# Data-Driven Methods for Aviation Safety: from Data to Knowledge

Irene Buselli[1], Luca Oneto[2], Carlo Dambra[1],
Christian Verdonk Gallego[3], and Miguel Garcia Martinez[4]

[1] ZenaByte s.r.l., Via Cesarea 2, 16121, Genova, Italy
{irene.buselli,carlo.dambra}@zenabyte.com
[2] University of Genoa, Via Opera Pia 11a, 16145, Genova, Italy
luca.oneto@unige.com
[3] Crida, Juan Ignacio Luca de Tena 14, 28027, Madrid, Spain
{ceverdonk,mgmartinez}@e-crida.enaire.es

**Abstract.** Demand upon the future Air Traffic Management (ATM) systems is expected to grow to possibly exceed available system capacity, pushing forward the need for automation and digitisation to maintain safety while increasing efficiency. This work focuses on a manifestation of ATM safety, the Loss of Separation (LoS), exploiting safety reports and ATM-system data (e.g., flights information, radar tracks, and Air Traffic Control events).
Current research on Data-Driven Models (DDMs) is rarely able to support safety practitioners in the process of investigation of an incident after it happened. Furthermore, integration between different sources of data (i.e., free-text reports and structured ATM data) is almost never exploited.
To fill these gaps, the authors propose (i) to automatically extract information from Safety Reports and (ii) to develop a DDM able to automatically assess if the Pilots or the Air Traffic Controller (ATCo) or both contributed to the incident, as soon as the LoS happens.
The LoSs' reported in the public database of the Comisión de Estudio y Análisis de Notificaciones de Incidentes de Tránsito Aéreo (CEANITA) support the authors' proposal.

**Keywords:** ATM; Safety; Digitisation; Data-Driven Models; Random Forests; Machine Learning; Loss of Separation; Safety Reports.

## 1 Introduction

In the last decades, air transportation has seen a considerable increase in demand, and pressure over Air Traffic Management (ATM) system is predicted to grow to possibly exceed the currently available system capacity [21]. The Single European Sky ATM Masterplan [22] defines the modernisation of the European ATM system as a crucial process to maintain safety while increasing efficiency. A cornerstone of the Masterplan is to further deploy automation and digitisation tools, leading to an integration of human and technical systems [16].

This work is framed within the context of H2020 FARO project - saFety And Resilience guidelines for aviatiOn - which focuses on a manifestation of ATM safety, the loss of separation (LoS). There are two main sources of data which can inform about what happened during a LoS:

- the Safety Reports produced by states' Civil Aviation Authorities and ANSPs after investigating the safety-related events;
- the Automatic Safety Monitoring Tools (ASMTs), which allow the monitoring and recording of safety-related events. These tools are usually augmented with ATM system data, which gather surveillance (e.g., flight tracks) and operational data (e.g., ATC events) [4].

In particular, thanks to the new ASMTs, there are a number of safety-related occurrences - previously unnoticed by the old systems - which can now be identified. As a consequence, probably many more LoSs will need to be investigated and studied in the future. Since human review of incidents is an extensive process, the objective of this work is to provide the ability to partially estimate the results of these investigations timely (i.e., a few minutes after the LoS). This would facilitate the safety practitioners in prioritising the investigations and in understanding potential precursors of these LoS events.

To reach this objective, it is vital to connect the historical information included in the reports with the one contained in the structured data. Manual analysis of safety reports is complex and requires considerable resources, so the approach proposed in this work is twofold:

- First, a simple mining of free-text safety reports is implemented, with the purpose of identifying the information needed to connect reports to structured data;
- Then, an Automatic Contribution Assessment model was developed, able to leverage data to assess if the Pilots or the Air Traffic Controller (ATCo) or both contributed directly to the incident, almost immediately after the LoS and before investigation.

The LoS events considered in this study are those reported in the public database of the Comisión de Estudio y Análisis de Notificaciones de Incidentes de Tránsito Aéreo (CEANITA).

