Advanced Statistical Learning and Prediction of Complex Runway Incursion

I. Song[1]; I. Cho, Ph.D., M.ASCE [2*]; T. Tessitore[3]; T. Gurcsik[4]; and H. Ceylan, Ph.D., M.ASCE[5]

[1]Dept. of Civil, Construction and Environmental Engineering, Iowa State Univ., Ames, IA 50010. E-mail: isong@iastate.edu
[2]Dept. of Civil, Construction and Environmental Engineering, Iowa State Univ., Ames, IA 50010. E-mail: icho@iastate.edu; (* corresponding author)
[3]Aviation Research Division, FAA William J Hughes Technical Center, Atlantic City, NJ 08405. E-mail: tom.tessitore@faa.gov
[4]Aviation Research Division, FAA William J Hughes Technical Center, Atlantic City, NJ 08405. E-mail: tony.gurcsik@faa.gov
[5]Dept. of Civil, Construction and Environmental Engineering, Iowa State Univ., Ames, IA 50010. E-mail: hceylan@iastate.edu

Abstract

In 2015, 1,507 runway incursions capable of inducing collisions occurred at airports in the United States, so it is obviously very important to identify significant factors underlying such incursions, to predict potential runway incursion occurrences, and to prepare systematic programs for reducing the number of incursions and prevent runway collisions. Presence of a large volume of data, multiple variables, and complex interactions among them pose a significant challenge to resolving this problem. To tackle this challenge, we developed a data-driven prediction model using a component of advanced statistical theory, i.e., a generalized additive model (GAM). GAM can account for flexible modeling of multiple variables over a broad range of modeling distributions. We obtained, parsed, and transformed various predictor variables from many heterogeneous databases to create interpretable datasets for statistical modeling. We demonstrated promising performance of GAM while making systematic investigations into prediction accuracy of runway incursion at United States airports (including all types of commercial, military, and other general data). Results show that GAM can identify critical factors (airport complexity, number of operations, and visibility) in predicting a number of the runway incursions. Performance comparison of two popular GAM smoothers (i.e., cubic regression splines and thin plate regression splines) has demonstrated promising accuracy of both methods. These results imply that statistical predictions developed using GAM will help in better prediction of runway incursion when more data become available in the future.

INTRODUCTION

In 2015, 1,507 runway incursions (RIs) occurred at airports in the United States (FAA, 2016), and such events can lead to runway collisions., There have been practical efforts to address this problem and solve the RI issue, e.g., Direct alerting to the cockpit (DAC) (Ludwig, 2007), airport movement area safety system (AMASS) (Watnick and Ianniello, 1992), RI alert system (Jones et al., 2001), and RI prevention system (Schönefeld and Möller, 2012). Despite these efforts, the occurrence of runway incursion is reported to have increased in almost every year (FAA, 2015). In view of significant role of airport in US industries, development of reliable methods for prediction of RIs and solutions is imperative.

While some factors identified by previous research studies and Federal Aviation Administration (FAA, 2008) are: poor weather, low visibility, time of day, miscommunication with air traffic control (ATC), etc., it is difficult to elucidate the quantitative relevance and relative importance of such factors to RI. Prediction of future RI occurrence is more difficult because of many complex interrelations among contributing factors. Simple statistical methods could present a straightforward solution, but typical linear or nonlinear regression methods appear to have been unsuccessful, probably because of complex nonlinearity of factors (Figure 1). On the other hand, while machine learning-based approaches could be a successful remedy (Karlaftis and Vlahogianni, 2011), the small database of RI and lack of causal pathways among factors pose challenges to direct adoption of machine learning. To overcome the so-called “curse of dimensionality” the present study lacks a sufficient dataset, i.e., for some machine learning algorithms the current dataset is still insufficient for training, validation, and testing (Baesens, 2014), so to establish a foundation for data-driven solutions to the RI problem, this study focuses on advanced statistical learning and prediction with emphasis on comprehensibility of the RI database.

