Master Thesis

Thesis Title: Concise and Engaging Title

by

Mateusz Kędzia
(2666752)

Supervisor:	Ronald Siebes (VU Amsterdam)
Daily Supervisor:	Jiancheng Weng (Beijing University of Technology)
Internal Advisor:	Zhisheng Huang (VU Amsterdam)
External Advisor:	Shuai Wang (VU Amsterdam/Maastricht University)
Second Reader:	Name and Surname

June 4, 2025

Submitted in partial fulfillment of the requirements for
the VU degree of Master of Science in Artificial Intelligence

你好，世界！Contribution Title

Mateusz Kędzia[1][0009−0001−4296−4479]

[1] Vrije Universiteit Amsterdam, Amsterdam
[2] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands lncs@springer.com
http://www.springer.com/gp/computer-science/lncs
[3] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
{abc,lncs}@uni-heidelberg.de

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: First keyword · Second keyword · Another keyword.

1 Literature Review

The field of vehicle trajectory anomaly detection has evolved through distinct methodological paradigms, each addressing fundamental limitations of preceding approaches while introducing novel computational perspectives. This review traces this evolution, examining how different methodological schools have contributed to our understanding of trajectory anomalies and identifying opportunities for methodological synthesis that inform our proposed integration of isolation forests with knowledge graph representations.

Foundational Isolation-Based Approaches The seminal work of Zhang et al. [10] established isolation forests as a paradigm-shifting approach for trajectory anomaly detection through their iBat framework. Operating on the principle that anomalies require fewer random partitions for isolation than normal trajectories, their method demonstrated remarkable computational efficiency compared to distance-based alternatives while achieving superior detection accuracy on Beijing taxi data. This foundational contribution not only validated tree-based partitioning for spatial-temporal data but also established Beijing as a crucial testbed for trajectory anomaly research—a precedent that directly motivates our choice of dataset and geographic focus.

Evolution of Density-Based Methodologies Building upon earlier clustering paradigms, He et al. [6] advanced density-based anomaly detection by integrating pattern information mining with enhanced DBSCAN clustering. Their incorporation of multiple distance metrics, including DTW and Hausdorff distances, addressed a critical limitation identified in Zhang et al.'s work [10]—the need for more sophisticated similarity measures beyond simple geometric proximity. This methodological evolution suggests that while isolation forests provide computational advantages, they could benefit from the multi-metric similarity insights developed in density-based approaches.

Extending this density-based paradigm to urban taxi trajectories, Hu et al. [9] specifically targeted the spatial-temporal complexity that Zhang et al. [10] had highlighted as challenging for isolation methods. Their fusion of density and length outlier detection mechanisms provides complementary anomaly indicators that could enhance the feature engineering pipeline for isolation forest implementations. The convergence of their findings with Zhang et al.'s computational efficiency requirements suggests a promising avenue for hybrid methodological development.

The Deep Learning Revolution and Its Implications The emergence of deep learning approaches marked a significant departure from the tree-based and density-based methodologies established by Zhang et al. [10] and He et al. [6]. Huang et al. [8] pioneered the application of LSTM autoencoders with attention mechanisms, addressing the temporal sequence modeling limitations inherent in earlier static approaches. While computationally more demanding than Zhang et al.'s isolation forests, their attention mechanism insights reveal feature importance patterns that could inform more intelligent partitioning strategies in tree-based methods.

The recent breakthrough of Li et al. [5] with their DiffTAD framework represents the current pinnacle of generative model applications to trajectory anomaly detection. Their denoising diffusion approach fundamentally reframes anomaly detection as a reconstruction problem, achieving state-of-the-art performance that significantly surpasses both the classical isolation methods of Zhang et al. [10] and the density approaches of He et al. [6]. However, this performance comes at substantial computational cost, highlighting the persistent trade-off between accuracy and efficiency that has characterized the field since Zhang et al.'s original efficiency-focused contribution.

Zhao et al. [4] extended deep learning approaches to commercial vehicle monitoring using BeiDou GPS data, demonstrating scalability considerations that complement Zhang et al.'s efficiency focus [10]. Their multi-scale CNN architecture provides preprocessing insights that could enhance feature extraction for isolation forest applications, particularly for handling the large-scale urban datasets that both Zhang et al. and our proposed research target.

Traditional Methods and Feature Engineering Legacy Han et al. [1] contributed crucial preprocessing methodologies through their pathlet-based trajectory representation and edit distance calculations. Their offline mining approach provides trajectory segmentation strategies that address data preparation challenges implicit in Zhang et al.'s work [10] but not explicitly detailed. The pathlet representation offers a middle ground between the geometric simplicity favored by isolation methods and the complex feature spaces required by deep learning approaches like those of Huang et al. [8].

Multi-Modal and Real-Time Extensions While Yin et al. [2] focused on video analysis rather than GPS trajectories, their work illuminates multi-modal validation strategies that could strengthen the evaluation frameworks for both classical approaches like Zhang et al.'s iBat [10] and modern methods like Li et al.'s DiffTAD [5]. Their emphasis on behavioral pattern recognition provides conceptual bridges between visual and trajectory-based anomaly definitions.