The exploitation of Data-Driven Models (DDMs) in the ATM domain is quite extensive. Indeed, research has focused on a number of different fields, such as taxi-out time prediction [12, 17], trajectory prediction [1, 25], air traffic flow extraction [6, 26], and flight delay prediction [5, 24]. In the safety domain, some relevant applications of DDMs are proposed in literature to predict safety events or performance [7,19], or to provide safety metrics [2] or accident precursors [13]. However, there are very few references aiming at supporting safety practitioners in facilitating the investigation of an incident after it happened but before it is reported [18]. To the best of the authors' knowledge, the approach proposed in this work is a pioneering one in this field.

## 2 Scope of the work

The scope of this work is to develop an Automatic Contribution Assessment model able to leverage ATM-system data to assess whether the Pilots or the ATCos or both contributed to the incident. The model was able to assess contribution well before (i.e., 10 minutes after the incident) human evaluation (usually concluded even weeks after the incident).

## 3 Data Description

Two different data sources were exploited in this study: the CEANITA reports (see Section 3.1) and the structured data from ENAIRE-CRIDA data warehouse, containing contextual information about the LoSs together with ATC events (see Section 3.2).

### 3.1 CEANITA LoS reports

The considered CEANITA LoS reports consist of 70 safety reports, written in Spanish and published by Spanish Safety Aviation Agency (AESA), which is the Spanish Civil Aviation Authority, under the commission of CEANITA, covering safety-related occurrences in the Spanish airspace between January 2018 and July 2019.

Each report is written as a free text and contains the following information:

- *Initial situation*: the initial location and condition of the aircraft involved in the LoS is described with text and images.
- *Communications and radar tracks*: the communications of interest between ATCos and pilots.
- *Conclusions*: the dynamic of the LoS based on the main actions performed by the involved human actors, the main causes, and Pilots and ATCo contribution (classified as direct, indirect or none).

### 3.2 ENAIRE-CRIDA Contextual Information

The structured contextual information was provided by ENAIRE-CRIDA. In particular, they provided high-granularity ATM data such as flight plans, flight tracks, and ATM-processed information about the Spanish airspace. More precisely, two main sources were exploited:

- flight tracks and related contextual flight information (e.g., type, speed, and heading) and
- ATC events of the interactions between ATCos and the Controller Working Position (CWP).

The complete list of features used in this work can be seen in Figure 1.

## 4 Methods

This section presents the methods and tools exploited to achieve the scope of the work (see Section 2) leveraging the data described in Section 3. Data-driven predictive models are able to learn relations between inputs (e.g., ATC events) and outputs (e.g., incident direct contribution) based on a series of examples (i.e., historical data).

In this work, the main model exploited is Random Forests [3], a state-of-the-art solution in the field of Shallow Machine Learning algorithms. Even if, currently, Deep Learning approaches [9] were shown to outperform Shallow Learning models in many tasks, they require a huge amount of data to be trained, which was not available for this research.

Random Forests are one of the most effective approaches in the family of ensemble methods. It is a tree-based ensemble algorithm, combining bagging to random-subset feature selection. In bagging, each tree is independently constructed using a bootstrap sample of the dataset. Random Forests add a further layer of randomness to bagging, also changing how trees are constructed (the best split at each node of the tree is chosen among a subset of predictors randomly sampled at that node). Eventually, a simple majority vote is taken for prediction. Random Forests, while being not too influenced by their hyperparameters [15], still require to tune the number of trees (since the more trees the more accurate the model is, this number is chosen by trading off accuracy and computational complexity), and the number of predictors to be randomly sampled during trees construction.

So, as just described, the data-driven predictive models need to be tuned (by finding the optimal hyperparameters); however, at the same time, their performance needs to be estimated in a rigorous statistical way, in order to estimate their behaviour in production environment. Model Selection and Error Estimation deal exactly with this problem [14]. Resampling techniques like k-fold cross validation and non-parametric bootstrap are commonly exploited solutions, which work well in many situations [14]. These techniques rely on a simple idea: the original dataset is re-sampled once or more, without replacement, to build three independent datasets called learning, validation, and test set. The learning set is exploited to train the model. The validation set is exploited to find the optimal hyperparameters (namely the ones that lead to the optimal performance). The test set is exploited to estimate the performance of the final model (in this way, the test is independent from both the learning and the validation, so results are statically sound and no data snooping is allowed [27]). Performance measures strongly depend on the task to be solved. In this case, dealing with classification problems (see Section 5), Accuracy, Confusion Matrix, Area Under the Receiving Operating Characteristics (AUC), F1 score, Sensitivity, and Specificity are the most commonly used metrics [23].