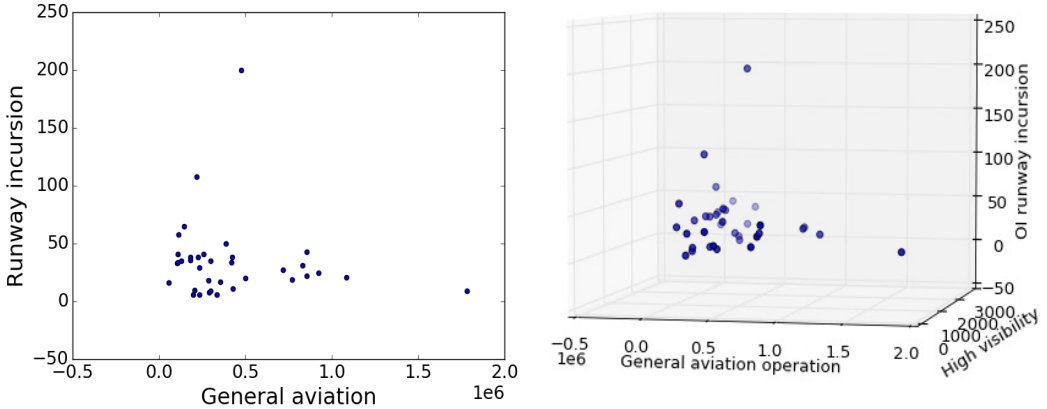


Figure 1. Scatter plot of variables: (a) runway incursion versus general aviation operation; (b) runway incursion versus general aviation operation and high visibility.

We adopted one of the most advanced and flexible statistical methods, the generalized additive model (GAM). GAM is a non-parametric statistical model developed by Hastie and Tibshirani (1990); it is highly flexible, being capable of embracing a large number of variables with substantial nonlinearity. GAM can cover a wide range of statistical distributions, and these favorable attributes of GAM are expected to enable us to learn and predict the future RI database and to make RI data more comprehendable. The detailed theory and advantages of GAM will be described in a later section.

A key challenge is the dispersed location of the RI database, i.e., key data pertaining to primary factors of RI are not located in a single location. We collected data from different databases, developed programs to extract required information from raw data, and transformed it into a suitable form for inclusion in the dataset. Another issue is computational cost that can be attributed to a number of factor variables. To determine key factors contributing to RI, multiple loop simulation is necessary, and this represents a very heavy computational load. We used a parallel strategy to solve this issue, and this will be described in a later section. The overall workflow is shown in Figure 2.

The outline of the paper is as follows: we address data structures used for building the RI dataset and GAM-based predictions. The central algorithms regarding how to collect, extract, and transform the raw data required for GAM modeling are then presented, followed by remarks related to a parallel strategy for finding the best combination of predictor variables.

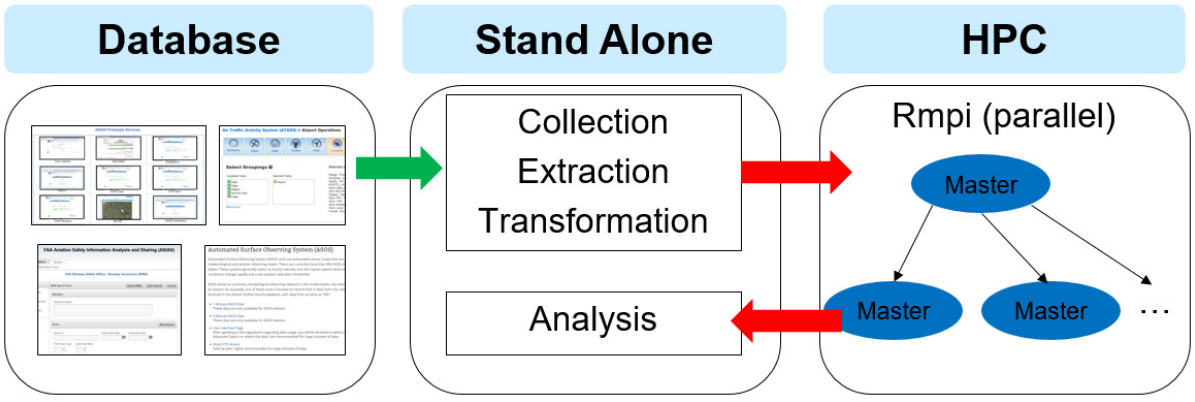


Figure 2. Workflow of runway incursion prediction using GAM: raw data is collected from various databases and transformed into suitable dataset form of for GAM modeling. Thereafter, the GAM is run on the dataset using a high-performance computing (HPC) system. Finally, runway incursion occurrence is predicted.

