Quality Issues in a Regulatory Data Collection System

Alan K. Jeeves
Bureau of Transportation Statistics, Research and Innovative Technology Administration,
Department of Transportation

Abstract[1]

The Federal Aviation Act requires all air carriers to report traffic and financial information to the DOT. Although these data collection systems originally supported airline regulation by the Civil Aeronautics Board (CAB), the government and private sector still use these data to monitor industry competition and financial condition. A new impetus for data quality occurred with the recent enactment of the Rural Service Improvement Act which established BTS airline data the basis for determining carrier eligibility and payment rates for intra-Alaskan mail transport. Our research showed that although the scope of the edit procedures could be broadened, the main problem was the number of records flagged exceeded the available resources to fix them. Maximizing data quality requires recognition that since not all errors can be fixed, the focus must be on the most egregious.

Key words: data quality, editing, imputation, regulatory data.

Introduction

Regulatory data collection systems operate like many other statistical data collection systems, but ensuring the quality of the data can be a significantly greater challenge. Ordinary statistical survey data can be cleaned through edit and consistency checks, and various imputation methods can be used to fill in missing data. Policy decisions ultimately ensue from the collective information contained in the filings. Despite certifications by filers, the data are still prone to the same types of errors as are common in any statistical survey. However, to change certified data filings can require the consent of the data provider. This is especially true when the information submitted is publicly attributed to the source and republished.

Bureau of Transportation Statistics Airline Information Programs

The Federal Aviation Act of 1958 established the Federal Aviation Administration (FAA) and redefined the role of the Civil Aeronautics Board (CAB). The CAB regulated airfares and decided how many and which airlines could fly between cities. The Act also established the Uniform System of Accounts and Reports for Large Certificated Air Carriers that mandates the reporting of airline traffic and financial data by air carriers that operate aircraft with a seating capacity of more than 60 seats or a maximum payload capacity of more than 18,000 pounds.

The CAB used this data for the economic regulation of the airline industry its sunset mandated in 1985 by the Airline Deregulation Act of 1978. However, the requirement for the airlines to file these reports remained and was transferred to the Department of Transportation (DOT). The Bureau of Transportation Statistics (BTS), a statistical agency within the DOT Research and Innovative Technology Administration (RITA), currently administers this program. Even in the current environment of airline deregulation, the government still uses these data to monitor industry competition and financial condition. The data are also widely used by private sector industry analysts.

The airline information programs are unique in that they provide a comprehensive view of an entire industry by individual entity. Each of these airlines is privately owned and the information they provide is useful to their competition.

[1] The views expressed in this article are those of the author and do not necessarily represent the views of the Bureau of Transportation Statistics or the Department of Transportation. The author thanks Eugene Burns for his invaluable suggestions on an earlier draft.

Uses of Airline Data

Congress, the General Accounting Office (GAO), and the Council of Economic Advisors (CEA) require national, regional and local airport, airline and passenger data to promote informed decision-making and serve constituent needs. The Federal Aviation Administration (FAA) uses the data to allocate airline safety inspection resources, determine control tower staffing levels, allocate Airport Improvement Program (AIP) grant funds, forecast activity levels in the air transportation sector, and monitor flight delays. DOT policymakers use the data to analyze airline competition, negotiate international air service agreements, set international and Alaskan mail rates, determine community eligibility for Essential Air Service subsidies, evaluate air carrier fitness and air carrier merger requests, conduct policy analyses, and advise the Secretary on major air transport industry issues.

Homeland Security allocates passenger screening resources and validates the collection of user fees for passenger security, customs, and animal and plant inspection services. The Internal Revenue Service (IRS) projects aviation trust fund and excise taxes. The Department of Defense (DOD) estimates the reimbursement rates for the Military Airlift Program. The Department of Commerce’s Bureau of Economic Analysis and Census Bureau estimate the aviation contribution to the Gross National Product. The Bureau of Labor Statistics uses airfare survey data in its Consumer Price Index (CPI) calculations. The Department of Justice (DOJ) analyzes the antitrust implications of proposed mergers, airline acquisitions, and airline applications for antitrust immunity. The Environmental Protection Agency (EPA) assesses the environmental noise impact of carrier operations and uses air carrier fuel consumption data to estimate the emission of greenhouse gases. The Department of Energy (DOE) monitors airline industry fuel consumption for emergency preparedness. The National Transportation Safety Board (NTSB) analyzes safety levels in commercial aviation.