Once the model is built and has been confirmed to be sufficiently effective, it can be of interest to investigate how this model is affected by the different input features [10, 11]. This procedure is called Feature Ranking and allows the user to detect if the features are appropriately taken into account by the learned models, according to their relevance from the perspective of the domain experts. In particular, Feature Ranking based on Random Forest via Mean Decrease in Accuracy (i.e., the importance of each feature is assessed by randomly permuting the values of the feature and measuring the resulting increase in error) is one of the most effective techniques [8, 20].

## 5 Experimental Results

This section shows how the methods presented in Section 4 were exploited to achieve the scope of the work (see Section 2) demonstrating the effectiveness of the proposed approach on the data described in Section 3. Specifically, Section 5.1 presents the results of automatic information extraction from the CEANITA reports (necessary to connect them with the relative structured data), while Section 5.2 reports the performance of the data-driven model in estimating who directly contributed to the incident before the actual human evaluation.

### 5.1 Automatic information extraction from free text

The information extraction from CEANITA reports was an ancillary process aimed at retrieving two sets of features:

- The features necessary to connect each (anonymised) report with the relative structured data: date, time, and position.
- The contribution of Pilots and ATCo, classified as direct or none in both cases - please note that both Pilots and ATCo may have directly contributed to the incident. In particular, contribution was assessed as direct in the 36% of cases for Pilots and in the 72% of cases for ATCo's.

This task was performed through a rule-based procedure, based on the automated search of keywords, characters, and punctuation signs. The connection with the ENAIRE-CRIDA data was successful and enabled the subsequent development of the Automatic Contribution Assessment.

### 5.2 Automatic Contribution Assessment

After the preliminary extraction of information (Section 5.1), a DDM was exploited to assess agents' contribution before (i.e., 10 minutes after the incident) human evaluation (which is a post-operation activity) based on the automatic analysis of ATC events and other contextual data (i.e., radar tracks of the aircraft and flight information). Furthermore, the analysis shows that this predictive model actually captured meaningful relations and not just spurious correlations from the data (see Section 4).

Specifically, for each incident, the goal was to predict:

- the Pilots' contribution, i.e., classified as direct or not and
- the ATCo's contribution, i.e., classified as direct or not.

The prediction was based on a total of 19 features covering:

- the flight type,
- the flight rule at the moment of the incident,
- the flight level at the moment of the incident,
- the airspace class at the moment of the incident, and
- for each of the 15 classes of ATC events recorded from 30 minutes before to 10 minutes after the incident, their number of occurrences. Considering this time window is fundamental since the contributions of ATCo and Pilots depend both on what was done to prevent the potential LoS and on how it was managed when it became an actual LoS;

Different white-box and grey-box models were tested on this problem (i.e., Decision Trees, Logistic Regression, and Random Forests). The choice of not testing black-box models was due to the necessity of identifying, at least partially, the underlying process, in order to provide safety practitioners with potential precursors and to verify how the model is affected by the different features. In the end, a Random Forest model was selected (see Section 4), as it was proven to outperform the other ones.

The model was trained on the 70 incidents for which recorded ATC events were available (the number of trees was set to 1000 and the number of predictors

to be randomly sampled during trees construction was searched in $\{5, 6, 7, 8, 9\}$ according to what was described in Section 4). Random Forests facilitate the generation of different optimal models changing the cut-off of the voting (i.e., how many trees need to agree to decide for a particular class). By doing so, it was possible to report different models, maximising respectively: the AUC, the F1 score, the Sensitivity, and the Specificity. Moreover, Random Forests provide the confidence of the prediction: this allows the user to trust the model only when its confidence is higher than a certain threshold.

