EMBRY-RIDDLE Aeronautical University SCHOLARLY COMMONS

International Journal of Aviation, Aeronautics, and Aerospace

Volume 2 | Issue 3 Article 4

8-1-2015

An Analysis of Airline Quality Rating Components Using Bayesian Methods

John H. Mott
Purdue University - Main Campus, jhmott@purdue.edu
Branden K. Avery
Delta Air Lines, branden.avery@delta.com

After the passage of Public Law 95-504, more commonly known as the Airline Deregulation Act of 1978, the U.S. Department of Transportation began to collect various metrics related to service quality of domestic air carriers. These records were not publicly available, leaving air carriers unable to access other carriers' metrics; consequently, carriers could not effectively determine their standing in relation to one another from an operational efficiency perspective. Similarly, airline passengers had little knowledge about which airlines performed the best, or which provided a higher quality of service (Mott & Avery, 2013). Nine years after the passage of the act, this situation changed with the publication of the first Air Travel Consumer Report (ATCR). The Report increased awareness of carrier performance on the part of the traveling public (Mott & Avery, 2013). The ATCR is a monthly product of the U. S. Department of Transportation that is "designed to assist consumers with information on the quality of services provided by the airlines" (USDOT, 2011, p. 2).

Since its inception, the ATCR has provided a foundation for researchers interested in evaluating service quality in the domestic airline industry. Bowen and Headley published the first Airline Quality Rating (AQR) in 1991 (Bowen & Headley, 2012). For the past 23 years, this study has ranked U. S. air carriers that account for at least 1% of the domestic passenger volume. The AQR provides a month-by-month measure of quality using a weighted average of metrics representing on-time arrivals (OT), involuntary denied boardings (DB), mishandled baggage (MB), and a combination of 12 customer complaint categories (CC). The 12 customer complaint categories are: flight problems, oversales, reservations, ticketing, and boarding, fares, refunds, baggage, customer service, disability, advertising, discrimination, animals, and other (Bowen & Headley, 2012).

The AQR quality measure is determined as follows:

$$Q = ((8.63 \times OT) + (-8.03 \times DB) + (-7.92 \times MB) + (-7.17 \times CC)) / ((8.63 + 8.03 + 7.92 + 7.17))$$

On-time arrivals are reported monthly, involuntary denied boardings are reported quarterly per 10,000 passengers, mishandled baggage is reported monthly per 1,000 passengers, and customer complaints are reported monthly per 100,000 passengers (Bowen & Headley, 2012).

Waguespack and Rhoades (2008) separated the safety, service, and financial performance data from the ATCR and used total departures to normalize safety and service data; this allowed the researchers to examine these parameters independently and explore their relationship to one another. Their resulting Service Disquality Index (SDI) has been used both to provide a 20-year perspective on service quality performance at major U.S. carriers (Rhoades & Waguespack, 2008), and to measure service quality issues

between carriers in different industry segments (Rhoades & Waguespack, 2000a, 2000b).

Most recently, a third group of researchers attempted to determine whether separate econometric models are appropriate for different carrier groups (Mott & Avery, 2013). These researchers showed that separate models are appropriate based on the differences between carrier groups shown in the following methodology: frequentist methods investigated the four primary components of the AQR over a six year period through a two-way analysis of variance with post-hoc difference testing. However, due to the limitations of the data and the requisite assumptions, the researchers suggested that future research be conducted using Bayesian statistical methods rather than the null hypotheses significance testing (NHST) that was employed in the study.

The present research is a continuation of the previous study by Mott and Avery (2013) and will attempt to determine whether separate economic models are appropriate for different carrier groups using Bayesian two-way analyses of variance instead of NHST. As explained in the previous study, the potential for predicting the AQR rankings using econometric modeling can be significant; the predictions could be utilized by managers to allocate resources in an effort to improve the metrics that comprise the overall measure (Mott & Avery, 2013).

Literature Review

The use of Frequentist (also known as and referred to here as "classical") methods in statistical analysis and the application of NHST are common across many scientific disciplines. While extant literature demonstrates the weaknesses inherent in these methods, NHST is still the predominant statistical methodology employed for research in the social sciences. Although Bayesian inference as a means of statistical analysis has made inroads in the scholarly literature in some social science disciplines, the use of Bayesian data analysis in the area of technology is limited. Nevertheless, this form of analysis has numerous advantages, as evidenced in the literature (Mott & Bowen, 2014). There are several fundamental differences between the Frequentist and Bayesian approaches; these are described below, and is provided for the use of the Bayesian methodology in the current study.