State and local governments and airports analyze traffic demand and patterns of service to promote tourism and perform planning and development for airport and air service improvements. The data also enable airports to prepare competition plans as required by the Aviation Investment and Reform Act for the 21st Century (AIR 21) legislation.

The Air Transport Association (ATA) and the Regional Airline Association (RAA), among others track the state of the industry and their member carriers in order to serve as an advocate for their respective constituencies. The air carriers perform market analysis of their competition. The airlines set their “high season” flight schedules. The Airframe and Engine Manufacturers plan the types and size of aircraft to be marketed to airlines, and evaluate the risk of financing sales to the airlines. The media use the aviation data regularly as the basis for consumer and investment stories. (RITA-BTS Aviation Data Stakeholders and Uses of the Data, BTS internal document, March 8, 2006)

Data Quality

Ensuring airline data quality is difficult. Even though many types of data error can be easily identified, the correction process is cumbersome because of the need to obtain air carrier concurrence. The BTS Office of Airline Information (OAI) brings major issues to the attention of the carriers for correction. However, the data is not changed unless the carrier resubmits it even when the original data is obviously wrong. Lesser errors which are not brought to the attention of the carriers are entered into the files as submitted.

Thus, a discrepancy can exist between what the carriers report and the actual truth. Statisticians and economic analysts want the truth, but some users may also be interested in what is actually reported. Policy makers are reluctant to base regulatory decisions on data that the carrier did not actually submit, particularly since most of the submitted data is published with the carrier identified.

Due to the sheer volume of data that is reported every month, it is impossible to correct all data issues. In addition, the carriers do not normally face significant penalties for failure to file reports, sloppy reporting, or even the falsification of records. The maximum fine per incident that can be imposed is less than $5,000 and requires a misdemeanor conviction. Adjudication requires a time consuming docket which makes enforcement cumbersome. For example, despite the documented data quality issues identified by the DOT Office of the

Secretary – Aviation (OSTX) and BTS Office of Statistical Quality (OSQ), there were only two consent orders issued in 2003.

Private Sector Use

The majority of airline information filed with the BTS is regularly released to the public. However, many private sector users prefer to obtain it for a fee from third party redistributors even though it could be downloaded gratis from the BTS website.

These repackagers are satisfied with the data quality insofar as they have their own edit procedures that clean any irregularities that they are concerned about. Since they are not the government and do not hold the airlines accountable for the data, they are free to treat it as though it were statistical survey data. They do not need to contact the airlines to make changes.

One repackager even advertises that they do not simply pass on raw DOT data, but instead they perform several edits on the integrity of the data. They claim to cross-validate the data several different ways to find inconsistencies. For errors that they can fix, they make a correction before publishing the data. They also say that they notify the DOT and sometimes the carriers directly of errors that they find so that the DOT can fix their published data to agree with theirs. Ironically, if BTS were able to more easily and comprehensively correct the data errors in its airline information files, the market for data repackager would be reduced.

New Quality Urgency

Airline data is filed by large certificated carriers which include large network carriers such as Northwest, United, or Delta; low-cost carriers such as Southwest or Jet Blue; as well as regional carriers, some of which have operating revenues of less than $20 million per year. Until 2002, the primary focus on overall system quality was concentrated on the larger major and national carriers rather than the smaller regional carriers. However, enactment of the Rural Service Improvement Act (RSIA) of 2002 expanded the focus of the airline data quality efforts to include the Alaskan air carriers who transport mail for the U. S. Postal Service.

The BTS supplies the USPS with individual air carrier market (origin to ultimate destination) and segment data (flight origin to destination) for air transportation within the State of Alaska. This enables the USPS to tender mail to those Alaskan bush carriers that meet the RSIA requirements. Without this data, the USPS would have to institute its own data collection. The USPS experimented with this alternative, when RSIA was passed, but decided their system was unsatisfactory and costly and elected to use BTS data.