DATA STRUCTURES

To facilitate prediction of RIs, the primary data are classified into three categories: (1) geometric information, (2) operational data, and (3) visibility data. An airport runway is a long stretch of pavement at an airport from which an aircraft can take off and land. The aviation safety information analysis and sharing (ASIAS) system developed by the FAA provides a wide range of data regarding safety. In this study, spatial and geometric information from 36 airports was obtained from the ASIAS system. Using both spatial and geometric data, numbers of runways, intersections between runways, and intersections between runways and taxiways were obtained by parsing XML data. Operational data related to aircraft in an airport are also important, and to collect and extract such data, we leveraged the air traffic activity system (ATADS). ATADS provides all activity information related to air traffic, including airport operation, tower operation, terminal operation, and so on. The data obtained includes airport name and operational history of air carriers, air taxis, general aviation, and military aviation over the past 15 years (from 2001 to 2015) from the major 36 airports. Visibility data was obtained from an automated surface-observing system (ASOS) developed as a joint work by the National Weather Service (NWS) that is a component of the National Oceanic and Atmospheric Administration (NOAA), the FAA, and the Department of Defense (DoD). The NOAA is a government agency that provides extensive information about weather, climate, and the ocean. ASOS provides meteorological and climatological observation measures from more than 900 ASOS sites, covering all major airports in the United States. Data can be obtained in the form of 1- , 5-minute, or 1-hour intervals. Visibility is described through three potential impact factors (i.e., slight, moderate, and high). The hours describing such factors were counted for 15 years.

Finally, the RI data are obtained from ASIAS, providing comprehensive information about RI from most airports. Three different types of runway incursion are considered: (1) pilot deviation (PD), (2) operational incident (OI), and (3) vehicle (driver) deviation (VD). A PD is defined as an incursion committed by a pilot of aircraft (e.g., by landing or taking off without clearance from ATC); An OI is associated with an ATC error (e.g., clearance of an aircraft onto a runway while another aircraft is on the runway); A VD is associated with passing a runway holding mark without ATC clearance (FAA, 2008). A summary of this data is given in Table 1. Here, “predictor” is defined as the factor used for GAM-based predictions of RI occurrence.

Table 1. Summary of predictors for GAM model

Data	Predictors	Types	Sources
Geometric information	Runway, intersection between runways, intersection between runway and taxiway	Count (integer)	ASIAS
Operation data	Air carrier, air taxi, general aviation, military aviation, total aviation	Count (integer)	ATADS
Visibility	High impact, moderate impact, slight impact	Hour (integer)	ASOS

DATA EXTRACTION AND TRANSFORMATION

Because of the dispersed nature of database locations, this study first investigated multiple heterogeneous databases to obtain the data that required in building a GAM model. We collected raw data from different databases, extracting only the required parts, and transformed them into a form suitable for the GAM.

First, the geometric data was obtained in the form of an XML file from AVIAS. An XML file contains polygon information related to runway, taxiway, and other structures in an airport, and the file contains coordinates of points (i.e., x and y coordinates) e connected to on another, making polygonal lines. We counted the number of runways, intersections between runways, and intersections between runways and taxiways based on the number of a keyword in a tag. For example, a tag with <Runway name="35L" id="8"> in the XML file means the polygon information about a new runway would be within the tag. We searched for the keyword “Runway” and counted it as the runway number whenever our program found it. Second, the operational information was downloaded from the ATADS in spreadsheet formt. We directly downloaded operational information from 36 airports over a 15-year interval and, thanks to various available download options of ATADS, further parsing process was not necessary.