Table 1 reports the confusion matrices of the developed predictive models (maximising AUC, F1 score, Sensitivity, and Specificity) for both ATCos' and Pilots' contributions.

Table 2, instead, reports the confusion matrices of the predictive models (maximising the AUC, since they appeared to be the most balanced ones) when predictions are considered only if their confidence is higher than 60% and 75%.

Table 1 shows that:

- When the AUC is maximised (i.e., assuming the user wants a balanced accuracy on both "Yes" and "No" classes), accuracy reaches ≈75% for Pilots contribution and ≈81% for ATCo; F1 score is ≈70% for Pilots and ≈86% for ATCo.
- When the F1 score is maximised (i.e.,assuming the user wants to maximise the accuracy for the "Yes" class without too many false positives), accuracy reaches ≈73% for Pilots contribution and ≈85% for ATCo; F1 score is ≈70% for Pilots and ≈91% for ATCo.
- When the Sensitivity is maximised, (i.e., assuming the user wants to be as sure as possible that if the Pilots/ATCo contribute to the LoS the algorithm classifies it as "Yes") the level of sensitivity reached is ≈100% for Pilots, with ≈70% of accuracy, and ≈98% for ATCo, with ≈85% of accuracy; F1 score is ≈70% for Pilots and ≈91% for ATCos.
- When the Specificity is maximised (i.e., assuming the user wants to be as sure as possible that if the Pilots/ATCo are not responsible, the algorithm classifies it as "No") the level of specificity reached is ≈100% for Pilots, with ≈80% of accuracy, and ≈96% for ATCo, at the price of a low accuracy, ≈54%. F1 score is ≈60% for Pilots and ≈56% for ATCos.

Furthermore, Table 2 shows that:

- When just predictions with confidence ≥75% are considered, the accuracy reaches ≈97% for Pilots contribution and ≈94% for ATCo. With this threshold, only 43% of the predictions are trusted when assessing Pilots contribution and 44% when considering the ATCo.
- When, instead, the accepted confidence level is decreased from 75% to 60%, the accuracy reaches ≈86% for both Pilots and ACTo contributions. With this new confidence level, 62% of observations are classified when assessing Pilots contribution and 70% when considering ATCo.

Finally, the ranking of the features (see Section 4) produced by the Random Forest algorithm is presented in Figure 1 to better understand what the predictive model actually learned from the data.

Figure 1 allows us to observe that, based on the experience of the domain experts, the models learned correctly the importance of features related to the separation responsibility, such as the Flight type, the Flight rules, or the Airspace Class. In addition, the models learned correctly the relevance of interactions

**Table 1:** Confusion matrices of the developed predictive models of contribution based on the ATC events (maximising AUC, F1 score, Sensitivity, and Specificity) for both ATCo and Pilots contributions.

**(a)** Pilots Contribution (Maximising AUC)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 46.0±0.3 | 18.3±0.3 |
| | Yes | 6.9±0.3 | 28.8±0.3 |

**(b)** ATCo Contribution (Maximising AUC)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 20.8±0.3 | 4.9±0.3 |
| | Yes | 14.1±0.3 | 60.2±0.3 |

**(c)** Pilots Contribution (Maximising F1 score)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 41.7±0.3 | 22.6±0.3 |
| | Yes | 4.0±0.3 | 31.7±0.3 |

**(d)** ATCo Contribution (Maximising F1 score)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 12.1±0.3 | 13.6±0.3 |
| | Yes | 1.6±0.3 | 72.7±0.3 |

**(e)** Pilots Contribution (Maximising Sensitivity)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 33.9±0.3 | 30.4±0.3 |
| | Yes | 0.1±0.2 | 35.6±0.2 |

**(f)** ATCo Contribution (Maximising Sensitivity)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 11.9±0.3 | 13.8±0.3 |
| | Yes | 1.4±0.2 | 72.9±0.2 |

**(g)** Pilots Contribution (Maximising Specificity)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 64.3±0.0 | 0.0±0.0 |
| | Yes | 20.3±0.2 | 15.4±0.2 |

**(h)** ATCo Contribution (Maximising Specificity)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 24.9±0.2 | 0.8±0.2 |
| | Yes | 45.2±0.3 | 29.1±0.3 |

between the ATC and the CWP, such as Radar Contact, ETO Over Fix or Action on Flight Level, in order to identify ATM contributions. These are promising results as the model presents room for improvement, such as the inclusion of more surveillance information or operational indicators such as traffic load.

