IMPUTING MISSING DATA OF ADS-B AND GPS DROPOUTS USING MACHINE LEARNING

Barathwaja Subash Chandar, Prakash Ranganathan
Data, Energy, Cyber and Systems (DECS) Laboratory
School of Electrical Engineering and Computer Science (SEECS), University of North Dakota

Abstract

This study investigates the risk and severity of ADS-B (automated dependent surveillance-broadcast) drop-out. These drop-out data points will be problematic for detecting and avoiding the purpose of other aircraft, resulting in poor decision-making and insufficient and undesirable outcomes. When learning from incomplete time series data, missing data estimation is a major problem. Experiments with different models were applied and compared with MAE, and RMSE metrics in imputing the missing data for discrete and continuous missing values at different rates – 10%, 20%, and 30%.

Dataset Overview

- Obtained from Opensky Network.
- Dataset Feature Columns are:
  - **time**
  - **icao24**
  - **lat**
  - **lon**
  - **geoaltitude**
- Considered 6 flights with multiple trips from 19th Feb 2022 – 27th Feb 2022.
- Overall record points ~ 60100 (dataset size).
- Average flight duration (6 flights) – 30 mins to 3 hours.

Background / Motivation

- An ADS-B receiver is set up to broadcast once every second. If there is a discontinuation of an update within one second is referred to as ADS-B Dropout [2].
- A fake ADS-B track could be used to persuade unmanned aircraft to fly towards controlled airspace, structures, terrain, and other targets for automated Detect and Avoid.
- Missing data points can lead to insufficient and undesirable outcomes. Figure 1 shows the sources of missing data.

Figure 1: Sources of missing data

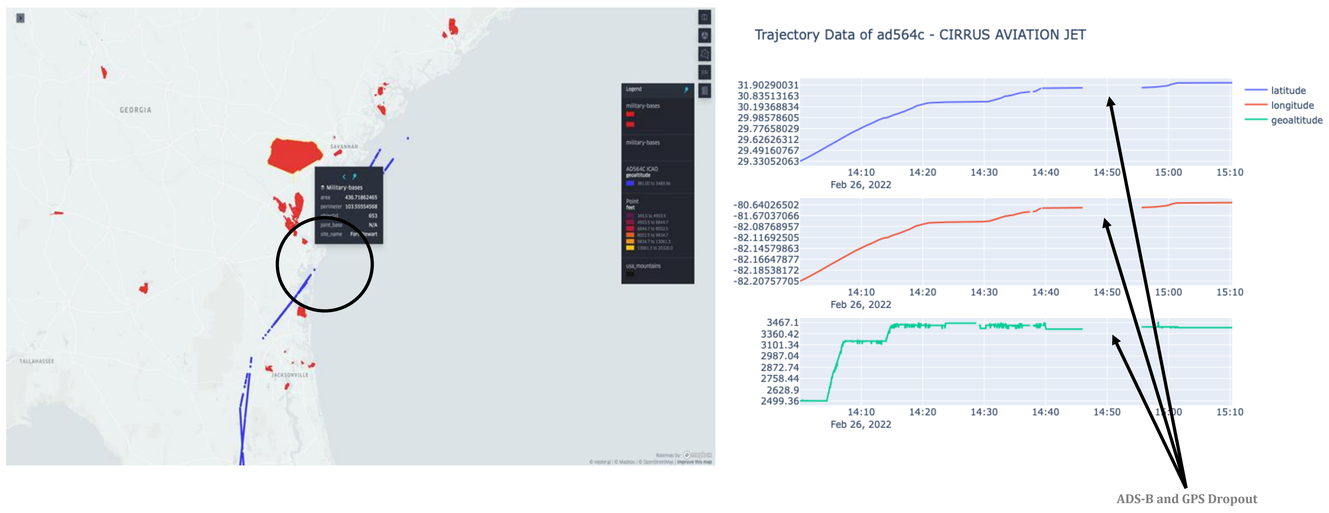

Figure 2: ad564c CIRRUS AVIATION JET with Data Discrepancies

Machine Learning Framework and Experiment Results

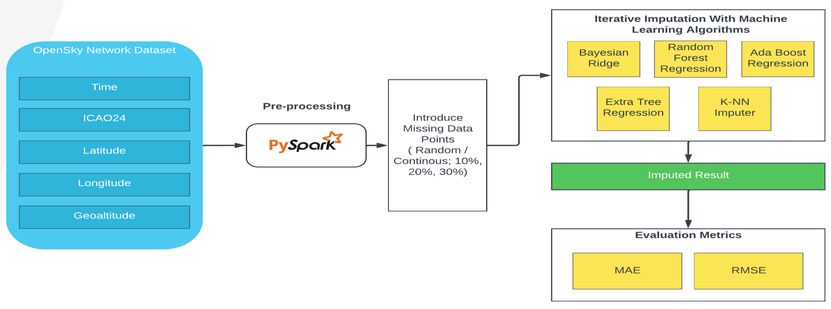

Figure 3: Overview of Machine Learning Framework for Imputing ADS-B and GPS Dropout Data

Imputation Methods	Hyperparameter Settings	
	Name	Parameter
Bayesian Ridge Regressor	-	
Random Forest, Adaboost Regressor, Extra Tree Regressor	n_estimators	(10,50,100)
k-NN	n_neighbors weights	(1,2,3,4,5,10,20,30,50) (uniform, distance)

Table 1: Overview of Hyperparameter Settings

Evaluation Metrics

In general, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are preferred metrics for determining the separation between two vectors: the vectors of prediction and target values.