Third, visibility information was the most difficult to obtain because it required multiple processing steps. A number of raw data files were downloaded from the NOAA file transfer protocol (FTP) sever (FAA) and they were then transformed into more interpretable form by using the JAVA program provided by NOAA. It should be noted that the time frame of weather data from NOAA is also used to describe each incursion incident (the generated dataset is available upon request). The transformed data includes United States Air Force (UASF) codes so the airports can be identified using these codes. The data contains information at one-hour intervals, including visibility, presented in mile units. The program counts hours of slight, moderate, and high visibility of 36 airports for the 15 years based on the meteorological terminal aviation routine weather report (METAR)] board FAA, (Table 2).

Table 2. Visibility criteria based on METAR board

Potential impact (mile)				
Threat	None	Slight	Moderate	High
Visibility	≥ 5.1	$5.1 > X \geq 3$	$3 > X \geq 1$	< 1

SUMMARY OF GAM

The generalized additive model (GAM) is a generalized linear model with substantial flexibility and general applicability. Rather than using pre-defined distributions or parameters, GAM is composed of multiple unspecified smoothing functions. Because of the nature of these unspecified smoothing functions, covariates do not need to have a set of parameters. For predicting RI occurrence of i^{th} airport (denoted by $Y_i \in \mathbb{R}$) with n predictors (denoted by $\boldsymbol{x}_i \in \mathbb{R}^n$), the general form of GAM can be represented as:

$$g(\mu_i) = f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + \cdots,$$

where g is a smooth link function; the expectation $\mu_i \equiv \mathbb{E}(Y_i|\boldsymbol{x}_i)$; Y_i is from some exponential family of distribution (e.g., normal, binomial, or gamma distribution); f_j are smooth functions of covariates x_{ji} (Wood, 2006). In particular, Y_i would be the number of RI at the *i*th airport and $\boldsymbol{x}_i$ represents of the numbers of runways, visibility, etc. In essence, GAM has a non-specified smoothing function per each predictor, and this fact imparts substantial flexibility to GAM. For brevity of explanation, the following description involves only a single covariate and normal distribution, but generalization to multiple variables is straightforward. Let the GAM be $\mathbb{E}(\mathrm{Y}|x) = f(x)$, then the smoothing function f can be represented as:

$$f(x) = \sum_{j=1}^{q} b_j(x)\beta_j,$$

where $b_j(x)$ is the j^{th} basis function and β_j is an unknown parameter. Model fitting can be done by maximizing the corresponding likelihood with a penalty term given as:

$$\lambda \int [f''(x)]^2 dx,$$

where λ is a *smoothing parameter*. λ is internally optimized by GAM to balance smoothness of regression and accuracy of prediction. The optimized λ value can be chosen in such a way to make the model fit accurate by minimizing generalized cross validation (GCV) scores (Golub et al., 1979).

There are two popular types of basis functions: (a) thin-plate regression splines (TPRS) (Wood, 2003) and (b) cubic regression spline (CRS) (Wood, 2006). A cubic spline is a curve formed by connecting a number of cubic polynomial sections (Gu, 2013). CRS is one of the smoothest interpolators but it requires "knot" location selection to connect disjoint cubic splines. In contrast, TPRS can be used for any number of covariates and is "knot-free", requiring no knot location selection

(Duchon, 1977). In general, the computational cost of TPRS is more expensive than that if CRS. It is instructive to note how the TPRS spans multi-dimensional variable space. In the TPRS, the $\boldsymbol{f}$ is determined by minimizing

$$\|\boldsymbol{y} - \boldsymbol{f}\|^2 + \lambda J_{md}(\mathrm{f}) \quad (1)$$

where $\boldsymbol{y}$ is the vector of y_i data and $\boldsymbol{f} = [f(x_1), f(x_2), \cdots, f(x_n)]^{\mathrm{T}}$. $J_{md}(\mathrm{f})$ is a penalty functional measuring the 'wiggliness' of $\boldsymbol{f}$, and controlled by a tradeoff between data-fitting accuracy and smoothness. One example of a thin-plate spline basis function with 2 covariates is shown in Figure 3, showing that TPRS can span multi-dimensional variable space with a smooth "thin" plate, thereby offering great flexibility.

