendārs palīdz nodrošināt pilnvērtīgāku ainu par to, kā plānot stropu pārvaldību dažādos gada posmos. Bišu aktivitātes līmeņi tabulā 4.2. norāda, kad bites ir aktīvākas un kad to aktivitāte samazinās līdz minimumam. Bišu aktivitātes dati parāda, ka vidējā aktivitāte mēnesī ir 4,92, bet vidējā bišu aktivitāte ir 50%, ar zemāku mediānu, kas ir 4 aktivitātei un 40% bišu aktivitātei, norādot, ka vairāk datu ir koncentrēti zemākā aktivitātes līmenī. Moda, kas ir visbiežāk sastopamā vērtība, aktivitātei ir 1, bet bišu aktivitātei – 10%, kas norāda uz to, ka viszemākā aktivitāte ir visbiežākā. Standartnovirze parāda, ka aktivitāte ir relatīvi stabila, ar standartnovirzi 2,87, savukārt bišu aktivitātei novirze ir plašāka – 29,28%, kas atspoguļo lielākas izmaiņas aktivitātes līmeņos gada griezumā. Minimālā un maksimālā vērtība aktivitātei ir attiecīgi 1 un 10, savukārt bišu aktivitātei – 10% un 100%, ar izkliedi 9 aktivitātei un 90% bišu aktivitātei, kas norāda uz lielu svārstību diapazonu. Dati skaidri parāda, ka bišu aktivitāte pieaug pavasara un vasaras mēnešos, sasniedzot maksimumu jūlijā, bet ziemas mēnešos tā ievērojami samazinās. Šī informācija ir īpaši svarīga, lai nodrošinātu, ka stropi tiek uzturēti atbilstoši, ņemot vērā bišu aktivitātes un barošanas vajadzības. Pamatdarbības periods ietver peru audzēšanu, spietošanu, aktīvās barības meklēšanas u. c. Ierasts kā vadlīniju izmantot atsauci uz vietējo biškopības kalendāru, jo šāds avots mēdz sniegt gan informāciju par bišu aktivitātēm, gan nepieciešamajām biškopības aktivitātēm (*Beekeeping Calendar*, n.d.). Katru darbību ir grūti novērtēt ar vērtību, tāpēc demonstrējumu nolūkos tika izvēlēta skala no 1 līdz 10.

4.2. tabula. **Bišu aktivitāte**

Mēnesis	**Aktivitāte**	**Bee_activity**
1	1	10
2	2	20
3	3	30
4	5	50
5	7	70
6	8	80
7	10	100
8	9	90
9	6	60
10	3	30
11	2	20
12	1	10

Apvienojot šo bišu aktivitātes informāciju ar nokrišņu datiem un augu ziedēšanas laikiem, kā arī biškopības kalendāru, iespējams izstrādāt visaptverošu metodoloģiju, kas ļauj biškopjiem optimizēt savas darbības. Piemēram, augi, kas zied vasarā, kad bišu aktivitāte ir augsta, būs vērtīgs barības avots, īpaši augiem kā *Grevillea robusta*, kas zied visa gada garumā, un *Coffea arabica*, kas rudenī sasniedz ziedēšanas maksimumu. Ja šie augi ziedēs laikā, kad bites ir aktīvas un laika apstākļi ir piemēroti (ar atbilstošu nokrišņu daudzumu), tas maksimizēs bišu barības savākšanas efektivitāti. Šī skala atspoguļo hipotētisko bišu aktivitāti konkrētajā mēnesī, ieskaitot iepriekšminēto notikumu iespējamību (sk. 4.6. att.).



4.6. att. **Bišu aktivitātes datu slānis.**

Mērogošanas nolūkos šīs vērtības tika pārveidotas procentos, un demonstrācijai tika pievienots bišu aktivitātes datu slānis.

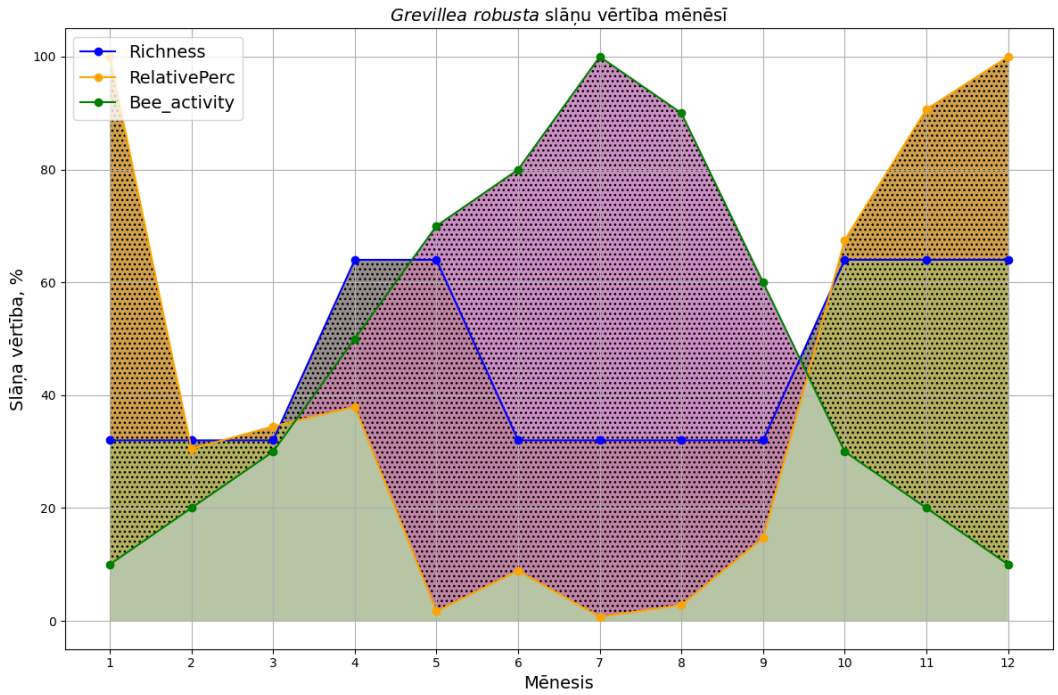
Tālāk ir norādītas to reģionu proporcijas, kas nepārklājas katram parametram (sk. 4.7. att.):

- bagātība: 19,35%;
- nokrišņi: 57,91%;
- bišu aktivitāte: 51,79%.

Šīs proporcijas atspoguļo katras līknes procentuālo daļu, kas nepārklājas ar pārējām divām līknēm:

- tikai 19,35% no bagātības līknes *Richness* nepārklājas ar pārējām divām līknēm, norādot, ka bagātības modelis būtiski pārklājas ar nokrišņu daudzumu un bišu procentuālajiem modeļiem pa mēnešiem;
- 57,91% no nokrišņu līknes *RelativePerc* un 51,79% no bišu aktivitātes līknes nepārklājas ar citām līknēm, norādot, ka šiem diviem parametriem ir unikālāki modeļi, *Bee_activity* salīdzinot ar auga bagātību.

No šīs analīzes var secināt, ka, lai gan pastāv zināma saistība starp augu bagātību un citiem parametriem, ievērojama daļa nokrišņu un bišu procentuālā daudzuma nav atkarīga no augu bagātības. Tas liecina, ka citi ārēji faktori var ietekmēt šos modeļus.



4.7. att. **Trīs slāņu pārklāšanās.**

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 $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$, lineārā interpolācija ir y vērtības noteikšanas process konkrētam x , izmantojot formulu (sk. (4.8.)):

$$y = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) \tag{4.8.}$$

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. (4.9.)):

$$\text{weighted\_value}_i = \sum_{j=1}^{n} \text{weight}_j \times \frac{\text{parameter}_j}{100} \tag{4.9.}$$

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:

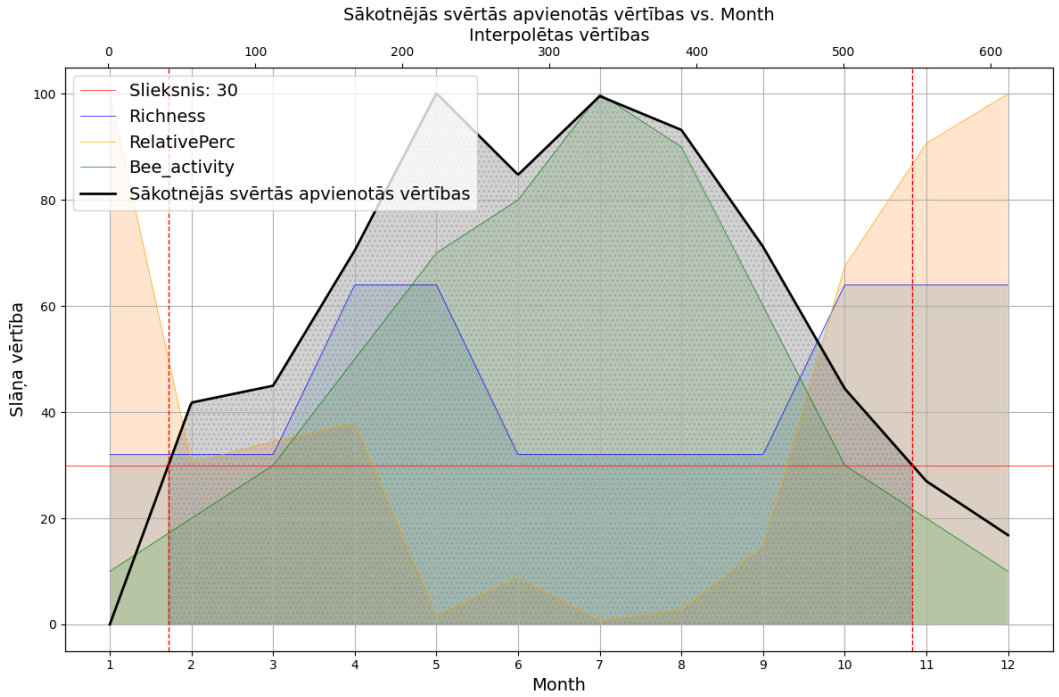
1. tiek atņemta vidējā vērtība: $X_{mean} = X - \bar{X}$;
2. tiek aprēķināta kovariācijas matrica Σ no X_{mean};
3. tiek aprēķināts Σ īpašvektors un īpašvērtība;
4. tiek izvēlēts pazīmju vektors, kas saistīts ar lielāko īpašvērtību.

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:

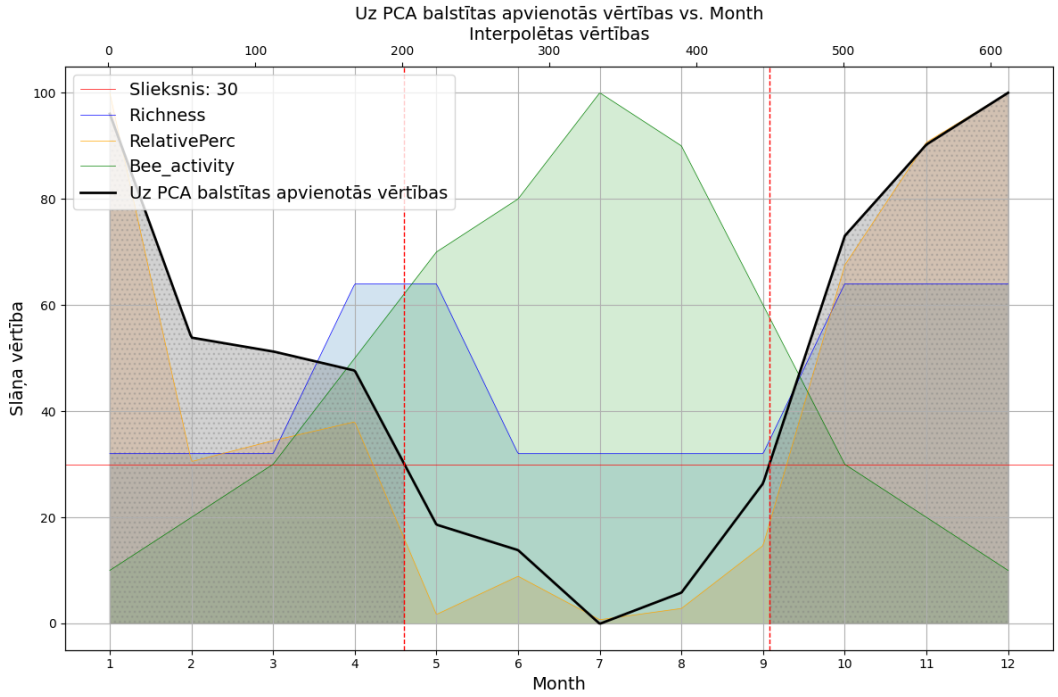
ja fused_value ≥ slieksnis, reģions tiek iezīmēts.

Rezultātā tiek iegūtas divas diagrammas. Sākotnējo apvienoto vērtību diagramma, kur (sk. 4.8. 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.



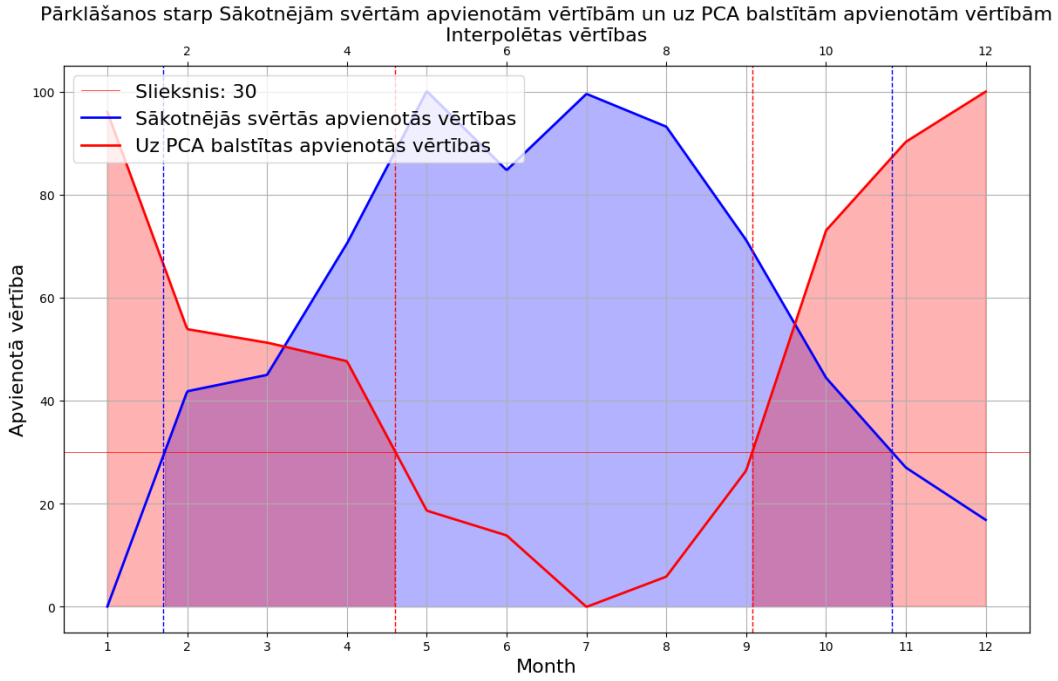
4.8. att. **Sākotnējo apvienoto vērtību diagramma.**

Uz *PCA* balstīto apvienoto vērtību diagrammā (sk. 4.9. 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.



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

Izmantojot šīs diagrammas, var vizuāli salīdzināt abas datu apvienošanas metodes un izprast abus nozīmīgos reģionus (tos, kas pārsniedz slieksni). Tas var palīdzēt identificēt mēnešus, kuros ir izpildīti vai pārsniegti noteikti nosacījumi (ko attēlo apvienotās vērtības). Turpmākai analīzei tiek izmantots abu apvienošanas metožu pārklāšanas diagramma (sk. 4.10. att.).



4.10. att. **Apvienoto vērtību pārklāšanās.**

Turpmāk var pielietot iepriekš aprakstīto 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.

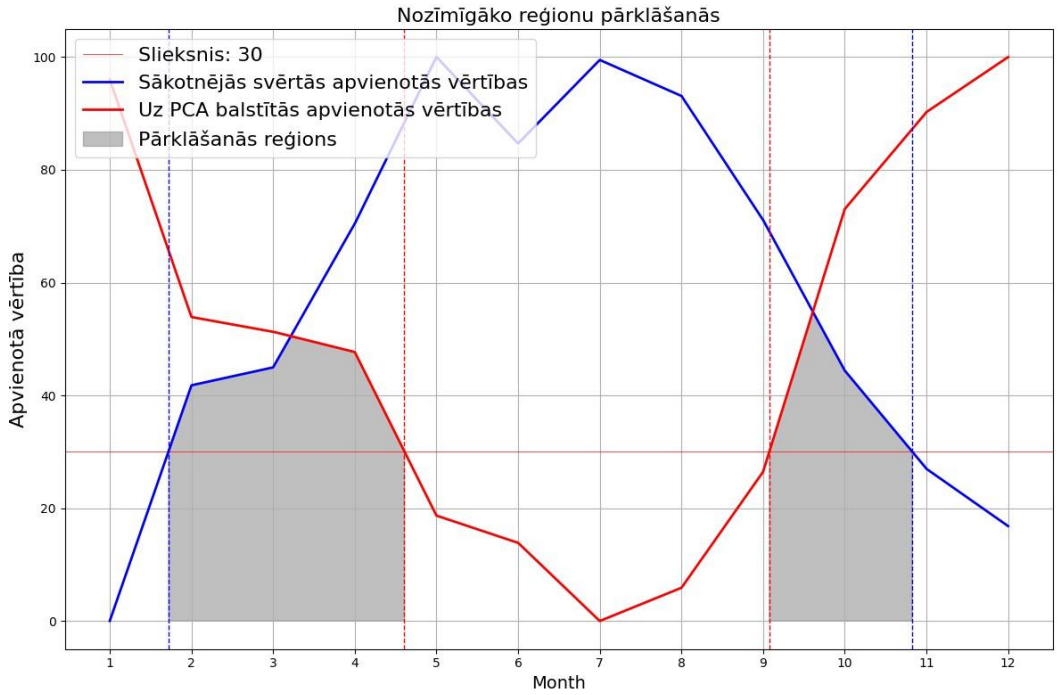
1. Tiek aprēķināta atšķirība starp divām datu kopām:
   - katram punktam x_i domēnā tiek aprēķināta atšķirība Δy_i starp abām datu kopām: $\Delta y_i = y1_i - y2_i$.
2. Tiek noteikti reģioni, kas pārklājas:
   - ja Δy_i un Δy_{i+1} ir viena un tā pati zīme, tad abas datu kopas nekrustojas starp x_i un 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. (4.10.)):
   - reģioniem, kas nepārklājas starp x_i un x_{i+1}:

$$Area = \frac{x_{i+1} - x_i}{2} \times (|y_{1i} + y_{1i+1}| - |y_{2i} + y_{2i+1}|) \qquad (4.10.)$$

   - 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. 4.11. 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.



4.11. att. **Apvienoto vērtību pārklājošie reģioni.**

4.2. Datu slāņošanas metodes pielietošana

Datu slāņošanas metode ir izstrādāta *Python* vidē, izmantojot pandas bibliotēku datu apstrādei, *numpy* skaitliskajiem aprēķiniem un *matplotlib* vizualizācijai. Metodes pamatā ir vairāku soļu process, kas ļauj apvienot un analizēt datus no dažādiem avotiem:

1. Datu sagatavošana. Pirmais solis ir datu kopu izvēle un sagatavošana. Šeit svarīgi, lai visām datu kopām būtu vismaz viena kopīga kolonna – parasti tā ir laika ass, piemēram, mēnesis. Lietotājs no katras datu kopas izvēlas interesējošās kolonnas, piemēram, nektāra daudzumu vai bišu aktivitāti.
2. Datu izlīdzināšana. Pirms apvienošanas dati tiek izlīdzināti, izmantojot lineāro interpolāciju. Lietotājs var izvēlēties izlīdzināšanas pakāpi – jo tā augstāka, jo vairāk starppunktu tiek aprēķināti. Tas palīdz iegūt vienmērīgāku un detalizētāku ainu par procesiem laikā.
3. Automātiskā izlīdzināšana. Ja precīza izlīdzināšanas pakāpe nav zināma, var izmantot automātisko režīmu. Tas meklē optimālo izlīdzināšanas pakāpi, balstoties uz noteikto slieksni – punktu, kur dati kļūst nozīmīgi. Šī pieeja ir īpaši noderīga, strādājot ar nezināmiem vai mainīgiem datiem.
4. Datu filtrēšana. Metode piedāvā iespēju atlasīt konkrētu datu apakškopu, piemēram, fokusēties uz vienu augu vai laika periodu. Tas ļauj detalizētāk izpētīt interesējošos aspektus.
5. Datu apvienošana. Apvienošana notiek divos veidos – ar svērto summu un ar galveno komponentu analīzi (PCA). Svērtā summa ļauj lietotājam kontrolēt katra parametra ietekmi, kamēr PCA automātiski atrod galvenās datu tendences.
6. Rezultātu attēlošana. Metode izveido pārskatāmas diagrammas, kurās redzami gan sākotnējie dati, gan to apvienotās vērtības. Īpaši tiek izcelti periodi, kad apvienotā vērtība pārsniedz noteikto slieksni.
7. Metodes pielāgošana. Visu procesu var atkārtot, mainot svarus vai citus parametrus, līdz tiek iegūts vēlamais rezultāts. Tas ir īpaši noderīgi, kad nepieciešams atrast optimālo parametru kombināciju konkrētam gadījumam.

4.2.1. Datu slāņošanas metodes parametri

Metodes centrālā funkcija `data_fusion_main` pieņem vairākus parametrus, kas kontrolē datu apvienošanas procesu. Šos parametrus var iedalīt trīs loģiskās grupās:

Datu avotu parametri:

- `DataFrames_dict_or_df`: satur ievades datus. Var būt vai nu vārdnīca ar vairākām datu tabulām, kur atslēgas ir failu nosaukumi, vai viena datu tabula. Abos gadījumos datiem jāsatur vismaz viena kopīga kolonna apvienošanai.
- `Data_params`: norāda, kuras kolonnas no katra faila izmantot analīzē. Katrs elements ir formātā "fails/kolonna" (piemēram, "Plants.xlsx/Richness"), kas ļauj precīzi identificēt vajadzīgos datus.
- `CommonColumn`: kopīgās kolonnas nosaukums (piemēram, "Month"), kas tiek izmantots datu tabulu apvienošanai.

Analīzes kontroles parametri:

- `Weights`: skaitlisku vērtību saraksts, kas nosaka katra parametra ietekmi gala rezultātā. Svaru summai nav jābūt 1, jo rezultāti tiek automātiski normalizēti diapazonā no 0 līdz 100.
- `Layering_threshold`: sliekšņa vērtība (pēc noklusējuma 30), kas tiek izmantota gan vizualizācijā kā horizontāla atskaites līnija, gan automātiskās izlīdzināšanas aprēķinos.

Datu apstrādes opcijas:

- `Smoothing`: izlīdzināšanas koeficients (pēc noklusējuma 1), kas nosaka, cik precīzi tiek veikta datu interpolācija. Lielākas vērtības rada vienmērīgāku līkni.
- `AutoSmooth`: ja ieslēgts (vērtība 1 vai 0), metode automātiski izvēlas izlīdzināšanas koeficientu, balstoties uz norādīto slieksni.
- FilterBy un `FilterValue`: ļauj atlasīt konkrētu datu apakškopu. Piemēram, var filtrēt pēc auga nosaukuma vai cita kvalitatīva parametra.

Šie parametri kopā nodrošina elastīgu datu apstrādes procesu, ļaujot pielāgot analīzi dažādām vajadzībām un datu veidiem. Īpaši svarīga ir iespēja kontrolēt gan datu izlīdzināšanas pakāpi, gan katra parametra ietekmi uz gala rezultātu.

4.2.2. Datu slāņošanas metodes funkciju darbība

Kā jau iepriekš minēts, metodes kodolu veido vairākas savstarpēji saistītas funkcijas, kur katra atbild par konkrētu datu apstrādes posmu. Galvenā funkcija `data_fusion_main` koordinē šo funkciju darbību, organizējot datu plūsmu starp tām:

Metodes pamatā ir vairākas savstarpēji saistītas funkcijas, kas kopā veido vienotu datu apstrādes ķēdi. Šo procesu vada galvenā funkcija data_fusion_main, kas darbojas kā dirigents, koordinējot pārējo funkciju darbu.

Datu sagatavošana sākas ar interpolāciju – procesu, kas līdzinās punktu savienošanai ar līniju, tikai daudz precīzāk. Par to rūpējas divas funkcijas:

- `interpolate_data` veic lineāro interpolāciju starp datu punktiem, palielinot to izšķirtspēju. Šī funkcija ir būtiska vienmērīgu un salīdzināmu datu iegūšanai;
- ja izvēlamies automātisko režīmu, `interpolate_data_auto_smoothing` meklē optimālo izlīdzināšanas pakāpi, pakāpeniski palielinot to, līdz dati pietiekami precīzi atspoguļo pāreju caur noteikto slieksni. Funkcija apstājas, kad atrasts piemērots izlīdzināšanas koeficients vai sasniegta maksimālā pieļaujamā vērtība.

Tālāk seko datu apvienošana, ko var veikt divos veidos:

- `weight_based_data_fusion` apvieno datus pa slāņiem, piešķirot katram slānim savu nozīmīguma pakāpi;

- `pca_based_data_fusion` izmanto galveno komponentu analīzi, lai identificētu dominējošās tendences datos un automātiski pielāgotu to orientāciju, saglabājot sākotnējo parametru savstarpējās attiecības.

Visbeidzot, `plot_fused_data_final` parāda rezultātus uzskatāmā veidā – kā stāstu ar attēliem, kur redzami gan sākotnējie dati, gan to apvienotā versija.

Viss process norit šādi:

1. vispirms tiek pieņemts lēmums par izlīdzināšanas pakāpi – vai ļaut datoram to izvēlēties automātiski;
2. tad notiek datu apvienošana pēc norādītajiem svariem, radot pirmo versiju;
3. paralēli tiek veidota otra versija, izmantojot PCA analīzi;
4. beigās abi rezultāti tiek parādīti blakus, ļaujot salīdzināt un izvērtēt, kurš labāk atbilst vajadzībām.

Šāda pieeja ļauj:

- viegli sekot līdzi tam, kas notiek ar datiem katrā solī;
- pielāgot procesu atbilstoši vajadzībām;
- salīdzināt dažādas pieejas;
- skaidri redzēt gala rezultātu.

4.3. Datu slāņošanas nodaļas kopsavilkums

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.

5. 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. 5.1. tabulā ir norādīti ikdienas parametri, kas novēroti Latvijas putnu fermā.

5.1. tabula. **Ražošanas laikā mērītie parametri**

Parametra nosaukums	**Apraksts**
Datums	Ražošanas dienas datums
Vistas vecums nedēļās	Putnu vecums nedēļās konkrētajā dienā
Vistas vecums dienās	Putnu vecums dienās konkrētajā dienā
Iekštelpu temperatūra	Vidējā iekštelpu gaisa temperatūra (ºC) konkrētajā dienā
Relatīvais mitrums	Vidējais iekštelpu gaisa mitrums (%) konkrētajā dienā
Barības patēriņš uz putnu	Barības patēriņš uz putnu (g) konkrētajā dienā
Ūdens patēriņš uz putnu	Ūdens patēriņš uz putnu (L) konkrētajā dienā
CO_2	Vidējā oglekļa dioksīda koncentrācija (ppm) konkrētajā dienā
NH_3	Vidējā amonjaka koncentrācija (ppm) konkrētajā dienā
Dējēju-vistu olu ražošana	Dējēju-vistu olu ražošana konkrētajā dienā (%)
Vidējais olas svars	Vidējais olas svars konkrētajā dienā (g)
Barības pārveides koeficients	Barības pārveides koeficients konkrētajā dienā
Neapstrādāts proteīns	Ieteicamais neapstrādātā proteīna daudzums (%) dējējvistām atkarībā no putnu vecuma nedēļās
Neapstrādāts tauki	Ieteicamais neapstrādāto tauku daudzums (%) dējējvistām atkarībā no putnu vecuma nedēļās
Neapstrādāts šķiedrvielas	Ieteicamais neapstrādāto šķiedrvielu daudzums (%) dējējvistām atkarībā no putnu vecuma nedēļās
Kalcijs	Ieteicamais kalcija daudzums (%) dējējvistām atkarībā no putnu vecuma nedēļās

Viens no svarīgākajiem precīzās putnkopības jautājumiem ir olu īpatsvara prognozēšana. Lai prognozētu olu īpatsvaru, var izmantot dažādas metodes, sākot no modeļiem, kas izmanto nelielu ieejas parametru skaitu, līdz modeļiem, kas prasa sarežģītu parametru kopumu. Lai labāk izprastu iegūtos prognozēšanas rezultātus un atšķirības starp dažādām metodēm, tiks pielietota izstrādātā datu slāņošanas metode. Šī pieeja ļaus dziļāk analizēt dažādu parametru mijiedarbību un to ietekmi uz prognozēšanas precizitāti. Šajā nodaļā vispirms tiks apskatīti prognozēšanas uzdevuma rezultāti, un pēc tam, izmantojot datu slāņošanas metodi, tiks veikta to padziļināta analīze.

5.1. Nelineārās regresijas modeļi

Nelineārās regresijas modeļu izmantošana putnu olu īpatsvara līknes piemērošanai ir bijusi populāra vairākas desmitgades – ir piedāvāti dažādi nelineāri modeļi (nepilnīgais gamma modelis pēc *Wood* (Wood, 1967), modificētais gamma modelis pēc *McNally* (McNally, 1971), modelis pēc *Adams un Bell* (Adams & Bell, 1980), nodalījuma modelis pēc *McMillan* (McMillan, 1981), modificētais nodalījums (arī saukts par *Logistic-Curvilinear* (A. Safari-Aliqiarloo, F. Faghih-Mohammadi, M. Zare, A. Seidavi, V. Laudadio, M. Selvaggi, & V. Tufarelli, 2017; Cason, 1990)), modelis pēc *Yang, Wu, & McMillan* (Yang et al., 1989).