Alaskan Air Cargo Operations

Among the all of the states, Alaska has the most unique transportation system. Due to its low population density, large area, and polar climate, it does not have a significant road network within the state outside the metropolitan areas of its major cities. The major highways run only along the southern Pacific coast, between Anchorage and Fairbanks, and between Fairbanks and the Canadian border. Many places are inaccessible except by aircraft. Therefore, most cargo traffic within Alaska is carried by air.

Air freight rates are more expensive than surface rates. However, mailing rates are considerably cheaper than air freight rates. Although Parcel Post mail is normally carried by surface transportation in the lower 48 states, it goes by air when no ground transportation is available. Consequently, any mailable cargo—which includes almost anything other than alcohol, tobacco, and firearms that weighs less than 70 pounds—is transported by the U. S. Postal Service using intra-Alaska Bush Air Carrier Service.

Some Alaskan carriers had been deriving virtually all of their business from the transport of mail, providing virtually no passenger or freight service. With the passage the Rural Service Improvement Act (RSIA) in 2002, Congress directed that only air carriers that provide specified levels of regular passenger and freight service to a market would eligible to receive mail tender for that market. The basis for determining carrier eligibility is the airline traffic and financial data submitted by the carriers to BTS.

Airline Traffic Filed

The T-100 Air Carrier Traffic data are filed by flight segment and market. All flights are covered including diversions (non-scheduled stop), flag stops (demand service stop), tech-stops (service and refueling), and emergency landings. Non-stop segment data detail the total passengers, freight, and mail transported between a single pair of points regardless of any other segments that they may have also used the completion of a trip. On-flight Market data report the total passengers, freight, and mail enplaned between two points regardless of the number of segments used provided that the flight number does not change. The data reported are aggregates for a month and not information by individual flight.

To illustrate the difference between a market and segment, consider a flight from Los Angeles to New York with a stopover in Chicago. The segments are Los Angeles to Chicago and Chicago to New York. These segment records reflect only that transported between each pair of points. The market is Los Angeles to New York, and the market record would reflect only the passengers, freight, and mail transported over both segments.

For non stop flights, the passengers, freight, and mail reported would be the same for both the segment and market record. Every passenger on a flight leaving an airport is a segment passenger, but they would not be market passengers of that airport if they did not board there.

The Non-Stop Segment data elements reported include:

- Service Class (passenger, cargo, both)
- Aircraft Type
- Departures Performed and Scheduled
- Available Capacity / Payload (passenger and cargo weight)
- Available Seats (number of seats for sale)
- Passengers (number transported)
- Freight (weight)
- Mail (weight)
- Ramp to Ramp Time (begin take-off to end landing)
- Airborne Hours (wheels off to wheels on time in the air)

The On-Flight Market data elements include the following:

- Service Class (passenger, cargo, both)
- Passengers (number enplaned at origin and deplaned at destination)
- Freight (weight enplaned at origin and deplaned at destination)
- Mail (weight enplaned at origin and deplaned at destination)

(T-100 Traffic Reporting Guide, February 2004, internal document).

Initial Review of Alaskan Air Carrier Data

The initial review of these data by the Aviation and International Affairs revealed that the quality was insufficient to meet the requirements of RSIA. Among the issues were:

- missing reports,
- identical data reports filed in consecutive months
- data elements that translated into out of range estimates for cruise speed, payload, seat capacity, and flight time,
- airborne flight hours in excess of the departure-arrival intervals,
- inconsistencies in airline passenger and cargo load factors (proportion of capacity utilized).

In addition, the data reported for the markets and segments were not generally compared for consistency. Consequently, there were many instances where the carrier system market data exceeded the system segment data. The system market data for aggregate passenger and cargo data must be less than or equal to the system segment data as well as for each origin airport within the system.

The Office of Airline Information acknowledged that there had been persistent problems with collecting required data from the Alaskan air carriers. For example, some air carriers frequently (as many as 7 or 8 times) re-filed their data, at times covering period of a year or more. Other air carriers would consistently submit data of poor quality or in an untimely manner. In each case, the BTS was required to attend to the identified issues, ranging from direct contact with the air carriers, formal notice and warning letters, to referrals to the Office of General Counsel for enforcement actions.