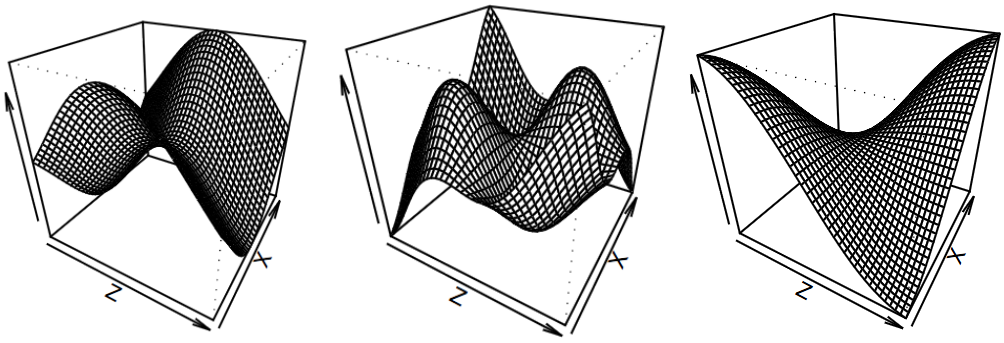


Figure 3. Example of thin-plate spline basis function using 2 covariates (cited from Wood, 2006).

METHOD AND METRICS FOR PREDICTION ACCURACY

In this study, three metrics were used to compare the GAM-based prediction performance: (1) $\mathrm{CVE_b}$/CVE = the ratio between base cross-validation error ($\mathrm{CVE_b}$) and cross-validation error (CVE); (2) the Pearson correlation, ρ; (3) the coefficient of determination, R^2. CVE and $\mathrm{CVE_b}$ are defined as

$$\mathrm{CVE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_{ex}^{i} - y_{pr}^{i}\right)^2;\ \mathrm{CVE}_b = \frac{1}{N}\sum_{i=1}^{N}\left(y_{ex}^{i} - y_{mean,pr}\right)^2$$

where N is number of data points, y_{ex}^{i} is the i^{th} real-world measured response, y_{pr}^{i} is the i^{th} predicted response in the cross-validation procedure, and $y_{mean,pr}$ is the mean of predicted values. ρ and R^2 are defined as

$$\rho = \frac{COV(y_{pr}, y_{ex})}{\sigma_{y_{pr}} \times \sigma_{y_{ex}}};\quad R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_{ex}^{i} - y_{pr}^{i}\right)^2}{\sum_{i=1}^{N}\left(y_{ex}^{i} - y_{mean,pr}\right)^2}$$

This choice has been made following a comparable study on machine-learning comparisons of Kamdar et al. (2016); In essence, the higher the metrics, the more

accurate the predictions. Throughout this study, we seek to achieve the GAM model that exhibits the highest scores of these metrics.

SELECTION OF BEST GAM MODEL USING PARALLEL COMPUTING

GAM can be built upon arbitrary combinations of many predictors. A prudent choice of predictors is critical for accurate GAM modeling. To avoid artificial bias in the selection of predictors, this study objectively refers to the aforementioned three metrics of prediction performance (CVE_b/CVE, ρ and R^2), and seeks to find the best combination of predictors. We depart from all possible combinations of predictors. In total, 14 variables are taken from raw data without any prejudice related to relations or *a priori* knowledge with respect to the relative significance of predictors. The 14 variables are: number of runways, number of intersections of runways, number of intersections of runway and taxiway, air carrier operation, air taxi operation, general aviation operation, military operation, total operation, average visibility, low-visibility hours, moderate-visibility hours, high-visibility hours, sum of high and moderate visibility hours, and sum of all visibility hours. The prediction target response is the number of RI occurrence.

