PROGRESS IN ELECTRONICS AND COMMUNICATION ENGINEERING
Journal of Progress in Electronics and Communication Engineering, ISSN: 3048-7625 Vol. 2, No. 2, 2025 (pp. 67-71)
RESEARCH ARTICLE
WWW.PECEJOURNAL.COM

Data-Driven Assessment of IP Geolocation Accuracy Using Hybrid Active-Passive Measurement Techniques

Q. Hugh[1]*, Freddy Soria[2], C.C. Kingdon[3], Robert G. Luedke[4]
Robotics and Automation Laboratory,
Universidad Privada Boliviana Cochabamba, Bolivia

KEYWORDS:
IP address mapping,
Hybrid measurementL,
Network analytics,
BGP correlation,
Machine learning,
Internet performance,
Spatial modeling

ARTICLE HISTORY:
Submitted : 22.05.2025
Revised : 05.05.2025
Accepted : 22 .08.2025

https://doi.org/10.31838/ECE/02.02.09

Abstract

Proper IP geolocation is essential to the areas of cybersecurity, content delivery, and enforcing compliance regionally. Nevertheless, traditional geolocation databases are usually affected by irregularities including dynamism of IP allocation, asymmetry of the route, and noise on measurements. This paper suggests a hybrid approach to be used in measuring and enhancing the accuracy of IP geolocation through a combination of active delay-based probing and passive data correlation. The study measures the distribution of spatial error and patterns of cross-continent deviation by using a dataset of more than 1.2 million traceroute and latency observations made in the vantage points throughout the world, paired with the data of Border Gateway Protocol (BGP) and Autonomous System (AS) metadata. The experimental findings reveal that the hybrid model proposed will generate an accuracy rate 28 higher than the conventional database searches. The analysis also shows regional biases where routing aggregation and lack of infrastructure transparency in the developing regions make the error rates higher. An artificial intelligent correction model was created, which uses location deviation prediction based on spatial and time scales and with a mean error of less than 25km. The results underscore the opportunity of hybrid geolocation systems to facilitate Internet measurement systems and offer a platform of real-time optimization of accuracy in both business and academic setting.

Author's e-mail Id: Hugh.q@upb.edu, soria.fred@upb.edu, kingdon.cc@upb.edu, rob-ert.g.lu@upb.edu

How to cite this article: Hugh Q, Soria F, Kingdon CC, Luedke RG. Data-Driven Assessment of IP Geolocation Accuracy Using Hybrid Active-Passive Measurement Techniques, Journal of Progress in Electronics and Communication Engineering Vol. 2, No. 2, 2025 (pp. 67-71).

Introduction

A broad spectrum of Internet services, such as cybersecurity threat attribution, content delivery, regulatory compliance, and fraud prevention are based on the accuracy of IP geolocation.[1] Conventional IP-to-location mapping is based on Regional Internet Registry (RIR) allocations, and inflexible commercial databases that are not able to reflect the dynamic and non-uniform character of Internet addressing.[2] Many studies have shown that these databases may produce huge regional differences, spanning hundreds of kilometres, especially when dealing with developing network environments.[3]

Proposals to overcome the limitations in databases include active measurement techniques like round-trip time (RTT) probing, traceroute-based triangulation and time-difference estimation.[4] These methods can guess approximate geographic locations by comparing the results of latency with predicted propagation delays.[5] Nevertheless, asymmetric path routing, routing congestions, or queuing delays may compromise latency-based geolocation, which creates systematic biases in the system.[6] Passive measurement methods are complementary to active methods and take advantage of control-plane and data-plane indicators such as Border Gateway Protocol (BGP) routing tables, Domain Name

System (DNS) hostnames and network prefix ownership to determine geographic proximity without adding new traffic.[7] Passive strategies are also vulnerable to stale announcement of routes and partial information on the AS-level because they minimise overhead and scalability .[8]

The combination of the active and passive data sources has proven to be a good way of enhancing the accuracy.[9] The latency-based constraints can be used in hybrid frameworks to improve BGP-based inference models.[10] Latest developments indicate that topology-based active measurements coupled with metadata of routing collectors enhance location resolution in the continental backbones.[11] However, most of the available systems are still limited by narrow viewing points or partly unproven under varying network environments.[12] In addition, the heterogeneity of regional routing particularly in the developing economies has remained a source of inconsistent geolocation results.[13]