Vairāki autori ir analizējuši un salīdzinājuši tradicionālos nelineāros modeļus olu īpatsvara līknes piemērošanai.

- *Faridi, Mottaghitalab, Rezaee, & France* (Faridi et al., 2011) pētījumā parādīja trīs *Narushin-Takma* modeļu (*NT1, NT2, NT3*) pielietojamību, lai piemērotu vairākas ar putniem saistītas pazīmes (piemēram, olu īpatsvars, olu svars, barības pārveidošanas koeficients) broilera audzētāju pulkiem. Šo modeļu veiktspēju salīdzināja ar 5 citiem – *Gompertz, Modified Compartmental, Richards, Adams-Bell* un *Lokhorst*. Tika secināts, ka *NT3* modelis parādīja labāko veiktspēju, ko mēra ar 4 piemērotības kritērijiem (*MSE, R2, AIC* un *BIC*).
- Savegnago u. c. (Savegnago et al., 2012) savā pētījumā izvēlējās 7 modeļus, lai noteiktu to piemērotību iknedēļas olu īpatsvara līknēm ražošanas salīdzināšanai starp dažādām izlases līnijām dējējvistām. Šī pētījuma rezultāti liecina, ka vairāki nelineāri modeļi – *Logistic*, *Persistency, Segmented Polynomial* un *Modified Compartmental* modelis – ir atbilstošas kvalitātes pielāgošanā un tāpēc tos var piemērot olu īpatsvara līknei, kamēr pārējo vērtēto modeļu, proti, nodalījuma (I un II) un *McNally*, veiktspēja bija nepietiekama salīdzinājumā.
- *Safari-Aliqiarloo* u. c. (A. Safari-Aliqiarloo, F. Faghih-Mohammadi, M. Zare, A. Seidavi, V. Laudadio, M. Selvaggi, V. Tufarelli, et al., 2017) salīdzināja trīs nelineāros modeļus (*Compartmental, Modified Compartmental* un *Adams-Bell*) ar mākslīgā neironu tīkla pieeju, lai piemērotu broilera audzētāju olu īpatsvara līkni. Pamatojoties uz izvērtēšanas kritērijiem, autori secināja, ka ANN modelis bija ar labāko veiktspēju, bet no izvēlētajiem tradicionālajiem nelineārajiem modeļiem *Modified Compartmental* modelis pārspēja pārējos, piemērojot olu īpatsvara līkni. Tas arī palielina Savegnago u. c. publicēto rezultātu uzticamību (Savegnago et al., 2012).
- *Görgülü & Akilli* (Görgülü & Akilli, 2018) savā pētījumā analizēja četru tradicionālo nelineāro regresijas modeļu (*Adams-Bell, Compartmental, McNally* un *Logistic* modelis (Nelder, 1961)) veiktspēju kopā ar vismazāko kvadrātu atbalsta vektoru mašīnu (LSSVM) modeli. Tika secināts, ka LSSVM ir labākā veiktspēja, sekojot *Adams-Bell* modelim no izvēlētās nelineārās modeļu kategorijas, bet tā veiktspēja ir labāka tikai ar šauru robežu, un, kā autori norādīja, visi testētie modeļi labi piemēro dējējvistu ražošanas līkni.
- Emams (Emam, 2021) salīdzināja četru nelineāro modeļu (*Wood, McNally, Compartmental, Modified Compartmental*) veiktspēju, koncentrējoties uz to, cik labi tie piemēro *Fayoumi* slāņu ražošanas līkni. Pētījuma rezultāti liecināja, ka modificētais nodalījuma modelis pierādīja labāko piemērošanu.
- *Morales-Suárez* u. c. (Morales-Suárez et al., 2021) savā pētījumā izpētīja dažādu diētu ietekmi (koncentrējoties uz trīs būtisko aminoskābju koncentrāciju variācijām) uz dējējvistu olu ražošanu. Lai novērtētu olu ražošanu, autori izmantoja trīs modeļu grupas – tradicionālos nelineāros modeļus (*Adams-Bell, McNally, Logistic, Compartmental, Modified Compartmental, Modified Gompertz, Modified Logistic*), multivariātos polinomiālos modeļus (otrais un trešais kārtas) un *ANN* modeļus (ar barošanas uz priekšu un kaskādes uz priekšu arhitektūrām). Pamatojoties uz izvērtēšanas kritērijiem, kopumā *ANN* pārsniedza visus modeļus. Attiecībā uz nelineāro regresijas modeļu grupu autori norāda, ka šie modeļi labi piemēroja ražošanas līkni, bet modificētais loģistikas modelis bija ar visaugstākajiem piemērotības rādītājiem.
- *Sharifi, Patil, Yadav, & Bangar* (Sharifi et al., 2022) salīdzināja 8 modeļus (*Logistic, Morgqan Mercer Flodin, Polynomial Fit, Rational Function, Sinusoidal Fit, Quadratic Fit, Gompertz function, Modified Compartmental*) attiecībā uz to, cik labi šie modeļi pielāgo olu īpatsvara un olu svara līknes dējējvistām. Izvērtējums parādīja, ka *Rational Function, Modified Compartmental, Sinusoidal Fit un Polynomial Fit* modeļi bija labākie olu īpatsvara pielāgošanā.

Ir jāuzsver, ka šie nelineārie modeļi ir paredzēti tikai viena faktora apstrādei, proti, putnu vecumam (vai dēšanas nedēļai) kā ieejai, un pamatojoties uz to, pilnībā nosaka olu ražīgumu. Lai gan tas nozīmē, ka modeļus var piemērot gadījumos, kad pieejams tikai viens faktors, tas arī atklāj šo modeļu ierobežojumus – nav pielāgošanās vides izmaiņām. Vide tiek noteikta ar tādiem faktoriem kā ārējā un iekšējā temperatūra, mitrums, CO_2 un NH_3 koncentrācija, gaisa ventilācija utt. Šie faktori var arī ietekmēt olu ražošanu, vai nu individuāli, vai kombinācijā (Li et al., 2020).

5.2. Mašīnmācīšanās modeļi

Mākslīgā intelekta (*Artificial Intelligence, AI*) jomā gūtie panākumi ir noveduši pie vairāku *ML* modeļu izstrādes, kuriem piemīt nelinearitāte un kurus var arī izmantot olu īpatsvara prognozēšanai un ar ražošanu saistīto problēmu laikus identificēšanai. Daži no *ML* modeļu pielietojumiem putnkopības nozarē ir aprakstīti šādos pētījumos:

- *Morales, Cebrián, Fernandez-Blanco, & Sierra* (Morales et al., 2016) pētījumā testēja *SVM* modeli ar mērķi laikus atklāt problēmas olu ražošanā. Modeļa precizitāte bija augsta (≈ 0,985), prognozējot problēmas dienu iepriekš;
- pētījums, kas iepazīstināja ar *ANN* pieeju agrīnai problēmu atklāšanai olu ražošanā (Morales et al., 2016). Salīdzinot ar *SVM* modeli, *ANN* tehnika izrādījās ļoti precīza, prognozējot olu īpatsvara problēmas dienu iepriekš. Turklāt tā parādīja uzlabojumus vairākos veiktspējas izvērtēšanas rādītājos salīdzinājumā ar LSSVM modeli;
- *Ahmad* (Ahmad, 2011) savā pētījumā salīdzināja trīs dažādas *ANN* arhitektūras un tradicionālos modeļus olu īpatsvara prognozēšanai. No eksperimenta rezultātiem autors secināja, ka *ANN* modeļi pārsniedz tradicionālos nelineāros modeļus.

Atskatoties uz šo pētījumu rezultātiem, papildus daļēji saistītiem darbiem (A. Safari-Aliqiarloo, F. Faghih-Mohammadi, M. Zare, A. Seidavi, V. Laudadio, M. Selvaggi, & V. Tufarelli, 2017; Görgülü & Akilli, 2018; Morales-Suárez et al., 2021) secinājums ir tāds, ka *AI* tehnikas ir piemērojamas olu īpatsvara prognozēšanai un arī var novest pie veiktspējas uzlabojumiem (mērot pēc vairākiem piemērotības kritērijiem) salīdzinājumā ar tradicionālajiem nelineārajiem modeļiem.

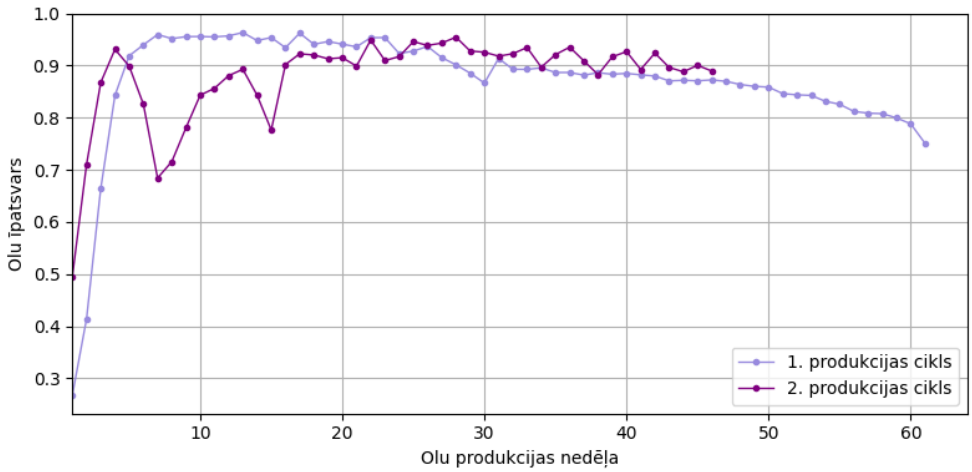
Pamatojoties uz literatūras pārskatu par nelineāro modeļu izmantošanu putnkopības nozarē, modificētā kompartmentālā modeļa veiktspēja pierādījās konsekventi atbilstoša olu īpatsvara izvērtēšanai, un tam ir salīdzinoši mazs parametru skaits, ko var bioloģiski un matemātiski interpretēt (Yang et al., 1989). Tāpēc modificētais kompartmentālais modelis tika izvēlēts praktiskajiem eksperimentiem. No *ML* grupas tika izvēlēti vairāki modeļi, jo to veiktspēja izcēlās, kā par to vēstīja vairāki raksti (aprakstīti iepriekšējā nodaļā). *HENCO2* pētījuma mērķiem tika izvēlēti šādi modeļi, ieskaitot gan tos, par kuriem citi autori jau ir ziņojuši, gan tos, kuri vēl nav izmantoti olu īpatsvara prognozēšanai:

- ilgtermiņa īstermiņa atmiņa (*Long Short-Term Memory, LSTM*) – ir modelis, kas balstīts uz atkārtotu neironu tīklu un izmantots precīzajā lauksaimniecībā (Shadrin et al., 2020), tieši putnkopības nozarē arī dēšanas īpatsvara prognozēšanā (Yin et al., 2021);
- konvolūcijas neironu tīkls (*Convolutional neural network, CNN)* – ir dziļās mācīšanās arhitektūra. Putnkopībā tas ir izmantots ar mērķi noteikt un klasificēt olas (Lubich et al., 2019);
- *Random forest* (*RF*) – tehnika, kas izveido vienu modeli, balstoties uz vairāku lēmumu koku kombināciju (Maindonald, 2012). *RF* izmantošana olu svara prognozēšanai jau ir aprakstīta (Pitesky et al., 2020);
- *XGBoost* – "koku palielināšanas sistēma" (T. Chen & Guestrin, 2016) ir balstīta uz lēmumu koku un gradienta palielināšanas pieeju. Šis modelis tika izmantots īpašību atlasei, lai prognozētu ūdensputnu dēšanas īpatsvaru (Yin et al., 2021).

Modeļu salīdzināšanai un novērtēšanai nepieciešamie olu dēšanas īpatsvara dati tika iegūti no novietnes, kurā tika audzētas vistas *(Lochman* brūnā šķirne) bagātinātos būros. Šajā novietnē ir ieviesta lentes tipa kūtsmēslu izvešanas sistēma, kuru katru dienu efektīvi darbina. Dējējvistas pārvieto uz būriem 16 nedēļu vecumā un tur tiek turētas līdz pat 80–90 nedēļu vecumam. No 20 nedēļām vistas uzsāk olu dēšanu. Diennakts olu produkcija jeb īpatsvars tiek aprēķināts kā dienā saražoto olu skaits attiecību pret kopējo vistu skaitu attiecīgajā dienā (Paura et al., 2022).

5.3. Modeļu struktūra un apmācība

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. 5.1. att.).



5.1. 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.

Izvēlēto modificēto nodalījumu (*Modified Compartmental*) modeli (Yang et al., 1989) var matemātiski izteikt ar šādu vienādojumu (sk. (5.1.)):

$$y(t) = \frac{ae^{-bt}}{1 + e^{-c(t-d)}} \tag{5.1.}$$

kur

$y(t)$ – atspoguļo olu īpatsvaru dēšanas nedēļā t;

a – skalas parametrs, kas, iespējams, atspoguļo maksimālo olu īpatsvaru,

b – atspoguļo ražošanas samazināšanās īpatsvaru pēc maksimuma. Augstākas b vērtības nozīmētu ātrāku lejupslīdi pēc maksimuma;

t – dēšanas nedēļu skaits, veselais skaitlis;

c – abpusējs dzimumbrieduma izmaiņu rādītājs. Augstāka vērtība c nozīmētu mazākas atšķirības vecumā, kurā vistas sasniedz dzimumbriedumu;

d – vistu vidējais vecums dzimumgatavībā. Tas novirzītu loģistikas funkciju pa kreisi vai pa labi, norādot vidējo laiku, kurā vistu dējība sāk ievērojami palielināties.

ae^{-bt} – apzīmē eksponenciālu samazinājuma funkciju. Palielinoties t, vistu dējība samazinās, fiksējot olu īpatsvara samazināšanos pēc maksimuma;

$1/\left(1+\mathrm{e}^{-\mathrm{c}(\mathrm{t}-\mathrm{d})}\right)$ – loģistikas funkcija. Tas modelē sākotnējo olu īpatsvara pieaugumu, kad vistas sasniedz dzimumbriedumu un sāk dēt olas. Vērtība d novirzīs šo līkni, lai tā atbilstu faktiskajam vistu dzimumbrieduma vidējam vecumam. Parametrs c nosaka šīs loģistikas līknes stāvumu, tādēļ, ja ir daudz atšķirību attiecībā uz to, kad vistas sasniedz dzimumbriedumu, šī līkne būs plakanāka (mazāks c), savukārt, ja lielākā daļa vistu sasniedz dzimumbriedumu apmēram vienlaikus, tas būtu stāvāks (lielāks c).

Precīzās putnkopības uzdevuma risināšanas 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). Katra modeļa hiperparametru meklēšanas telpa ir parādīta 5.2. – 5.4. tabulās.

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.

5.2. tabula. **Hiperparametri un to aplūkotās vērtības *LSTM* modelim**

Hiperparametrs	**Apsvērto vērtību diapazons**
Ievades izmērs	32, 48, 64, 80, 96, 112, 128
Atkritumu līmenis	0, 0.1, 0.2, 0.3, 0.4, 0.5
Optimizētājs	Adam
Mācību līmenis	0.01, 0.001, 0.0001
Aktivizācijas funkcija	ReLU, sigmoid
Agrīna pacietības pārtraukšana	10
Zaudējuma funkcija	Mean Squared Error

5.3. tabula. **Hiperparametri un to aplūkotās vērtības *CNN* modelim**

Hiperparametrs	**Apsvērto vērtību diapazons**
Filtra izmērs	32, 48, 64, 80, 96, 112, 128
Kodola izmērs	2, 3, 5
Maksimālais pūles izmērs	2
Aktivizācijas funkcija	ReLU, sigmoid
Agrīna pacietības pārtraukšana	10
Zaudējuma funkcija	Mean Squared Error
Filtra izmērs	32, 48, 64, 80, 96, 112, 128

5.4. tabula. **Hiperparametri un to aplūkotās vērtības *XGBoost* modelim**

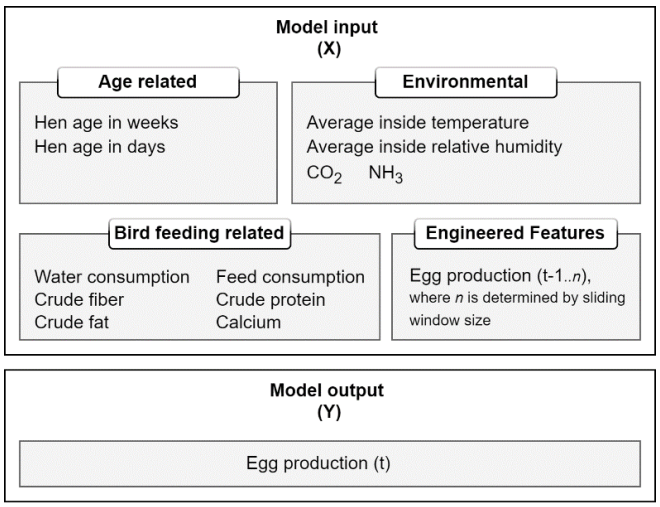
Hiperparametrs	**Apsvērto vērtību diapazons**
Maksimālais dziļums	3, 6, 10
Mācību līmenis	0.01, 0.05, 0.1
Novērtētāju skaits	100, 500, 1000
Colsample bytree	0.3, 0.7

5.5. tabula. **Hiperparametri un to aplūkotās vērtības *RF* modelim**

Hiperparametrs	Apsvērto vērtību diapazons
Bootstrap	True
Maksimālais dziļums	90, 100, 110
Min paraugi lapu	4, 5, 6
Minimālais paraugu sadalījums	8, 10, 12
Novērtētāju skaits	100, 500, 1000

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 (sk. 5.2. att.) (un papildus vēsturiskajiem ražošanas datiem atkarībā no bīdāmā loga izmēra). Modeļu veiktspēja tika novērtēta pēc statistikas kritērijiem.



5.2. att. **Ieejas un izejas parametri *ML* modeļiem** (Bumanis et al., 2023)**.**

5.4. Prognozēšanas modeļu precizitātes novērtēšana

Modificētā nodalījuma modeļa parametri (sk. 5.6. tabulu) tika novērtēti, izmantojot R programmēšanas valodu, lai atbilstu olu īpatsvara līknei (1. cikls). No 5.6. tabulas visi parametri ir nozīmīgi ($p < 0{,}001$) un ir piemērojami šim prognozēšanas uzdevumam.

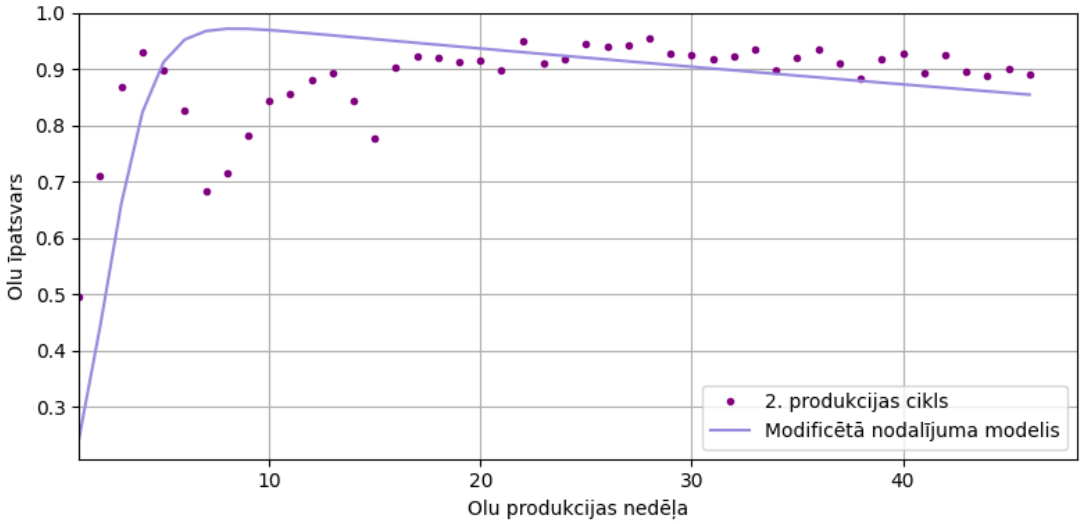
5.6. tabula. **Modificētā nodalījuma modeļa aprēķinātās parametru vērtības**

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 ***

Pēc tam modeļa parametrus var aizpildīt ar aprēķinātajām vērtībām, kur vienīgais ievades parametrs ir *t*, kas atspoguļo dēšanas nedēļu skaitu (sk. (5.2.)):

$$\hat{y}(t) = \frac{0.13099e^{-0.90414t}}{1 + e^{0.90766(t-2.24435)}} \quad (5.2.)$$

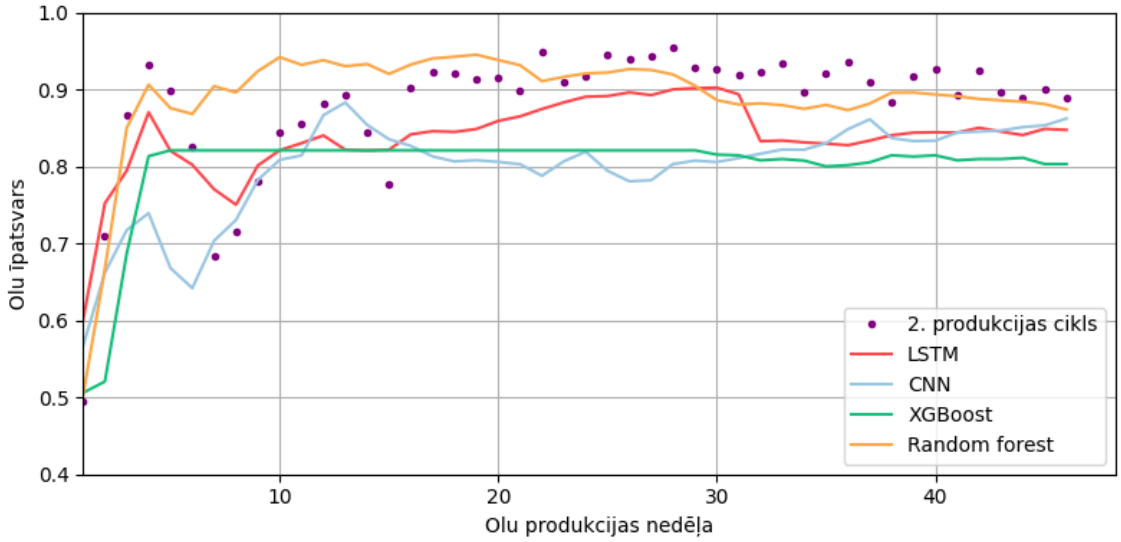
Pielāgotā līkne testa olu īpatsvara datu kopai, izmantojot modificēto nodalījumu modeli, ir šāda:



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

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.

ML modeļi mēdz sekot nenormālam ražošanas samazinājumam (sk. 5.4. 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:



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

Attiecībā uz *ML* modeļiem tika pārbaudīti vairāki logu izmēri, lai noteiktu, kurš nodrošina vislabākos prognozēšanas rezultātus. Tika ņemti vērā logu izmēri 1, 2, 3, 5, 7 un 14. Modeļa veiktspējas rezultāti dažādiem logu izmēriem, lai prognozētu olu ražošanu nākamajai dienai, ir parādīti 5.7, 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.

5.7. 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 5.8 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*.

5.8. 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.

Vēl viens aspekts, kas būtu jāpēta vairāk, ir faktoru skaits, kas nepieciešams, lai nodrošinātu ticamu prognozi. Tas varētu būt svarīgi mazākām saimniecībām, kurās var nebūt vairāku sensoru aprīkojuma, lai izmērītu dažādus parametrus, kas ietekmē vistu olu ražošanu (piemēram, CO_2, NH_3). Kāda pētījuma (Yin et al., 2021) mērķis bija noteikt vides faktorus, kas ietekmē dējējputnu (lauvas galvas zosu) olu ražošanu. Autori atklāja, ka optimālais parametru skaits ir vienāds ar 4 (t. i., dēšanas īpatsvars, oglekļa dioksīds, temperatūra un putekļi), prognozēšanā izmantojot *LSTM* un *RF* kombināciju. Lai gan tai var būt dažas atšķirības, vispārīgās vadlīnijas par ierobežotiem parametriem, kuras var novērot konsekventi, var attiecināt arī uz dējējvistām, taču, lai to apstiprinātu, joprojām ir nepieciešami turpmāki pētījumi (lielākas datu kopas) un praktisks pierādījumu novērtējums.

Optimālas parametru kopas atrašanas procesā var tikt ieviesti daudzi lieki un nesvarīgi faktori, novedot pie nekorekti apmācīta modeļa, līdz ar to var pieņemt, ka vairāku parametru modeļa konvolūcija ne tikai palielina datu vākšanas soļa grūtības, bet var arī samazināt kopējo veiktspēju attiecībā uz precīzu prognozēšanu.

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.

5.5. Datu slāņošanas metodes pielietošana

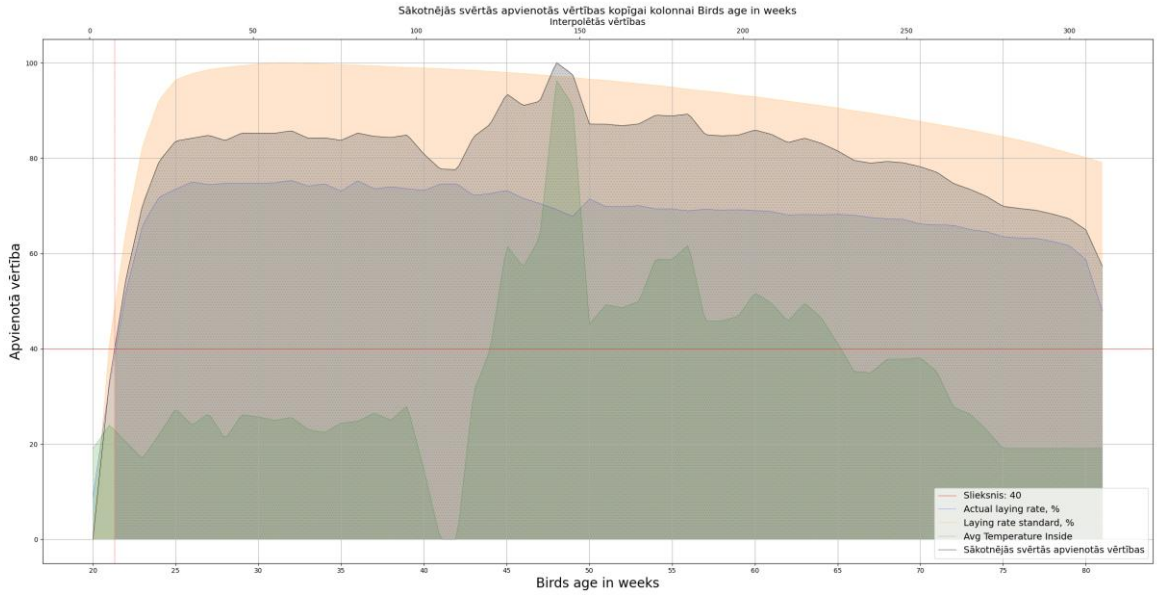
Prognozēšanas modeļu rezultātu analīzei un potenciālo problēmu identificēšanai tika izmantota datu slāņošanas metode. Šī metode nodrošina sistemātisku pieeju vairāku parametru mijiedarbības izpētei un to ietekmei uz olu ražošanas procesu. Pētījumā tika analizēti trīs fundamentāli parametri, kas raksturo 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 ar vienādojumu (5.3.), 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.

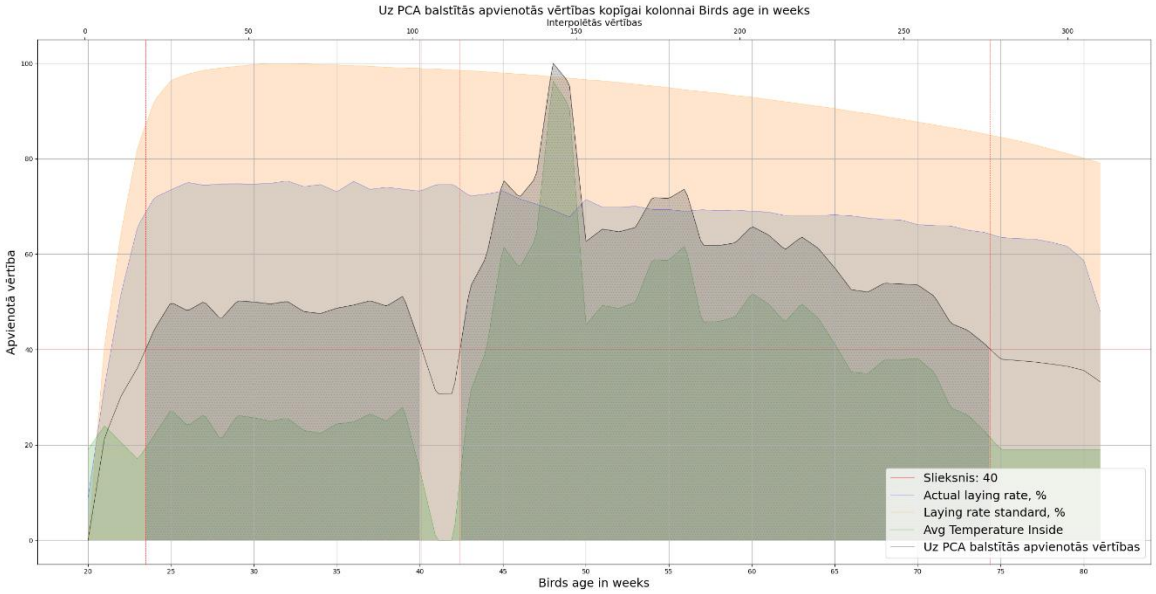
$$normalized\_data = \left(\frac{data - \min\_val}{\max\_val - \min\_val}\right) \times (range\_max - range\_min) + range\_min \qquad (5.3.)$$

Metodes pielietošanas rezultāti atspoguļoti trīs attēlos (sk. 5.5. att., 5.6. att. un 5.7. att.). Sākotnējo svērto apvienoto vērtību analīze (sk. 5.5. 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.