The proposed approach for searching for the best predictor combination is straightforward, yet computationally expensive, viz., it involves the total number of combinations of 7 variables selected from 14 total variables = $\frac{14!}{7!(14-7)!}$ = 3,432. To reduce the computation time we developed an algorithm-oriented parallel computing algorithm using *Rmpi* Lu et al., 2013. It should be noted that, while feature-space reduction algorithms such as Principal Component Analysis (PCA) (Jolliffe, 2002) might also be an efficient remedy, this would necessitate additional tasks such as axis rotation of multivariate space and selection process, and such extensions will thus be considered in a future research study. The *Rmpi* is controlled by only one master, but a number of slaves can be spawned. Since the computation load decreases as the size of interwoven loops decreases, so-called "cyclic allocation" of tasks is used to ensure load balance on the slave processors. With cyclic allocation, a successful parallelization can be achieved by cyclically allocating jobs to available slaves. It has been shown that, as the problem size increases, the cyclic allocation approaches the optimal parallel load balancing (Kam et al., 2011). Finally, we performed tests on our parallel algorithm, producing the results summarized in Figure 4. The best speedup was achieved with 56 slaves.

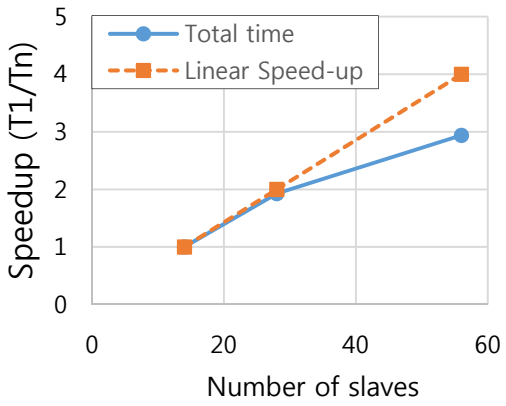
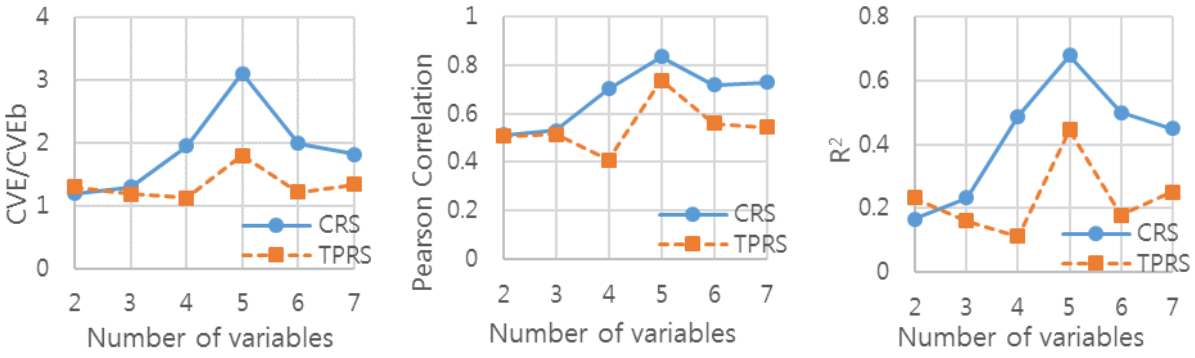


Figure 4. Parallel computing performance of *R & Rmpi* code for finding the 7-variable combination out of 3,432 total combinations.

By leveraging the harmonious use of parallel computing and an advanced statistical program, this study obtained the best combination consisting of five predictors: (1) number of taxi operations, (2) number of general operations, (3) hours of high impact visibility, (4) hours of slight impact visibility, and (5) sum of hours of high, moderate, and slight impact visibility. Table 3 summarizes the obtained metrics of the best combination of predictors and Figure 5 shows the performance comparison between CRS and TPRS on this study. CRS appears to perform better than TPRS based on the metrics values

Table 3. Metrics used for the best combination of predictor variables (GAM-CRS)

Number of variables	CVE_b/CVE	Correlation of determination	Coefficient of determination
2	1.2	0.512	0.1667
3	1.302	0.529	0.232
4	1.952	0.701	0.488
5	**3.115**	**0.835**	**0.679**
6	1.995	0.719	0.499
7	1.818	0.729	0.45