The real-time trajectory prediction work [3] directly addresses deployment scalability concerns that Zhang et al. [10] anticipated but did not fully explore. By combining prediction with anomaly detection, this approach suggests how isolation forest efficiency advantages could be leveraged in production systems where the computational constraints highlighted by the deep learning methods of Li et al. [5] become prohibitive.

Hybrid Methodological Convergence Recent developments in spatial and feature mixed outlier detection [7] exemplify the methodological synthesis that characterizes the field's current evolution. These hybrid approaches integrate the computational efficiency principles established by Zhang et al. [10] with the multi-metric sophistication developed by He et al. [6], while incorporating feature complexity insights from deep learning approaches like those of Huang et al. [8].

Synthesis and Future Directions The trajectory of research from Zhang et al.'s foundational isolation forests [10] through He et al.'s density enhancements [6] to Li et al.'s diffusion models [5] reveals a persistent tension between computational efficiency and detection sophistication. While deep learning approaches have achieved remarkable performance gains, they have simultaneously highlighted the enduring value of Zhang et al.'s efficiency principles for large-scale deployment scenarios. The preprocessing insights from Han et al. [1] and the multi-metric approaches of He et al. [6] suggest pathways for enhancing isolation forest performance without sacrificing computational advantages.

Our proposed integration of isolation forests with knowledge graph representations addresses this efficiency-sophistication trade-off by leveraging semantic knowledge to enhance feature discrimination while maintaining the computational scalability that Zhang et al. [10] established as crucial for urban-scale applications. This approach synthesizes the efficiency legacy of isolation methods, the feature sophistication insights from density-based approaches [6], and the semantic reasoning capabilities that neither classical methods nor current deep learning approaches [5] adequately address. The convergence of hybrid methodologies [7] with real-time deployment requirements [3] further validates this direction as both theoretically sound and practically viable for next-generation trajectory anomaly detection systems.

1.1 Data and Preprocessing

1.2 Data

1.3 Data Preprocessing

Report with examples on why there is a need for preprocessing. And how these exceptional cases are dealt with.

Report all the statistics before and after the preprocessing.

See Appendix A for the reasoning and choices of all the parameters.

2 Methodology

Below are several ways you can obtain the list of abnormal traj.

2.1 Isolation Tree

Isolation Tree method. And the choice of parameters. - anything that is too detailed goes into the Appendix

2.2 Ratio...

- anything that is too detailed goes into the Appendix

2.3 Improve the results

When examining the real data, we noticed that simply applying the abovementioned traj. detection algorithm is not enough. Some exceptions should be taken into consideration.

- anything that is too detailed goes into the Appendix

Explain exception 1, 2, 3.

3 Evaluation

Baseline of your naive traj. section algorithm: simple and imperfect.

Then the Isolation tree algorithm

Fine-tuned/improve isolation tree algorithm

I want to see a table of Precision

	Precision	Parametric Setting	Comments
Baseline			
Iso Tree			
Improved Iso Tree			

3.1 Conversion to Knowledge Graph????

...

3.2 Synthetic Knowledge Graph Generation

...

[Fix Chinese chars not displaying]

References

1. Boyang, H., Zhaoyang, W., Beihong, J.: 一种基于轨迹大数据离线挖掘与在线实时监测的出租车异常轨迹检测算法. **46**(3), 247–252 (2016). https://doi.org/10.3969/j.issn.0253-2778.2016.03.010
2. Hongpeng, Y.: . Master's thesis, Beijing Jiaotong University (2014)
3. Hu, W., Li, M., Kwan, M.P., Luo, H., Chen, B.: Real-time taxi spatial anomaly detection based on vehicle trajectory prediction. Travel Behaviour and Society **34**, 100698 (2024). https://doi.org/https://doi.org/10.1016/j.tbs.2023.100698
4. Jian-dong, Z., Xiao-dong, Z., Yu, L., Wei-wei, W.: . **22**(1), 284–291 (2021)
5. Li, C., Feng, G., Li, Y., Liu, R., Miao, Q., Chang, L.: Difftad: Denoising diffusion probabilistic models for vehicle trajectory anomaly detection. Knowledge-Based Systems **286**, 111387 (2024). https://doi.org/https://doi.org/10.1016/j.knosys.2024.111387, https://www.sciencedirect.com/science/article/pii/S0950705124000224
6. Ming, H., Yuting, C., Qiang, L., Bo, Z., Gongda, Q.: . (6), 49–54 (2019)
7. Name, A.: Safety: A spatial and feature mixed outlier detection. Journal Name (2024), based on file: Safety A spatial and feature mixed outlier detecti_250430_111524.pdf
8. Shi-chen, H., Chun-fu, S., Juan, L., Zong-tao, D.: . **21**(3), 47–54 (2021)
9. Yuan, H.: . In: . pp. 54–58. No. 6 (2019)
10. Zhang, D., Li, N., Zheng, Y., Ramamohanarao, K., Zhao, Z.: ibat: Detecting anomalous taxi trajectories from gps traces. In: UbiComp'11: Proceedings of the 13th international conference on Ubiquitous computing. pp. 99–108. ACM (2011). https://doi.org/10.1145/2060091.2060106

A Data Preprocessing Details