Public Transp
DOI 10.1007/s12469-015-0107-6

ORIGINAL PAPER

Identification of transit farebox data errors: impacts on transit planning

Shu Yang[1] · Yao-Jan Wu[1] · Bernadette Marion[2] · Isaac E. Moses[3]

Abstract Transit agencies require a constant stream of operations performance data to support standard planning, scheduling and operations management activities. Ridership and revenue statistics play a critical role in strategic system design, policy development, and budgeting decisions at all levels of transit management. Many agencies rely on electronic fare collection devices as a primary source for ridership and revenue data. The quality of this data will greatly affect transit-related reporting and decision making. This study proposes a systematic, data-driven approach to process revenue and ridership data pulled off electronic farebox equipment installed on a bus fleet operating in the St. Louis region. Three major farebox data errors are identified and impacts of these data errors are further evaluated and discussed at the system and trip level. Results indicate ridership and revenue may be overestimated by up to 8.05 and 9.95 %, respectively, due to farebox data errors. The results of this development effort offer a range of low-cost error identification and processing techniques that transit staff could easily and quickly implement. Even though the St.

✉ Shu Yang
shuyang@email.arizona.edu

Yao-Jan Wu
yaojan@email.arizona.edu

Bernadette Marion
bmmarion@metrostlouis.org

Isaac E. Moses
iemoses@wmata.com

[1] Department of Civil Engineering and Engineering Mechanics, The University of Arizona, 1209 E 2nd St. Room 324F, Tucson, AZ 85721, USA

[2] Metro Transit, St. Louis, St. Louis, MO 63102-2595, USA

[3] Office of Applications Development and Operations, Department of Information Technology, Washington Metropolitan Area, Transit Authority, Washington, DC 20001, USA

Published online: 17 June 2015

Louis Metro Transit data was used for analysis, these proposed approaches can be considered as a general framework and used by other transit agencies.

Keywords Farebox data · Data quality assurance · Transit planning · Minimum covariance determinant · Ridership

1 Introduction

Transit agencies require a constant stream of operations performance data to support standard planning, scheduling and operations management activities, to monitor trends, and to provide accountability to funding partners and the public. In particular, ridership statistics play a critical role in strategic system design, policy development, and budgeting decisions at all levels of transit management. For these reasons, system-level ridership statistics are routinely collected, analyzed and reported to higher-level management, local governing bodies, and the Federal Transit Administration. Increasingly, decision-makers and system designers demand that staff produce timely reports of vehicle trips using the route, trip, or stop level as the primary unit of analyses.

The availability of data inputs at a more local level of analysis facilitates service design plans that serve community livability goals more efficiently and effectively. However, the cost and staff-resource burden dedicated to data collection and management of large data streams presents a significant challenge. Such data can be collected through traditional electronic farebox equipment, which is used nationwide to collect transit fares on buses. Many transit agencies are replacing traditional farebox equipment (e.g., GFI CENTSaBill Electronic Registering Farebox) with more sophisticated fare collection (AFC) systems. Nevertheless, traditional farebox equipment is still widely used to collect transit fares and ridership data (Metro St. Louis 2010; Greater Cleveland Regional Transit Authority 2005) and many of the same challenges are still inherent in the updated replacement systems.

All data collected during standard service operations are subject to equipment limitations and malfunctions. For example, in the field of Intelligent Transportation Systems (ITS), the data is mainly collected from the existing ITS infrastructures (e.g. single-loop sensor, radar-based sensor and probe vehicles). The failures of sensors or network communication lead to the “missing data issues”; while, the sensors with reliability and stability issues could produce and report “dirty data”. Thus, a large volume of research tackles data management and data quality issues associated with technologies (see, for example, Corey et al. 2012; Lee and Coifman 2012; Wu et al. 2010).