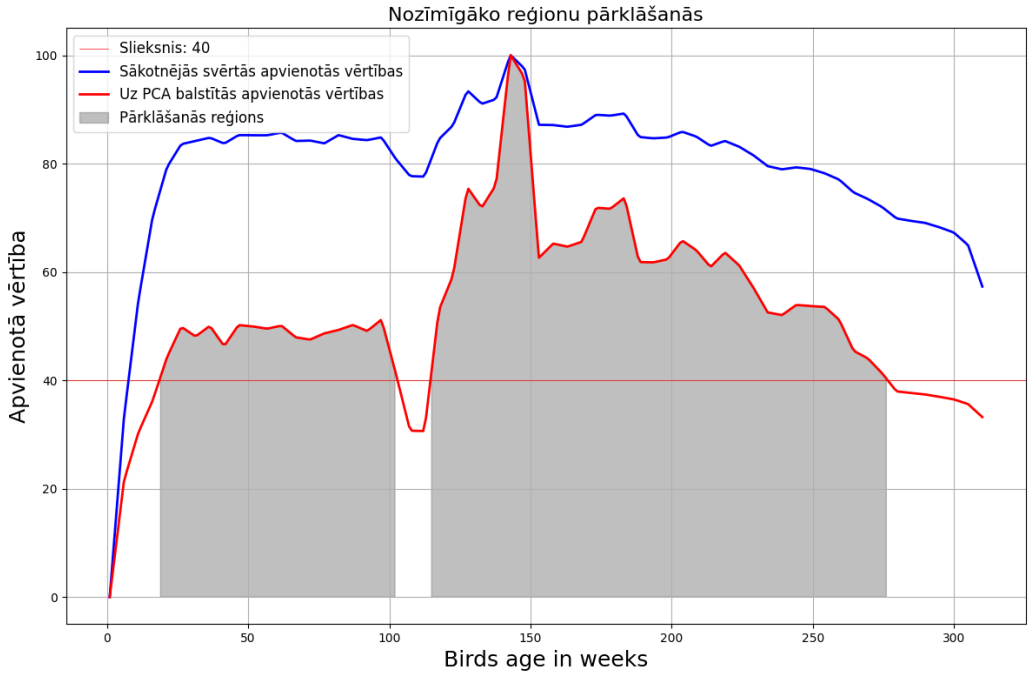
5.5. att. **Sākotnējās svērtās apvienotās vērtības.**

Uz PCA balstītā analīze (sk. 5.6. 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.



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

Nozīmīgāko reģionu pārklāšanās analīze (sk. 5.7. 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.



5.7. 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.

Šie novērojumi pamato nepieciešamību izstrādāt specializētas metodes datu kvalitātes problēmu identificēšanai un risināšanai precīzās lauksaimniecības kontekstā. Īpaši būtiska ir automātiska anomāliju detektēšana un koriģēšana reālā laika datos, kas ļautu uzlabot gan tradicionālo, gan ML modeļu prognozēšanas precizitāti.

5.6. Prognozēšanas uzdevuma nodaļas kopsavilkums

Dažādi nelineārie modeļi (piemēram, *Wood*, *McNally*, *Adams-Bell* un *Compartmental* modeļi) ir bijuši populāri gadu desmitiem, lai pielāgotu olu īpatsvara līknes. Vairākos pētījumos tika salīdzināti šie modeļi, secinot, ka modificēto nodalījumu un ANN modeļi uzrāda labākos rezultātus.

Šajā pētījumā tika analizēti vairāki mašīnmācīšanās modeļi: ilgtermiņa īstermiņa atmiņa (*LSTM*), konvolūcijas neironu tīkli (*CNN*), izlases mežs (*RF*) un *XGBoost*. Šie modeļi tika apmācīti, izmantojot datus no putnu fermas, un pārbaudīti dažādos logu izmēros. Modificētais nodalījumu modelis uzrādīja MAPE 9,134% un RMSPE 14,809%, savukārt LSTM modelis uzrādīja labāko veiktspēju ar bīdāmo loga izmēru 2 (MAPE 5,390%, RMSPE 7,751%).

Olu īpatsvara analīzei tika pielietota datu slāņošanas metode, kas apvieno trīs faktorus: vecumu, dēšanas īpatsvaru un vidējo temperatūru iekštelpās. Rezultātā tika identificēti trīs laika periodi ar atšķirīgām parametru vērtībām.

Datu slāņošanas metodes *PCA* komponents darbojas ar pilnām datu kopām, tāpēc pirms tā izmantošanas tika pielietota modificētā standarta vidējā svērtā metode trūkstošo vērtību aizpildīšanai. Šāda pieeja ļāva apvienot gan datu kvalitātes uzlabošanu, gan datu apvienošanu.

PCA balstītās analīzes rezultāti, īpaši 45.-50. nedēļas periodā identificētās novirzes, norāda uz iespējamām datu kvalitātes problēmām. Modeļu veiktspējas atšķirības šajā periodā liecina par nepieciešamību pēc uzlabotām datu apstrādes metodēm.

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 ceturtajā nodaļā izstrādāto datu slāņošanas metodi un tās praktisko pielietojumu piektās nodaļas 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

Ceturtajā 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 (sk. 4.7. att.). Metodes interaktīvais raksturs nodrošina tās spēju identificēt un vizualizēt datu pārklāšanās zonas, kas izpaužas trapecveida metodes

pielietojumā reģionu analīzē (sk. 4.10. att.). Šī pieeja ļauj kvantitatīvi novērtēt gan pārklāšanās zonas, gan atšķirību reģionus, tādējādi sniedzot skaidru priekšstatu par datu mijiedarbību.

Metodes praktiskā pielietojamība tiek apstiprināta piektās nodaļas precīzās putnkopības pētījumā, kur attēlos 5.5., 5.6. un 5.7. redzamā vizualizācija parāda metodes daudzpusīgo analītisko potenciālu. Attēlos 5.5. un 5.6. 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 5.7. 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.

6. DATU KVALITĀTES UZLABOŠANA

Datu kvalitātes kritēriji ir būtiski datu apstrādes procesos, jo tie ietekmē datu pielietošanu analīzē, prognozēšanā un lēmumu pieņemšanā. Divi no nozīmīgākajiem kvalitātes kritērijiem ir datu pilnīgums un datu precizitāte. Datu pilnīgums raksturo pieejamo mērījumu īpatsvaru no visiem nepieciešamajiem mērījumiem, savukārt datu precizitāte nosaka, cik lielā mērā mērījumi atbilst patiesajām vērtībām. Iepriekšējā nodaļā veiktā analīze parādīja, ka abi šie kritēriji ir savstarpēji saistīti un var būtiski ietekmēt prognozēšanas modeļu veiktspēju.

Datu pilnīgums attiecas uz to, cik lielā mērā dati ir pieejami apstrādei, t. i., cik daudz ierakstu vai datu kopu ir pieejamas, un vai šīs datu kopas satur visus nepieciešamos elementus. Tas nozīmē, ka datu kopām jābūt pietiekami izsmeļošām, lai nodrošinātu to izmantojamību konkrētajā pielietojuma scenārijā. Ja datu kopā trūkst kritisku datu, analīze un secinājumi var būt kļūdaini vai neprecīzi. Piemēram, ja vides monitoringa sistēmā, kur tiek kontrolēti tādi faktori kā temperatūra, mitrums un CO_2 līmenis, viens no šiem komponentiem netiek mērīts vai trūkst datu, var būt ļoti sarežģīti pilnvērtīgi novērtēt apkārtējās vides ietekmi uz ražošanas procesiem vai veselības rādītājiem. Datu pilnīgumu var novērtēt divos veidos. Pirmkārt, nepieciešamo datu kopu esamība, kad tiek skatīts, vai visas nepieciešamās datu grupas vai kategorijas ir pieejamas. Kāda datu veida trūkums (piemēram, nav pieejami CO2 dati putnkopības pārvaldības sistēmā) parasti ir jāatrisina jau datu vākšanas vai sensoru sistēmas izstrādes fāzē. Otrkārt, datu ierakstu pilnīgums, kas attiecas uz atsevišķu datu punktu esamību katrā datu kopā. Piemēram, ja kādā brīdī netiek reģistrēti temperatūras vai mitruma dati, šo trūkumu iespējams risināt, piemērojot tādas metodes kā interpolācija vai citu līdzīgu ierakstu izmantošana trūkstošo datu aizvietošanai. Trūkstošie dati ir tipisks datu problēmu scenārijs, kuru jārisina pirms apstrādes vai analīzes sākšanas. Nepilnīgu datu problēmas ir diezgan izplatītas reālos datos, kur bieži vien dažādi tehniski faktori var radīt datu ierakstu neesamību. Piemēram, ja sensoru tīklā notiek kļūme, daži mērījumi var tikt neierakstīti. Šādos gadījumos ir nepieciešams izmantot trūkstošo datu aizvietošanas metodes, piemēram, prognozēšanu, interpolāciju vai citu līdzīgu tehnoloģiju, lai nodrošinātu trūkstošo vērtību adekvātu aizvietošanu.

Datu precizitāte ir svarīgs elements analītiskajā izpētē, kas kalpo kā stūrakmens uzticamiem zinātniskiem atklājumiem. Datu precizitāte var definēt kā pakāpi, kādā iegūtie mērījumi atspoguļo reālo situāciju. Šis rādītājs ir izšķirošs datu lietderības un kvalitātes novērtēšanā. Neprecīzu datu radītās sekas var būt nozīmīgas visdažādākajās nozarēs. Datu precizitāti galvenokārt ietekmē divi aspekti: tehniskās neprecizitātes, kas saistītas ar mēraparatūras un sensoru darbību, un cilvēciskais elements – kļūdas, kas rodas manuālā datu apstrādē. Tehniskās neprecizitātes var rasties gan nepareizas kalibrācijas rezultātā, gan ierīču tehnoloģisko ierobežojumu dēļ. Īpaši kritiska ir datu precizitāte *IoT* risinājumos un mākslīgā intelekta lietojumos, kur neprecīzi ievaddati var būtiski ietekmēt modeļu apmācības procesu, radot novirzes to darbībā. Kvalitatīvu datu nodrošināšanai nepieciešama sistemātiska pieeja un regulāra pārbaude.

Datu pilnīgums un precizitāte ir savstarpēji saistīti kvalitātes rādītāji. Datu kopa var būt pilnīga, taču neprecīza, vai otrādi – precīza, bet nepilnīga. Abos gadījumos datu izmantošanas iespēja ir ierobežota. Trūkstošo vērtību aizpildīšanai bieži izmanto līdzīgu datu kopu informāciju, taču šīs metodes efektivitāte ir tieši atkarīga no bāzes datu precizitātes. Piemēram, CO_2 koncentrācijas novērojumos neprecīzi sensoru dati, kas tiek izmantoti trūkstošo vērtību aprēķināšanai, var radīt maldinošu priekšstatu par gaisa kvalitāti. Šādi neprecīzi dati var novest pie kļūdainiem secinājumiem par vides stāvokli. Tāpēc īpaši svarīgi ir nodrošināt gan datu pilnīgumu, gan precizitāti mašīnmācīšanās sistēmās, kur datu kvalitāte tieši ietekmē modeļu sniegumu un to radīto prognožu precizitāti.

Datu pilnīguma uzlabošanai pastāv vairākas efektīvas stratēģijas. Galvenā no tām balstās uz esošo precīzo datu izmantošanu kā pamatu trūkstošo vērtību aprēķināšanai. Šī pieeja ir īpaši noderīga situācijās, kad daļa parametru ir precīzi un pilnīgi dokumentēti. Matemātiskās

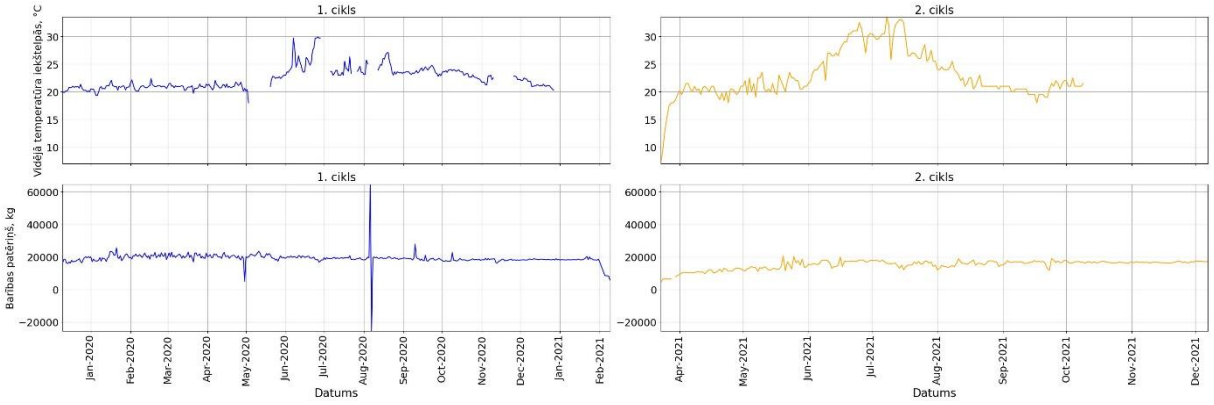
metodes, tādas kā interpolācija vai ekstrapolācija, ļauj ģenerēt trūkstošās vērtības, balstoties uz zināmajiem datiem. Šo metožu izvēle ir atkarīga no datu rakstura un trūkstošo vērtību īpatnībām. Interpolācija ir īpaši piemērota gadījumos, kad trūkstošās vērtības atrodas starp zināmiem datiem. Precizitātes nodrošināšanas process ietver sistemātisku datu elementu pārbaudi. Šī pārbaude palīdz identificēt trūkstošo datu cēloņus un novērst tos. Datu kvalitātes novirzes bieži atklājas tieši trūkstošo datu analīzes procesā. Sistēmas darbības traucējumi var izpausties kā noteiktu parametru nereģistrēšana. Tas var norādīt uz tehniskām problēmām mērīšanas iekārtās vai nepilnībām datu vākšanas procesā. Šādu problēmu savlaicīga identificēšana ļauj veikt koriģējošas darbības un uzlabot datu kvalitāti.

Efektīva datu apstrāde prasa kompleksu pieeju kvalitātes nodrošināšanai. Datu pilnīgums un precizitāte ir divi galvenie kvalitātes kritēriji, kuru nodrošināšanai nepieciešamas specializētas metodes un rīki. Trūkstošo datu aizvietošanas metodēm jābalstās uz padziļinātu izpratni par datu struktūru. Šīs metodes izmanto sakarības starp dažādām datu kopām, lai ģenerētu ticamas aizvietojošās vērtības. Datu precizitātes nodrošināšana savukārt ietver noviržu identificēšanu un koriģēšanu. Sensoru kļūdas un cilvēkfaktors rada novirzes, kas var būtiski ietekmēt datu interpretāciju. Šo noviržu noteikšana un labošana ir izšķiroša precīzu secinājumu izdarīšanai. Nozares zināšanas un iepriekšējā pieredze sniedz vērtīgu kontekstu datu pielāgošanas procesā. Datu kvalitātes uzlabošanas process ir cikliski saistīts. Trūkstošo vērtību aizvietošana un noviržu korekcija darbojas kā vienots mehānisms. Precizitāte bez pilnīguma nenodrošina pietiekamu datu kvalitāti, un pilnīgums bez precizitātes nesniedz ticamus rezultātus. Šāda integrēta pieeja datu kvalitātes nodrošināšanai ļauj izveidot uzticamu pamatu tālākai datu analīzei un lēmumu pieņemšanai. Tas ir īpaši svarīgi sarežģītās sistēmās, kur katram datu elementam ir būtiska nozīme kopējā analīzes procesā.

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. 6.1. att.), vidējiem temperatūras datiem ir nepilnības, bet barības patēriņam – novirzes.

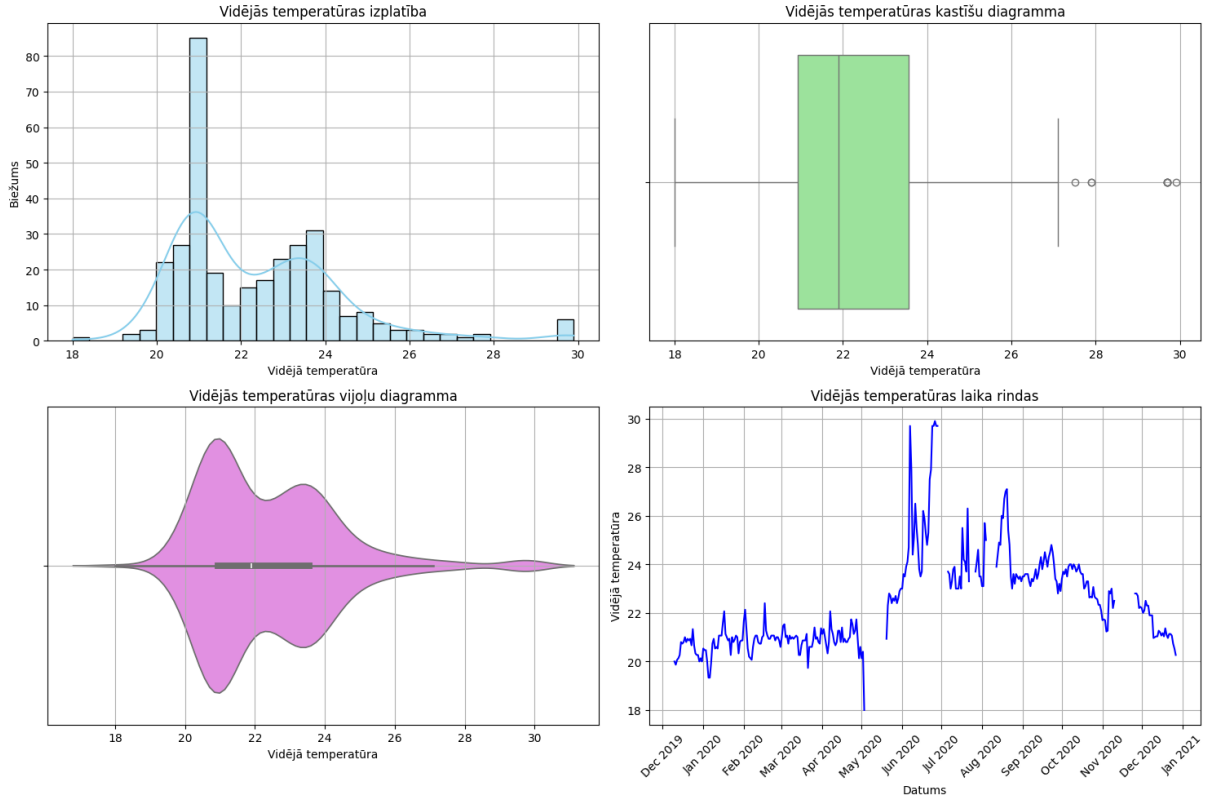


6.1. att. **Divu ražošanas ciklu datu kvalitātes reprezentācija.**

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

6.1. Trūkstošo vērtību aizvietošana

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. Rezultātā iegūtā kolonna skatāma 6.2. att.



6.2. att. **Aprēķinātās vidējās temperatūras statistikas datu vizualizācija.**

Histogramma parāda vidējās temperatūras vērtību frekvences sadalījumu. Attiecīgi, lielākā daļa temperatūras vidējo vērtību ir koncentrētas ap 21−23°C, norādot, ka šis diapazons ir visizplatītākais. Sadalījums ir nedaudz šķībs, kas nozīmē, ka ir daži gadījumi, kad temperatūra ir augstāka par parasto diapazonu.

Pēc kastīšu diagrammas lielākā daļa temperatūras vidējo vērtību ir sagrupētas salīdzinoši nelielā diapazonā, kā redzams pēc kastes izmēra. Potenciālo noviržu klātbūtne virs augšējās ūsas liecina, ka ik pa laikam ir novērojamas temperatūras svārstības.

Vijoļu diagramma, kas apvieno kastes diagrammas elementus un kodola blīvuma novērtējuma elementus, tās platākajās daļās attēlo lielāku blīvumu (vairāk datu punktu), bet šaurākajās – mazāku blīvumu. No šī var secināt, ka vijoles diagramma atkārto histogrammas un kastes diagrammas atklājumus. Tas sniedz detalizētāku priekšstatu par to, kur datu blīvums ir vislielākais, vēl vairāk uzsverot, ka visizplatītākā temperatūra ir aptuveni 21–23 °C diapazonā. Augšējā pusē esošā aste parāda augstāku temperatūru sadalījumu, apstiprinot histogrammā novēroto nelielo novirzi pa labi.

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.

6.1.1. *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. (6.1.)):

$$\mathrm{AR(p)}: \mathrm{Y_t} = \mathrm{c} + \phi_1 \mathrm{Y_{t-1}} + \phi_2 \mathrm{Y_{t-2}} + \cdots + \phi_\mathrm{p} \mathrm{Y_{t-p}} + \epsilon_\mathrm{t} \qquad (6.1.)$$

kur

(Y_t) – sērijas vērtība laikā, t;
c – konstante, vesels vai decimālais skaitlis;
$\phi_1, \phi_2, \ldots, \phi_p$ – 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. (6.2.)):

$$\mathrm{I(d)}: \nabla^\mathrm{d} \mathrm{Y_t} \qquad (6.2.)$$

- *MA* (q): mainīgais vidējais parametrs. Modelis, kas izmanto atkarību starp novērojumu un atlikušo kļūdu no slīdošā vidējā modeļa, ko izmanto novēlotiem novērojumiem (sk. (6.3.)):

$$\mathrm{MA(q)}: \mathrm{Y_t} = \mathrm{c} + \epsilon_\mathrm{t} + \theta_1 \epsilon_\mathrm{t-1} + \theta_2 \epsilon_\mathrm{t-2} + \cdots + \theta_\mathrm{q} \epsilon_\mathrm{t-q} \qquad (6.3.)$$

Apvienojot *AR*, *I* un *MA* parametrus, *ARIMA* modelis tiek izteikts šādi (sk. (6.4.)):

$$ARIMA(\mathrm{p,d,q}): \nabla^\mathrm{d} \mathrm{Y_t} = \mathrm{c} + \phi_1 \mathrm{Y_{t-1}} + \cdots + \phi_\mathrm{p} \mathrm{Y_{t-p}} + \theta_1 \epsilon_\mathrm{t-1} + \cdots + \theta_\mathrm{q} \epsilon_\mathrm{t-q} + \epsilon_\mathrm{t} \qquad (6.4.)$$

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 y_t *ADF* pārbauda nulles hipotēzi (sk. (6.5.)):

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \cdots + \epsilon_t \quad (6.5.)$$

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;

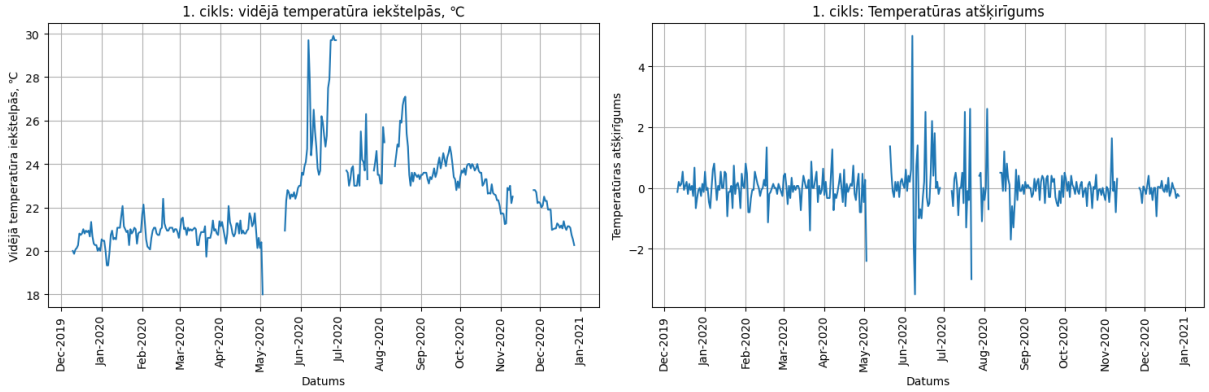
δ_i – atkarīgā mainīgā aizkavēto pirmo starpību koeficienti;

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.

Pārbaudīt, vai dati ir stacionāri, ir iespējams arī vizuāli (sk. 6.3. att.):



6.3. att. **Vidējās temperatūras atšķirības attēlošana.**

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;

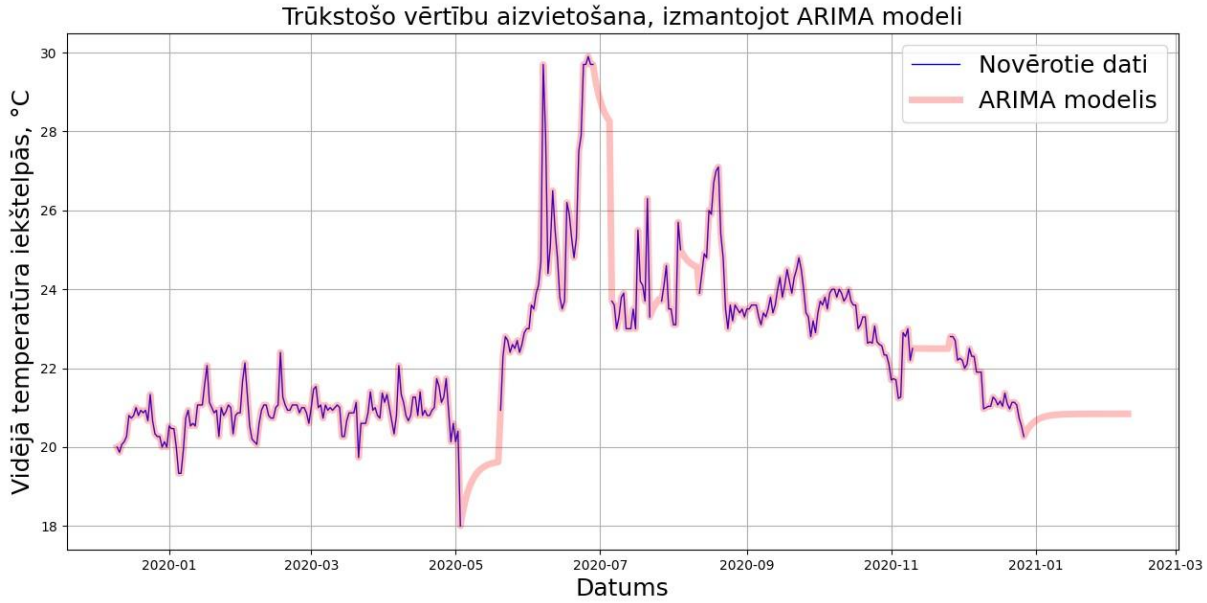
- optimālais d ir tas, kas padara sēriju nekustīgu un samazina *AIC*.

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. 6.4. att.):



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

6.1.2. 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. (6.6.)):

$$y = \beta_0 + \beta_1 x \tag{6.6.}$$

kur

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

2. Tiek prognozēti turpmākie indeksi x′, izmantojot apmācīto modeli (sk. (6.7.)):

$$\hat{y} = \beta_0 + \beta_1 x' \tag{6.7.}$$

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. (6.8.)):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (6.8.)$$

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. (6.9.)):

$$L_{avg} = \frac{\sum_{i=j-n}^{j+n} e^{-0.1i} y_i}{\sum_{i=j-n}^{j+n} e^{-0.1i}} \quad (6.9.)$$

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. (6.10.)):

$$S_D = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i - \overline{D}}{\sigma_D}\right)^3 \quad (6.10.)$$

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. (6.11.)):

$$L_{avg_{adj}} = L_{avg} + \alpha \times S_D \quad (6.11.)$$

kur

α = 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. (6.12.), (6.13.)):

$$O_f = 0.5, \text{ja } \frac{dy_{back}}{dx} \cdot \frac{dy_{forw}}{dx} < 0 \quad (6.12.)$$

$$O_f = 1.0, \text{ja } \frac{dy_{back}}{dx} \cdot \frac{dy_{forw}}{dx} \geq 0 \quad (6.13.)$$

Kur izmaiņu ātrumi tiek aprēķināti:

$$\frac{dy_{back}}{dx} = \frac{y_{j-1} - y_{j-n}}{n-1} \quad (6.14.)$$

$$\frac{dy_{forw}}{dx} = \frac{y_{j+n} - y_{j+1}}{n-1} \quad (6.15.)$$

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. (6.16.)):

$$w = e^{-0.15 \cdot \min(d_{left}, d_{right})} \cdot O_f \quad (6.16.)$$

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. (6.17.)):

$$V_{imp}(i) = w \times L_{avg_{adj}}(i) + (1 - w) \times T(i) \quad (6.17.)$$

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. (6.19.):

$$V_{imp}^{final}(i) = 0.7 \cdot V_{imp}(i) + 0.3 \cdot V_{imp}(i-1) \quad (6.18.)$$

kur

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

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

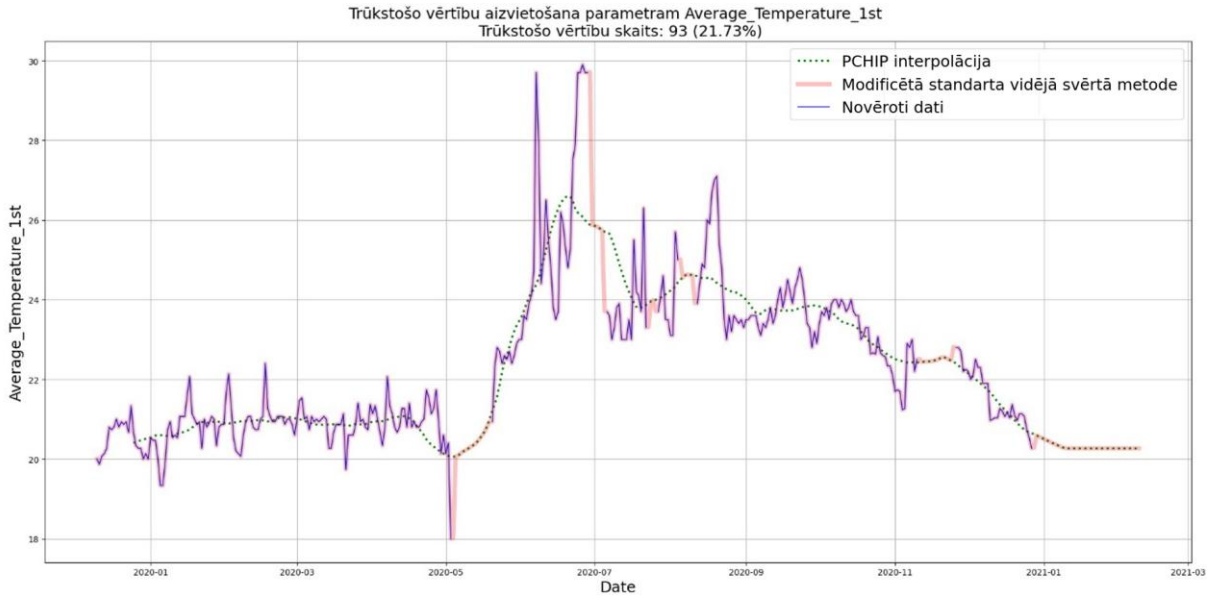
11. Gala rezultāta izlīdzināšanai tiek pielietots trīs punktu slīdošais vidējais (sk. (6.19.)(6.20.)):

$$V_{smooth}(i) = \frac{1}{3} \sum_{k=i-1}^{i+1} V_{imp}^{final}(k) \quad (6.19.)$$

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 6.5. attēlā.