(a) **(b)** **(c)**

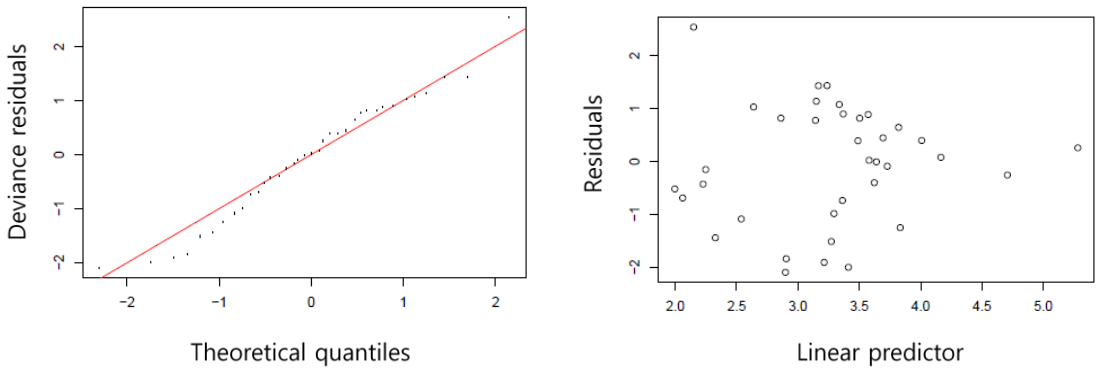
Figure 5. Comparison of performance between CRS and TPRS on this study: (a) ratio of CVE_b/CVE; (b) Pearson correlation; (c) coefficient of determination.

PREDICTION RESULTS

Per the least requirements of GAM, we chose a *logarithmic* link function that can easily incorporate the multiplicative relations of the engineering variables. Since all the quantified predictors and responses are positive and represented as counts (i.e., integer), a Poisson distribution is assumed. As recommended by Wood (2006), the parameter k (i.e. the number of basis dimensions in smooth functions) is set to 6; the smoothing parameter λ is readily optimized by the library of R that utilizes GCV.

To systematically evaluate the prediction capability, cross validations were applied. The prediction process mainly followed three steps: (1) exclusion of an airport, (2) construction of a GAM by learning the remaining airport data, and (3) prediction of runway incursion at the omitted airport. To construct the GAM, one airport is excluded while learning samples (i.e., other airport data) are used during cross validation. Thereafter, a series of runway incursions at the excluded airport is predicted using the GAM. These steps are repeated throughout all airport data. The difference between the predicted number of runway incursions from GAM and the original actual value for the excluded airport directly represents how precisely the constructed GAM can predict the target response.

To demonstrate the prediction results, so-called Q-Q plots were drawn to correlate the scaled response of real measured and predicted values (Figure 6a; a straight line corresponds to accurate prediction). Note that all statistical predictions in Figure 6 are drawn for the best GAM model that only uses 5 predictors. Remarkably, the predicted responses show good correlation with real-world measured data even though there was no prejudice with respect to the statistical models. Figure 6b shows that residuals are scattered evenly and the proposed model fitting thus appears to be acceptable.



(a) (b)

Figure 6. (a) Q-Q plot of real-world measured data and predicted data; (b) Residuals plot showing that residuals are evenly scattered.

CONCLUSION

In this study, we applied a generalized additive model (GAM) that can facilitate solely data-driven studies to predict a valuable target value in the aviation field. The datasets were retrieved from databases containing various factors that appear to be closely tied to runway incursion. All the processed data of the 36 target airports will be made available upon request. The proposed evaluation process of the prediction models using GAM shows promising applicability of such advanced statistical approaches to aviation data. Notably, accurate predictions were made without any prejudices with respect to relationships or *a priori* knowledge of the raw data. Our method and results suggest that all variables were not always necessary for making the best prediction, implying that there exist significant relationships among a few manageable factors that appear to govern RIs. In future extensions, with increases in accessibility and in the number of aviation databases (e.g., national transportation safety board, aviation safety reporting system, etc.), another validation, using additional airports' data and variables such as angle and number of taxiways merging to an interaction and visibility of markings, can be addressed in addition to a cross validation procedure. Such future extensions with new data and independent validations should be straightforward since the proposed framework establishes comprehensive procedures from data gathering, processing, learning, prediction, and validation. The proposed statistical learning and prediction approaches will thus complement new data-driven discoveries in the aviation field and also facilitate machine learning-based approaches.