Transit fare data, which has been used in many research efforts related to performance and metrics, can be subject to similar data quality issues. For instance, Ma et al. (2013) use smartcard data to estimate transit riders’ daily travel patterns in Beijing, China. Lee and Hickman (2014) use smart card transaction data to derive trip activity information in an urban area of Minneapolis, USA. More applications of the smartcard data can be found in Pelletier et al. (2011) work. Moore (2002) investigates the expected relationship between revenue and ridership; Lu and Reddy

(2012) and Navick and Furth (2007) develop vehicle and passenger revenue miles estimation techniques; and Hofmann and O'Mahony (2005) suggest an approach for predicting passenger transfer travel behavior. These applications indicate the usefulness and necessity of smartcard and farebox data in the fields of transit operation and planning. However, few studies primarily focus on comprehensive investigation on these transit data to ensure the data quality before use. Recently, Robinson et al. (2014) generally identify four potential sources of problems that can result in erroneous smartcard data, namely software, hardware, data and user. Robinson et al. define rules for three issues, including identifying trips by grouping a sequence of rides, identifying the errors in tap-ins and tap-outs records, and identifying malfunctioning bus fleets and data supply. Wong (1997) states that the farebox data errors mainly resulted from drivers' behaviors by comprehensively reviewing the uses of farebox data in transit agencies in the USA based on the experience of the Massachusetts Bay Transportation Authority (MBTA). But the author does not list the potential types of errors in farebox data and no data-driven approaches proposed to assist in identifying the errors. Furth (1995, 1996) also points out drivers' behaviors were the primary potential causes of the erroneous farebox data. Again, systematic approaches are not provided to identify these errors.

The approach outlined in this investigation seeks to build on research conducted by Wong (1997) and Furth (1995). Both publications investigate farebox datasets and draw conclusions about potential sources of error. Our research differs in that we seek to develop functional data analysis techniques that transit professionals can readily implement and use to improve reports that are subject to sensitive time and deadline constraints. This study explores the development of a range of techniques that can be applied by data management transit staff in the production of time-sensitive reports. Ultimately, the goal of the larger case study is to support transit staff efforts to build a cost-effective, comprehensive, integrated, and responsive data system.

The remainder of this paper is organized as follows. The data used in this study is first introduced. Four types of farebox data errors are described, followed by the quantitative measures and potential causes. The impacts of the data errors on transit planning are discussed in the end.

2 Study data

Raw farebox data used in this study is provided by the Bi-State Development Agency dba Metro Transit. Metro Transit, Saint Louis, the largest public transportation provider to the public in the Greater St. Louis region. Fixed-route buses, light-rail and demand response paratransit vans connect travelers across three independent funding jurisdictions in two states (City of Saint Louis, MO; Saint Louis County, MO; Saint Clair County, IL). In fiscal year 2014, Metro reports an estimated 30 million passenger trips were transported on a 385-bus vehicle fleet.

Farebox equipment installed on the revenue bus vehicle fleet can automatically accept, tally, and register a limited set of fare types, including the default base fare for 1-Ride purchases and any pass product distributed on magnetic stripe cards. All

Table 1 Basic statistical descriptions of revenue and ridership per vehicle trip

Statistics	Revenue ($)	Ridership (person)	Revenue ($)-zero-reset records[a] removed	Ridership (person)-zero-reset records[a] removed
Mean	6.50	14.1	9.50	21
Median	0.0	4.0	4	12
90th percentile value	20	40	25	49
95th percentile value	29	56	34	65
Minimum	0	0	0	0
Maximum	346	3463	346	3463
Standard Deviation	13.20	25.38	15.07	28.44
Total Number of Records	12,594,181	27,444,724	12,594,181	27,444,724

[a] Zero-reset records are defined as "Large numbers of records with values of zero for key attributes were found, mostly two times per day", defined in Sect. 3.1.1

other fare products require a manual driver registration via a nine-digit button keypad also mounted on the box. Drivers are also responsible for manually entering key service attributes associated with collected fares—such as the route, operator badge and shift, and trip—via the nine-digit keypad at the start of each work shift and the start of each trip.