6.5. 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.

6.2. 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. 6.1. 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.

6.1. 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	$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$

6.1. tabulas turpinājums

Kritērijs	Vienādojums
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.

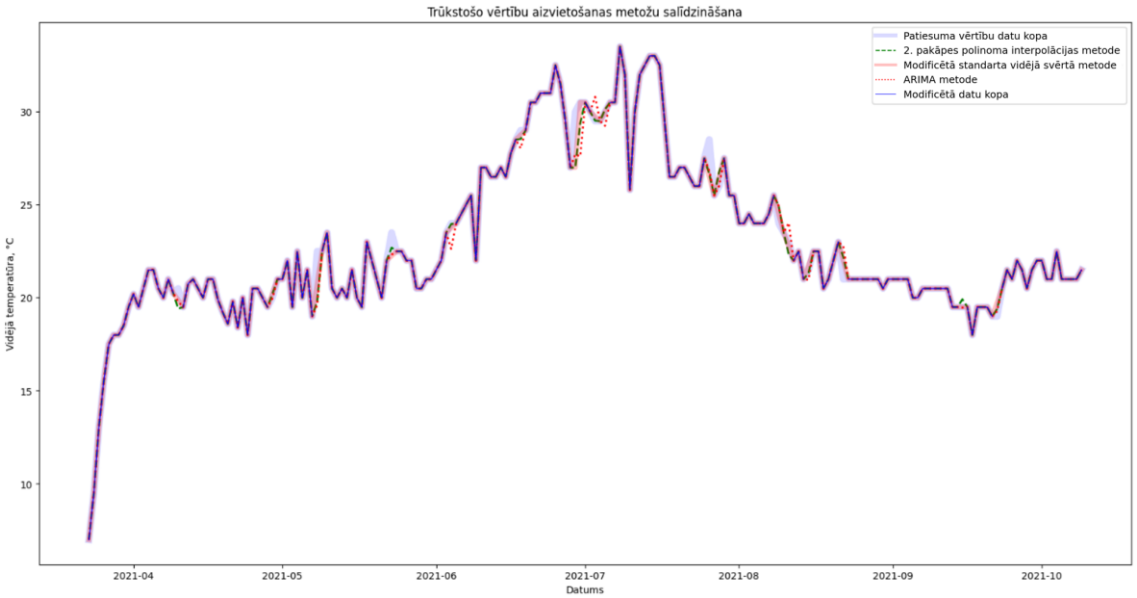
Novērotās kopas tiek izveidotas, ieviešot trūkstošās vērtības ar šādiem nosacījumiem:

1. gadījuma trūkstošās vērtības, līdz 10% no kopējā skaita;
2. gadījuma trūkstošās vērtības, līdz 20% no kopējā skaita (sk. 1);
3. gadījuma trūkstošās vērtības, līdz 40% no kopējā skaita;
4. gadījuma trūkstošo vērtību virknes, divi gabali ar garumu 10 elementi katra (sk. 2);
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.

Scenārijs 1: gadījuma trūkstošās vērtības, līdz 10% no kopējā skaita.

Pēc rezultātiem (sk. 6.2. tabulu, 6.6. att.) MSVSM ir visprecīzākā metode trūkstošo vērtību aizvietošanai šim scenārijam, jo tai ir viszemākās vērtības visās trīs metrikās.



6.6. att. **Gadījuma trūkstošās vērtības, līdz 10% no kopējā skaita.**

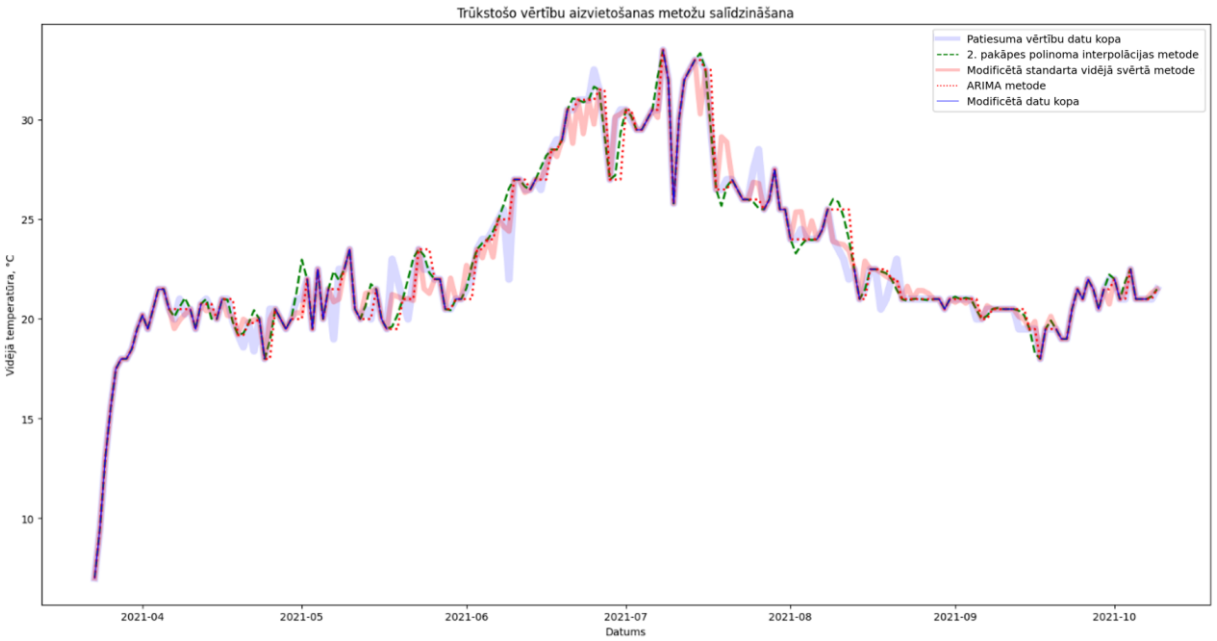
6.2. tabula. **Metriku rezultāti scenārijam ar gadījuma trūkstošām vērtībām, līdz 10% no kopējā skaita**

Metode	*MSE*	*RMSE*	*MAE*
Polinoma interpolācija	0.133012	0.364708	0.079453
MSVSM	0.101508	0.318604	0.064097
ARIMA	0.185362	0.430537	0.102718

Polinomu interpolācija ir otrs labākais rādītājs, un kļūdu rādītāji ir vidēji zemi, bet augstāki nekā MSVSM. *ARIMA* ir visaugstākā kļūdu metrika, kas norāda, ka tā var neatbilst šai konkrētajai datu kopai, kā arī citām aizvietošanas uzdevumu metodēm. Kopumā visām metodēm ir zemas metriku kļūdas.

Scenārijs 3: gadījuma trūkstošās vērtības, līdz 40% no kopējā skaita.

Polinomu interpolācijas gadījumā *MSE*, *RMSE* un *MAE* pieaugums ir būtisks (sk. 6.3. tabulu, 6.7. att.), kas liecina, ka to veiktspēja ievērojami pasliktinās, palielinoties trūkstošo datu īpatsvaram.



6.7. att. **Gadījuma trūkstošās vērtības, līdz 40% no kopējā skaita.**

Lai gan MSVSM metodes *MSE* un *RMSE* pieaugums ir manāms, tas ir mazāks nekā polinomu interpolācijas pieaugums. Interesanti, ka *MAE* palielinājās nedaudz vairāk nekā polinomu interpolācija. Šķiet, ka šī metode ir noturīgāka, lai palielinātu trūkstošās vērtības, salīdzinot ar citām.

6.3. tabula. **Metriku rezultāti scenārijam ar gadījuma trūkstošām vērtībām, līdz 40% no kopējā skaita**

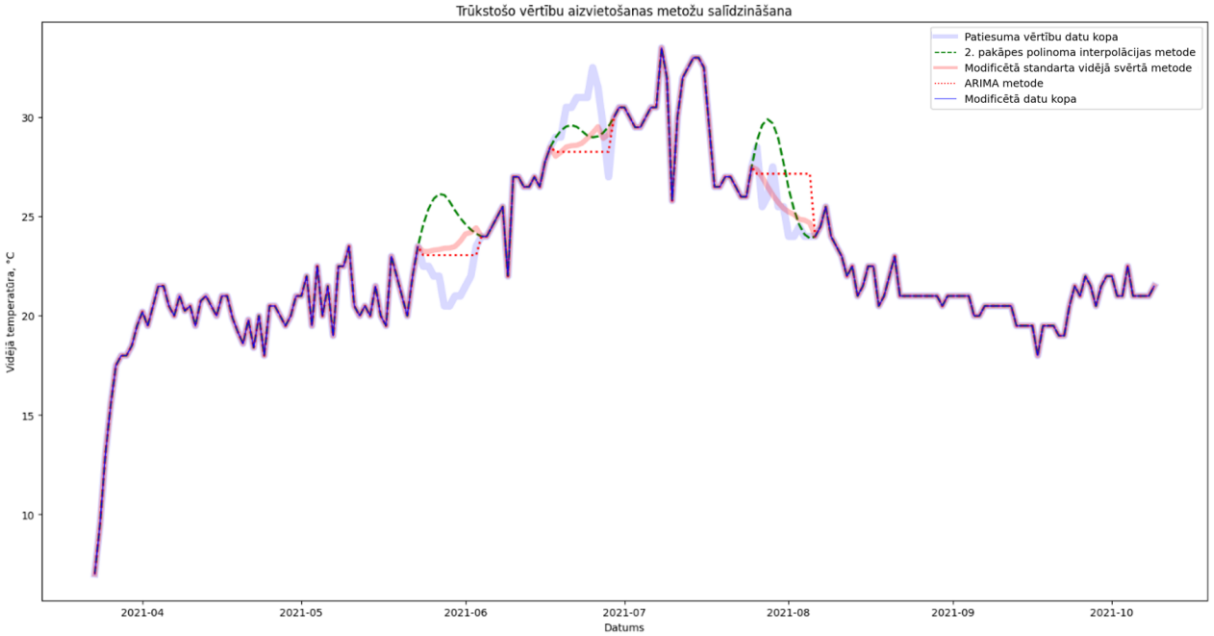
Metode	*MSE*	*RMSE*	*MAE*
Polinoma interpolācija	0.630742	0.794193	0.327112
MSVSM	0.459900	0.678159	0.331093
ARIMA	0.678010	0.823414	0.341294

ARIMA uzrāda lielu *MSE* un *RMSE* pieaugumu, līdzīgi kā polinomu interpolācija, un arī lielāko *MAE* pieaugumu. Tas liecina, ka *ARIMA* darbību būtiski ietekmē arī lielāks trūkstošo datu īpatsvars. Visas metodes negatīvi ietekmē trūkstošo datu īpatsvara pieaugums, ko atspoguļo augstāka kļūdu metrika.

MSVSM paliek labāka no trim metodēm trūkstošo vērtību palielināšanai, saglabājot zemāko *MSE* un *RMSE*. Polinomu interpolācija un *ARIMA* ir jutīgākas pret palielinātajiem trūkstošajiem datiem, kas liecina par būtiskāku veiktspējas pasliktināšanos. Aizvietošanas metodes izvēli varētu ietekmēt trūkstošo datu īpatsvars, un MSVSM, iespējams, ir labākā izvēle datu kopām ar augstāku trūkstošo vērtību līmeni. Bet līdz šim trūkstošās vērtības tika iejauktas nejauši, bez bloku veidošanas.

Scenārijs 5: gadījuma trūkstošo vērtību virknes, trīs gabali ar garumu 10 elementi katra.

Polinomu interpolācijai ir visaugstākais *MSE* un *RMSE* (sk. 6.4. tabulu, 6.8. att.), kas norāda, ka polinoma interpolācijas metodei ir vidēji lielākas kļūdas salīdzinājumā ar citām metodēm. *MAE* ir arī visaugstākā, kas liecina, ka vidēji absolūtās novirzes no faktiskajām vērtībām ir lielākas.



6.8. att. **Gadījuma trūkstošo vērtību virknes, trīs gabali ar garumu 10 elementi katra.**

MSVSM aizvietošana uzrāda zemāko *MSE*, *RMSE* un *MAE* starp trim metodēm, norādot, ka tai ir mazākās vidējās kļūdas gan kvadrātiska, gan absolūtā izteiksmē. Tas liek domāt, ka MSVSM aizvietošana ir salīdzinoši precīza un konsekventa, prognozējot trūkstošās vērtības šai konkrētajai datu kopai. Jo īpaši zemais *RMSE* liecina, ka aprēķinātajos datos ir mazāk lielu kļūdu, kas nozīmē, ka šī metode var efektīvāk apstrādāt datu novirzes vai atšķirības.

6.4. tabula. **Metriku rezultāti scenārijam ar gadījuma trūkstošo vērtību virknēm, trīs gabali ar garumu 10 elementi katra**

Metode	***MSE***	***RMSE***	***MAE***
Polinoma interpolācija	1.247475	1.116904	0.370906
MSVSM	0.507510	0.712398	0.250721
ARIMA	0.754821	0.868804	0.312186

ARIMA metodes kļūdu metrika ieņem vidējo pozīciju starp divām citām metodēm – polinomu interpolāciju un MSVSM aizvietošanu. Šī metode uzrāda labākus rezultātus nekā polinomu interpolācija, tomēr nesasniedz MSVSM aizvietošanas precizitātes līmeni. Šāds sniegums norāda uz *ARIMA* metodes spēju sabalansēt dažādus faktorus. Tā nodrošina pietiekamu precizitāti, vienlaikus saglabājot noturību pret datu novirzēm. Šī īpašība padara *ARIMA* metodi par piemērotu izvēli specifiskos gadījumos, īpaši situācijās, kur nepieciešams līdzsvars starp precizitāti un elastību datu apstrādē. Metodes izvēle būtu jābalsta uz konkrētā

lietojuma prasībām un datu īpatnībām. *ARIMA* var būt optimāla izvēle gadījumos, kad augstākā precizitāte nav galvenā prioritāte, bet svarīga ir metodes spēja pielāgoties dažādām datu variācijām.

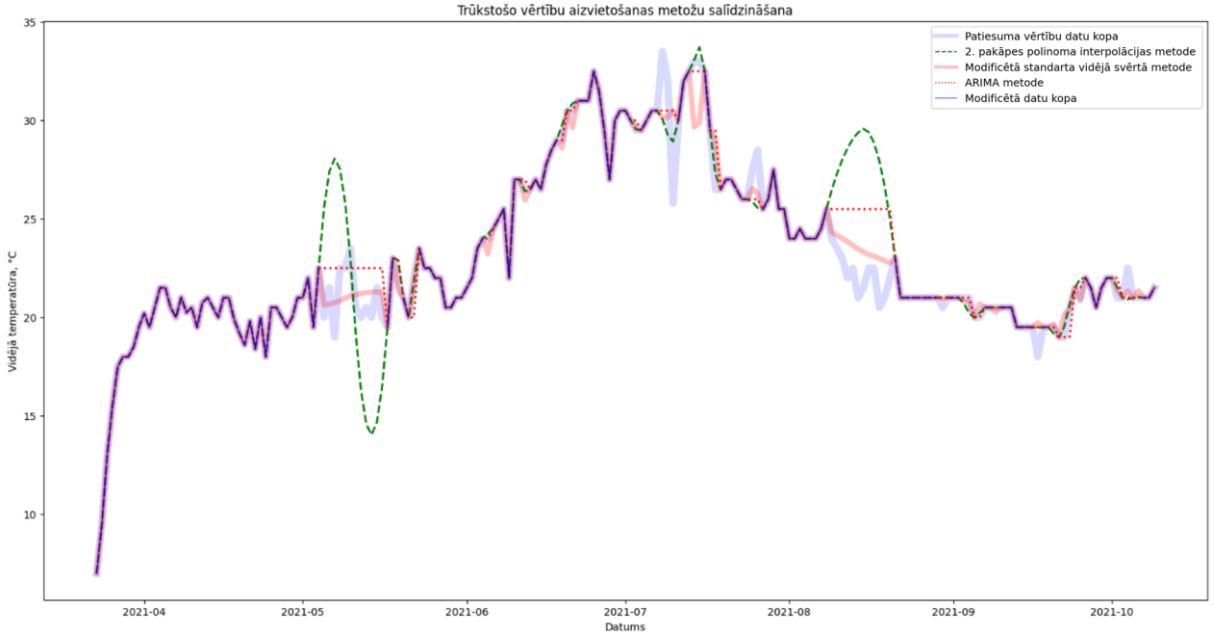
Scenārijs 6: gadījuma trūkstošās vērtības, līdz 20% no kopējā skaita, un gadījuma trūkstošo vērtību virknes, divi gabali ar garumu 10 elementi katra.

Šajā gadījumā gabalu ģenerēšana tiek realizēta jau uz "apgrieztās" datu kopas, kas attiecīgi nozīmē, ka trūkstošo vērtību īpatsvars var būt stipri mainīgs starp dažādām iterācijām.

Polinomu interpolācijai ir ievērojami augstāka *MSE*, *RMSE* un *MAE* vērtība nekā citām metodēm (sk. 6.5. tabulu, 6.9. att.), kas norāda, ka šajā scenārijā tā ir mazāk precīza. Augstās *MSE* un *RMSE* vērtības liecina, ka dažās prognozēs ir lielas kļūdas, kas varētu būt saistītas ar pārmērīgu pielāgošanu trūkstošajiem gabaliem. 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.



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

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.

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

Metode	***MSE***	***RMSE***	***MAE***
Polinoma interpolācija	4.145397	2.036025	0.758791
MSVSM	0.630528	0.794058	0.310932
ARIMA	1.287761	1.134796	0.438806

ARIMA ir precīzāka par polinomu interpolāciju, bet ne tik efektīva kā MSVSM. Tās mērenā veiktspēja varētu būt saistīta ar spēju apstrādāt laika rindu datus, kur trūkstošās vērtības

atbilst noteiktam modelim (piemēram, gabalos), taču nejaušas trūkstošās vērtības var radīt papildu sarežģītību, ko *ARIMA* apstrādā mazāk efektīvi.

Polinomu interpolācija šajā gadījumā ir vismazāk efektīva metode. Lielā kļūdu metrika norāda, ka tā var nebūt labākā izvēle, ja datos ir trūkstošo vērtību daļas, kas ir izplatīta problēma reālajā pasaulē.

6.2.1. Visaptverošā testēšana

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.

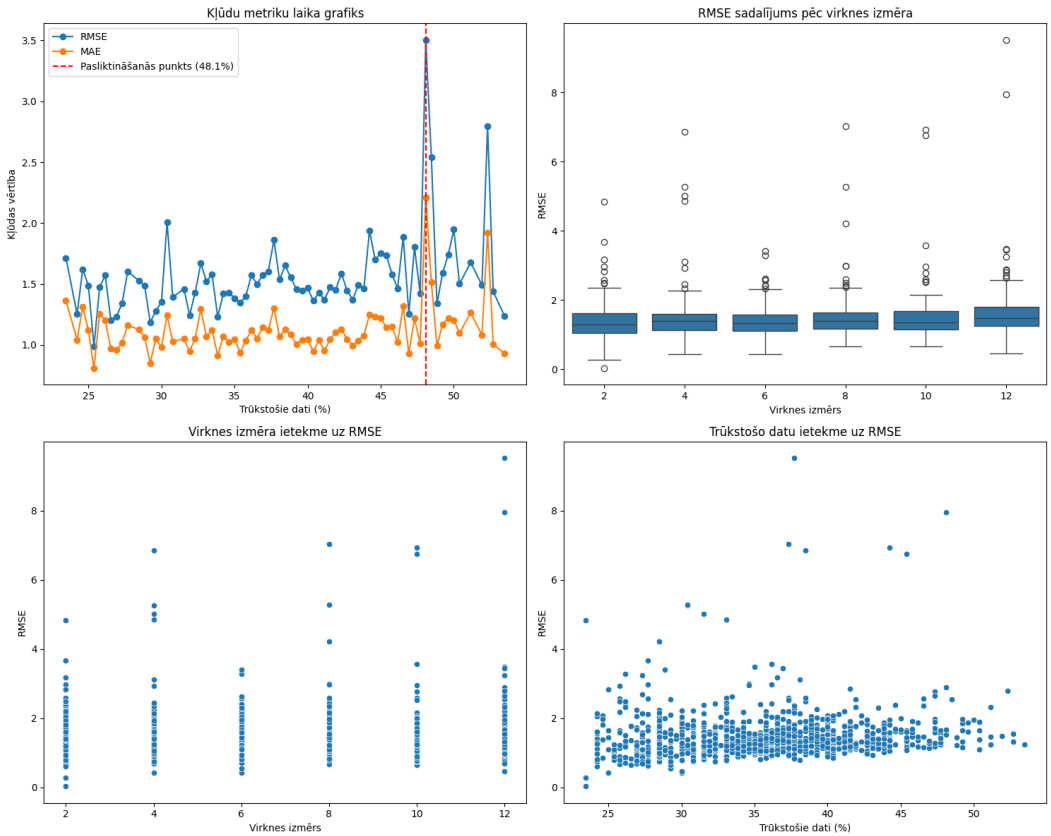
Kopumā tiek veikti 990 testi, kas iegūti no 6 virkņu izmēriem, 5 virkņu skaitiem, 11 proporcijām un 3 atkārtojumiem. Katra testa procedūra sastāv no diviem posmiem. Pirmajā posmā tiek veikta trūkstošo datu ievietošana, kur vispirms tiek ievietotas trūkstošo vērtību virknes nejaušās pozīcijās, pēc tam 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.

Pēc veiktās analīzes ARIMA metodei pasliktināšanās punkts tiek noteikts pie 48,1% trūkstošo datu, modificētai standarta vidējās svērtās metodei pie 52,3% trūkstošo datu, un polinoma interpolācijas metodei pie 25,0% trūkstošo datu. Statistiskās ticamības nodrošināšanai katra konfigurācija tiek testēta trīs reizes, kas ļauj mazināt nejaušo variāciju ietekmi trūkstošo datu izvietojumā. Galīgajai analīzei tiek izmantoti vidējie rādītāji no šiem atkārtojumiem, kas nodrošina ticamākus rezultātus.

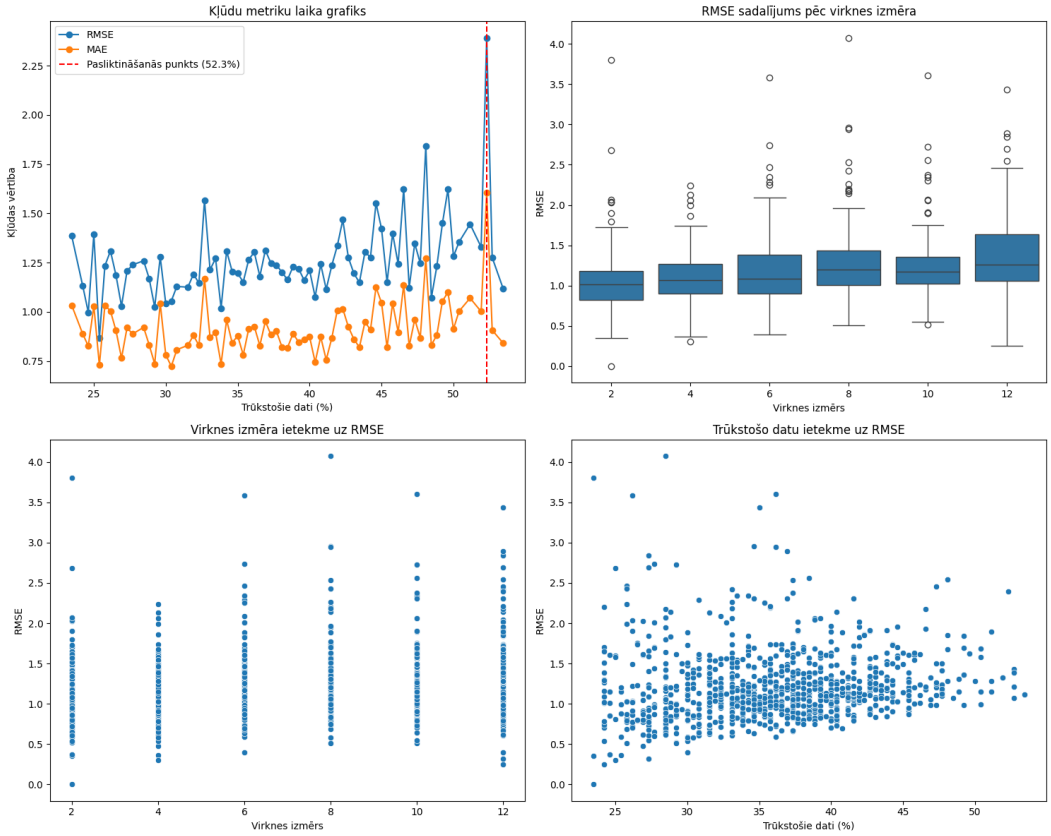
ARIMA metode uzrāda stabilu veiktspēju līdz 48,1% trūkstošo datu apjomam, kur tiek novērots pasliktināšanās punkts (sk. 6.10. att.).

Pirms šī sliekšņa ARIMA metode uztur relatīvi stabilu RMSE vērtību 1,48, kas norāda uz konsekventu prognozēšanas precizitāti. Pēc pasliktināšanās punkta tiek novērots būtisks MSE pieaugums par 91,4% un RMSE pieaugums par 27,9%, kas liecina par metodes precizitātes samazināšanos. Analīze parāda, ka metode vislabāk darbojas ar mazākiem virkņu izmēriem (10–20 elementi), kur RMSE vērtības ir robežās no 1,2 līdz 1,4. Veiktspēja pakāpeniski pasliktinās, palielinoties gan virknes izmēram, gan trūkstošo datu procentuālajam daudzumam. Visstabilākā veiktspēja tiek novērota ar virkņu izmēriem no 10 līdz 30 elementiem un trūkstošo datu apjomu zem 40%.



6.10. att. **ARIMA metodes veiktspēja.**

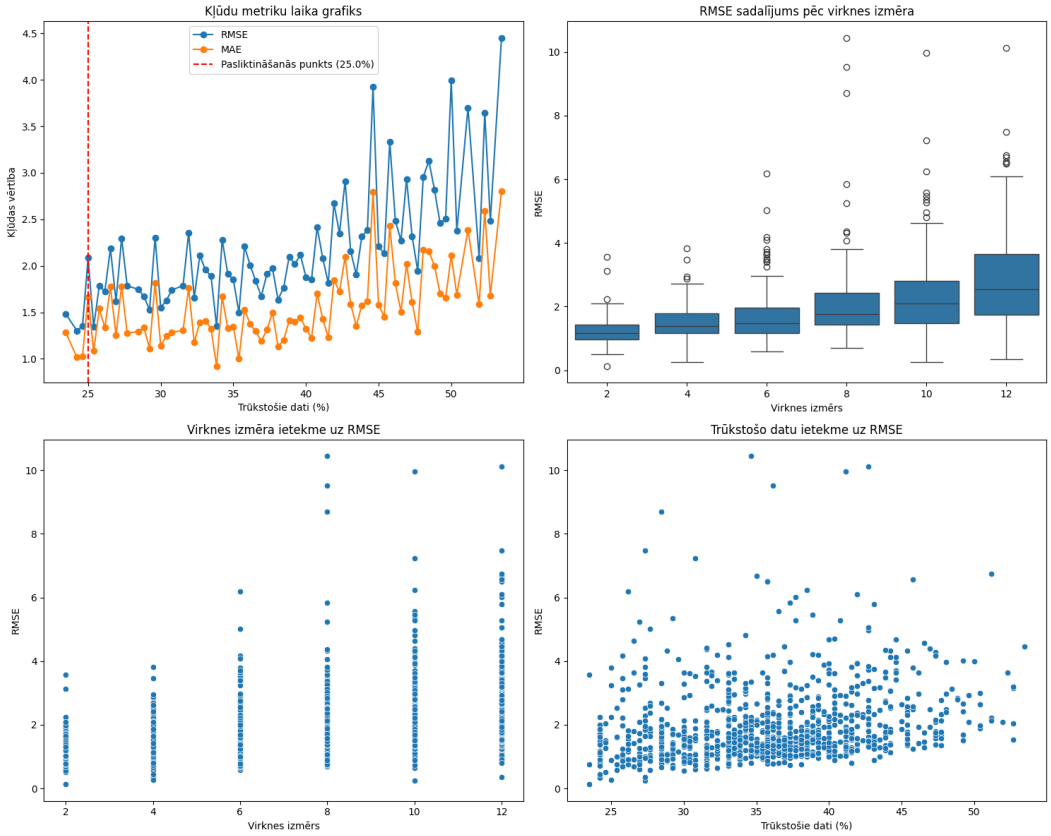
MSVSM uzrāda (sk. 6.11. att.) vislabāko kopējo veiktspēju starp visām trim metodēm, ar pasliktināšanās punktu pie 52,3% trūkstošo datu – augstākais slieksnis starp visām metodēm.