The first difference between the two approaches is found in their respective definitions of probability. In the classical approach, probability concepts express uncertainty about events that can be regarded as having been generated by some random process (Kruschke, 2010). On the other hand, Bayesian probability statements may be made regarding any potentially verifiable proposition, whether or not a random process affecting that proposition can be imagined (Kruschke, 2010). The classical approach can be

expressed in the form of conditional probabilities as *Pr*(outcome|model), and the Bayesian formulation can be expressed similarly as *Pr*(model|outcome).

The second difference is the incorporation of prior information. When using the classical approach, all prior knowledge is omitted, as that information is considered bias. Consequently, the inclusion of which is traditionally considered conceptually inappropriate in research methodology. A fair hypothesis test is based only on data at hand, assumed to be representative of the population from which it is sampled (Krushke, 2010). Unlike the classical approach, Bayesian inference uses prior knowledge about a problem to the extent incorporated by the researcher. This generates a prior distribution that is combined with a likelihood function to result in a posterior distribution; the result completely describes all known information related to the problem (Krushke, 2010).

A third difference between the two approaches regards statistical inference. In the classical approach, the *p*-value (the probability that the test statistic would have been more extreme than the actual observed value of the statistic provided that the null hypotheses is true) is based on the sampling distribution (Kinney, 2002). The *p*-value is compared with an acceptable error rate [$\alpha = Pr$(Type I error)] to decide whether to reject the null hypothesis. In contrast, Bayesian inference, assigns prior subjective probability to each statistical hypothesis. The hypothesis with the greatest posterior probability is deemed to be the most likely to be true. Researchers with different prior beliefs may reach different conclusions through this approach (Krushke, 2010). In addition, the fact that the posterior distribution is available to the Bayesian researcher implies that a wide and rich variety of post-hoc testing can be conducted. Relative frequency analyses, as a result of their limited availability of post-hoc tests, suffer from derogation at the hand of Bayesian approaches (Mott & Bowen, 2014).

The motivation for conducting the Bayesian analog to the classical analysis of variance (ANOVA) follows because such models do not depend on corrections to ensure that test assumptions are met; instead, Bayesian methods rationally mitigate statistical error based on the data itself (Kruschke, 2010). In addition, when using Bayesian hierarchical modeling, any concern regarding the robustness of the analysis technique is reduced. Bayesian inference treats all effects as random, avoiding the otherwise required distinction between fixed vs. random effects. It also provides estimates of effects and variance components with corresponding uncertainty. Bayesian inference can also easily handle research designs that are unbalanced, and data that is non-normal or missing values (Gill & Swartz, 2007). Because of these differences, Bayesian hierarchical modeling reduces any concerns regarding the strength of the analysis technique.

Method

This study was conducted using the four metrics that were part of the original ATCR, and which comprise the AQR measure (OT, MB, DB, and CC). The data spans a six-year period from 2006 to 2011 (Mott & Avery, 2013). A challenge in making longitudinal comparisons across the AQR dataset is that incumbent carriers are occasionally dropped from the rankings. This occurs either because they have merged with another carrier or because they fail to meet the required 1% domestic volume criterion for inclusion in the AQR. Because of these inconsistencies, the present research focused on the fourteen airlines that have been included in the annual AQR reports over the entire period. These carriers have been classified according to their business models into one of three service groups, and are presented in Table 1.

Table 1
Carrier Groupings

Legacy Carriers	**Regional Carriers**	**Low-Cost Carriers**
Alaska Airlines	American Eagle Airlines	Air Tran Airways
American Airlines	Atlantic Southeast Airlines	Frontier Airlines
Continental Airlines	Mesa Airlines	Jet Blue Airlines
Delta Airlines	Sky West Airlines	Southwest Airlines
United Airlines		
US Airways		

Legacy carriers typically operate hub-and-spoke networks. These serve both international destinations and medium-to-large domestic cities using diverse fleets of aircraft carrying approximately100 to 300 passengers (Erstad, Jednachowski, Bowen, Meehan, & Bowen, 2013). Regional carriers generally complement legacy carrier operations by operating flights from smaller cities to their respective partners' hubs. They do this under code-sharing agreements using smaller, more efficient aircraft (Forbes & Lederman, 2006). Finally, low-cost carriers often employ both point-to-point and hub-and-spoke models, serving medium-to-large cities with larger aircraft that have approximately 125 to 175 seats.