**Table 2:** Confusion matrices of the developed predictive models based on the ATC events (maximising AUC) for both ATCo and Pilots contributions when predictions are trusted only if their confidence is higher than 60% and 75%.

**(a)** Pilots contribution (Confidence ≥60%)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 58.1±0.3 | 9.3±0.3 |
| | Yes | 4.7±0.3 | 27.9±0.3 |

**(b)** ATCo contribution (Confidence ≥60%)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 20.4±0.3 | 6.1±0.3 |
| | Yes | 8.2±0.3 | 65.3±0.3 |

**(c)** Pilots contribution (Confidence ≥75%)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 60.0±0.0 | 00.0±0.0 |
| | Yes | 3.3±0.1 | 36.7±0.1 |

**(d)** ATCo contribution (Confidence ≥75%)

| | | Pred. No | Pred. Yes |
|---|---|---|---|
| Truth | No | 29.0±0.1 | 3.2±0.1 |
| | Yes | 3.2±0.2 | 64.6±0.2 |

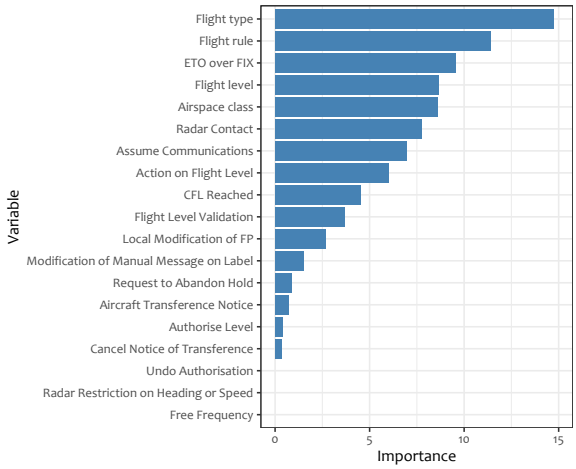


**Fig. 1:** Average variable importance ranking of the models (metric: mean decrease in accuracy).

## 6 Conclusions

The objective of this work was to facilitate the automatic extraction of meaningful and actionable information from LoS reports and to investigate how the information recorded by the systems can help estimating contribution assessment. For this purpose, the authors proposed a twofold approach based on (i) an automatic extraction of quantitative features from free text and (ii) an automatic contribution assessment model based solely on the information recorded by the systems and available a few minutes after the ASMTs' identification of the LoS. The approach was tested on the LoSs reported in the CEANITA public database and the related ATC events.

Different performance metrics were considered to evaluate the validity of the result. In particular, the results show that when only high-confidence predictions are considered, the model output reaches approximately 97% of accuracy for pilots' contribution and 94% for ATCo.

Future work could validate these techniques on other databases of reports (e.g., UKAB AirProx Board, NTSF Board, etc.). Moreover, integrating other sources of structured data (e.g., about weather phenomena, STCA or TCAS activation, or traffic load) to develop richer models could lead to further insights in the estimation of contributors and precursors.

## Acknowledgements

This project has received funding from the SESAR Joint Undertaking (JU) trough EU-H2020-ICT Project FARO - saFety And Resilience guidelines for aviatiOn (G.A. 892542). The dissemination reflects only the authors' view and the SJU is not responsible for any use that may be made of the information it contains.