Why These Errors were a Problem

Alaskan air carriers, like all airlines, report financial expenses for their operations, however, they do not differentiate between their passenger and cargo or scheduled and non-scheduled services. The aggregate ramp-to-ramp times on the segment files for each of the service classes are used to allocate the expenses among them. These ramp-to-ramp times are used to further allocate expenses by aircraft category: Part 121 (19 seats and up), Part 135 (smaller planes), and Amphibious.

The allocation of these expenses to passenger, freight, and mail service is done using the Revenue Ton-Miles (RTM) for passenger, freight, and mail. These are computed using the passenger, freight, and mail weight carried; and distance traveled. These data must be accurate for both the market and the segment data. The Postal Service required that air carrier payment rates be set on a market to market basis based on the great circle distance. However, the data used to cost by aircraft type are available only by segment. To resolve this, the ratios of the segment RTMs to market RTMs would be used as an adjustment factor to the convert rates developed using segment data into rates based on markets. This is why the Office of the Secretary Aviation (OST-X) was concerned about situations where the aggregate market amounts exceed the aggregate segment amounts.

The OST-X also planned to set territorial rates to compensate for differences in available payload capacity due to fuel load requirements. Payload is intended to be a net space available measure that does not include fuel. The amount of fuel required is determined largely by the distance and the proximity of the next nearest airport to the intended destination. Destinations in the Aleutian Islands not only further, but the airports are more widely spread apart. Thus, within an aircraft type, the reported payload should decrease as the segment distance increases.

The Root of the Problem

Originally, the quality issues were assumed to be the result of incomplete edit procedures. However, our research showed that although the scope of the edit procedures could be broadened, the main problem was that more records were being flagged than there were resources to fix them. Insofar as the air carriers are accountable for their submissions, DOT cannot unilaterally alter their data by editing or imputation.

One of the continuing challenges with the edit process is resolving warning and error flags. The Office of Airline Information has dealt with these flags through a largely manual process requiring review by the data base administrator, who often follows-up with the carrier, and must make decisions on how to handle ambiguous situations. Adding new edit checks can actually compound the problem if they introduce more false-positive warnings. There are often situations that arise that outside the normal realm of carrier operations resulting in edit flag overrides to accept data that falls outside normal ranges.

Since not every flag could be corrected, it became clear that many of the edit parameters were set at levels that were inappropriate considering the time and resources available to get them corrected. Other parameters were simply inaccurate. In addition, traffic data were submitted in two separate files for market and segment records, but there were no edit checks comparing the data in the two files.

New Strategy

Not every record error can be corrected since it is not possible to take every issue to the carriers for adjudication. Thus, not all issues identified in the edit checks can be resolved prior to the creation of the final accepted file. There are thousands of records on the accepted file with unresolved edit flags. If these records had been fixed, many of the problems identified by OST-X would have been corrected. Therefore, the focus needed to be on consistently identifying the most egregious errors and getting them corrected.

Three of the primary edit checks involved derived aircraft speed, aircraft seats, and aircraft capacity. These parameters were aircraft specific. They were set years before based on manufacturer specifications. We analyzed the three most recent years data and discovered that the range of reported values was frequently far inside the parameter values. Thus, many records were being flagged that should not have been. There were also cases where a very large number of records were just outside an incorrectly set parameter range, thereby generating extraneous flags. The net effect was that the real problems were buried among the incorrect flags and less important issues. This is being resolved by resetting these parameters to more appropriate values.

Through the addition of cross-file edit checks and refinement of existing edit checks, the BTS has been able to improve the data quality of the entire airline traffic and financial data systems including both the Alaskan air carriers and all others. There have also been new checks added to identify instances of identical report submissions in consecutive months.

Future Considerations

Even under the best circumstances, carriers will make unintentional errors in their submissions. BTS feedback to the carriers regarding errors found in data processing should be more direct and interactive. Once carriers get the message that the data is being carefully reviewed and that quality matters, a "Hawthorne" effect can be expected to take hold and the number errors should diminish.

Nonetheless, even after all of these efforts, the file will still not be perfect. This presents with Bureau of Transportation Statistics with an important policy decision. As a statistical bureau, is it our job to report what is submitted, or is it our job to treat these regulatory data in the same manner as statistical survey data and make final corrections through the application of recognized statistical processing procedures? To put it another way, are we here to report data or information.