In this case study, we utilize all data recorded on electronic farebox equipment from all fixed route bus during the 2011 calendar year. Each record in the dataset is a summary tally of passenger boardings, fare types, and cash revenue received on a single route trip. Revenue and ridership are two primary factors in transit planning and are the focus of this study. Table 1 presents a basic statistics summary for revenue and ridership using the entire dataset. Several unreasonable numbers in Table 1 motivated the study of farebox data error identification. For example, the median value of the revenue collected per vehicle trip remains zero, meaning at least 50 % of revenue records are recorded as zero; while the average revenue per trip is $6.50. Moreover, the maximum of ridership per vehicle trip reaches 3463. Therefore, the potential data errors causing these biased results will be investigated in this study.

3 Farebox data error

After investigating the entire study dataset, "zero-reset" records and three types of suspect records are identified in the dataset. The suspect records are categorized as (1) Potential Duplicate Records; (2) Simultaneous Records; (3) Outlier Records. The impact on ridership estimations are dependent on the type of record problem and mechanism that likely produced the problematic trip record. It should be noted that "zero-reset records" is not considered as a type of data error because neither revenue nor ridership are positively or negatively offset by the zero-reset records, as shown by the Total numbers in Table 1. Since the causes of the three types of suspect records are independent, the corresponding methods and rules proposed to identify the errors are also independent and will be explained individually in this section. In practice, "zero-reset" records and three types of suspect records can be examined without a particular order.

3.1 Zero-reset records

3.1.1 Definition

Large numbers of records with values of zero for key attributes were found, mostly at two times per day.

The farebox executes a function test at midnight that results in a record full of zeroes. This explains the high count volume at that time of day. As Fig. 1 illustrates, zero-reset records are also generated around peak times for pull-ins to the garage, indicating that these zero-reset records were an artifact of the farebox log-off process. Because they do not contain ridership information, zero-reset records do not pose a risk to aggregate ridership analyses. Nevertheless, flagging and removing these records from the dataset is strongly recommended to minimize any confounding interpretations related to true zero-boarding revenue trips and faulty records generated due to equipment failure.

Two rules are mathematically defined to flag zero-reset records.

(A) Rule 1: 0 Resets related to Log-Off
If (run = 0 and trip = 0 and route = 0 and ridership = 0 and revenue = 0)

(B) Rule 2: Resets related to timestamp resets at midnight
If (hh:mm = 00:00 and ridership = 0 and revenue = 0)

3.1.2 Results

A total of 621,469 records are flagged as zero-reset records and removed from the dataset. Figure 1 below depicts two trend lines, squared dot line and solid triangle line, with respect to the number of unmarked records after applying Rule 1 and Rule 2, respectively. The resulting count of records is more consistent with expected

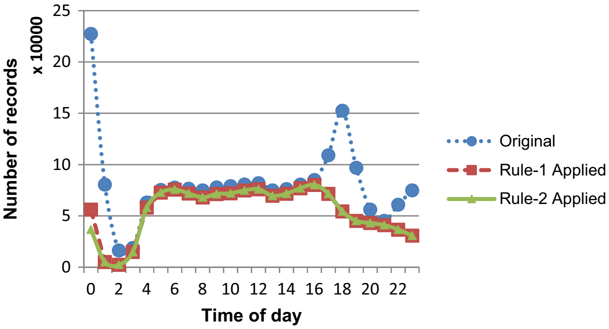


Fig. 1 The number of records by time of day

patterns by the time of day. The large reduction percentages more likely occur in the early mornings and at dusk. The records without being flagged as zero-reset records, 1,320,128 records in total, are used in the following analysis.

3.1.3 Potential causes

Two major reasons may produce the zero-reset records.

- When the farebox equipment executes function test at mid-night sharp.
- When the farebox equipment is shut off or logged out.

The nature of the equipment mechanism may account for the zero-reset records.

3.2 Potential duplicate records

3.2.1 Definition

Two or more records are defined as potential duplicates if the values of revenue, ridership, and key service identifiers such as route, run, vehicle, operator, and trip are mirrored, and the two records share the same timestamp.