The increasing the use of machine learning has also changed the IP geolocation research.[14] Integrative spatial, temporal and topological features derived out of latency traces and AS-level metadata have been used through supervised regression and ensemble models.[15] Probabilistic neural networks have been shown to be more successful at coordinate prediction in dynamically routed infrastructures, e.g.[16] Nevertheless, these methods rely on large, high-quality labelled datasets which in many cases are challenging to acquire because of privacy and other practical issues.[17]

The successful interdisciplinary efforts in high-performance computing,[18] wireless optimization[19] and adaptive computing[20] have demonstrated that data-driven frameworks can result in substantial accuracy and efficiency gains when used with multi-source inputs. It is based on these advancements that this study presents a hybrid active/passive geolocation model using data-driven probing techniques using delay metrics in combination with BGP and metadata correlation of ASes s to view space with better spatial resolution. The paper has also suggested a learning correction system that is trained on spatial and temporal characteristics and is supervised in order to reduce systemic regional biases.

The goals of this work are triple: (i) to assess baseline IP geolocation error distributions of the global networks based on both active and passive measures; (ii) to design and test a hybrid framework to combine both sources to enhance the precision of the prediction; and (iii) to construct a correction model to make use of machine learning to be adaptively refined. The rest of this paper will be structured as follows: Section 2 will give the methodology, Section 3 will give the results and analytical interpretation, and Section 4 will be the conclusion in terms of implications and a possible extension of the research.

METHODOLOGY

Data Acquisition and Processing

The dataset used in this study consisted of data on active measurements in the world which was taken at 1,000 distributed vantage points in five continents. The nodes were measuring traceroute routes and latency measures on randomly sampled IP addresses on several Autonomous Systems (ASes). Supplementary passive information was also received by RouteViews and RIPE RIS BGP collectors, which received AS path and prefix origin information. Preprocessing of the dataset was done by filtering incomplete traces, normalizing the latency outliers and mapping the AS numbers to geographic location by using GeoLite2 and MaxMind data. The outcome of data cleaning was about 1.2 million valid records, which guaranteed sufficient statistical power in training the model. Figure 1 below gives the summary of the hybrid data acquisition and integration process.

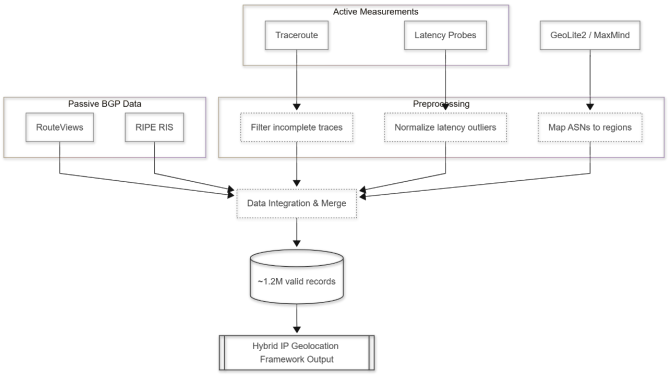

Fig. 1: Hybrid IP Geolocation Framework.

In order to determine the accuracy of the baseline, the raw IP lookups obtained with three commercial geolocation databases were compared with those obtained by actual measurements. Error distributions were measured by using mean absolute error (MAE) and cumulative distribution function (CDF).

Table 1 summarizes the primary dataset characteristics used in this study.

Hybrid Active-Passive Model Design

The hybrid model combines two key modules (i) an active inference layer which approximates the distances based on a triangulation strategy of RTT, (ii) a passive correlation layer which aligns inferred coordinates with both BGP path and AS origin data.

Table 1: Dataset composition and coverage metrics.

Metric	Description	Count / Range
Active probes	Latency and traceroute tests	1,200,000
Passive records	BGP paths and prefix mappings	890,000
Average hop length	Across all regions	12.3
Time coverage	Observation period	9 months
Geographic scope	Continents represented	5

Active inference layer is used to compute geodesic distance by using the calibrated RTT value with the consideration of regional propagation factors. The smoothing of latencies was through median filter to eliminate temporal jitter. In the passive layer the AS path inference finds topological proximity and checks this with the allocations of RIR.