A Bayesian two-way analysis of variance (BANOVA), as adjusted for assumptions of non-normality and lack of sphericity, was conducted on the data using the R Studio integrated development environment. This analysis was used to determine whether any credible differences existed between carrier group means for the four AQR factors (on-time arrivals, mishandled baggage, denied boardings, and customer complaints). This particular model was utilized based on prior recommendations from the researchers' previous study using conventional null hypothesis testing.

A Shapiro-Wilk test indicated that the data was not normally distributed for any of the carrier groups. Moreover, Mauchly's test of sphericity indicated that such assumptions were violated for each carrier. Using the JAGS package within R, an appropriate model was created to fit the AQR data. The model was designed to accommodate outliers by using heavy-tailed distributions to describe data within cells, and to accommodate heteroscedasticity by using distinct variance parameters in each cell.

Each of the four metrics was treated as a separate dependent variable, and the independent variables were the carrier service level groups and the year groupings. To eliminate zeroes within the data set, the study performed an affine transformation on the denied boarding and customer complaints categories.

The hypotheses that were used for this study are as follows:

- H_o: There is no statistically credible difference between the means at the given α level.
- H_a: There is a statistically credible difference between the means at the given α level.

A significance level of $\alpha = 0.05$ was applied in all tests.

Results

Four Bayesian two-way, non-normal, non-homogeneous analyses of variance as previously described were conducted to determine whether there were statistically significant differences between carrier group means for the four AQR factors (on-time arrivals, mishandled baggage, denied boardings, and customer complaints) over the course of six years.

On-time Performance

Regarding on-time performance, a test of between-subjects effects for the carrier service level variable indicated a credible difference between regional and legacy carriers (X1.2 v. X1.1), with a mean of -0.0183 and a 95% Highest Density Interval (HDI) of (-0.0288, -0.00786). Additionally, there was a credible difference between low-cost and regional carriers (X1.3 v. X1.2), with a mean of 0.0215 and a 95% HDI of (0.00944, 0.0339). However, there was no credible difference between low-cost and legacy carriers (X1.3 v. X1.1), with a mean of 0.00328 and a HDI of (-0.00675, 0.0134). Figure 1 delineates these comparisons.

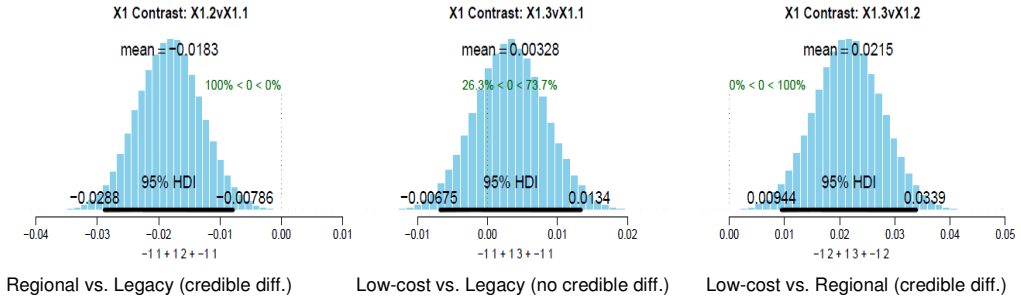


Figure 1. Comparison of on-time performance.

Mishandled Baggage

Secondly, mishandled baggage was analyzed; similar to the analysis of on-time performance, the analysis indicated a credible difference between carriers. The first credible difference was between low-cost and legacy carriers (X1.3 v. X1.1), with a mean of -2.01 and a 95% HDI of (-2.5, -1.51). The other credible difference was between low-cost and regional carriers (X1.3 v. X1.2), with a mean of -1.94 and a 95% HDI of (-2.42, -1.49). There was no credible difference between regional and legacy carriers (X1.2 v. X1.1), with a mean of -0.0646 and a 95% HDI of (-0.587, 0.456). Figure 2 shows these comparisons.