6.11. att. **MSVSM veiktspēja.**

Salīdzinošā analīze parāda pakāpeniskāku RMSE pieaugumu salīdzinājumā ar citām metodēm. Analīze atklāj 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 ar tikai 36,7% MSE pieaugumu un 18,6% RMSE pieaugumu.

Polinomu interpolācijas metode uzrāda visagrāko pasliktināšanās punktu pie 25% trūkstošo datu, kas padara to visjutīgāko pret datu iztrūkumiem. Salīdzinošā analīze atklāj strauju RMSE pieaugumu pēc šī punkta. Analīze parāda, ka vislabākā veiktspēja tiek sasniegta ar maziem virkņu izmēriem (10–15 elementi) un zemiem trūkstošo datu procentuālajiem daudzumiem (zem 20%) (sk. 6.12. att.).



6.12. att. **Polinomu interpolācijas metodes veiktspēja.**

Metode uzrāda visaugstāko svārstīgumu starp visām trim pieejām, ar RMSE vērtībām, kas dramatiski mainās dažādās konfigurācijās. Pēc pasliktināšanās punkta tiek novērots dramatisks MSE pieaugums par 137,6% un RMSE pieaugums par 49%, kas ir visaugstākā pasliktināšanās starp visām metodēm. Metode kļūst īpaši neuzticama, kad gan virknes izmērs, gan trūkstošo datu procentuālais daudzums ir augsts, ar RMSE vērtībām, kas pārsniedz 2,0.

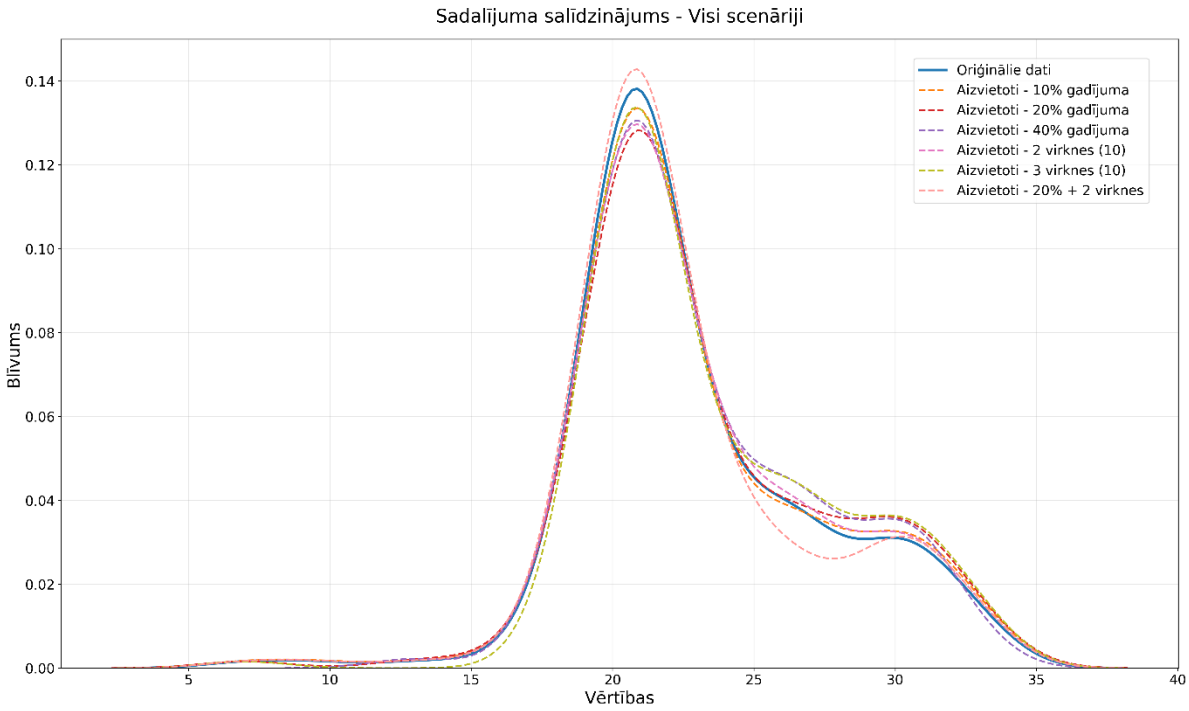
Visu trīs metožu salīdzinošā analīze parāda būtiskas atšķirības to veiktspējā un piemērotībā dažādiem trūkstošo datu scenārijiem. MSVSM izceļas kā visefektīvākā, uzrādot augstāko pasliktināšanās slieksni (52,3%) un viszemāko veiktspējas pasliktināšanos pēc šī sliekšņa (MSE pieaugums 36,7%, RMSE pieaugums 18,6%). Šī metode ir īpaši efektīva ar maziem līdz vidējiem virkņu izmēriem (10–30 elementi), nodrošinot stabilus rezultātus pat pie augsta trūkstošo datu īpatsvara. ARIMA metode uzrāda vidēju veiktspēju ar pasliktināšanās slieksni pie 48,1% trūkstošo datu. Lai gan tās veiktspējas pasliktināšanās ir ievērojamāka nekā modificētajai standarta vidējās svērtās metodei (MSE pieaugums 91,4%), tā joprojām saglabā pietiekamu precizitāti līdz pasliktināšanās punktam. Savukārt polinomu interpolācijas metode uzrāda viszemāko noturību pret trūkstošajiem datiem, ar agrīnu pasliktināšanās slieksni pie 25% un dramatisku veiktspējas kritumu pēc tā (MSE pieaugums 137,6%). Balstoties uz šiem

rezultātiem, MSVSM ir ieteicamā izvēle lielākajai daļai praktisko pielietojumu, īpaši gadījumos, kur sagaidāms augsts trūkstošo datu īpatsvars vai nepieciešama augsta precizitāte. ARIMA metode var būt piemērota kā alternatīva situācijās ar zemāku trūkstošo datu īpatsvaru, savukārt polinomu interpolācijas metode būtu izmantojama tikai specifiskos gadījumos ar ļoti zemu trūkstošo datu īpatsvaru un maziem virkņu izmēriem.

6.2.2. MSVSM metodes stabilitāte un precizitāte

Trūkstošo vērtību aizvietošanas metožu novērtēšana ir būtiska datu analīzes procesa sastāvdaļa, jo tā ļauj izprast, cik lielā mērā izvēlētā metode spēj saglabāt datu oriģinālās īpašības. Iepriekšējos scenārijos labākus rezultātus uzrādīja MSVSM, tomēr ir jāsaprot, cik stabili tā strādā.

Sadalījuma salīdzinājuma tests (sk. 6.13. att.) ir fundamentāls rādītājs aizvietošanas kvalitātes novērtēšanā, jo tas parāda, cik labi tiek saglabātas datu statistiskās īpašības.



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

Scenārijā ar 10% gadījuma trūkstošajām vērtībām metode uzrāda izcilus rezultātus – vidējās vērtības novirze ir tikai 0,02%, kas praktiski nozīmē, ka aizvietotie dati gandrīz perfekti atbilst oriģinālajam datu kopas līmenim. Standartnovirzes izmaiņas 0,40% apmērā norāda uz ļoti labu datu izkliedes saglabāšanu. Sadalījuma forma vizuāli praktiski neatšķiras no oriģinālās, kas ir būtiski tālākai statistiskai analīzei.

Pārejot pie sarežģītākiem scenārijiem, novērojam pakāpenisku kvalitātes samazināšanos, tomēr tā joprojām saglabājas pieņemamā līmenī. Divu secīgu trūkstošo vērtību virkņu gadījumā (katra 10 elementu gara) vidējās vērtības novirze pieaug līdz 0,11%, bet standartnovirzes izmaiņas sasniedz 4,88%. Šādas izmaiņas jau var ietekmēt datu analīzes rezultātus, tomēr tās joprojām ir pietiekami mazas, lai metodi uzskatītu par uzticamu. Interesanti, ka sadalījuma asimetrija mainās minimāli, kas liecina par metodes spēju saglabāt datu kopas pamatstruktūru.

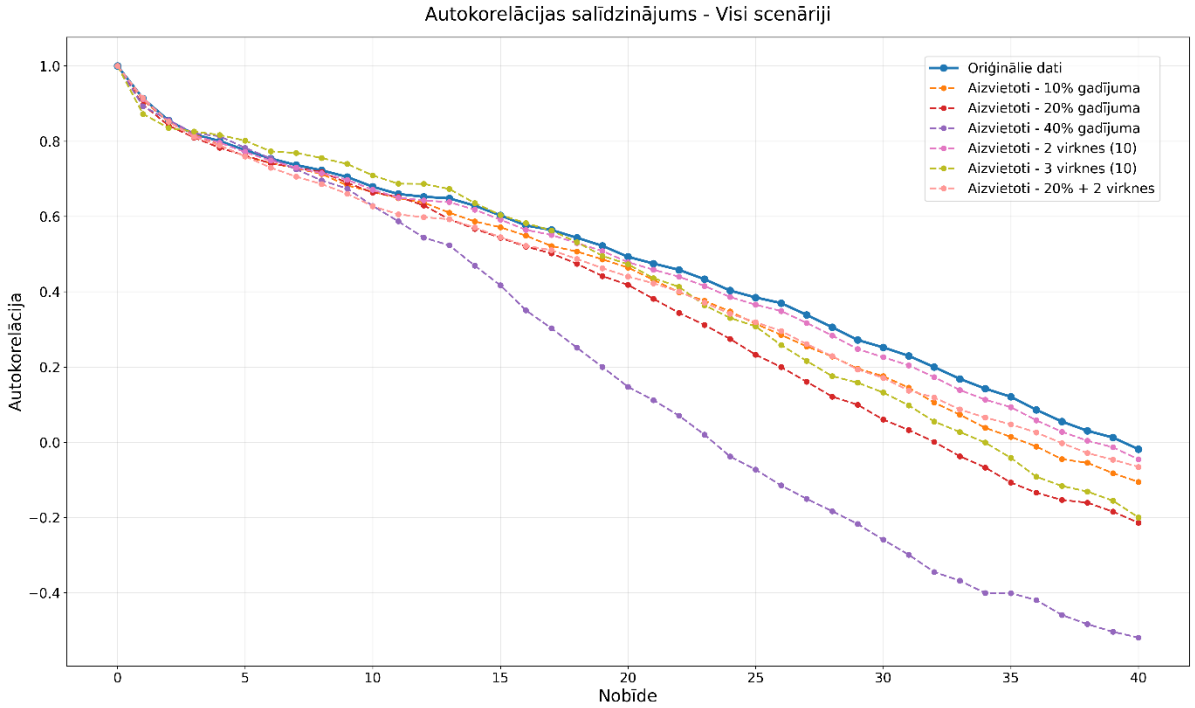
Vislielākās izmaiņas novērojamas 40% gadījuma trūkstošo vērtību scenārijā. Šeit vidējās vērtības novirze sasniedz 1,58%, bet standartnovirzes izmaiņas – 9,20%. Šādas novirzes jau var būtiski ietekmēt secinājumus, kas tiek izdarīti no datiem. Sadalījuma forma uzrāda ievērojamas

nobīdes, īpaši “astes” daļās, kas norāda uz grūtībām precīzi rekonstruēt ekstremālās vērtības. Šis ir būtisks ierobežojums gadījumos, kad tieši ekstremālie novērojumi ir analīzes fokusā.

Autokorelācijas tests ir īpaši nozīmīgs laika rindu analīzē, jo tas parāda, cik labi tiek saglabātas secīgo vērtību savstarpējās sakarības (sk. 6.14. att.).

Pārsteidzoši labākos rezultātus uzrāda kombinētais scenārijs (20% gadījuma trūkstošās vērtības kopā ar divām 10 elementu virknēm), kur vidējā autokorelācijas atšķirība ir tikai 0,009, bet maksimālā – 0,036. Šāds rezultāts, iespējams, skaidrojams ar to, ka dažādu trūkstošo vērtību veidu kombinācija ļauj metodei labāk “uztvert” datu periodiskumu un tendences.

Vidēja līmeņa rezultātus uzrāda scenārijs ar divām 10 elementu trūkstošo vērtību virknēm, kur vidējā autokorelācijas atšķirība ir 0,025, bet maksimālā – 0,076. Šie rādītāji liecina, ka metode spēj saglabāt laika rindas pamatstruktūru pat gadījumos, kad trūkst secīgu vērtību posmi. Tomēr jau ir novērojama neliela ietekme uz periodiskuma struktūru, īpaši ap trūkstošo vērtību virkņu robežām.



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

Vissliktākos rezultātus uzrāda 40% gadījuma trūkstošo vērtību scenārijs, kur vidējā autokorelācijas atšķirība sasniedz 0,180, bet maksimālā – 0,442. Šādas novirzes jau būtiski ietekmē laika rindas struktūru un var novest pie kļūdainiem secinājumiem par datu periodiskumu vai tendencēm. Īpaši problemātiskas ir augstākas kārtas autokorelācijas, kas norāda uz grūtībām saglabāt ilgtermiņa atkarības struktūras datos.

Veiktā analīze sniedz vispusīgu priekšstatu par aizvietošanas metodes spējām un ierobežojumiem. Metode uzrāda ļoti labus rezultātus situācijās ar nelielu trūkstošo vērtību īpatsvaru (līdz 10%), kur tiek saglabātas gan statistiskās, gan autokorelācijas īpašības. Tā spēj efektīvi tikt galā arī ar vidēja apjoma trūkstošajām vērtībām (līdz 20%) un ierobežota garuma secīgu vērtību iztrūkumiem. Tomēr pie lielāka trūkstošo vērtību apjoma (40%) novērojama būtiska kvalitātes pasliktināšanās, kas izpaužas gan kā statistisko rādītāju novirzes, gan kā autokorelācijas struktūras izmaiņas. Šis ir būtisks ierobežojums, kas jāņem vērā, plānojot metodes izmantošanu reālās situācijās. Interesanti, ka metode uzrāda labākus rezultātus kombinētos scenārijos nekā vienkāršos, kas varētu būt noderīgi reālās situācijās, kur bieži sastopamas dažādu trūkstošo vērtību veidu kombinācijas.

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.

6.3. Noviržu noteikšana un pielāgošana

Noviržu noteikšana un pielāgošana ir būtisks datu priekšapstrādes posms, kas ietekmē analīzes kvalitāti. Šim nolūkam izmanto vairākas metodes, katra ar savām priekšrocībām un ierobežojumiem. *Z-Score* metode ir vienkārša pieeja, kas balstās uz standartnovirzes aprēķinu. Tā identificē novirzes, nosakot to attālumu no vidējās vērtības standartnoviržu vienībās. Tomēr metode ir jutīga pret ekstremālām vērtībām, kas var izkropļot gan vidējo vērtību, gan standartnovirzi (Yaro et al., 2024). Starpkvartiļu diapazona (*IQR*) metode piedāvā robustāku pieeju, jo izmanto datu sadalījuma kvartiles. Tā ir efektīva asimetrisku datu gadījumos, taču var izrādīties pārāk konservatīva situācijās, kur būtiska ir datu dinamika (El Hairach, Tmiri, & Bellamine, 2024). Vinzorizācija ne tikai identificē, bet arī pielāgo novirzes, aizstājot tās ar tuvākajām "normālajām" vērtībām. Šī pieeja saglabā datu kopas pamatstruktūru, vienlaikus mazinot ekstremālo vērtību ietekmi (Yang. L. et al., 2024). Tomēr mainīgu datu gadījumā vinzorizācija viena pati var būt nepietiekama. Slīdošā loga (*Rolling Window*) metode papildina iepriekšējās pieejas, analizējot datus to lokālajā kontekstā. Tā ir īpaši piemērota laika rindu datiem, jo spēj pielāgoties mainīgām datu īpašībām dažādos laika posmos.

Ņemot vērā katras metodes priekšrocības un trūkumus, tika izstrādāta kombinēta pieeja, kas apvieno vinzorizācijas un slīdošā loga metodes priekšrocības. Slīdošā loga pieeja tiek īstenota, izmantojot trīs dažāda izmēra logus – 9, 19 un 39 punktus, kas ļauj aptvert dažāda mēroga datu izmaiņas. Šāda dažādu izmēru izvēle nav nejauša – tā ļauj aptvert gan īstermiņa, gan vidēja termiņa datu izmaiņas, vienlaikus saglabājot pietiekamu jutību pret lokālām novirzēm. Mazākais logs (9 punkti) ir jutīgs pret īstermiņa izmaiņām un spēj precīzi identificēt atsevišķas novirzes. Vidējais logs (19 punkti) nodrošina stabilitāti īstermiņa svārstību gadījumā, bet saglabā pietiekamu jutību pret būtiskām novirzēm. Lielākais logs (39 punkti) palīdz noteikt ilgtermiņa tendences un identificēt novirzes, kas varētu būt grūti pamanāmas mazākos laika posmos. Katrā logā tiek aprēķināti lokālie statistiskie rādītāji, kas ļauj precīzāk novērtēt konkrētā datu segmenta īpašības. Vinzorizācija tiek pielietota katra loga ietvaros, izmantojot z-vērtības slieksni 3.0, kas ir empīriski noteikts kā optimāls kompromiss starp jutību pret novirzēm un viltus pozitīvo rezultātu skaitu. Šāda dažādu izmēru logu un vinzorizācijas kombinācija nodrošina labu pamatu noviržu noteikšanai, tomēr reālos datos bieži novērojamas arī ilgtermiņa izmaiņas un sezonāli efekti. Tieši šī iemesla dēļ metodē tika ieviesta papildus tendences komponente.

Tendences komponenta ieviešana būtiski uzlaboja noviržu noteikšanas precizitāti, risinot fundamentālu problēmu – atšķirību starp īstām novirzēm un dabiskām datu izmaiņām. Šī problēma ir īpaši aktuāla temperatūras datos, kur dabiskās svārstības var būt ievērojamas, bet likumsakarīgas. Tendences aprēķināšanai izmanto slīdošā vidējā vērtību ar pielāgotu loga izmēru, kas ļauj identificēt ilgtermiņa izmaiņas, saglabājot jutību pret īstermiņa novirzēm. Komponents darbojas gan noviržu noteikšanā, gan to pielāgošanas procesā, tādējādi saglabājot datu dabisko dinamiku. Piemēram, pakāpeniska temperatūras pieauguma gadījumā tas nodrošina dabiskā trenda saglabāšanu. Lai pilnībā izprastu metodes darbību, svarīgi atsevišķi aplūkot gan noviržu noteikšanu, gan to pielāgošanu.

Noviržu apstrādes process ir sadalīts divās galvenajās fāzēs – noteikšanā un pielāgošanā. Noteikšanas fāzē katram datu punktam tiek veikta virkne aprēķinu visos trijos izvēlētajos logos. Tiek aprēķinātas lokālās statistiskās vērtības, kas ietver vidējo vērtību un standartnovirzi, kā arī z-vērtības attiecībā pret šiem rādītājiem. Vienlaikus tiek novērtēta punkta atbilstība aprēķinātajai tendencei. Lēmums par punkta klasificēšanu kā novirzi tiek pieņemts, balstoties uz vairākiem kritērijiem. Pirmkārt, punkta z-vērtībai jāpārsniedz noteiktais slieksnis (3.0) vismaz divos no trim logiem. Šāda pieeja samazina viltus pozitīvo rezultātu skaitu. Otrkārt, tiek

ņemta vērā punkta atbilstība tendencei – ja punkts būtiski atšķiras no aprēķinātās tendences vērtības, tas palielina varbūtību, ka punkts tiks klasificēts kā novirze. Pielāgošanas fāze ir tikpat būtiska kā noteikšana. Katrai identificētajai novirzei tiek aprēķināta jauna vērtība, izmantojot informāciju no visiem trim logiem. Šajā procesā tiek izmantota vinzorizācija, kas nodrošina, ka jaunās vērtības nepārsniedz "normālo" vērtību diapazonu attiecīgajā logā. Būtiski, ka pielāgošanas process ņem vērā gan lokālo datu struktūru, gan tendences komponentu, gan arī iepriekš pielāgotās vērtības, lai nodrošinātu rezultātu saskaņotību.

6.3.1. Multi-skalas integrētā noviržu analīzes metode

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. (6.20.)):

$$rolling\_med_i = median(x_{i-k}, \dots, x_i, \dots, x_{i+k}) \tag{6.20.}$$

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. (6.21.)):

$$trend_i = \sum_{j=-k}^{k} c_j \cdot rolling\_med_{i+j} \tag{6.21.}$$

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. (6.22.)):

$$MAD_i = median\left(\left|x_j - median(x)\right|\right) \tag{6.22.}$$

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. (6.23.)):

$$local\_var_i = \frac{MAD_i}{median(MAD)} \tag{6.23.}$$

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. (6.24.)):

$$\Delta x_i = |x_i - x_{i-1}| \tag{6.24.}$$

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. (6.25)):

$$spike\_threshold_i = \left(median(\Delta x) + 2 \cdot median(|\Delta x - median(\Delta x)|)\right) \cdot max(1, local\_var_i) \tag{6.25.}$$

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. (6.26.)):

$$\mu_{i,w} = \frac{1}{w} \sum_{j=i-k}^{i+k} x_j \tag{6.26.}$$

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. (6.27.)):

$$\sigma_{i,w} = \sqrt{\frac{1}{w} \sum_{j=i-k}^{i+k} \left(x_j - \mu_{i,w}\right)^2} \tag{6.27.}$$

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. (6.28.)):

$$z_{i,w} = \frac{|x_i - \mu_{i,w}|}{\sigma_{i,w}} \tag{6.28.}$$

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. (6.29.)):

$$confidence_i = \sum_w weight_w \cdot [z_{i,w} > z_{threshold}] + 1.5 \cdot [rapid\_recovery_i] + 1.0 \cdot [trend\_outlier_i] + 1.2 \cdot [consecutive\_deviation_i] \tag{6.29.}$$

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. (6.30.)):

$$outlier_i = 1 \text{ ja } confidence_i \geq 2.0 \text{ vai } |\Delta x_i| > spike\_threshold_i \text{ vai } |x_i - trend_i| > 2.0 \cdot \sigma_{trend} \text{ citādi } 0 \tag{6.30.}$$

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. (6.31.)):

$$x_i' = trend_i + sign(x_i - trend_i) \cdot min(|x_i - trend_i|, 2 \cdot \sigma_{trend}) \ ja \ outlier_i = 1 \ citādi \ x_i \quad (6.31.)$$

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.

6.3.2. Ieejas parametru optimizācija

MINA parametru optimizācijas galvenais mērķis ir 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, kas ietver dažādus noviržu scenārijus. Šie scenāriji tika rūpīgi izvēlēti, lai atspoguļotu reālās situācijās sastopamās problēmas:

- izolētas novirzes (pīķi) – īslaicīgas ekstremālas vērtības;
- secīgas novirzes – vairāku secīgu punktu nobīdes no normālā līmeņa;
- tendences izmaiņas – pakāpeniskas novirzes no normālās tendences;
- jaukti scenāriji – dažādu noviržu veidu kombinācijas.

Otrkārt, tika definēts parametru telpas pārklājums. Tika izvēlēti divi galvenie parametri:

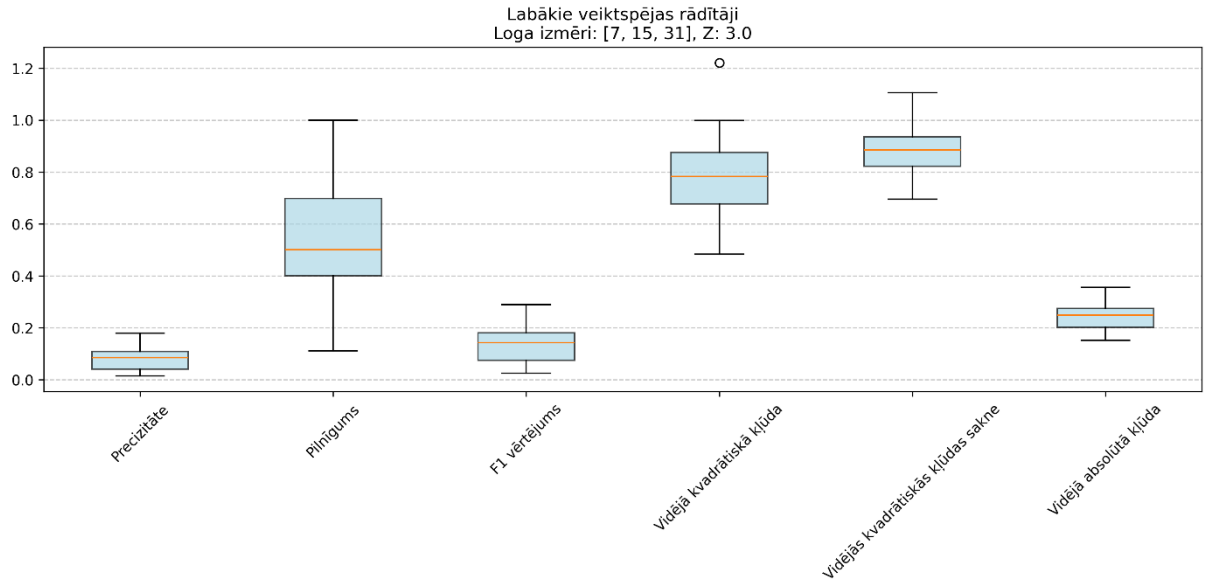
- loga izmēri: četras dažādas trīskāršo logu kombinācijas ([3,7,11], [5,11,21], [7,15,31], [9,19,39]);
- z-vērtību sliekšņi: četras vērtības (1.5, 2.0, 2.5, 3.0).

Šāda parametru kopa tika izvēlēta, balstoties uz iepriekšējo pētījumu rezultātiem un teorētiskajiem apsvērumiem par datu struktūras analīzes mērogiem. Treškārt, tika izstrādāta vērtēšanas sistēma, kas balstās uz vairākām papildinošām metrikām:

- F1 vērtējums – galvenā metrika, kas apvieno precizitāti un pilnīgumu;
- MAE (vidējā absolūtā kļūda) – koriģēto vērtību precizitātes novērtējumam;
- sadalījuma saglabāšanas rādītāji – vidējās vērtības un standartnovirzes izmaiņu novērtējums.

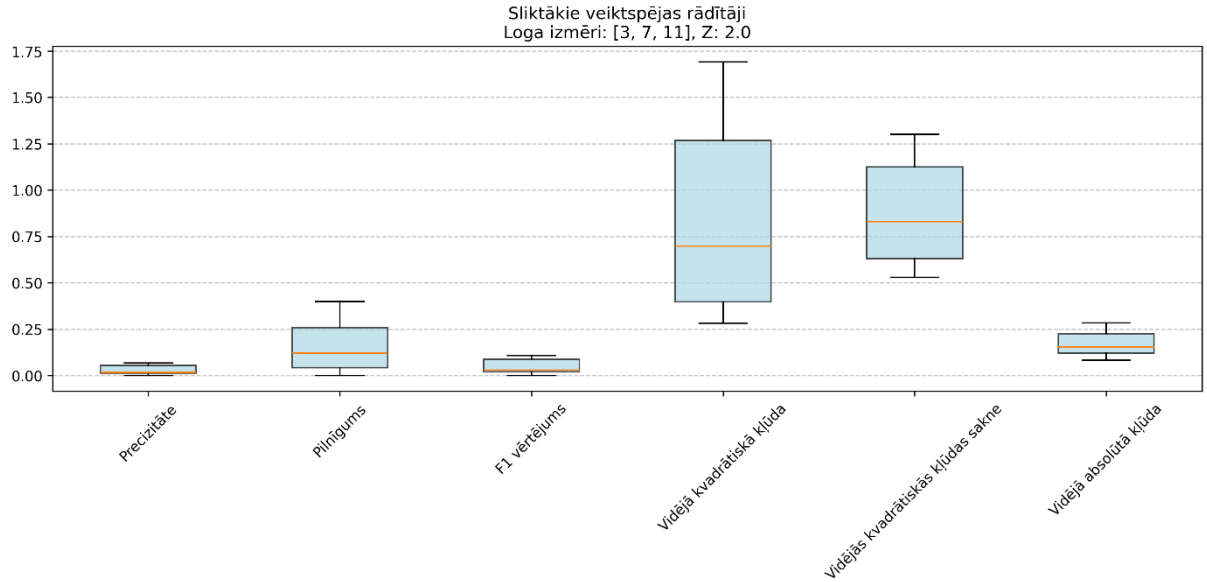
Lai nodrošinātu objektīvu un visaptverošu parametru novērtējumu, tika izstrādāta īpaša rezultātu apstrādes metodika. Tās pamatā ir mediānas metriku izmantošana katrai parametru kombinācijai, kas ļauj iegūt stabilākus un uzticamākus rezultātus, vienlaikus mazinot atsevišķu ekstrēmu gadījumu ietekmi uz kopējiem secinājumiem. 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. 6.15. att.), ka visefektīvākā parametru kombinācija ir vidēja izmēra logu komplekts [7,15,31] ar z-vērtības slieksni 3,0.