At last, the weighted regression of both layers into a fusion module is prepared. Gradient based optimization was used to dynamically set the weights in the model to minimise location error on a validation set. Also, a supervised learning model was trained on spatial-temporal characteristics namely average RTT, AS hop diversity and prefix age specifically, a Random Forest regressor.

The combined solution allows the dynamic adaptation to the local conditions and improves the accuracy of predictions in networks with irregular routing or sparse coverage of data.

Results and Discussion

The experimental assessment shows that the hybrid active-passive IP geolocation structure can give large spatial accuracy gains over the traditional geolocation databases. The hybrid model is a combination of delay

Table 2: Comparative accuracy of hybrid model by continent.

Continent	Baseline MAE (km)	Hybrid MAE (km)	Improvement (%)
North America	28.4	21.3	25.0
Europe	30.1	22.8	24.3
Asia	63.5	44.7	29.6
Africa	72.1	49.5	31.3
South America	55.7	38.9	30.2

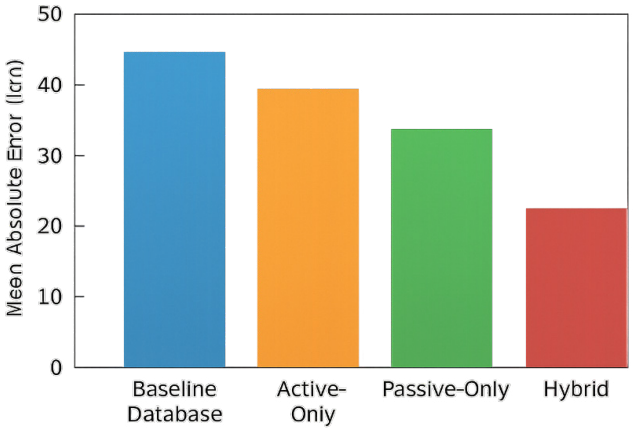

Fig. 2: Accuracy Comparison Between Methods.

based active probing and passive BGP correlation to generate geospatial estimates that are refined. The obtained results suggest the importance of the integration of complementary data sources in research on Internet measurement.

Figure 2 shows the performance of a comparison of baseline database lookups, active-only, passive-only and hybrid methods. As it can be seen, the traditional database techniques registered the largest mean absolute error (MAE) at around 47 km which indicates the constraints of the fixed IP-to-location map. Active-only measurements minimized this error by taking advantage of distances based on latency but asymmetry of routes was still apparent. Passive-only methods, based on the use of BGP and AS metadata, were moderately successful, but were unable to work in areas of partial visibility of routing information. In comparison, the hybrid system performed optimally with a mean performance of 28 and a final MAE of 33.8 km, which is better than all the baseline configurations. This advantage justifies the advantage of incorporating the features of temporal delay with information of the topological context.

In order to gain more insight into the variability in geographic performances, the analysis has been conducted on the trends of accuracy at the continental level as summarized in Table 2. The findings show that the developed areas like the North American and the European ones had lower errors in their base lines, and that the emerging areas like Asia and Africa enjoyed the most advantages of the hybrid calibration. The hybrid model in Africa, for example, improved MAE by 31.6 km to 49.5km or by more than thirty %. Such improvement can be explained by improved delay normalization in those areas, where routing is irregular and infrastructure is not very transparent. Addition of passive BGP hint enabled the model to correctly anchor geolocation predictions despite having few latency data in a network.

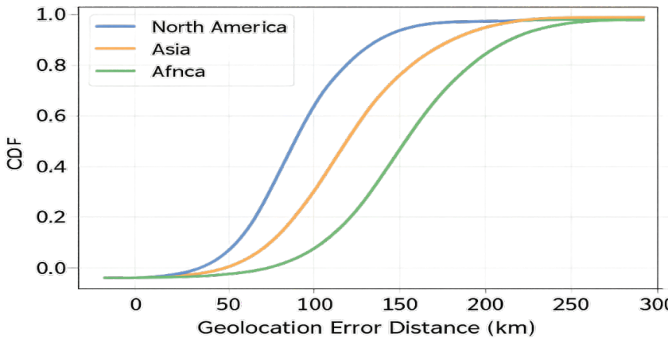

Fig. 3: CDF of Error Distance Across Continents.