3.2.2 Keep vs. delete

To evaluate the impact of inclusion of these records in the dataset, a "Keep vs. Delete" procedure is presented. Since it is challenging to determine whether a duplicate record is a true duplicate, potential duplicate records are annotated and ridership estimations are compared in both scenarios of "Keep" and "Delete". The "Keep" scenario shows the case where the entire set of potential duplicates is kept, while the "Delete" scenario shows the case where all the potential duplicates are removed.

3.2.3 Results

This data error appears in the dataset infrequently. Revenue estimations in 2011 could be overestimated by $503 and by 2344 vehicle trips. Noteworthy is that most of these errors occurred as a result of an equipment malfunction on one vehicle over a handful of service days, thus disproportionately impacting route and trip-level ridership estimations. Table 2 compares the results of estimation of revenue and ridership before and after deleting the potential duplicate records.

3.2.4 Potential causes

Potential causes are currently unknown. One guess may be that the raw data is abstracted from the farebox system more than once. In addition, a mechanical farebox failure could potentially generate duplicate records.

Table 2 Estimation of revenue and ridership by applying "Keep vs. Delete" procedure

Estimation of revenue	Total ($)	Percentage (%) of the year's value	Estimation of ridership	Total (person)	Percentage (%) of the year's value
Keep	893	0.01	Keep	3788	0.01
Delete	390	0.00	Delete	1444	0.00
Offset	503	0.00	Offset	2344	0.01

3.3 Simultaneous record

3.3.1 Definition

"Simultaneous records" are defined as multiple records that share the vehicle identifier and route-trip start timestamp. These records are similar to duplicated records discussed above in that they share values for some key attributes, but "Simultaneous Records" may have different values for other attributes, ridership, or revenue, indicating that they are unlikely to be a result of the system simply recording the same record multiple times. Therefore, these records are more likely to contain meaningful transit information. Table 3 shows examples of simultaneous records.

3.3.2 Keep vs. pick procedure

Because the similarity of these records makes it unclear whether they represent actual transit information or are strictly the result of an error, a "Keep vs. Pick" procedure is proposed to compare the impacts on ridership and revenue estimations under "Keep" and "Pick" scenarios. In the "Keep" scenario, all the records are considered as valid records, the summation of both revenue and ridership therefore can be calculated by accumulating these records. The "Pick" step assumes that only one of each group of simultaneous records is real and picks a value from either "Revenue" or "Ridership" column in one record in each group. As the groups of simultaneous records each tend to contain one record with significantly higher revenue or ridership values than the others, in the "Pick" scenario, the maximum value of "Revenue" or "Ridership" per record for each group is picked. In the example presented in Table 3, the "Keep" results in total values of revenue and ridership for this group of records of $2 and 12 people, respectively, while "Pick" changes the totals to $2 and 9 people, respectively.

Table 3 Examples of Simultaneous Records

Trip-date time	Bus	Operator	Route	Run	Trip	Revenue	Ridership
2011-10-19 6:09 PM	3505	9625	3002	27	7	0	1
2011-10-19 6:09 PM	3505	9625	3002	27	8	0	1
2011-10-19 6:09 PM	3505	9625	3002	27	9	0	1
2011-10-19 6:09 PM	3505	9625	3002	27	10	2	9

3.3.3 Results

5717 groups of records are identified as simultaneous records, with recorded revenue and ridership totals respectively of $56,755 and 151,587 boardings. In the “Pick” scenario, the chosen records would total $54,856 and 129,167 boardings. The difference to the year’s totals of choosing the “Keep” or “Pick” scenario is about 0.08 %. The detailed results are tabulated in Table 4.

3.3.4 Potential causes

A potential source of simultaneous record errors may be related to incorrect driver interaction with the farebox. Qualitative analysis of most of these records suggests drivers may attempt to correct for neglecting to increment trip numbers during a trip by incrementing all trips at the end of the shift and registering at least one boarding on each trip. We recommend monitoring this data group where warranted, correct operators who repeatedly generate simultaneous records.