ACKNOWLEDGEMENT

This study is supported by the Partnership to Enhance General Aviation Safety, Accessibility and Sustainability (PEGASAS) Center of Excellence (COE) fellowship program of the Federal Aviation Administration (FAA). Regarding the data acquisition and working environment, the generous support of the FAA technical center is appreciated.

REFERENCES

Baesens, B. (2014). *Analytics in a big data world: The essential guide to data science and its applications.* John Wiley & Sons.

Duchon, J. (1977). *Splines minimizing rotation-invariant semi-norms in Sobolev spaces. Constructive theory of functions of several variables* (pp. 85-100) Springer.

FAA. "METAR board." <https://www.aviationweather.gov/metar/help?page=board> (July 15, 2016)

FAA. "NOAA FTP server." <ftp://ftp.ncdc.noaa.gov/pub/data/noaa/> (July 15, 2016)

FAA. (2008). *Pilot's Handbook of Aeronautical Knowledge.* US Department of Transportation-Federal Aviation Administration-Flight Standards Service, Oklahoma City, OK, USA.

FAA. (2015). *National runway safety report 2013-2014*.

FAA. (2016). "Runway Incursion Totals by quarter FY2016 vs. FY2015." <https://www.faa.gov/airports/runway_safety/statistics/year/?fy1=2016&fy2=2015>

Golub, G. H., Heath, M., and Wahba, G. (1979). "Generalized cross-validation as a method for choosing a good ridge parameter." *Technometrics,* 21(2), 215-223.

Gu, C. (2013). *Smoothing spline ANOVA models.*(Vol. 297) Springer Science & Business Media.

Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized additive models.*(Vol. 43) CRC Press.

Jolliffe, I. (2002). *Principal component analysis.* Wiley Online Library.

,Jones, D. R., Quach, C. C., and Young, S. D. (2001). "*Runway incursion prevention system-demonstration and testing at the dallas/fort worth international airport".* Proc., Digital Avionics Systems, 2001. DASC. 20th Conference.

Kam, W. Y., Pampanin, S., and Elwood, K. (2011). "Seismic performance of reinforced concrete buildings in the 22 February Christchurch (Lyttelton) earthquake." *Bulletin of the New Zealand Society for Earthquake Engineering,* 44(4), 239-278.

Kamdar, H. M., Turk, M. J., and Brunner, R. J. (2016). "Machine learning and cosmological simulations–I. Semi-analytical models." *Monthly Notices of the Royal Astronomical Society,* 455(1), 642-658.

Karlaftis, M. G., and Vlahogianni, E. I. (2011). "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights." *Transportation Research Part C: Emerging Technologies,* 19(3), 387-399.

Lu, X., Lu, X., Guan, H., and Ye, L. (2013). "Collapse simulation of reinforced concrete high-rise building induced by extreme earthquakes." *Earthquake Engineering & Structural Dynamics,* 42(5), 705-723.

Ludwig, D. (2007). "*Direct alerting to the cockpit for runway incursions".* Proc., Digital Avionics Systems Conference, 2007. DASC'07. IEEE/AIAA 26th.

Schönefeld, J., and Möller, D. (2012). "Runway incursion prevention systems: A review of runway incursion avoidance and alerting system approaches." *Progress in Aerospace Sciences,* 51, 31-49.

Watnick, M., and Ianniello, J. W. (1992). "*Airport movement area safety system".* Proc., 11th Digital Avionics Systems Conf., IEEE/AIAA.

Wood, S. (2006). *Generalized additive models: an introduction with R.* CRC press.

Wood, S. N. (2003). "Thin plate regression splines." *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 65(1), 95-114.