**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%.

**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 craft 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.

**Dataset Overview**

- Obtained from Opensky Network.
- Dataset Feature Columns are:
  - icao24
  - lat
  - lon
  - geocoeil
- Overall record points = 60,100 (dataset size).
- Average flight duration (6 flights) – 30 mins to 3 hours.

**Machine Learning Framework and Experiment Results**

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Speed</th>
<th>Height</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.02752316</td>
<td>0.06861313</td>
<td>0.625000018</td>
<td>0.5249491</td>
<td>0.09938746</td>
</tr>
<tr>
<td>2</td>
<td>0.02745436</td>
<td>0.06850885</td>
<td>0.64191088</td>
<td>0.53913108</td>
<td>0.09938952</td>
</tr>
<tr>
<td>3</td>
<td>0.02741683</td>
<td>0.06832009</td>
<td>0.40553296</td>
<td>0.51750597</td>
<td>0.49753589</td>
</tr>
<tr>
<td>4</td>
<td>0.02730005</td>
<td>0.06841767</td>
<td>0.85509115</td>
<td>0.4762244</td>
<td>0.8943182</td>
</tr>
<tr>
<td>5</td>
<td>0.02730533</td>
<td>0.06841767</td>
<td>0.85509115</td>
<td>0.4762244</td>
<td>0.8943182</td>
</tr>
<tr>
<td>6</td>
<td>0.02730005</td>
<td>0.06841767</td>
<td>0.85509115</td>
<td>0.4762244</td>
<td>0.8943182</td>
</tr>
<tr>
<td>7</td>
<td>0.02723389</td>
<td>0.06832009</td>
<td>0.64191088</td>
<td>0.53913108</td>
<td>0.09938952</td>
</tr>
<tr>
<td>8</td>
<td>0.02730005</td>
<td>0.06841767</td>
<td>0.85509115</td>
<td>0.4762244</td>
<td>0.8943182</td>
</tr>
<tr>
<td>9</td>
<td>0.02730005</td>
<td>0.06841767</td>
<td>0.85509115</td>
<td>0.4762244</td>
<td>0.8943182</td>
</tr>
<tr>
<td>10</td>
<td>0.02730005</td>
<td>0.06841767</td>
<td>0.85509115</td>
<td>0.4762244</td>
<td>0.8943182</td>
</tr>
</tbody>
</table>

**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.

**Imputation Methods**

- Bayesian Ridge Regressor
- Random Forest, Adaboost Regressor, Extra Tree Regressor
- k-NN

**Hyperparameter Settings**

<table>
<thead>
<tr>
<th>Name</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Ridge Regressor</td>
<td>-</td>
</tr>
<tr>
<td>Random Forest, Adaboost Regressor, Extra Tree Regressor</td>
<td>n_estimators (10,50,100)</td>
</tr>
<tr>
<td>k-NN</td>
<td>n_neighbors weights (1,2,3,4,5,10,20,30,50) (uniform, distance)</td>
</tr>
</tbody>
</table>

**Conclusions and Future Work**

Experiments were performed for different percentages of missing rates - 10%, 20%, and 30% and 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 geocoeil 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).

**References**