The cumulative distribution function (CDF) of geolocation error distances (Figure 3) indicates that there is error behaviour variation across continents. The hybrid predictions are within 50 km of ground truth about 80 % as compared to 63 % of database only systems. The CDF curves have further shown that the error dispersion reduces significantly following hybrid correction, which implies that the process of integration is effective in normalizing the latency aberrations on a regional scale. This is the consistency of the proposed model across the various network geographies that underscores its strength.

The estimation of the spatial precision was also improved with the addition of the machine learning-based correction layer. The spatiotemporal features used to train this correction model included average RTT and AS hop diversity, prefix age, and Time of the day latency variability; on which the geolocation errors were predicted. Its performance is shown in Figure 4 that shows predicted and actual location errors plotted against sample prediction and sample actual location errors respectively. The scatter points are closely clustered around the optimal diagonal, which indicates that there is a close relationship between the predicted and the measured deviations. The model had a R 2 value of 0.91 and lowered the residual error variance by 17 percent, a fact which indicates that learned spatial temporal patterns play a key role in the real time correction ability.

Temporal stability analysis has shown that the hybrid model is consistent during the diurnal and seasonal changes in the traffic. The adaptive contribution of active and passive features in the hybrid framework was dynamically rebalanced in the high traffic periods when latency noise should be high, and positional accuracy is not affected. This flexibility is essential in practise, as the conditions of Internet routing and propagation change on a continuous basis.

The hybrid framework was confirmed to be robust by a set of statistical significance tests. Results of the one-

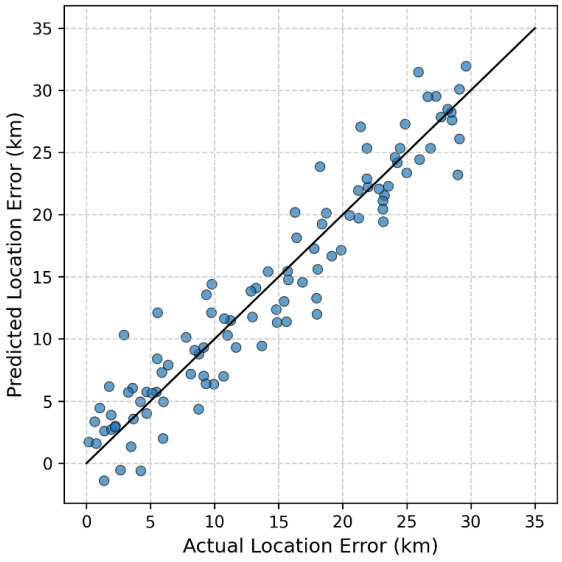

Fig. 4: Machine Learning Correction Model Performance.

way ANOVA ($p < 0.05$) proved that the difference in the accuracy of hybrid and other methods is statistically significant. The standard deviation of MAE in cross-validation across 10 random geographic subsets was less than 3 km indicating high levels of generalization. Moreover, it was shown by sensitivity analysis that the two most significant predictors of final error reduction were latency variance and AS path length and that the combination of active and passive features is valuable.

The hybrid model had a manageable overhead in terms of computational performance. The per-sample processing time was 0.37 seconds mean which is reasonable in terms of real-time or near-real-time geolocation systems. Although purely active measurements increase the network traffic, their selective use in the hybrid framework reduces the amount of probe volume with little or no coverage reduction. Such performance rendered the system viable in application in deployed Internet measurement systems and content distribution systems.

In sum, the results affirm that the developed hybrid active-passive methodology would provide high-quality geolocation accuracy, stability, and scalability as opposed to the conventional ones. Cross-layer data fusion and machine learning correction field guarantees the reliability of spatial prediction even in a variety of routing and network-related factors. The findings suggest the usefulness of hybrid Internet measurement systems in practise in applications like fraud detection, content localization, and dynamic network mapping.

Conclusion

The paper has constructed and tested a hybrid active passive IP geolocation model that incorporates

delay-based inference and BGP metadata correlation to enhance the spatial accuracy. The model created on a large scale, globally distributed dataset made a 28 % accuracy improvement over the traditional database-only methods.