## References

1. Ayhan, S., Samet, H.: Aircraft trajectory prediction made easy with predictive analytics. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
2. Bati, F., Withington, L.: Application of machine learning for aviation safety risk metric. In: IEEE/AIAA Digital Avionics Systems Conference (2019)
3. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
4. CANSO: Incidents investigation toolbox. `https://canso.fra1.digitaloceanspaces.com/uploads/2021/04/CANSO-Incidents-Investigation-Toolbox.pdf` (2021)
5. Choi, S., Kim, Y.J., Briceno, S., Mavris, D.: Prediction of weather-induced airline delays based on machine learning algorithms. In: IEEE/AIAA Digital Avionics Systems Conference (2016)
6. Conde Rocha Murca, M., DeLaura, R., Hansman, R.J., Jordan, R., Reynolds, T., Balakrishnan, H.: Trajectory clustering and classification for characterization of air traffic flows. In: AIAA Aviation Technology, Integration, and Operations Conference (2016)
7. Di Gravio, G., Mancini, M., Patriarca, R., Costantino, F.: Overall safety performance of Air Traffic Management system: Forecasting and monitoring. Safety science **72**, 351–362 (2015)

8. Genuer, R., Poggi, J.M., Tuleau-Malot, C.: Variable selection using random forests. Pattern recognition letters **31**(14), 2225–2236 (2010)
9. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning. MIT press Cambridge (2016)
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys **51**(5), 1–42 (2018)
11. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research **3**(Mar), 1157–1182 (2003)
12. Lee, H., Malik, W., Jung, Y.C.: Taxi-out time prediction for departures at Charlotte airport using machine learning techniques. In: AIAA Aviation Technology, Integration, and Operations Conference (2016)
13. Nazeri, Z., Barbara, D., De Jong, K., Donohue, G., Sherry, L.: Contrast-set mining of aircraft accidents and incidents. In: Industrial Conference on Data Mining (2008)
14. Oneto, L.: Model Selection and Error Estimation in a Nutshell. Springer (2020)
15. Orlandi, I., Oneto, L., Anguita, D.: Random forests model selection. In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) (2016)
16. Performance Review Comission, EUROCONTROL: Performance review report. an assessment of Air Traffic Management in Europe during the calendar year 2019. `https://www.eurocontrol.int/sites/default/files/2020-06/eurocontrol-prr-2019.pdf` (2020)
17. Ravizza, S., Chen, J., Atkin, J.A.D., Stewart, P., Burke, E.K.: Aircraft taxi time prediction: comparisons and insights. Applied Soft Computing **14**, 397–406 (2014)
18. Robinson, S.D., Irwin, W.J., Kelly, T.K., Wu, X.O.: Application of machine learning to mapping primary causal factors in self reported safety narratives. Safety science **75**, 118–129 (2015)
19. Rodríguez-Sanz, Á., Gómez, F., García, J.M.C., Meler, L.: Analysis of saturation at the airport-airspace integrated operations. In: USA/Europe Air Traffic Management Research and Development Seminar (2017)
20. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Machine Learning and Knowledge Discovery in Databases. pp. 313–325 (2008)
21. SESAR Joint Undertaking: European ATM master plan - executive view, 2015 edition. `https://www.sesarju.eu/node/2865` (2015)
22. SESAR Joint Undertaking: European ATM master plan - executive view, 2020 edition. `https://op.europa.eu/en/publication-detail/-/publication/8afa1ad9-aac4-11ea-bb7a-01aa75ed71a1` (2020)
23. Shalev-Shwartz, S., Ben-David, S.: Understanding machine learning: From theory to algorithms. Cambridge university press (2014)
24. Takeichi, N., Kaida, R., Shimomura, A., Yamauchi, T.: Prediction of delay due to air traffic control by machine learning. In: AIAA Modeling and Simulation Technologies Conference (2017)
25. Verdonk Gallego, C.E., Gómez Comendador, V.F., Amaro Carmona, M.A., Arnaldo Valdés, R.M., Sáez Nieto, F.G., García Martínez, M.: A machine learning approach to air traffic interdependency modelling and its application to trajectory prediction. Transportation Research Part C: Emerging Technologies **107**, 356–386 (2019)
26. Verdonk Gallego, C.E., Gómez Comendador, V.F., Saez Nieto, F.J., García Martinez, M.: Discussion on density-based clustering methods applied for automated identification of airspace flows. In: IEEE/AIAA Digital Avionics Systems Conference (2018)
27. White, H.: A reality check for data snooping. Econometrica **68**(5), 1097–1126 (2000)