Table 2 shows the 10% random imputed results for the k-NN algorithm. After 30 iterations, the minimum score was obtained.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i|}$$

where: x_i = actual value,
$\hat{x}_i$ = predicted value
n = no. of observations / rows

Equation 1: MAE and RMSE

	Iterations	Latitude	Longitude	Geoaltitude	Latitude	Longitude	Geoaltitude
		MAE	MAE	MAE	RMSE	RMSE	RMSE
	K-NN (Uniform)						
10%	1	0.027532166	0.066613315	9.662930198	0.52249312	0.999307847	118.8085095
	2	0.027487276	0.066568082	9.641451088	0.520611007	0.993298562	123.0517479
	3	0.027419963	0.066353669	9.455552998	0.51735079	0.987595268	119.6967444
	4	0.027383505	0.064974578	8.651886915	0.47622242	0.898431802	112.2230434
	5	0.027399683	0.064794468	8.562564323	0.47224573	0.890847045	109.9124322
	10	0.027428692	0.065292114	8.464688216	0.45582701	0.86521706	104.773125
	20	0.027161867	0.066208489	8.398535101	0.42147702	0.81622559	97.32200922
	30	0.027063793	0.065744485	8.312910983	0.40031532	0.773790004	91.99387661
	50	0.02743398	0.06683715	8.498259174	0.3881106	0.737222914	88.03382786
	K-NN (Distance)						
	1	0.027532166	0.066613315	9.662930198	0.5224931	0.999307847	118.8085095
	2	0.027484898	0.066566884	9.640831459	0.5206109	0.99329856	123.0526078
	3	0.027416004	0.066351613	9.457178016	0.5173507	0.987595263	119.6973065
	4	0.027350133	0.064931037	8.646724263	0.4759822	0.898158033	112.1622027
	5	0.027368772	0.064759	8.559351435	0.4721132	0.890699001	109.8792518
	10	0.027403484	0.065253267	8.471833034	0.4568031	0.865817011	104.8702989
	20	0.027172885	0.06628078	8.455892749	0.4294899	0.828011282	99.48241624
	30	0.027038094	0.065566315	8.344993921	0.4165358	0.796425152	96.07001966
	50	0.027027155	0.066206696	8.404728221	0.4074265	0.774048722	93.69566042

Table 2: Overall MAE and RMSE Score for Random 10% Imputed Results for k-NN

Conclusions and Future Work

Experiments were performed for different percentages of missing rates - 10%, 20%, and 30% randomly and continuously for five different machine learning models (Bayesian Ridge, Random Forest, Adaboost Regressor, Extra Tree Regressor, and k-NN). The k-NN performed better compared to other machine learning models to impute latitude, longitude, and geoaltitude as it imputes the missing values by grouping similar records based on distance functionality [8]. These perform well with smaller dataset compared to other tree-based structure. Future works is to expand the imputation to other columns – velocity and heading, etc. and applying other deep learning models such as – GAIN (Generative Adversarial Imputation Networks), MIDAS (Multiple Imputation with Denoising Autoencoders) and LSTM (Long Short-Term Memory).

Contact Information

Name – Barathwaja Subash Chandar
Degree – M.S in Computer Science
Email – b.subashchandar@und.edu

Acknowledgements

This research work is supported by FAA grant award under UND0026156.

References

[1] M. Richards, J. Scheer and W. Holm, Principals of Modern Radar, Edison: SciTech Publishing, 2010.

[2] Tabassum, A.; Semke, W. UAT ADS-B Data Anomalies and the Effect of Flight Parameters on Dropout Occurrences. Data 2018, 3, 19. https://doi.org/10.3390/data3020019.

[3] Ming Hua and Jian Pei. Cleaning disguised missing data: A heuristic approach. InKDD, KDD '07, page 950–958, New York, NY, USA, 2007

[4] Gabriel L Schlomer, Sheri Bauman, and Noel A Card. Best practice for missing data management in counseling psychology.J. Couns. Psychol., 57(1):1, 2010.

[5] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, second ed.,John Wiley and Sons, 2002

[6] W.-C. Lin, C.-F. Tsai, Missing value imputation: a review and analysis ofthe literature (2006–2017), Artif. Intell. Rev. 53 (2020) 1487–1509.

[7] S.Nikfalazar, C.-H. Yeh, S. Bedingfield, H.A. Khorshidi, Missing data imputation using decision trees and fuzzy clustering with iterative learning,Knowl. Inf. Syst. 62 (2020) 2419–2437

[8] A. Jadhav, D. Pramod, and K. Ramanathan, "Comparison of Performance of Data Imputation Methods for Numeric Dataset," Applied Artificial Intelligence, vol. 33, no. 10, pp. 913–933, 2019, doi: 10.1080/08839514.2019.1637138.

[9] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values.Sci. Rep.,8(1):6085, 2018.

[10] Zachary C Lipton, David Kale, and Randall Wetzel. Directly modeling missing datain sequences with rnns: Improved classification of clinical time series. InProc. Mach.Learn. Healthcare Conf., pages 253–270, 2016

[11] Jinsung Yoon, William R. Zame, and Mihaela van der Schaar. Multi-directionalrecurrent neural networks : A novel method for estimating missing data. 2017.

School of Electrical Engineering & Computer Science
UNIVERSITY OF NORTH DAKOTA
DATA, ENERGY, CYBER AND SYSTEMS
DECS
C2SR
CENTER FOR CYBER SECURITY RESEARCH
OPENSKY NETWORK
FEDERAL AVIATION ADMINISTRATION