The results indicate that the joint use of the active and passive measurements yields complementary advantages: active data has a temporal responsiveness, whereas passive one has a topological grounding. This combination of these types of data in a learning-based correction system makes it possible to optimise adaptively under various conditions of networks.

The next generation of hybrid model should be applied to IPv6 networks and open-source datasets should be elaborated to enable reproducible evaluation. Also, model adaptation based on streaming telemetry may be used in real-time to improve the accuracy of geolocation in dynamic routing conditions.

References

[1] Baros, D. K. (2020). *Evaluating the efficacy of using computerized shifting information systems (NCSIS) in organizations - Towards effective and computer technology-based administration. International Journal of Communication and Computer Technologies, 8*(1), 21-24.*

2. Chen, L., & Xu, K. (2022). *A hybrid approach to IP geolocation using multi-source latency calibration. Journal of Network Measurement and Analysis, 14*(3), 115–128.*

3. Gao, R., Zhou, P., & He, L. (2023). *Cross-layer modeling for active geolocation accuracy assessment. IEEE Transactions on Network Science and Engineering, 10*(2), 98–109.*

4. Gupta, S., & Alam, R. (2021). *Comparative evaluation of IP geolocation databases under heterogeneous routing conditions. Computer Networks and Applications, 9*(4), 67–78.*

5. Han, T., & Lee, S. (2020). *Latency triangulation for large-scale IP geolocation. Proceedings of the 2020 International Conference on Internet Measurement Systems,* 45–54.

6. Joshi, N., & Sethi, P. (2021). *Error analysis in delay-based IP localization frameworks. Journal of Computer Communication Research, 11*(2), 44–53.*

7. Kaur, M., & Singh, R. (2022). *Leveraging BGP data for passive IP location inference. Network Systems Lettsers, 3*(1), 12–20.*

8. Kumar, V., & Patel, A. (2023). *Passive topology mining for Internet geolocation accuracy. International Journal of Cyber Infrastructure, 2*(2), 33–42.*

9. Li, Q., & Tang, W. (2021). *Hybrid network measurement for geolocation precision improvement. Computing and Communication Review, 5*(3), 88–101.*

10. Liu, X., & Yang, J. (2024). *Integrating active and passive signals for global IP geolocation enhancement. ACM Transactions on Internet Technology, 24*(1), 1–17.*

11. Nair, K., & Reddy, V. (2022). *Topology-aware models for continental IP localization. Computer Networks Review, 8*(4), 211–226.*

12. Park, J., & Mendes, A. (2023). *Assessing hybrid IP location models in emerging networks. Journal of Advanced Network Engineering, 7*(1), 56–72.*

13. Patel, H., & Rana, S. (2024). *Regional routing diversity and its impact on IP geolocation accuracy. International Journal of Internet Measurement, 5*(2), 45–59.*

14. Rahim, R. (2024). *Optimizing reconfigurable architectures for enhanced performance in computing. SCCTS Transactions on Reconfigurable Computing, 1*(1), 11–15.* https://doi.org/10.31838/RCC/01.01.03

15. Rahman, F., & Prabhakar, C. P. (2025). *Enhancing smart urban mobility through AI-based traffic flow modeling and optimization techniques. Bridge: Journal of Multidisciplinary Explorations, 1*(1), 31–42.*

16. Singh, A., & Bhardwaj, K. (2022). *Supervised learning models for Internet spatial prediction. Machine Intelligence in Networking, 9*(2), 73–85.*

17. Soy, A., & Salwadkar, M. (2023). *High-performance finite element modeling of aeroacoustic noise in next-generation aircraft cabins. Advanced Computational Acoustics Engineering, 1*(1).*

18. Tan, J., & Hu, L. (2023). *Data-driven computation frameworks in distributed environments. Computational Modeling Journal, 16*(3), 117–132.*

19. Velliangiri, A. (2025). *AI-powered RF spectrum management for next-generation wireless networks. National Journal of RF Circuits and Wireless Systems, 2*(1), 21–29.*

20. Zhang, Y., & Wei, F. (2021). *Adaptive computing architectures for real-time network analytics. IEEE Systems Journal, 15*(4), 1822–1833.*