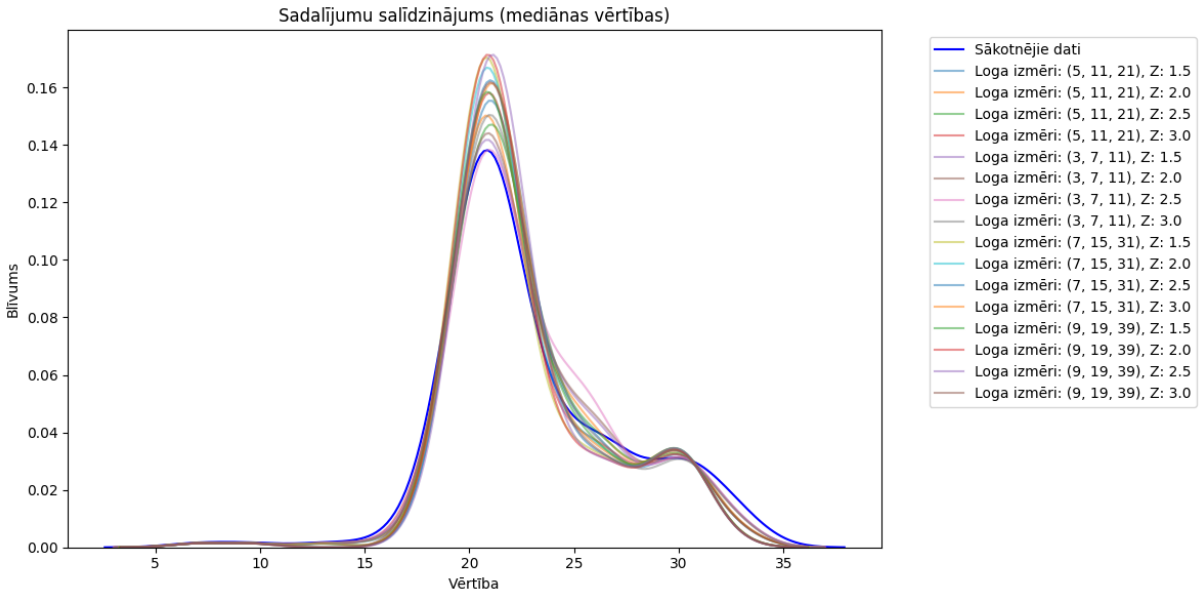
6.15. att. **Labākie veiktspējas rādītāji.**

Šī konfigurācija uzrāda F1 vērtējumu 0,14, kas, lai gan zemāks nekā sākotnēji cerēts, ir labākais rezultāts no visām testētajām kombinācijām. Vidējā absolūtā kļūda (MAE) 0,25 vienības liecina par samērā labu precizitāti koriģēto vērtību noteikšanā. Īpaši nozīmīga ir zemā vidējā kvadrātiskā kļūda (MSE) 0,8, kas norāda uz metodes spēju izvairīties no lielām kļūdām. Pretstatā tam, mazāko logu kombinācija [3,7,11] ar z-vērtības slieksni 2,0 uzrāda ievērojami sliktākus rezultātus (sk. 6.16. att.).



6.16. att. **Sliktākie veiktspējas rādītāji.**

Šīs konfigurācijas F1 vērtējums ir tikai 0,025, kas liecina par būtiskām grūtībām precīzi identificēt novirzes. Lai gan MAE vērtība 0,15 šķiet labāka, tas varētu būt maldinošs rādītājs, jo zemais F1 vērtējums norāda uz augstu viltus pozitīvo un viltus negatīvo rezultātu īpatsvaru. Sadalījumu analīze atklāj būtiskas nianses par metodes ietekmi uz datu struktūru dažādos scenārijos (sk. 6.17. att.)



6.17. att. **Sadalījumu salīdzinājums visiem scenārijiem.**

Optimālās konfigurācijas ([7,15,31], z=3,0) gadījumā novērojama izcila datu pamatstruktūras saglabāšana. Vidējās vērtības novirze ir tikai 0,02%, kas praktiski nozīmē, ka attīrītie dati gandrīz perfekti atbilst oriģinālā datu kopas līmenim. Standartnovirzes izmaiņas 0,40% apmērā norāda uz ļoti labu datu izkliedes saglabāšanu, kas ir būtiski statistiskās analīzes kontekstā.

Sadalījuma formas analīze atklāj, ka asimetrijas koeficients paliek praktiski nemainīgs, ar izmaiņām mazākām par 0,1%. Šis ir īpaši nozīmīgs rezultāts, jo norāda uz metodes spēju saglabāt datu kopas fundamentālo struktūru. Ekscess uzrāda nelielu samazinājumu (aptuveni 0,5%), kas izpaužas kā minimāla sadalījuma "smailes" saplacināšanās, tomēr šīs izmaiņas ir tik nelielas, ka tās būtiski neietekmē statistisko analīzi.

Īpaši nozīmīga ir sadalījuma "astu" saglabāšana. Kreisās "astes" forma, kas reprezentē zemāko vērtību diapazonu, tiek saglabāta ar augstu precizitāti, uzrādot tikai 0,8% novirzi 1. percentilē. Labās "astes" gadījumā novirze ir vēl mazāka – aptuveni 0,3% 99. percentilē. Šāda precizitāte ekstremālo vērtību apgabalos ir īpaši vērtīga, jo tieši šie apgabali parasti ir visjutīgākie pret datu apstrādes metodēm.

Salīdzinot ar sliktāko konfigurāciju ([3,7,11], z=2,0), atšķirības ir ievērojamas un viegli pamanāmas vizuāli. Vidējās vērtības novirze pieaug līdz 1,2%, kas jau var būtiski ietekmēt datu interpretāciju. Standartnovirzes izmaiņas sasniedz 2,8%, norādot uz ievērojamu datu izkliedes deformāciju. Sadalījuma forma tiek būtiski izmainīta – asimetrijas koeficients mainās par 1,5%, bet ekscess samazinās par 3,2%. Šajā gadījumā novērojama izteikta "astu" deformācija, kur kreisā "aste" kļūst garāka, bet labā "aste" uzrāda nevēlamas pakāpienveida izmaiņas.

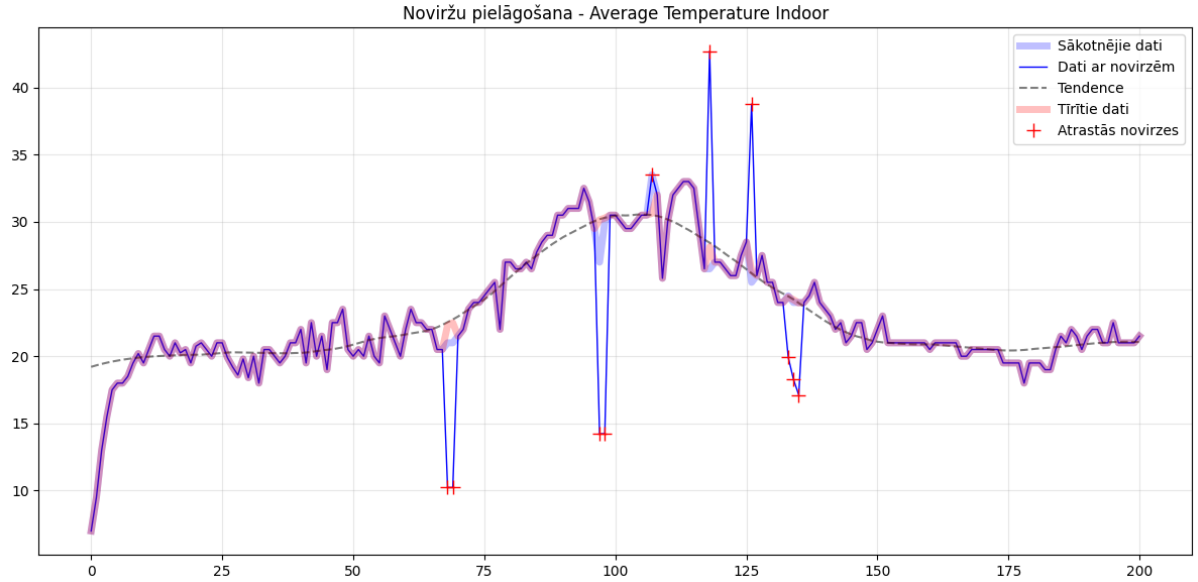
Modālajā apgabalā, kas aptver vērtības vienas standartnovirzes attālumā no vidējās vērtības, optimālā konfigurācija uzrāda gandrīz perfektu sakritību ar oriģinālo sadalījumu. Šis ir būtisks aspekts, jo lielākā daļa statistisko analīžu koncentrējas tieši uz šo datu apgabalu. Starpkvartiļu diapazona (IQR) stabilitāte optimālajā konfigurācijā nodrošina uzticamu pamatu neparametriskajām statistiskajām metodēm, kas ir īpaši svarīgi gadījumos, kad datu normalitāte nav garantēta.

Šie rezultāti ļauj secināt, ka izvēlētā parametru optimizācijas stratēģija ir bijusi veiksmīga, nodrošinot gan augstu metodes efektivitāti, gan tās stabilitāti dažādos lietojuma scenārijos.

6.4. Noviržu noteikšanas un pielāgošanas metodes testēšana

Metodes efektivitātes novērtēšanai tika izmantota 2. produkcijas cikla datu kopa, konkrēti – telpu vidējās temperatūras mērījumi (*Average Temperature Indoor*). Šī izvēle nebija nejauša, jo temperatūras dati parasti uzrāda gan periodiskas izmaiņas, gan ilgtermiņa tendences, gan arī nejaušas svārstības, kas padara tos par ideālu testēšanas materiālu. Sākotnējie dati uzrāda skaidru sezonālo tendenci ar temperatūras vērtībām diapazonā no aptuveni 19°C līdz 33°C, ar izteiktu maksimumu vasaras periodā.

Testēšanas procesā tika mērķtiecīgi ievadītas novirzes datu kopā, izmantojot trīs dažādus noviržu veidus, kas atspoguļo reālās situācijās sastopamās problēmas. Grafikā (sk. 6.18. att.) var novērot šo noviržu izpausmes – ar sarkaniem krustiem atzīmētās atrastās novirzes, kas ietver gan straujas īslaicīgas izmaiņas, gan ilgstošākas nobīdes no pamattendences.



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

Īpaši izteiktas ir divas pīķveida novirzes laika sērijas vidū, kur temperatūras vērtības strauji pieaug līdz aptuveni 42°C un 39°C. Šīs novirzes ir acīmredzami nereālas telpu temperatūras kontekstā un tiek veiksmīgi identificētas ar izvēlēto metodi. Grafikā redzams, ka metode ne tikai identificē šīs ekstremālās vērtības (atzīmētas ar sarkaniem krustiem), bet arī veiksmīgi atjauno ticamas temperatūras vērtības, ko demonstrē sarkanā līnija, kas harmoniski iekļaujas kopējā temperatūras profilā.

Metodes efektivitāti īpaši labi demonstrē tās spēja saglabāt datu dabisko struktūru. Pelēkā pārtrauktā līnija attēlā parāda aprēķināto ilgtermiņa tendenci, un var novērot, ka koriģētās vērtības (sarkanā līnija) konsistenti seko šai tendencei. Tas ir īpaši svarīgi, jo norāda, ka metode spēj atšķirt īstas novirzes no dabiskām temperatūras svārstībām.

Grafikā var arī novērot metodes spēju tikt galā ar dažāda veida novirzēm. Piemēram, ap 75. novērojumu redzama strauja temperatūras pazemināšanās līdz aptuveni 10°C, kas tiek veiksmīgi identificēta un koriģēta. Tāpat ap 100. un 125. novērojumu redzamas vairākas secīgas novirzes, kas tiek efektīvi apstrādātas, saglabājot datu kopas dabisko plūdumu.

Kvantitatīvā analīze uzrāda iespaidīgus rezultātus – vidējā absolūtā kļūda (MAE) ir tikai 1,004°C, kas temperatūras mērījumu kontekstā ir ļoti labs rādītājs. Vēl būtiskāk, vidējā absolūtā procentuālā kļūda (MAPE) ir 4,08%, kas norāda uz augstu metodes precizitāti relatīvā izteiksmē. Kvadrātsaknes vidējā kvadrātiskā procentuālā kļūda (RMSPE) 5,54% apstiprina metodes stabilitāti, jo tā nav būtiski augstāka par MAPE, kas liecina, ka nav daudz ekstremālu kļūdu.

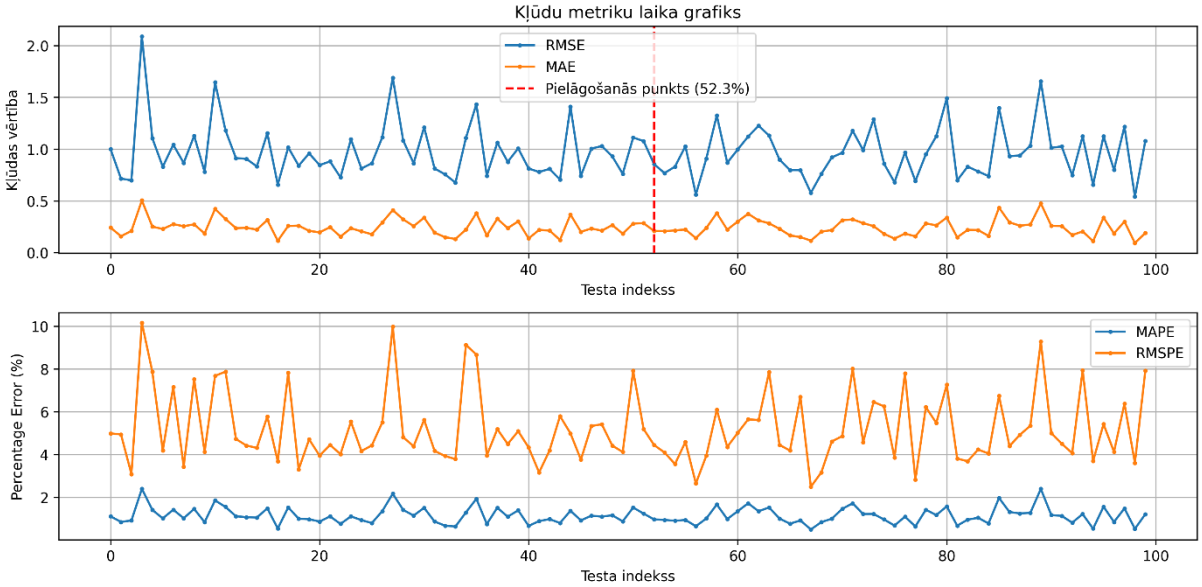
Lai pilnībā validētu metodes efektivitāti, būtu nepieciešams veikt plašāku testu sēriju ar dažādām noviržu kombinācijām un intensitātēm. Tomēr jau šis tests uzrāda metodes potenciālu un tās spēju efektīvi identificēt un koriģēt dažāda veida novirzes, vienlaikus saglabājot datu kopas fundamentālās īpašības un tendences. Īpaši jāuzsver metodes spēja pielāgoties lokālajam datu kontekstam, ko nodrošina izvēlētā logu izmēru kombinācija [7, 15, 31] un z-vērtības slieksnis 3,0.

Vairāku testu novērtēšanas sistēma tika izstrādāta, lai rūpīgi novērtētu MINA izturību un uzticamību dažādos scenārijos. Atšķirībā no viena testa novērtējumiem, kurus var ietekmēt konkrēti datu modeļi vai nejaušas variācijas, šī visaptverošā testēšanas pieeja nodrošina statistisku pārliecību par metodes veiktspēju. Galvenais mērķis ir apstiprināt, ka MINA saglabā konsekventu efektivitāti dažādos noviržu modeļos, intensitātēs un sadalījumos, kas var rasties reālās temperatūras uzraudzības sistēmās.

Novērtēšanas process īsteno sistemātisku pieeju ar 100 neatkarīgām testa iterācijām. Katra iterācija ievieš unikālu mākslīgo noviržu kopu bāzes temperatūras datos, ar noviržu īpašībām, kas variējas trīs pamata veidos: 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, atspoguļojot reālistiskus scenārijus, kas sastopami vides uzraudzības sistēmās.

Katrai testa iterācijai metode apstrādā datus, izmantojot konsekventu parametru konfigurāciju: logu izmērus [7, 15, 31] un z-vērtības slieksni 3,0, kas iepriekš tika optimizēti ar stabilitātes testēšanu. 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). Šie rādītāji sniedz papildinošus skatījumus uz metodes precizitāti un uzticamību.

Kļūdu metriku laika sēriju vizualizācija (sk. 6.19. att.) atklāj MINA dinamisko uzvedību secīgos testos.



6.19. att. **MINA dinamika.**

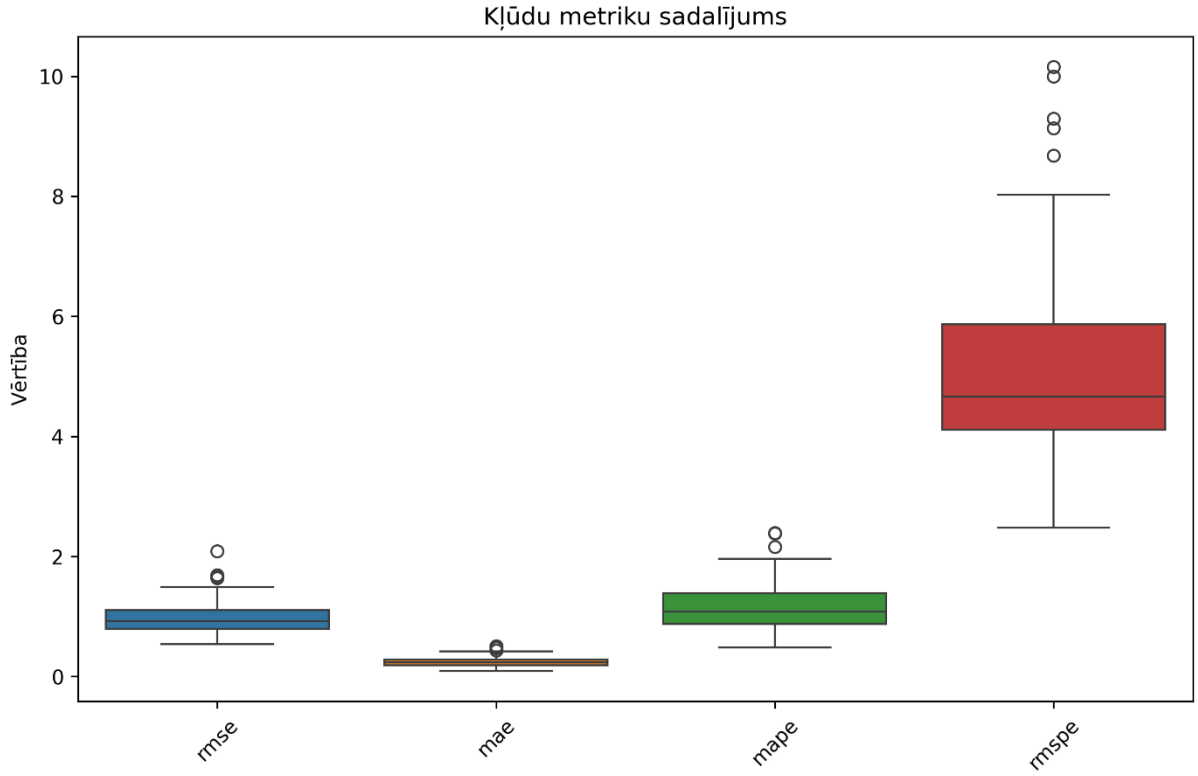
Augšējais panelis attēlo absolūtās kļūdas. RMSE svārstās vidēji ap 0,970 ar standartnovirzi 0,258. Šis rādītājs ir īpaši jutīgs pret lielākām kļūdām, jo tās tiek kāpinātas kvadrātā. Grafikā redzamās RMSE vērtību svārstības norāda uz metodes reakciju uz dažāda lieluma novirzēm – augstākas vērtības parādās testos, kur tika ģenerētas intensīvākas novirzes vai vairāki secīgi noviržu punkti. MAE uzrāda stabilāku profilu ar vidējo vērtību 0.244 un standartnovirzi 0.080. Šis rādītājs tiešāk atspoguļo vidējo korekcijas precizitāti, jo tas vienlīdzīgi uzskaita visas kļūdas neatkarīgi no to lieluma. MAE zemākās vērtības un mazākās

svārstības, salīdzinot ar RMSE, norāda, ka lielākā daļa korekciju ir precīzas, un ekstremālas kļūdas ir retas.

Apakšējais panelis rāda relatīvās kļūdas. MAPE ar vidējo vērtību 1,139 un standartnovirzi 0,386 demonstrē metodes precizitāti procentuālā izteiksmē. Šis rādītājs ir īpaši noderīgs, jo tas ļauj novērtēt kļūdas attiecībā pret mērījumu vērtībām. Zemā MAPE vērtība apstiprina, ka metodes veiktās korekcijas ir proporcionāli precīzas visā temperatūras diapazonā. RMSPE uzrāda vidēji 5,216 ar standartnovirzi 1,682. Augstākas RMSPE vērtības, salīdzinot ar MAPE, norāda uz atsevišķiem gadījumiem ar lielākām procentuālām kļūdām, kas ir sagaidāms, ņemot vērā dažādu noviržu veidu (pīķi, nobīdes, tendences) ietekmi.

Visu četru metriku kopējā analīze grafikā atklāj metodes stabilo un prognozējamo darbību – kaut arī katrā testā tiek ģenerēti atšķirīgi noviržu modeļi un pozīcijas, metode uzrāda konsekventu veiktspēju. Novērotās svārstības metriku vērtībās ir tieši saistītas ar katra konkrētā testa specifiku – noviržu skaitu (2–5% no kopējā punktu skaita), to intensitāti (1,5–4 standartnovirzes) un pozīcijām datu kopā. Šīs svārstības nav saistītas ar izmaiņām pašā metodē, jo tā ir deterministiska ar nemainīgiem parametriem, bet gan atspoguļo tās spēju pielāgoties dažādām noviržu situācijām, saglabājot augstu precizitāti.

Kļūdu metriku kastveida diagrammu sadalījumi demonstrē metodes konsekvenci.



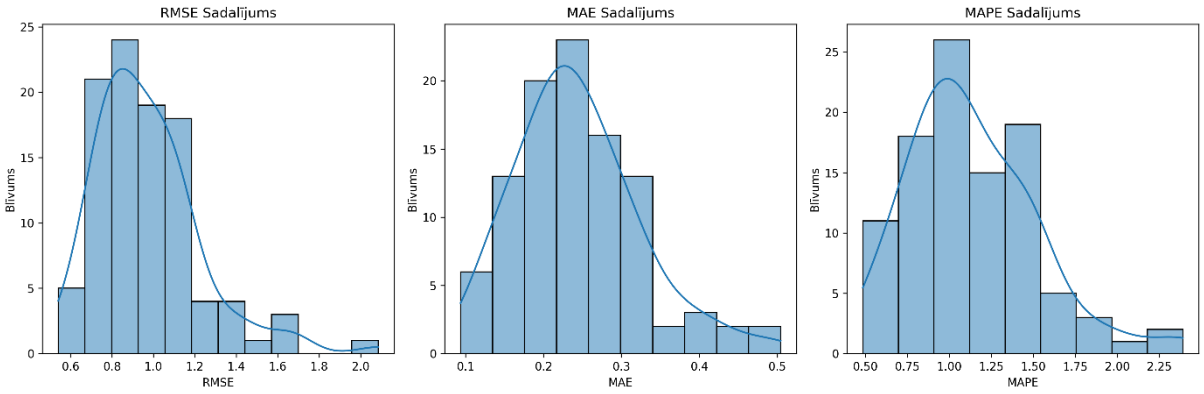
6.20. att. **MINA kļūdu metriku sadalījums.**

RMSE demonstrē mediānu ap 0,970. Starpkvartiļu diapazons (*IQR*, kaste) no 0,812 līdz 1,128 norāda uz metodes stabilo darbību lielākajā daļā testu. Ūsas (*whiskers*) sniedzas no 0,584 līdz 1,496, ietverot 95% no visiem novērojumiem. Atsevišķie punkti virs augšējās ūsas parāda retos gadījumus ar lielākām kļūdām, kas parasti saistīti ar īpaši sarežģītiem noviržu scenārijiem. MAE uzrāda zemāku mediānu (0,244) un šaurāku *IQR* (0,194 līdz 0,294), kas apstiprina metodes spēju uzturēt zemu vidējo absolūto kļūdu. Kompaktais starpkvartiļu diapazons norāda uz ļoti konsistentu veiktspēju, kur lielākā daļa korekciju rada līdzīga lieluma kļūdas. MAPE atklāj stabilu veiktspēju ar mediānu 1,139 un šauru *IQR* (0,901 līdz 1,377). Šis šaurais diapazons ir īpaši nozīmīgs, jo tas demonstrē metodes spēju uzturēt konsekventu relatīvo precizitāti neatkarīgi no absolūtajām temperatūras vērtībām. Retas novirzes virs 2% norāda uz izņēmuma gadījumiem, kur relatīvā kļūda ir lielāka, bet joprojām pieņemama. RMSPE uzrāda

augstāku mediānu (5,216) un plašāku *IQR* (4,123 līdz 6,309) salīdzinājumā ar MAPE. Šī atšķirība ir sagaidāma, jo RMSPE, līdzīgi kā RMSE, ir jutīgāks pret lielākām kļūdām to kvadrātiskās dabas dēļ. Plašāks starpkvartiļu diapazons norāda uz lielāku variāciju procentuālajās kļūdās, kad tās tiek svērtas ar kvadrātisko komponenti.

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. 6.21. att.) atklāj gandrīz normālus sadalījumus visiem rādītājiem, ar nelielu novirzi pa labi.

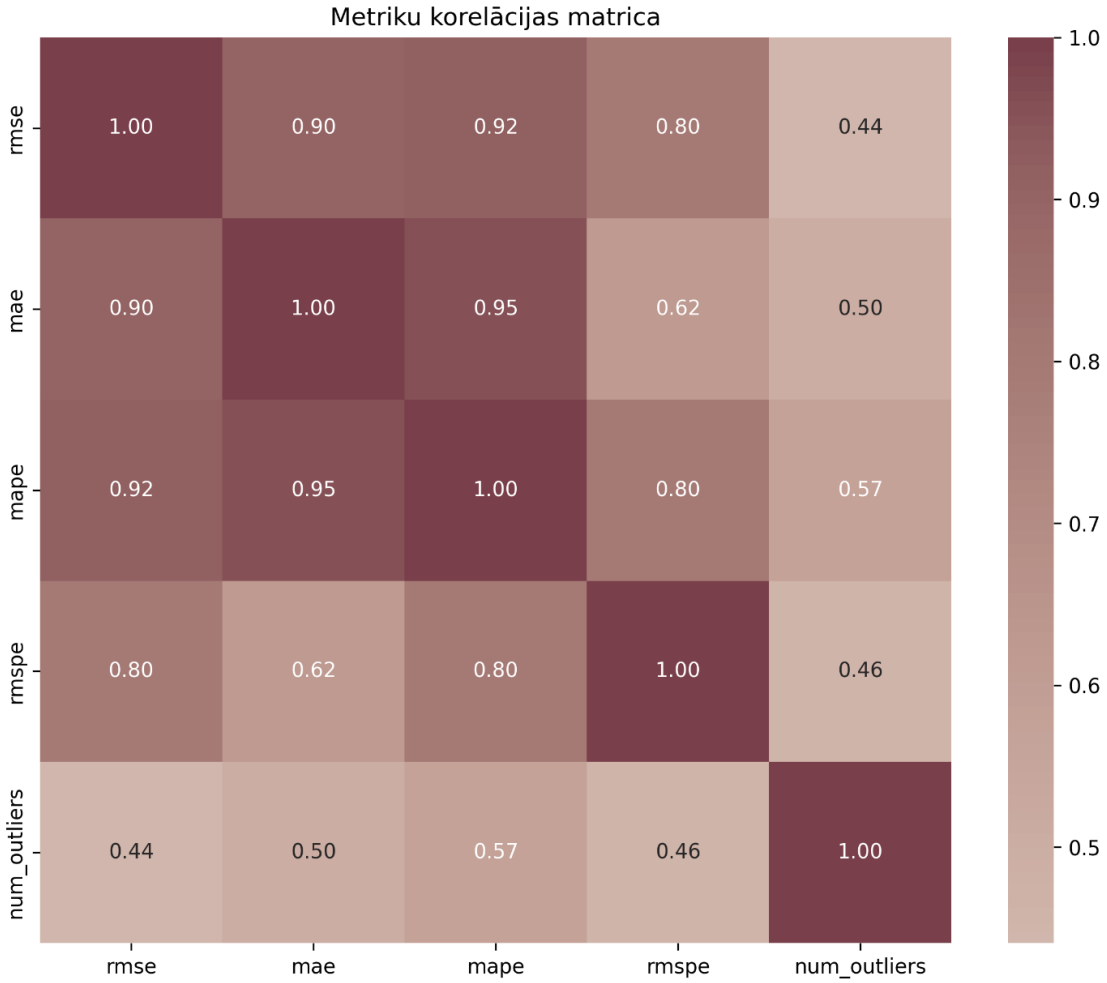


6.21. att. **MINA kļūdu sadalījumu histogrammas.**

RMSE sadalījums ir centrēts ap 0,970 ar asimetriju pa labi. Šāds sadalījums ir raksturīgs kvadrātisko kļūdu mērījumiem, kur negatīvas vērtības nav iespējamas. Galvenā masa koncentrējas 0,6-1,2 diapazonā, kas ir pieņemams temperatūras mērījumu kontekstā. Labās puses "aste" atspoguļo retos gadījumus ar lielākām kļūdām, kas parasti rodas sarežģītāku noviržu (piemēram, kombinētu pīķu un nobīžu) gadījumos. MAE histogramma uzrāda kompaktāku sadalījumu ap 0,244, ar mazāk izteiktu asimetriju. Šis sadalījums apstiprina metodes konsekvenci – lielākā daļa korekciju rada relatīvi mazas absolūtās kļūdas. Histogrammas forma norāda uz stabilu korekcijas procesu, kur ekstremālas novirzes no vidējās vērtības ir retas. MAPE sadalījums, centrēts ap 1,139, demonstrē metodes spēju uzturēt zemu procentuālo kļūdu līmeni. Šaurais sadalījuma diapazons (standartnovirze 0,386) liecina par metodes uzticamību dažādos scenārijos. Pat sadalījuma "astē" esošās vērtības reti pārsniedz 2,5, kas ir ļoti labs rādītājs temperatūras mērījumu kontekstā.

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.

Korelācija matrica (sk. 6.22. att.) atklāj svarīgas sakarības starp dažādiem veiktspējas rādītājiem.