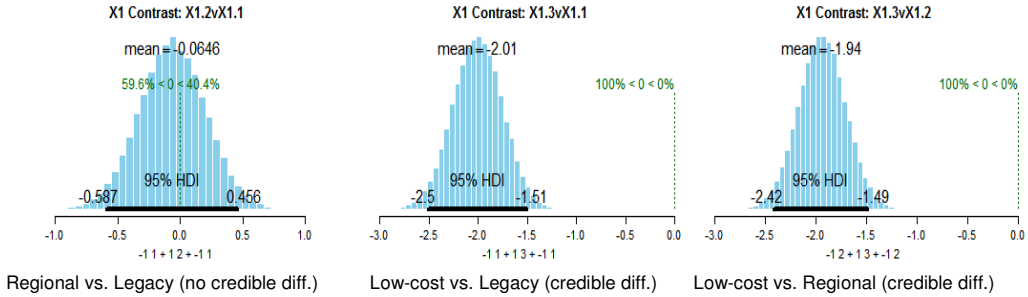
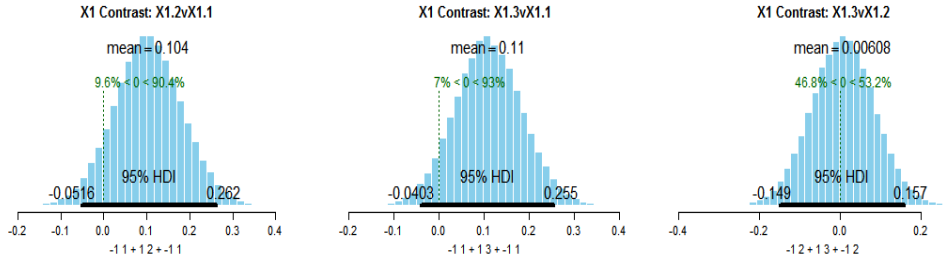


Figure 2. Comparison of mishandled baggage.

Denied Boardings

Unlike the on-time performance and mishandled baggage metrics, denied boardings did not appear to exhibit a credible difference between any of the three different carrier groups. None of the differences between regional and legacy carriers [(X1.2 v. X1.1), with a mean of 0.104 and an HDI of (-0.0516, 0.262)], low-cost and regional carriers [(X1.3 v. X1.2), with a mean of 0.00608 and an HDI of (-0.149, 0.157)], or low-cost and legacy carriers [(X1.3 v. X1.1) with a mean of 0.11 and an HDI of (-0.0403, 0.255)] was significant. Figure 3 summarizes these comparisons.

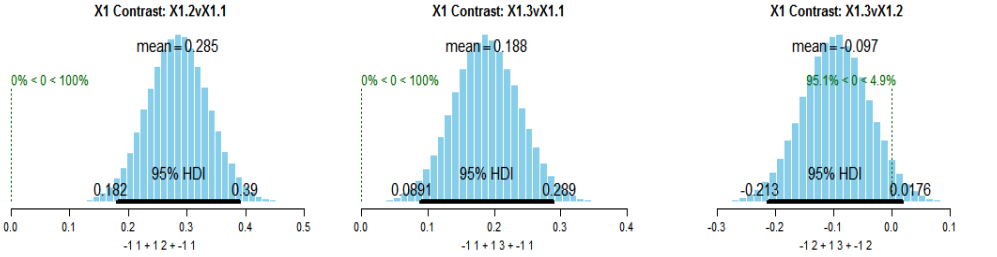


Regional vs. Legacy (no credible diff.) Low-cost vs. Legacy (no credible diff.) Low-cost vs. Regional (no credible diff.)

Figure 3. Comparison of denied boardings.

Customer Complaints

The analysis of customer complaints indicated a credible difference between the carrier groups. The first credible difference was between regional and legacy carriers (X1.2 v. X1.1), with a mean of 0.285 and an HDI of (0.182, 0.39). The next credible difference was between low-cost and legacy carriers (X1.3 v. X1.1), with a mean of 0.188 and an HDI of (0.0891, 0.289). There was not a credible difference between low-cost and regional carriers (X1.3 v. X1.2), with a mean of -0.097 and an HDI of (-0.213, -0.0176). Figure 4 displays these comparisons.



Regional vs. Legacy (credible diff.) Low-cost vs. Legacy (credible diff.) Low-cost vs. Regional (no credible diff.)

Figure 4. Comparison of customer complaints.

Discussion

The goal of this research was to further examine the differences between legacy, regional, and low-cost carriers to determine whether a significant difference between the groups exists, relative to the four primary AQR components (OT, DB, MB, CC) over the six year study period using Bayesian statistical methods. Based on the results of the Bayesian two-way analysis of variance, credible differences in on-time arrivals, mishandled baggage, and customer complaints were indicated. These results imply a rejection of the null hypothesis, H_o, for those data sets. There were no credible differences indicated between legacy, regional, and low-cost carriers regarding the denied boarding data set, implying a failure to reject the null hypothesis in that particular case.