Internal Revenue Service

The Internal Revenue Service (IRS) publishes statistics based on taxpayer filed returns in their Statistic of Income (SOI) reports. Taxpayers certify these results. However, unlike the airline data programs, no information by individual filer is every published. SOI data are collected only from non-amended and non-audit returns. These sample returns are subject to additional data abstraction for SOI by specially trained technicians. Due to substantial penalties for misreporting, the income and expenditure data reported on tax returns have proven to be more reliable than comparable survey data. Even so, IRS employees go to great lengths to protect against nonsampling errors, such as those due to taxpayer reporting variations or inconsistencies, or data processing errors. In order that final statistics are consistent and reliable, IRS economists develop extensive on-line tests and error resolution procedures that are applied to each sampled return. Missing data problems arise infrequently—less than 1 percent of the time. These missing items can be obtained through direct contact with taxpayers, or estimated through imputations based on other return data, prior-year data for the same taxpayer, or same-year data from a "statistically similar" return (Internal Revenue Service Information Quality Guidelines).

Securities and Exchange Commission

Publicly traded companies that are required to file financial statements with the SEC, do so through the EDGAR (Electronic Data Gathering, Analysis, and Retrieval System). These data are available by individual filer. However, the SEC filings are not really a statistical data collection and the data included are only summaries of larger financial statements. Most of the quality review on the information filed is in the form of Procedural checks performed by EDGAR to determine whether a filing meets certain minimum filing requirements. These requirements relate to the composition and completeness of the submission package, as well as to the particular type of filing being made. Financial data are included in these filings as part of a submission that consists mostly of text data. There are no edit checks to review these financial data, although any included data should have been reviewed by their accounting firm.

In January 2006, the SEC announced that it would offer expedited reviews of registration statements and annual reports to companies that volunteer to participated in the SECs interactive data initiative. Interactive data holds the promise of transforming the static, text-only documents companies file with the SEC into dynamic financial reports than can be quickly and easily accessed and analyzed. In April 2005, the SEC began a voluntary program for receiving financial information using eXtensible Business Reporting Language (XBRL) (SEC Offers Incentives for Companies to file Financial

Reports with Interactive Data, 2006-7, January 11, 2006).

Conclusion

The BTS Airline Information Programs are unique. They are regulatory data collections resulting in the release of reports that specifically identify the respondent air carrier. For this reason, edit corrections of the data must be cleared with the carriers. However, due to the additional procedural requirements that result, it is not possible to fully correct all errors. Instead only the most critical issues can be identified and corrected. However, as the use of interactive data entry technology increases, it may be possible to assist the carriers in getting the data right the first time as they enter it.

References

Airline Deregulation Act of 1978, Public Law 95-504, 92 Stat. 1705, October 24, 1978, http://amelia.db.erau.edu/reports/misc/PL95-504.pdf

Aviation Investment and Reform Act for the 21st Century (AIR 21), Public Law 106-181, 114 Stat. 61, 49 USC Section 42121, April 5, 2000 http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=106_cong_public_laws&docid=f:publ181.106.pdf

Internal Revenue Service Information Quality Guidelines, http://www.irs.gov/irs/article/0,,id=131181,00.html

Rural Service Improvement Act (RSIA), Public Law 107-206, Sec. 3002, 116 Stat. 910, 2002 Supplemental Appropriations Act, 116 Stat. 820, August 2 2002 http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=107_cong_public_laws&docid=f:publ206.107.pdf

Securities and Exchange Commission, EDGAR Filer Manual I: General Information (Version 2) February 2006, http://www.sec.gov/info/edgar/edgarfm-vol1-v2r1.pdf

Securities and Exchange Commission, EDGAR Filer Manual II: Edgar Filing (Version 3) February 2006, http://www.sec.gov/info/edgar/edgarfm-vol2-v3r1.pdf .

Securities and Exchange Commission, SEC Offers Incentives for Companies to file Financial Reports with Interactive Data, 2006-7, January 11, 2006) http://www.sec.gov/news/press/2006-7.htm .