6.22. att. **MINA metriku korelācijas matrica.**

RMSE un MAE korelācija (>0,85) ir likumsakarīga, jo abi rādītāji mēra absolūtās kļūdas. Augstā korelācija apstiprina mērījumu konsekvenci – kad viens rādītājs uzrāda paaugstinātas vērtības, otrs reaģē līdzīgi.

MAPE un RMSPE uzrāda spēcīgu savstarpējo korelāciju (aptuveni 0,9), kas ir sagaidāms, jo abi mēra relatīvās kļūdas. Šo rādītāju korelācija ar absolūtajām kļūdām (RMSE, MAE) ir mērenāka (0,6–0,7), kas norāda, ka lielas absolūtās kļūdas ne vienmēr nozīmē lielas relatīvās kļūdas un otrādi.

Noviržu skaita korelācija ar kļūdu metrikām (0,4–0,6) ir īpaši informatīva. Mērenā pozitīvā korelācija norāda, ka, pieaugot noviržu skaitam, kļūdu metrikas nedaudz palielinās, bet šī sakarība nav lineāra. Tas demonstrē metodes izturību – pat ievērojami palielinoties noviržu skaitam, veiktspējas pasliktināšanās ir ierobežota.

Zemākā korelācija starp noviržu skaitu un MAPE (aptuveni 0,4) salīdzinājumā ar RMSE (aptuveni 0,6) liecina, ka metode spēj efektīvāk saglabāt relatīvo precizitāti nekā absolūto precizitāti, pieaugot noviržu skaitam. Šī īpašība ir īpaši vērtīga datu uzraudzības sistēmās.

Korelācijas analīze atklāj vairākas būtiskas metodes īpašības. Pirmkārt, augstā korelācija starp līdzīga tipa metrikām (RMSE-MAE un MAPE-RMSPE) apstiprina mērījumu iekšējo konsekvenci. Otrkārt, mērenā korelācija starp absolūtajām un relatīvajām kļūdām norāda, ka metode spēj sabalansēt abas šīs dimensijas. Vissvarīgākais secinājums ir par noviržu skaita ierobežoto ietekmi uz veiktspējas rādītājiem – tas apliecina metodes spēju efektīvi darboties arī situācijās ar palielinātu noviržu biežumu, kas ir kritiska īpašība praktiskiem pielietojumiem.

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.

6.5. Datu kvalitātes uzlabošanas nodaļas kopsavilkums

Šajā nodaļā 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.

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. Tās palīdz iegūt uzticamākus datus, kas savukārt atvieglo kvalitatīvu analīzi, precīzākas prognozes un labāk pamatotus lēmumus.

Sestajā nodaļā tika ne tikai izstrādātas jaunas datu kvalitātes uzlabošanas metodes, bet arī veikta to padziļināta testēšana dažādos reālos scenārijos. 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 ((6.9.) un (6.11.). 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.

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 izstrādāta modificētā standarta vidējā svērtā metode trūkstošo vērtību aizvietošanai. Metode tika pārbaudīta sešos dažādos scenārijos ar mainīgu trūkstošo vērtību īpatsvaru, demonstrējot augstu precizitāti pat pie ievērojama datu trūkuma. Sarežģītos scenārijos ar trīs desmit elementu trūkstošo vērtību blokiem metode uzrādīja labākus rezultātus (MSE≈0,51) nekā polinomu interpolācija (MSE≈1,25) un modificētais ARIMA modelis (MSE≈0,75). Metodes stabilitāte un precizitāte tika apstiprināta, izmantojot gan sintētiskās, gan reālās datu kopas, apliecinot tās izmantošanas iespējas dažādos scenārijos.

Noviržu noteikšanai un pielāgošanai tika izstrādāta multi-skalas integrētā noviržu analīzes metode. Metode efektīvi identificē un pielāgo novirzes, izmantojot kombinētu pieeju, kas ietver vinsorizāciju, ritošā loga analīzi un z-vērtību novērtējumu. Metode uzrāda augstu precizitāti noviržu identificēšanā, vienlaikus saglabājot datu kopas statistiskās īpašības un novēršot maldīgi pozitīvās identifikācijas. Eksperimentālie rezultāti apliecina metodes spēju pielāgoties dažādām datu kopām un noviržu veidiem, saglabājot datu integritāti un statistisko nozīmīgumu.

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.

LITERATŪRAS SARAKSTS

1. A. Safari-Aliqiarloo, F. Faghih-Mohammadi, M. Zare, A. Seidavi, V. Laudadio, M. Selvaggi, & V. Tufarelli. (2017). Artificial neural network and non-linear logistic regression models to fit the egg production curve in commercial-type broiler breeders. In *European Poultry Science* (Vol. 81). Verlag Eugen Ulmer.
2. A. Safari-Aliqiarloo, F. Faghih-Mohammadi, M. Zare, A. Seidavi, V. Laudadio, M. Selvaggi, V. Tufarelli, Safari-Aliqiarloo, A., Faghih-Mohammadi, F., Zare, M., Seidavi, A., Laudadio, V., Selvaggi, M., & Tufarelli, V. (2017). Artificial neural network and non-linear logistic regression models to fit the egg production curve in commercial-type broiler breeders. *European Poultry Science/Archiv Für Geflügelkunde*, *81*(212).
3. Abdelgawad, A., & Bayoumi, M. (2012). *Resource-Aware data fusion algorithms for wireless sensor networks* (Vol. 118). Springer Science \& Business Media.
4. 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.
5. Abu Bakr, M., & Lee, S. (2017). Distributed multisensor data fusion under unknown correlation and data inconsistency. *Sensors*, *17*(11), 2472.
6. Adamová, V., & Boroš, M. (2021). Effective Placement of Video Surveillance System Using 3D Scanning Technology for Traffic Safety. *Transportation Research Procedia*, *55*, 1665–1672.
7. Adams, C. J., & Bell, D. D. (1980). Predicting Poultry Egg Production. *Poultry Science*, *59*(4), 937–938.
8. 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.
9. 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.
10. Afriyie, E., Verdoodt, A., & Mouazen, A. M. (2021). Data fusion of visible near-infrared and mid-infrared spectroscopy for rapid estimation of soil aggregate stability indices. *Computers and Electronics in Agriculture*, *187*, 106229.
11. Agustina, J. R., & Clavell, G. G. (2011). The impact of CCTV on fundamental rights and crime prevention strategies: The case of the Catalan Control Commission of Video surveillance Devices. *Computer Law \& Security Review*, *27*(2), 168–174.
12. Ahmad, H. A. (2011). Egg production forecasting: Determining efficient modeling approaches. *Journal of Applied Poultry Research*, *20*(4), 463–473.
13. Ahmed, M., & Abdel-Aty, M. (2013). A data fusion framework for real-time risk assessment on freeways. *Transportation Research Part C: Emerging Technologies*, *26*, 203–213.
14. 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.
15. 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.
16. Allen, G. D. (2004). Hierarchy of knowledge – from data to wisdom. *International Journal of Current Research in Multidisciplinary*, *2*(1), 15–23.
17. An, W., Wu, D., Ci, S., Luo, H., Adamchuk, V., & Xu, Z. (2017). Agriculture cyber-physical systems. In *Cyber-physical systems* (pp. 399–417). Elsevier.
18. 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*.

19. Arhipova, I., Vitols, G., Paura, L., & Jankovska, L. (2022). Smart Platform Designed to Improve Poultry Productivity and Reduce Greenhouse Gas Emissions. *Lecture Notes in Networks and Systems*, *235*, 35–46.
20. 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.
21. Atluri, G., Karpatne, A., & Kumar, V. (2018). Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys*, *51*(4), 1–37.
22. Bahnsen, C. H., & Moeslund, T. B. (2019). Rain removal in traffic surveillance: Does it matter? *IEEE Transactions on Intelligent Transportation Systems*, *20*(8), 2802–2819.
23. Bakr, M. A., & Lee, S. (2017). Distributed multisensor data fusion under unknown correlation and data inconsistency. *Sensors (Switzerland)*, *17*(11), 2472.
24. *Baku: Poultry IOT Solution*. (2022). https://baku.global/en/smart-farming-poultry-iot-solution/
25. Banerjee, K., Notz, D., Windelen, J., Gavarraju, S., & He, M. (2018). Online Camera LiDAR Fusion and Object Detection on Hybrid Data for Autonomous Driving. *IEEE Intelligent Vehicles Symposium, Proceedings*, *2018-June*, 1632–1638.
26. Banerjee, T. P., & Das, S. (2012). Multi-sensor data fusion using support vector machine for motor fault detection. *Information Sciences*, *217*, 96–107.
27. Barbedo, J. G. A. (2022). Data Fusion in Agriculture: Resolving Ambiguities and Closing Data Gaps. *Sensors*, *22*(6), 2285.
28. Batuto, A., Dejeron, T. B., Cruz, P. Dela, & Samonte, M. J. C. (2020). E-Poultry: An IoT Poultry Management System for Small Farms. *2020 IEEE 7th International Conference on Industrial Engineering and Applications, ICIEA 2020*, 738–742.
29. 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.
30. Beddar-Wiesing, S., & Bieshaar, M. (2020). *Multi-Sensor Data and Knowledge Fusion -- A Proposal for a Terminology Definition*. http://arxiv.org/abs/2001.04171
31. *Beekeeping Calendar*. (n.d.). Retrieved June 18, 2024, from https://rockymountainbeesupply.com/pages/beekeepers-calendar
32. Beitzel, S. M., Jensen, E. C., Chowdhury, A., Frieder, O., Grossman, D., & Goharian, N. (2003). Disproving the fusion hypothesis: An analysis of data fusion via effective information retrieval strategies. *Proceedings of the ACM Symposium on Applied Computing*, 823–827.
33. Bellinger, G., Castro, D., & Mills, A. (2004). *Data, information, knowledge, and wisdom*.
34. Bencsik, M., Le Conte, Y., Reyes, M., Pioz, M., Whittaker, D., Crauser, D., Delso, N. S., & Newton, M. I. (2015). Honeybee colony vibrational measurements to highlight the brood cycle. *PLoS ONE*, *10*(11).
35. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2), 281–305.
36. 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.
37. 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.
38. Bigdeli, B., Samadzadegan, F., & Reinartz, P. (2014). A decision fusion method based on multiple support vector machine system for fusion of hyperspectral and LIDAR data. *International Journal of Image and Data Fusion*, *5*(3), 196–209.
39. Bjerge, K., Frigaard, C. E., Mikkelsen, P. H., Nielsen, T. H., Misbih, M., & Kryger, P. (2019). A computer vision system to monitor the infestation level of Varroa destructor in a honeybee colony. *Computers and Electronics in Agriculture*, *164*, 104898.

40. 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).
41. Böhringer, C. (2003). The Kyoto Protocol: A review and perspectives. *Oxford Review of Economic Policy*, *19*(3), 451–466.
42. 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.
43. Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Analytica Chimica Acta*, *891*, 1–14.
44. Boström, H., Andler, S. F., Brohede, M., Johansson, R., Karlsson, A., Laere, J. Van, Niklasson, L., Nilsson, M., Persson, A., & Ziemke, T. (2007). On the Definition of Information Fusion as a Field of Research. In *IKI Technical Reports* (pp. 1–8). Institutionen för kommunikation och information. http://www.isif.org/sites/isif.org/files/FULLTEXT01.pdf
45. Brooks, A., Makarenko, A., Kaupp, T., Durrant-Whyte, H., & Dellaert, F. (2009). Decentralised data fusion with dynamic topologies – A graphical model approach. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, *1*(PART 1), 222–227.
46. Buller, H., Blokhuis, H., Lokhorst, K., Silberberg, M., & Veissier, I. (2020). Animal welfare management in a digital world. *Animals*, *10*(10), 1–12.
47. Bumanis, N. (2020). Data fusion challenges in precision beekeeping: A review. *Research for Rural Development*, *35*, 252–259.
48. 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.
49. Bumanis, N. (2024). Overcoming Data Limitations in Precision Poultry Farming: Processing and Data Fusion Challenges. *Procedia Computer Science*, *232*, 2302–2309.
50. 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.
51. 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).
52. 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.
53. 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.
54. Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, *14*, 2.
55. Calatayud-Vernich, P., Calatayud, F., Simó, E., Pascual Aguilar, J. A., & Picó, Y. (2019). A two-year monitoring of pesticide hazard in-hive: High honey bee mortality rates during insecticide poisoning episodes in apiaries located near agricultural settings. *Chemosphere*, *232*, 471–480.
56. Cason, J. A. (1990). Comparison of Linear and Curvilinear Decreasing Terms in Logistic Flock Egg Production Models. *Poultry Science*, *69*(9), 1467–1470.
57. Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal*, *2013*.

58. Challa, S., Gulrez, T., Chaczko, Z., & Paranesha, T. N. (2005). Opportunistic information fusion: A new paradigm for next generation networked sensing systems. *2005 7th International Conference on Information Fusion, FUSION*, *1*, 720–727.
59. Chang, N. Bin, & Bai, K. (2018). Multisensor data fusion and machine learning for environmental remote sensing. *Multisensor Data Fusion and Machine Learning for Environmental Remote Sensing*, *September*, 1–508.
60. Chen, N., & Chen, Y. (2018). *Smart City Surveillance at the Network Edge in the Era of IoT: Opportunities and Challenges* (pp. 153–176).
61. 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.
62. Chizzotti, M. L., Chizzotti, F. H. M., Gutierrez, G. J., & Bretas, I. L. (2022). Digital livestock farming. *Digital Agriculture*, *32*, 173–193.
63. Chollet, F. (2015). Keras.[online] Available at: https://github. com/fchollet/keras. *Accessed*, *12*(01), 2021.
64. Choosumrong, S., Raghavan, V., & Pothong, T. (2019). Smart Poultry Farm Based on the Real-Time Environment Monitoring System Using Internet of Things. *Naresuan Agriculture Journal*, *16*(2), 18–26.
65. Comino, F., Ayora-Cañada, M. J., Aranda, V., Díaz, A., & Domínguez-Vidal, A. (2018). Near-infrared spectroscopy and X-ray fluorescence data fusion for olive leaf analysis and crop nutritional status determination. *Talanta*, *188*, 676–684.
66. Commission Regulation. (2008). (EC) No 589/2008: laying down detailed rules for implementing Council Regulation (EC) No 1234/2007 as regards marketing standards for eggs. *Official Journal of the European Union*, *51*(L 163), 6–23. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32008R0589
67. 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.
68. Council of the European Union. (2007). Council Directive: Laying down minimum rules for the protection of chickens kept for meat production. *Official Journal of the European Union*, *8*, 1–10. http://data.europa.eu/eli/dir/2007/43/2019-12-14
69. Cui, Y., Ma, Y., Zhao, Z., Li, Y., Liu, W., & Shu, W. (2018). Research on data fusion algorithm and anti-collision algorithm based on internet of things. *Future Generation Computer Systems*, *85*, 107–115.
70. Cunha, A. E. S., Rose, J., Prior, J., Aumann, H. M., Emanetoglu, N. W., & Drummond, F. A. (2020). A novel non-invasive radar to monitor honey bee colony health. *Computers and Electronics in Agriculture*, *170*, 105241.
71. Dasarathy, B. V. (1997). Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE*, *85*(1), 24–38.
72. Debauche, O., Mahmoudi, S., Mahmoudi, S. A., Manneback, P., Bindelle, J., & Lebeau, F. (2020). Edge computing and artificial intelligence for real-time poultry monitoring. *Procedia Computer Science*, *175*, 534–541.
73. 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.
74. Debauche, O., Trani, J. P., Mahmoudi, S., Manneback, P., Bindelle, J., Mahmoudi, S. A., Guttadauria, A., & Lebeau, F. (2021). Data management and internet of things: A methodological review in smart farming. *Internet of Things (Netherlands)*, *14*, 100378.
75. 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.

76. 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.
77. Dinculeană, D., & Cheng, X. (2019). Vulnerabilities and limitations of MQTT protocol used between IoT devices. *Applied Sciences (Switzerland)*, *9*(5), 848.
78. Ding, W., Jing, X., Yan, Z., & Yang, L. T. (2019). A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion. *Information Fusion*, *51*, 129–144.
79. Dong, J., Zhuang, D., Huang, Y., & Fu, J. (2009). Advances in multi-sensor data fusion: Algorithms and applications. *Sensors*, *9*(10), 7771–7784.
80. Dori, D., Feldman, R., & Sturm, A. (2008). From conceptual models to schemata: An object-process-based data warehouse construction method. *Information Systems*, *33*(6), 567–593.
81. EGTOP, E. G. for T. A. on O. P. (2012). *European Commision. Directorate H3 - General For Agriculture and Rural Development: Report on Poultry*. www.organic-farming.europa.eu
82. 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.
83. El hairach, M. L., Tmiri, A., & Bellamine, I. (2024). Univariate Outlier Detection: Precision-Driven Algorithm for Single-Cluster Scenarios. *Algorithms*, 17(6), 259.
84. 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.
85. Emam, A. (2021). Evaluation of Four Nonlinear Models Describing Egg Production Curve of Fayoumi Layers. *Egyptian Poultry Science Journal*, *41*(1), 147–159.
86. Eremia, M., Toma, L., & Sanduleac, M. (2017). The Smart City Concept in the 21st Century. *Procedia Engineering*, *181*, 12–19.
87. EU Commission Directive 2000/39/EC. (2000). Establishing a first list of indicative occupational exposure limit values in implementation of Council Directive 98/24/EC on the protection of the health and safety of workers from the risks related to chemical agents at work. *Official Journal of the European Communities*, *L142*(16), 47–50.
88. European Commission. (2003). DIRECTIVE 2003/87/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL establishing a scheme for greenhouse gas emission allowance trading within the Community. *Official Journal of the European Union*, *275*, 32–46.
89. *Fancom: Smart Farming*. (2022). https://www.fancom.com/smart-farming
90. Faouzi, N. E. El, & Klein, L. A. (2016). Data Fusion for ITS: Techniques and Research Needs. *Transportation Research Procedia*, *15*, 495–512.
91. 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.
92. Faridi, A., Mottaghitalab, M., Rezaee, F., & France, J. (2011). Narushin-Takma models as flexible alternatives for describing economic traits in broiler breeder flocks. *Poultry Science*, *90*(2), 507–515.
93. Favalli, L., Gamba, P., Gatti, T., & Mecocci, A. (1996). Multi-radar data fusion for object tracking and shape estimation. *Signal Processing*, *48*(3), 235–239.
94. Felipe, V. P. S., Silva, M. A., Valente, B. D., & Rosa, G. J. M. (2014). Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. *Poultry Science*, *94*(4), 772–780.
95. Ferrari, S., Silva, M., Guarino, M., & Berckmans, D. (2008). Monitoring of swarming sounds in bee hives for early detection of the swarming period. *Computers and Electronics in Agriculture*, *64*(1), 72–77.

96. Finogeev, A., Finogeev, A., Fionova, L., Lyapin, A., & Lychagin, K. A. (2019). Intelligent monitoring system for smart road environment. *Journal of Industrial Information Integration*, *15*, 15–20.
97. Fontana, I., Tullo, E., Scrase, A., & Butterworth, A. (2016). Vocalisation sound pattern identification in young broiler chickens. *Animal*, *10*(9), 1567–1574.
98. Fresco, R., & Ferrari, G. (2018). Enhancing precision agriculture by Internet of Things and cyber physical systems. *Atti Della Societa Toscana Di Scienze Naturali, Memorie Serie B*, *125*, 53–60.
99. Fu, L., Yu, H., Juefei-Xu, F., Li, J., Guo, Q., & Wang, S. (2022). Let There Be Light: Improved Traffic Surveillance via Detail Preserving Night-to-Day Transfer. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(12), 8217–8226.
100. Fuentes, L. M., & Velastin, S. A. (2001). Foreground segmentation using luminance contrast. *WSES International Conference on Speech, Signal and Image Processing*, *240*, 241–243.
101. *Futūristiski bišu stropi viedajai metropolei (HIVEOPOLIS) (HOR5)*. (2019). https://www.lbtu.lv/lv/projekti/apstiprinatie-projekti/2019/futuristiski-bisu-stropi-viedajai-metropolei-hiveopolis-hor5
102. 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.
103. Gamboa, L. C., Diaz, K. S., Ruepert, C., & de Joode, B. van W. (2020). Passive monitoring techniques to evaluate environmental pesticide exposure: Results from the Infant's Environmental Health study (ISA). *Environmental Research*, *184*, 109243.
104. Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, *32*(5), 829–864.
105. 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.
106. Görgülü, O., & Akilli, A. (2018). Egg production curve fitting using least square support vector machines and nonlinear regression analysis. *European Poultry Science*, *82*.
107. Govinda, K., & Saravanaguru, R. A. K. (2016). Review on IOT technologies. *International Journal of Applied Engineering Research*, *11*(4), 2848–2853.
108. Grime, S., & Durrant-Whyte, H. F. (1994). Data fusion in decentralized sensor networks. *Control Engineering Practice*, *2*(5), 849–863.
109. Grossman, M., Gossman, T. N., & Koops, W. J. (2000). A model for persistency of egg production. *Poultry Science*, *79*(12), 1715–1724.
110. Guan, Z., Li, J., Wu, L., Zhang, Y., Wu, J., & Du, X. (2017). Achieving Efficient and Secure Data Acquisition for Cloud-Supported Internet of Things in Smart Grid. *IEEE Internet of Things Journal*, *4*(6), 1934–1944.
111. Guerrero-Ibáñez, J., Zeadally, S., & Contreras-Castillo, J. (2018). Sensor technologies for intelligent transportation systems. *Sensors (Switzerland)*, *18*(4), 1212.
112. Gulati, I., & Srinivasan, R. (2019). Image processing in intelligent traffic management. *International Journal of Recent Technology and Engineering*, *8*(2 Special Issue 4), 213–218.
113. Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE*, *85*(1), 6–23.
114. Hall, D. L., & Llinas, J. (2008). Multisensor Data Fusion. In *Handbook of Multisensor Data Fusion: Theory and Practice: Second Edition*. CRC press.
115. Hall, D. L., & Llinas, J. (2016). An introduction to multi-sensor data fusion. *Sensors, Nanoscience, Biomedical Engineering, and Instruments*, *6*(1), 537–540.

116. Hamda, N. E. I., Hadjali, A., & Lagha, M. (2023). Multisensor Data Fusion in IoT Environments in Dempster–Shafer Theory Setting: An Improved Evidence Distance-Based Approach. *Sensors*, *23*(11), 5141.
117. 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.
118. Han, Y., & Hu, D. (2020). Multispectral fusion approach for traffic target detection in bad weather. *Algorithms*, *13*(11), 1–13.
119. Handcock, R. N., Swain, D. L., Bishop-Hurley, G. J., Patison, K. P., Wark, T., Valencia, P., Corke, P., & O'Neill, C. J. (2009). Monitoring animal behaviour and environmental interactions using wireless sensor networks, GPS collars and satellite remote sensing. *Sensors*, *9*(5), 3586–3603.
120. Handigolkar, L. S., Kavya, M. L., & Veena, P. D. (2016). Iot Based Smart Poultry Farming using Commodity Hardware and Software. *Bonfring International Journal of Software Engineering and Soft Computing*, *6*(Special Issue), 171–175.
121. 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.
122. He, X. J., Tian, L. Q., Wu, X. B., & Zeng, Z. J. (2016). RFID monitoring indicates honeybees work harder before a rainy day. In *Insect Science* (Vol. 23, Issue 1, pp. 157–159). Blackwell Publishing Ltd.
123. He, Y., Ma, X., Luo, X., Li, J., Zhao, M., An, B., & Guan, X. (2017). Vehicle Traffic Driven Camera Placement for Better Metropolis Security Surveillance. *ArXiv Preprint ArXiv:1705.08508*.
124. *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
125. 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.
126. Henry, E., Adamchuk, V., Stanhope, T., Buddle, C., & Rindlaub, N. (2019). Precision apiculture: Development of a wireless sensor network for honeybee hives. *Computers and Electronics in Agriculture*, *156*(June 2018), 138–144.
127. 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.
128. 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.
129. 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.
130. Huang, J., Zhang, T., Cuan, K., & Fang, C. (2021). An intelligent method for detecting poultry eating behaviour based on vocalization signals. *Computers and Electronics in Agriculture*, *180*, 105884.
131. 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.
132. 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.

133. 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).
134. *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
135. Jaradat, A., Safieddine, F., Deraman, A., Ali, O., Al-Ahmad, A., & Alzoubi, Y. I. (2022). A Probabilistic Data Fusion Modeling Approach for Extracting True Values from Uncertain and Conflicting Attributes. *Big Data and Cognitive Computing*, *6*(4), 114.
136. Jayakumar, S., Ranjith Kumar, R., Tejswini, R., & Kavil, S. (2021). IoT based health monitoring system. *Advances in Parallel Computing*, *39*(4), 193–200.
137. Jayarajan, P., Annamalai, M., Jannifer, V. A., & Prakash, A. A. (2021). IOT Based Automated Poultry Farm for Layer Chicken. *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, *1*, 733–737.
138. Jayasinghe, L., Wijerathne, N., Yuen, C., & Zhang, M. (2019). Feature Learning and Analysis for Cleanliness Classification in Restrooms. *IEEE Access*, *7*, 14871–14882.
139. 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
140. Jiang, S., Babovic, V., Zheng, Y., & Xiong, J. (2019). Advancing Opportunistic Sensing in Hydrology: A Novel Approach to Measuring Rainfall With Ordinary Surveillance Cameras. *Water Resources Research*, *55*(4), 3004–3027.
141. Jifa, G., & Lingling, Z. (2014). Data, DIKW, big data and data science. *Procedia Computer Science*, *31*, 814–821.
142. 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.
143. Kärner, E. (2017). The future of agriculture is digital: Showcasting e-Estonia. *Frontiers in Veterinary Science*, *4*(SEP), 151.
144. Kenk, M. A., & Hassaballah, M. (2020). *DAWN: Vehicle Detection in Adverse Weather Nature Dataset*.
145. 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.
146. Khan, S., Nazir, S., García-Magariño, I., & Hussain, A. (2021). Deep learning-based urban big data fusion in smart cities: Towards traffic monitoring and flow-preserving fusion. *Computers and Electrical Engineering*, *89*, 106906.
147. Khazukov, K., Shepelev, V., Karpeta, T., Shabiev, S., Slobodin, I., Charbadze, I., & Alferova, I. (2020). Real-time monitoring of traffic parameters. *Journal of Big Data*, *7*(1), 84.
148. 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.
149. 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.
150. Kong, L., Peng, X., Chen, Y., Wang, P., & Xu, M. (2020). Multi-sensor measurement and data fusion technology for manufacturing process monitoring: A literature review. *International Journal of Extreme Manufacturing*, *2*(2), 22001.
151. Kratkiewicz, K. J., White Bear, J., & Miller, B. A. (2019). *Survey of Data Fusion in IoT*.

152. 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.
153. Kridi, D. S., De Carvalho, C. G. N., & Gomes, D. G. (2014). A predictive algorithm for mitigate swarming bees through proactive monitoring via wireless sensor networks. *PE-WASUN 2014 - Proceedings of the 11th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks*, 41–47.
154. 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.
155. 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.
156. 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*.
157. Kviesis, A., Komasilovs, V., Komasilova, O., & Zacepins, A. (2020). Application of fuzzy logic for honey bee colony state detection based on temperature data. *Biosystems Engineering*, *193*, 90–100.
158. Kviesis, A., & Zacepins, A. (2015). System architectures for real-time bee colony temperature monitoring. *Procedia Computer Science*, *43*(C), 86–94.
159. 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.
160. Kwon, K.-H., Kim, J.-S., & Lee, H.-B. (2019). Forecast of Bee Swarming using Data Fusion and LSTM. *Journal of Digital Contents Society*, *20*(1), 1–6.
161. Lakshmanarao, A., & Shashi, M. (2020). A survey on machine learning for cyber security. *International Journal of Scientific and Technology Research*, *9*(1), 499–502.
162. 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.
163. 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.
164. *Layer Farm Manager – Poulty Layer Farm Management Software*. (2021). http://www.layerfarm.com/
165. Levene, M., & Loizou, G. (2003). Why is the snowflake schema a good data warehouse design? *Information Systems*, *28*(3), 225–240.
166. Li, D., Tong, Q., Shi, Z., Zheng, W., Wang, Y., Li, B., & Yan, G. (2020). Effects of cold stress and ammonia concentration on productive performance and egg quality traits of laying hens. *Animals*, *10*(12), 1–14.
167. 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.
168. Liou, J. J. H., Hsu, C. C., & Chen, Y. S. (2014). Improving transportation service quality based on information fusion. *Transportation Research Part A: Policy and Practice*, *67*, 225–239.
169. 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
170. 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).

171. 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.
172. Liu, Yin, & Zhang, Y. (2022). A Weighted Evidence Combination Method for Multisensor Data Fusion. *Journal of Internet Technology*, *23*(3), 553–560.
173. 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.
174. 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.
175. Lohmann. (2013). *Management Guide for Cage Housing: Brown-Extra Layers*.
176. Lubich, J., Thomas, K., & Engels, D. W. (2019). Identification and Classification of Poultry Eggs: A Case Study Utilizing Computer Vision and Machine Learning. *SMU Data Science Review*, *2*(1), 24. https://scholar.smu.edu/datasciencereviewhttp://digitalrepository.smu.edu.Availableat:https://scholar.smu.edu/datasciencereview/vol2/iss1/20
177. Luo, R. C., Yih, C. C., & Su, K. L. (2002). Multisensor fusion and integration: Approaches, applications, and future research directions. *IEEE Sensors Journal*, *2*(2), 107–119.
178. Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., & Kim, T. K. (2021). Multiple object tracking: A literature review. In *Artificial Intelligence* (Vol. 293).
179. Luque-Vega, L. F., Michel-Torres, D. A., Lopez-Neri, E., Carlos-Mancilla, M. A., & González-Jiménez, L. E. (2020). Iot smart parking system based on the visual-aided smart vehicle presence sensor: SPIN-V. *Sensors (Switzerland)*, *20*(5), 1476.
180. MacLeod, M. J., Vellinga, T., Opio, C., Falcucci, A., Tempio, G., Henderson, B., Makkar, H., Mottet, A., Robinson, T., Steinfeld, H., & Gerber, P. J. (2018). Invited review: A position on the Global Livestock Environmental Assessment Model (GLEAM). *Animal*, *12*(2), 383–397.
181. Maindonald, J. H. (2012). Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery by Graham Williams. *International Statistical Review*, *80*(1), 199–200.
182. 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.
183. Masiza, W., Chirima, J. G., Hamandawana, H., & Pillay, R. (2020). Enhanced mapping of a smallholder crop farming landscape through image fusion and model stacking. *International Journal of Remote Sensing*, *41*(22), 8736–8753.
184. Mavrogiorgou, A., Kiourtis, A., Perakis, K., Pitsios, S., & Kyriazis, D. (2019). IoT in Healthcare: Achieving Interoperability of High-Quality Data Acquired by IoT Medical Devices. *Sensors (Basel, Switzerland)*, *19*(9), 1978.
185. McMillan, I. (1981). Compartmental Model Analysis of Poultry Egg Production Curves. *Poultry Science*, *60*(7), 1549–1551.
186. McNally, D. H. (1971). 315. Note: Mathematical Model for Poultry Egg Production. *Biometrics*, *27*(3), 735.
187. Meikle, W. G., & Holst, N. (2015). Application of continuous monitoring of honeybee colonies. *Apidologie*, *46*(1), 10–22.
188. Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A Data Quality in Use model for Big Data. *Future Generation Computer Systems*, *63*, 123–130.