Conclusion

The previous study conducted by Mott and Avery (2013) utilized a two-way classical ANOVA with Tukey-Kramer post-hoc difference testing. Results from that study showed significant differences between the mishandled baggage, denied boardings, and customer complaints data sets, indicating rejection of the null hypothesis, H_o. The researchers contend the study provided sufficient evidence to support the premise that separate econometric predictive models are needed to facilitate air carrier service quality improvements. In the study using Bayesian data analysis techniques, the model that was developed, and which provides for violations of both the normality and sphericity assumptions that are implicit in Frequentist tests, indicates that the earlier results are valid. Therefore, the assertion that separate predictive models are appropriate, based on the differences between the carrier groups, is validated. As suggested by Mott and Avery (2013), airline managers will, by utilizing the predictive econometric models developed separately for each carrier group, be able to more accurately forecast AQR components, and as a result, the overall quality rating. Consequently, they will have the ability to refine resource allocation methods in an effort to improve quality of service.

References

Bowen, B. D., & Headley, D. E. (2007). *Airline quality rating 2007* (Report No. 20). West Lafayette, IN: Purdue e-Pubs. Retrieved from http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1017&context=aqrr

Bowen, B. D., & Headley, D. E. (2008). *Airline quality rating 2008* (Report No. 21). West Lafayette, IN: Purdue e-Pubs. Retrieved from http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1018&context=aqrr

Bowen, B. D., & Headley, D. E. (2009). *Airline quality rating 2009* (Report No. 22). West Lafayette, IN: Purdue e-Pubs. Retrieved from http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1019&context=aqrr

Bowen, B. D., & Headley, D. E. (2010). *Airline quality rating 2010* (Report No. 4). West Lafayette, IN: Purdue e-Pubs. Retrieved from http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1020&context=aqrr

Bowen, B. D., & Headley, D. E. (2011). *Airline quality rating 2011* (Report No. 2). West Lafayette, IN: Purdue e-Pubs. Retrieved from http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1021&context=aqrr

Bowen, B. D., & Headley, D. E. (2012). *Airline quality rating 2012* (Report No. 2012). West Lafayette, IN: Purdue e-Pubs. Retrieved from http://www.airlinequalityrating.com/reports/2012aqr.pdf

Bowen, B. D., Headley, D. E., & Luedtke, J. R. (1991). *Airline quality rating 1991* (Report No. 3). West Lafayette, IN: Purdue e-Pubs. Retrieved from http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1001&context=aqrr

Gill, P. S., & Swartz, T. B. (2007). Bayesian analysis of dyadic data. *American Journal of Mathematical and Management Sciences*, *27*, 73–92.

Erstad, A., Jednachowski, M., Bowen, B. D., Meehan, R., & Bowen, E. (2013). *Modeling organizational performance: Differentiating factors in economic periods prior, during, and post recession* (Report No. 2). Retrieved from http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1017&context=atgrads

Forbes, S. J., & Lederman, M. (2007). The role of regional airlines in the US airline industry. *Advances in Airline Economics*, *2*, 193-208.

Kinney, J. J. (2002). *Statistics for science and engineering*. Boston, MA: Addison Wesley.

Kruschke, J. K. (2010). *An open letter to editors of journals, chairs of departments, directors of funding programs, directors of graduate training, reviewers of grants and manuscripts, researchers, teachers, and students.* Retrieved from http://www.indiana.edu/~kruschke/AnOpenLetter.htm

Mott, J. H., & Avery, B. K. (2013). An investigation of the effects of carrier groups on airline quality rating components using a two-way analysis of variance. *Collegiate Aviation Review, 31*(2), 103-121.

Mott, J. H., & Bowen, E. E. (2014). Teaching NHST vs. Bayesian inference in postsecondary technology programs. In *Proceedings of the 9th International Conference on Teaching Statistics.* Flagstaff, AZ: IASE.

Rhoades, D. L., & Waguespack, B. (2008). Twenty years of service quality performance in the US airline industry. *Managing Service Quality*, *18*(1), 20-33. doi: 10.1108/09604520810842821

U.S. Department of Transportation, Office of Aviation Enforcement and Proceedings, Aviation Consumer Protection Division. (2011). *Air travel consumer report*. Retrieved from http://airconsumer.ost.dot.gov/reports/2011/March/2011MarchATCR.pdf

U.S. Department of Transportation, Bureau of Transportation Statistics. (2013). *Airline on-time performance and causes of flight delays*. Retrieved from http://apps.bts.gov/help/aviation/#q9