189. Meuter, M., Nunn, C., Görmer, S. M., Müller-Schneiders, S., & Kummert, A. (2011). A decision fusion and reasoning module for a traffic sign recognition system. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1126–1134.
190. Minnikhanov, R., Dagaeva, M., Anikin, I., Bolshakov, T., Makhmutova, A., & Mingulov, K. (2020). Detection of traffic anomalies for a safety system of smart city. *CEUR Workshop Proceedings*, *2667*, 337–342.
191. Mirkouei, A. (2020). A Cyber-Physical Analyzer System for Precision Agriculture. *Environmental Science Current Research*, *3*(1), 1–8.
192. Morales-Suárez, W., Ospina-Rojas, I. C., Méndez-Arteaga, J. J., Ferreira, A. H. D. N., & Váquiro-Herrera, H. A. (2021). Multivariate modeling strategies to predict nutritional requirements of essential amino acids in semiheavy second-cycle hens. *Revista Brasileira de Zootecnia*, *50*, 1–20.
193. Morales, I. R., Cebrián, D. R., Fernandez-Blanco, E., & Sierra, A. P. (2016). Early warning in egg production curves from commercial hens: A SVM approach. *Computers and Electronics in Agriculture*, *121*, 169–179.
194. Mota de Carvalho, N., Oliveira, D. L., Saleh, M. A. D., Pintado, M. E., & Madureira, A. R. (2021). Importance of gastrointestinal in vitro models for the poultry industry and feed formulations. In *Animal Feed Science and Technology* (Vol. 271, p. 114730). Elsevier B.V.
195. *Multiobjektu detektēšana un izsekošana transportlīdzekļu satiksmes novērošanai: 3D-LiDAR un kameras datu apvienošana*. (2020). https://www.itkc.lv/services
196. 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.
197. 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.
198. Murillo, A. C., Abdoli, A., Blatchford, R. A., Keogh, E. J., & Gerry, A. C. (2020). Parasitic mites alter chicken behaviour and negatively impact animal welfare. *Scientific Reports*, *10*(1), 8236.
199. Naehev, V. G., Titova, T. P., & Kosturkov, R. D. (2019). Instrumental data fusion for food analysis application. *IFAC-PapersOnLine*, *52*(25), 58–63.
200. Nair, K., Kulkarni, J., Warde, M., Dave, Z., Rawalgaonkar, V., Gore, G., & Joshi, J. (2016). Optimizing power consumption in iot based wireless sensor networks using Bluetooth Low Energy. *Proceedings of the 2015 International Conference on Green Computing and Internet of Things, ICGCIoT 2015*, 589–593.
201. Narinc, D., Uckardes, F., & Aslan, E. (2014). Egg production curve analyses in poultry science. *World's Poultry Science Journal*, *70*(4), 817–828.
202. 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.
203. Nelder, J. A. (1961). The Fitting of a Generalization of the Logistic Curve. *Biometrics*, *17*(1), 89.
204. 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.
205. Ngo, T. N., Wu, K. C., Yang, E. C., & Lin, T. Te. (2019). A real-time imaging system for multiple honey bee tracking and activity monitoring. *Computers and Electronics in Agriculture*, *163*, 104841.

206. 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*.
207. Noh, S. (2020). Intelligent data fusion and multi-agent coordination for target allocation. *Electronics (Switzerland)*, *9*(10), 1–13.
208. 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
209. Okafor, N. U., Alghorani, Y., & Delaney, D. T. (2020). Improving Data Quality of Low-cost IoT Sensors in Environmental Monitoring Networks Using Data Fusion and Machine Learning Approach. *ICT Express*, *6*(3), 220–228.
210. Okinda, C., Nyalala, I., Korohou, T., Okinda, C., Wang, J., Achieng, T., Wamalwa, P., Mang, T., & Shen, M. (2020). A review on computer vision systems in monitoring of poultry: A welfare perspective. *Artificial Intelligence in Agriculture*, *4*, 184–208.
211. Omomule, T. G., Ajayi, O. O., & Orogun, A. O. (2020). Fuzzy prediction and pattern analysis of poultry egg production. *Computers and Electronics in Agriculture*, *171*, 105301.
212. Orakwue, S. I., Al-Khafaji, H. M. R., & Chabuk, M. Z. (2022). IoT Based Smart Monitoring System for Efficient Poultry Farming. *Webology*, *19*(1), 4105–4112.
213. 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.
214. Parish, C. M., & Edmondson, P. D. (2019). Data visualization heuristics for the physical sciences. *Materials and Design*, *179*, 107868.
215. 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).
216. Patil, P. W., Dudhane, A., Chaudhary, S., & Murala, S. (2022). Multi-frame based adversarial learning approach for video surveillance. *Pattern Recognition*, *122*.
217. 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*, *21*(1), 694–701.
218. Pawar, K., & Attar, V. (2022). Deep learning based detection and localization of road accidents from traffic surveillance videos. *ICT Express*, *8*(3), 379–387.
219. 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
220. Peng, W., Deng, H., & Chen, A. (2020). Using Hellinger and Bures metrics to construct two-dimensional quantum metric space for weather data fusion. *Information Fusion*, *55*(September 2019), 199–206.
221. Pepinsky, T. B. (2018). A Note on Listwise Deletion versus Multiple Imputation. *Political Analysis*, *26*(4), 480–488.
222. Pereira, D. F., Lopes, F. A. A., Filho, L. R. A. G., Salgado, D. D. A., & Neto, M. M. (2020). Cluster index for estimating thermal poultry stress (gallus gallus domesticus). *Computers and Electronics in Agriculture*, *177*, 105704.

223. Pereira, W. F., Fonseca, L. da S., Putti, F. F., Góes, B. C., & Naves, L. de P. (2020). Environmental monitoring in a poultry farm using an instrument developed with the internet of things concept. *Computers and Electronics in Agriculture*, *170*, 105257.
224. Piechnicki, F., Dos Santos, C. F., De Freitas Rocha Loures, E., & Dos Santos, E. A. P. (2021). Data fusion framework for decision-making support in reliability-centered maintenance. *Journal of Industrial and Production Engineering*, *38*(1), 1–17.
225. 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.
226. Pitesky, M., Gendreau, J., Bond, T., & Carrasco-Medanic, R. (2020). Data challenges and practical aspects of machine learning-based statistical methods for the analyses of poultry data to improve food safety and production efficiency. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, *15*(49), 1–11.
227. 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.
228. *Poultry ERP Software for Profitable Poultry Business | PoultryCare ERP*. (2021). https://www.poultry.care/
229. *PoultryWorld - Temperature and relative humidity critical in climate control*. (n.d.). Retrieved March 10, 2021, from https://www.poultryworld.net/Health/Articles/2017/1/Temperature-and-relative-humidity-critical-in-climate-control-78825E/
230. Pramanik, A., Sarkar, S., & Maiti, J. (2021). A real-time video surveillance system for traffic pre-events detection. *Accident Analysis and Prevention*, *154*.
231. Pramir Maharjan, & Yi Liang. (2020). Precision Livestock Farming: The Opportunities in Poultry Sector. *Journal of Agricultural Science and Technology A*, *10*(2), 45–53.
232. Putra, A. S., & Warnars, H. L. H. S. (2019). Intelligent Traffic Monitoring System (ITMS) for Smart City Based on IoT Monitoring. *1st 2018 Indonesian Association for Pattern Recognition International Conference, INAPR 2018 - Proceedings*, 161–165.
233. Qin, X., Luo, Y., Tang, N., & Li, G. (2020). Making data visualization more efficient and effective: a survey. *VLDB Journal*, *29*(1), 93–117.
234. Radak, J., Ducourthial, B., Cherfaoui, V., & Bonnet, S. (2015). Detecting road events using distributed data fusion: Experimental evaluation for the icy roads case. *IEEE Transactions on Intelligent Transportation Systems*, *17*(1), 184–194.
235. 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.
236. Raid R. Al-Nima, Fawaz S. Abdullah, A. N. H. (2021). Design a Technology Based on the Fusion of Genetic Algorithm, Neural network and Fuzzy logic. *ArXiv*, *2102.08035*.
237. Raina, A., & Palaniswami, M. (2021). The ownership challenge in the Internet of things world. *Technology in Society*, *65*, 101597.
238. Rangesh, A., Yuen, K., Satzoda, R. K., Rajaram, R. N., Gunaratne, P., & Trivedi, M. M. (2017). A Multimodal, Full-Surround Vehicular Testbed for Naturalistic Studies and Benchmarking: Design, Calibration and Deployment. *ArXiv Preprint ArXiv:1709.07502*. http://arxiv.org/abs/1709.07502
239. Rashid, H. T., & Ali, I. H. (2021). Multi-Camera Collaborative Network Experimental Study Design of Video Surveillance System for Violated Vehicles Identification. *Journal of Physics: Conference Series*, *1879*(2), 22090.
240. 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.

241. Ravindran, R., Santora, M. J., & Jamali, M. M. (2021). Multi-Object Detection and Tracking, Based on DNN, for Autonomous Vehicles: A Review. *IEEE Sensors Journal*, *21*(5), 5668–5677.
242. Revanth, M., Sanjeev Kumar, K., Srinivasan, M., Stonier, A. A., & Vanaja, D. S. (2021). Design and Development of an IoT Based Smart Poultry Farm. *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation, ICAECA 2021*.
243. Rowe, E., Dawkins, M. S., & Gebhardt-Henrich, S. G. (2019). A systematic review of precision livestock farming in the poultry sector: Is technology focussed on improving bird welfare? *Animals*, *9*(9), 614.
244. Sadiku, M., Wang, Y., Cui, S., & Musa, S. M. (2017). Cyber-physical systems: A literature review. *European Scientific Journal*, *13*(36), 52–58.
245. Sakomura, N. K., De Paula Reis, M., Ferreira, N. T., & Gous, R. M. (2019). Modeling egg production as a means of optimizing dietary nutrient contents for laying hens. *Animal Frontiers*, *9*(2), 45–51.
246. Sanyal, S., & Zhang, P. (2018). Improving quality of data: IoT data aggregation using device to device communications. *IEEE Access*, *6*, 67830–87840.
247. Savegnago, R. P., Cruz, V. A. R., Ramos, S. B., Caetano, S. L., Schmidt, G. S., Ledur, M. C., El Faro, L., & Munari, D. P. (2012). Egg production curve fitting using nonlinear models for selected and nonselected lines of White Leghorn hens. *Poultry Science*, *91*(11), 2977–2987.
248. 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.
249. Scientific Committee on Animal Health and Animal Welfare. (2000). The Welfare of chickens kept for meat production (broilers). In *European Union Commission* (Issue March).
250. Shadrin, D., Menshchikov, A., Somov, A., Bornemann, G., Hauslage, J., & Fedorov, M. (2020). Enabling Precision Agriculture through Embedded Sensing with Artificial Intelligence. *IEEE Transactions on Instrumentation and Measurement*, *69*(7), 4103–4113.
251. Sharifi, M. A., Patil, C. S., Yadav, A. S., & Bangar, Y. C. (2022). Mathematical modeling for egg production and egg weight curves in a synthetic white leghorn. *Poultry Science*, *101*(4), 101766.
252. Shendryk, Y., Sofonia, J., Garrard, R., Rist, Y., Skocaj, D., & Thorburn, P. (2020). Fine-scale prediction of biomass and leaf nitrogen content in sugarcane using UAV LiDAR and multispectral imaging. *International Journal of Applied Earth Observation and Geoinformation*, *92*, 102177.
253. Sheng, X., Chen, Y., Guo, L., Yin, J., & Han, X. (2018). Multitarget tracking algorithm using multiple GMPHD filter data fusion for sonar networks. *Sensors (Switzerland)*, *18*(10).
254. 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.
255. 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.
256. Silva-Ramírez, E. L., Pino-Mejías, R., & López-Coello, M. (2015). Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Applied Soft Computing*, *29*, 65–74.

257. 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.
258. 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.
259. *SmartBird Poultry Manager Features | SmartBird*. (2021). https://smartbirdapp.com/smartbird-poultry-manager-features/
260. So-In, C., Poolsanguan, S., & Rujirakul, K. (2014). A hybrid mobile environmental and population density management system for smart poultry farms. *Computers and Electronics in Agriculture*, *109*, 287–301.
261. 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.
262. 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.
263. 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.
264. Su, G., & Kensek, K. (2021). Fault-detection through integrating real-time sensor data into BIM. *Informes de La Construccion*, *73*(564).
265. Su, X., Fan, K., & Shi, W. (2019). Privacy-Preserving Distributed Data Fusion Based on Attribute Protection. *IEEE Transactions on Industrial Informatics*, *15*(10), 5765–5777.
266. Sun, Z., Liu, X., Li, L., & Yang, X. (2022). Military-Civil fusion and optimisation of urban industrial structure—an evidence from China. *Cogent Economics and Finance*, *10*(1), 2101221.
267. Tan, L., Haley, R., & Wortman, R. (2015). Cloud-based harvest management system for specialty crops. *Proceedings - IEEE 4th Symposium on Network Cloud Computing and Applications, NCCA 2015*, *122*, 91–98.
268. Targowski, A. (2005). From Data to Wisdom. *Dialogue and Universalism*, *15*(5), 55–71.
269. Tavares, T. R., Molin, J. P., Hamed Javadi, S., de Carvalho, H. W. P., & Mouazen, A. M. (2021). Combined use of vis-nir and xrf sensors for tropical soil fertility analysis: Assessing different data fusion approaches. *Sensors (Switzerland)*, *21*(1), 1–23.
270. 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.
271. The Council of the European Union, & C E C (Commission of the European Communities). (1999). Council directive 1999/74/EC of 19 july 1999 laying down minimum standards for the protection of laying hens. *Official Journal of the European Communities, L203*(203), 53–57. http://data.europa.eu/eli/dir/1999/74/oj
272. *Tree Flowering Calendar*. (2020). https://wiki.sams-project.eu/index.php/Tree_flowering_calendar_Ethiopia
273. Tschopp, F., Riner, M., Fehr, M., Bernreiter, L., Furrer, F., Novkovic, T., Pfrunder, A., Cadena, C., Siegwart, R., & Nieto, J. (2020). VersaVIS—an open versatile multi-camera visual-inertial sensor suite. *Sensors (Switzerland)*, *20*(5), 1439.
274. Turk, M. (2014). Multimodal interaction: A review. *Pattern Recognition Letters*, *36*, 189–195.
275. Useya, J., & Chen, S. (2018). Comparative Performance Evaluation of Pixel-Level and Decision-Level Data Fusion of Landsat 8 OLI, Landsat 7 ETM+ and Sentinel-2 MSI for Crop Ensemble Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(11), 4441–4451.

276. Van Horne, P. L. M., & Achterbosch, T. J. (2008). Animal welfare in poultry production systems: Impact of EU standards on world trade. *World's Poultry Science Journal*, *64*(1), 40–51.
277. 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.
278. Varshney, P. K. (1997). Multisensor data fusion. *Electronics and Communication Engineering Journal*, *9*(6), 245–253.
279. 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.
280. 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. *Communications in Computer and Information Science*, *1455 CCIS*, 421–432.
281. Vu, A. K. N., Nguyen, N. D., Nguyen, K. D., Nguyen, V. T., Ngo, T. D., Do, T. T., & Nguyen, T. V. (2022). Few-shot object detection via baby learning. *Image and Vision Computing*, *120*.
282. Wang-Li, L., Xu, Y., Padavagod Shivkumar, A., Williams, M., & Brake, J. (2020). Effect of dietary coarse corn inclusion on broiler live performance, litter characteristics, and ammonia emission. *Poultry Science*, *99*(2), 869–878.
283. Wang, C. Y., Chen, Y. J., & Chien, C. F. (2021). Industry 3.5 to empower smart production for poultry farming and an empirical study for broiler live weight prediction. *Computers and Industrial Engineering*, *151*.
284. 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.
285. Wang, Y., Huang, Y., Yang, K., Chen, Z., & Luo, C. (2022). Generator Fault Classification Method Based on Multi-Source Information Fusion Naive Bayes Classification Algorithm. *Energies*, *15*(24), 9635.
286. Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, *6*(60), 3021.
287. 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.
288. Weng, X., Luan, X., Kong, C., Chang, Z., Li, Y., Zhang, S., Al-Majeed, S., & Xiao, Y. (2020). A Comprehensive Method for Assessing Meat Freshness Using Fusing Electronic Nose, Computer Vision, and Artificial Tactile Technologies. *Journal of Sensors*, *2020*, 1–14.
289. White, F. E., & Steinberg, A. N. (1998). *Community Status Report and Proposed Revisions to the JDL Data Fusion Model*.
290. Wiseman, L., Sanderson, J., Zhang, A., & Jakku, E. (2019). Farmers and their data: An examination of farmers' reluctance to share their data through the lens of the laws impacting smart farming. *NJAS - Wageningen Journal of Life Sciences*, *90–91*, 100301.
291. Wood, P. D. P. (1967). Algebraic model of the lactation curve in cattle. *Nature*, *216*(5111), 164–165.
292. Wu, T., Tsai, C., & Guo, J. (2018). LiDAR / Camera Sensor Fusion Technology for Pedestrian Detection. *9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, February*, 1675–1678.

293. Xia, R., Chen, Y., & Ren, B. (2022). Improved anti-occlusion object tracking algorithm using Unscented Rauch-Tung-Striebel smoother and kernel correlation filter. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), 6008–6018.
294. Yan, Y., Xu, M., Smith, J. S., Shen, M., & Xi, J. (2021). Multicamera pedestrian detection using logic minimization. *Pattern Recognition*, *112*, 107703.
295. Yan, Z., Liu, J., Yang, L. T., & Chawla, N. (2018). Big data fusion in Internet of Things. *Information Fusion*, *40*(40), 32–33.
296. Yang, N., Wu, C., & McMillan, I. (1989). New Mathematical Model of Poultry Egg Production. *Poultry Science*, *68*(4), 476–481.
297. 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.
298. 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.
299. Yeong, D. J., Barry, J., & Walsh, J. (2020). A Review of Multi-Sensor Fusion System for Large Heavy Vehicles off Road in Industrial Environments. *2020 31st Irish Signals and Systems Conference, ISSC 2020*, 1–6.
300. 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.
301. Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*(2), 22022.
302. Zacepins, A., Brusbardis, V., Meitalovs, J., & Stalidzans, E. (2015). Challenges in the development of Precision Beekeeping. *Biosystems Engineering*, *130*, 60–71.
303. 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.
304. 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)*.
305. Zaefarian, F., Cowieson, A. J., Pontoppidan, K., Abdollahi, M. R., & Ravindran, V. (2021). Trends in feed evaluation for poultry with emphasis on in vitro techniques. *Animal Nutrition*, *7*(2), 268–281.
306. Zhang, J. (2010). Multi-source remote sensing data fusion: Status and trends. *International Journal of Image and Data Fusion*, *1*(1), 5–24.
307. 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.
308. Zhang, Y. C., Zhang, S., Liu, M., Daly, E., Battalio, S., Kumar, S., Spring, B., Rehg, J. M., & Alshurafa, N. (2020). SyncWISE: Window induced shift estimation for synchronization of video and accelerometry from wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(3), 1–26.
309. Zhang, Y., Meratnia, N., & Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *IEEE Communications Surveys and Tutorials*, *12*(2), 159–170.
310. Zhao, G., Xiao, X., Yuan, J., & Ng, G. W. (2014). Fusion of 3D-LIDAR and camera data for scene parsing. *Journal of Visual Communication and Image Representation*, *25*(1), 165–183.

311. Zhao, Y., Wu, W., He, Y., Li, Y., Tan, X., & Chen, S. (2021). Practices and a strong baseline for traffic anomaly detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 3988–3996.
312. Zheng, H., Zhang, T., Fang, C., Zeng, J., & Yang, X. (2021). Design and implementation of poultry farming information management system based on cloud database. *Animals*, *11*(3), 1–15.
313. Zheng, Y. (2015). Methodologies for Cross-Domain Data Fusion: An Overview. *IEEE Transactions on Big Data*, *1*(1), 16–34.
314. Zhu, Y., Song, E., Zhou, J., & You, Z. (2005). Optimal dimensionality reduction of sensor data in multisensor estimation fusion. *IEEE Transactions on Signal Processing*, *53*(5), 1631–1639.
315. 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.
316. Martín, L., Sánchez, L., Lanza, J., & Sotres, P. (2023). Development and evaluation of Artificial Intelligence techniques for IoT data quality assessment and curation. *Internet of Things*, 22, 100779.
317. 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.
318. 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.

PIELIKUMI

1. pielikums

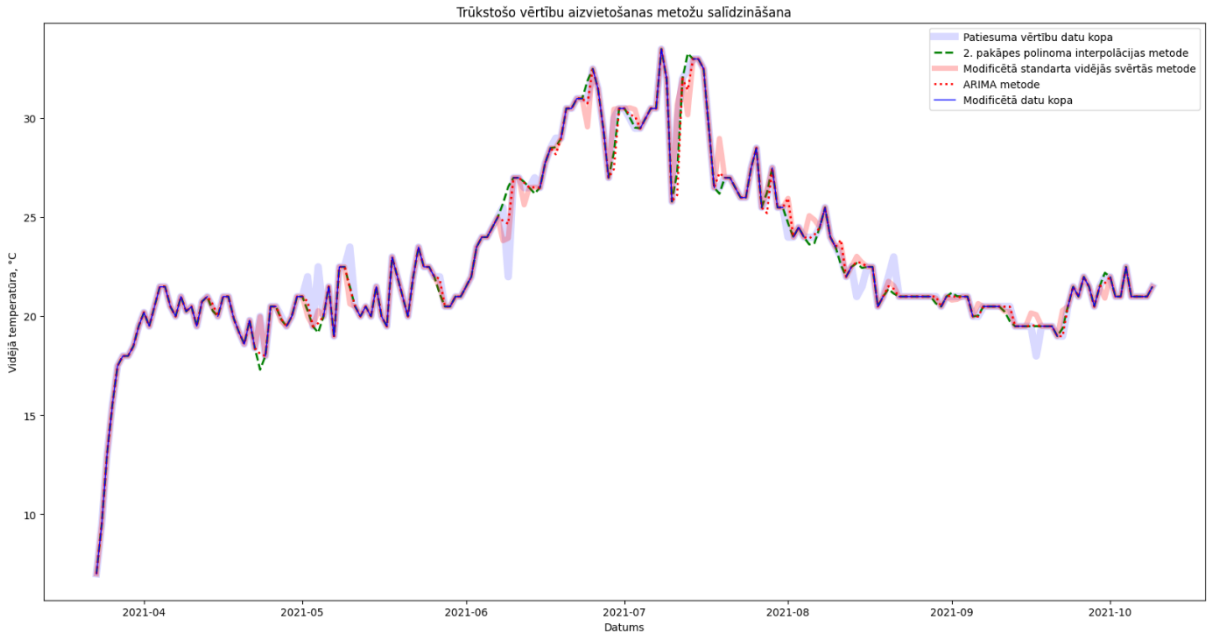
Četru augu atlase ar raksturīgām pazīmēm (*Tree Flowering Calendar*, 2020)

Plant	Nectar	Pollen	Month	Flowering	Flowering_value	Richness
Grevillea robusta	1.00	1.00	1	1	4	32
	1.00	1.00	2	1	4	32
	1.00	1.00	3	1	4	32
	1.00	1.00	4	2	8	64
	1.00	1.00	5	2	8	64
	1.00	1.00	6	1	4	32
	1.00	1.00	7	1	4	32
	1.00	1.00	8	1	4	32
	1.00	1.00	9	1	4	32
	1.00	1.00	10	2	8	64
	1.00	1.00	11	2	8	64
	1.00	1.00	12	2	8	64
Coffea arabica	1.50	1.00	1	1	5	40
	1.50	1.00	2	1	5	40
	1.50	1.00	3	1	5	40
	1.50	1.00	4	1	5	40
	1.50	1.00	5	1	5	40
	1.50	1.00	6	1	5	40
	1.50	1.00	7	1	5	40
	1.50	1.00	8	1	5	40
	1.50	1.00	9	2	10	80
	1.50	1.00	10	2	10	80
	1.50	1.00	11	2	10	80
	1.50	1.00	12	2	10	80
Eucalyptus citriodora	1.50	1.00	1	1	5	40
	1.50	1.00	2	0	0	0
	1.50	1.00	3	0	0	0
	1.50	1.00	4	0	0	0
	1.50	1.00	5	0	0	0
	1.50	1.00	6	0	0	0
	1.50	1.00	7	0	0	0
	1.50	1.00	8	0	0	0
	1.50	1.00	9	1	5	40
	1.50	1.00	10	1	5	40
	1.50	1.00	11	1	5	40
	1.50	1.00	12	1	5	40

Plant	Nectar	Pollen	Month	Flowering	Flowering_value	Richness
Dichrostachys cinerea	0.00	1.00	1	2	4	32
	0.00	1.00	2	2	4	32
	0.00	1.00	3	2	4	32
	0.00	1.00	4	2	4	32
	0.00	1.00	5	2	4	32
	0.00	1.00	6	1	2	16
	0.00	1.00	7	1	2	16
	0.00	1.00	8	2	4	32
	0.00	1.00	9	2	4	32
	0.00	1.00	10	2	4	32
	0.00	1.00	11	2	4	32
	0.00	1.00	12	2	4	32

2. pielikums

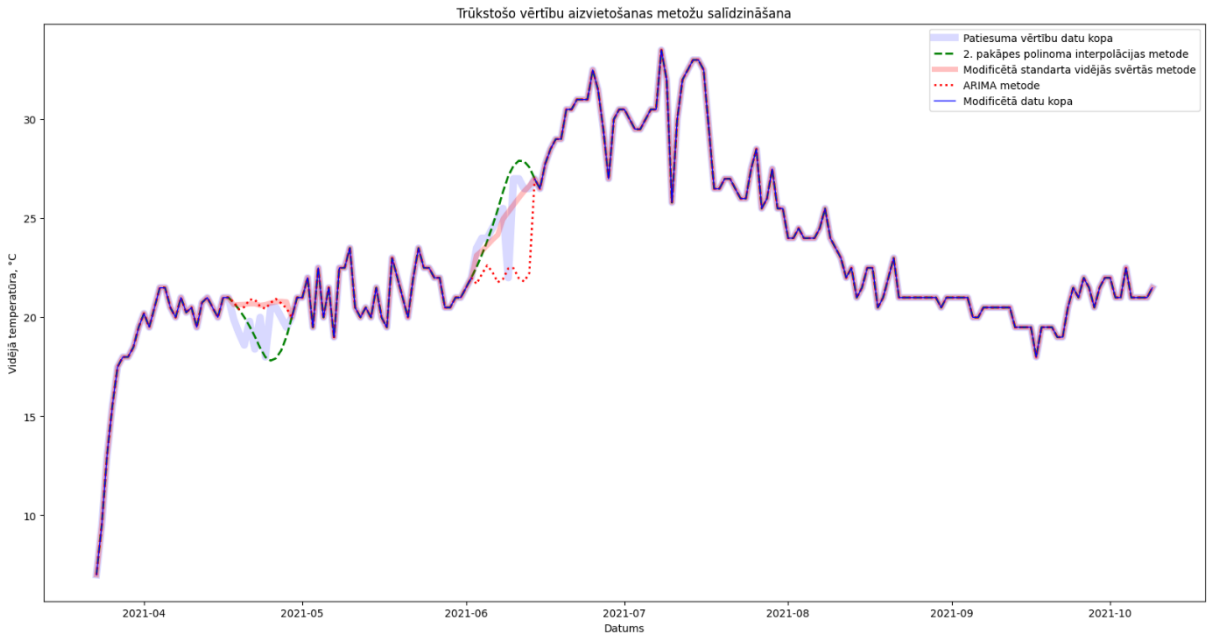
Trūkstošo vērtību metožu novērtēšanas scenārijs 2: gadījuma trūkstošās vērtības, līdz 20% no kopējā skaita



Metriku rezultāti scenārijam ar gadījuma trūkstošām vērtībām, līdz 20% no kopējā skaita

Metode	***MSE***	***RMSE***	***MAE***
Polinoma interpolācija	0.354109	0.595070	0.171887
MSVSM	0.314098	0.560444	0.193098
ARIMA	0.322169	0.567599	0.166225

Trūkstošo vērtību metožu novērtēšanas scenārijs 4: gadījuma trūkstošo vērtību virknes, divi gabali ar garumu 10 elementi katra



Metriku rezultāti scenārijam ar gadījuma trūkstošo vērtību virknēm, divi gabali ar garumu 10 elementi katra

Metode	***MSE***	***RMSE***	***MAE***
Polinoma interpolācija	0.283720	0.532654	0.128171
MSVSM	0.191233	0.437302	0.111985
ARIMA	0.715166	0.845675	0.228302