3.4 Outliers

3.4.1 Definition and examples

The basic statistical descriptions of revenue and ridership are listed in Table 1, clearly indicating that there are outliers of revenue and ridership in the dataset. Outlier records are defined as the records with unreasonable values of revenue, ridership or both. Three types of outliers are empirically defined on the basis of the dataset: (A) ratio outliers; (B) real outliers; and (C) suspect outliers. The corresponding examples are shown in Table 5. In the “ratio outliers” section of Table 5, the ratio of revenue and ridership is unrealistic according to the operational experiences, although the values of revenue and ridership fall into a reasonable range. The records marked in bold are detected as outliers. The records in the “real outliers” section of Table 5 featured large ridership and associated with the bus number zeros are named “real outlier”. Table 5 “suspect outlier” demonstrates a different scenario: the values of revenue and ridership in the record with time stamp “2011-01-03 10:19” are assigned with 226 and 336, respectively. By comparing with other records in the “suspect outliers” section of Table 5, the values of this pair exceed a reasonable range. The record containing the unreasonable values is therefore considered as an outlier.

Table 4 Estimation of revenue and ridership by executing “Keep vs. Pick” method

Estimation of revenue	Total ($)	Percentage (%) of the year’s value	Estimation of ridership	Total (person)	Percentage (%) of the year’s value
Keep	56,755	0.45	Keep	151,587	0.55
Pick	54,856	0.43	Pick	129,167	0.47
Offset	1899	0.02	Offset	22,420	0.08

Table 5 Examples of outliers

Trip-date time	Bus	Operator	Route	Run	Trip	Revenue	Ridership
Ratio outliers							
2011-07-14 10:30	3265	4721	3091	747	5	23	37
2011-07-14 10:40	3265	6195	3501	113	1	14	2
2011-07-14 10:57	3265	6195	3501	113	9	54	1
2011-07-14 11:48	3265	6195	3501	113	10	22	2
2011-07-14 11:49	3265	4721	3091	747	6	6	14
2011-07-14 12:43	3265	6195	3501	113	1	0	0
2011-07-14 12:43	3265	6195	3501	113	111	3	0
2011-07-14 12:44	3265	6195	3501	113	11	17	1
Real outliers							
2011-09-24 01:13	0	0	0	0	0	11	3463
2011-03-16 00:36	0	0	0	0	0	123	960
2011-10-16 01:12	0	0	0	0	0	123	912
2011-10-16 09:12	0	0	0	0	0	123	904
2011-10-16 17:12	0	0	0	0	0	125	904
2011-10-16 01:12	0	0	0	0	0	113	900
2011-10-14 09:12	0	0	0	0	0	123	898
Suspect outliers							
2011-01-03 03:43	3230	4261	3034	51	2	0	0
2011-01-03 04:02	3230	4261	3004	51	2	0	0
2011-01-03 04:25	3230	4261	3004	51	3	4	10
2011-01-03 05:20	3230	4261	3004	51	4	10	10
2011-01-03 06:51	3230	4261	3004	51	5	22	28
2011-01-03 07:49	3230	4261	3004	51	6	33	45
2011-01-03 09:20	3230	4261	3004	51	7	19	41
2011-01-03 10:19	3230	4307	3004	51	8	226	336
2011-01-03 20:25	3230	4307	3004	51	14	8	27
2011-01-03 21:29	3230	4307	3004	51	15	8	14
2011-01-03 22:02	3230	4307	3004	51	17	0	3
2011-01-03 22:18	3230	4307	3004	51	18	20	19

3.4.2 Outlier detection

Two methods described as follows are used to detect outliers.

1. *Fixed threshold value method* This method is based on a fixed threshold of the ratio of revenue and ridership collected during a vehicle trip. The ratio is chosen as four here, because the most costly fare paid at one time for a single vehicle trip is $4/person (Metro Transit-St. Louis 2010). Since this ratio equals revenue divided by ridership, but ridership could be zero, two cases are therefore generated to deal with this issue: [A-1] ridership does not equal to zero; [A-2] ridership equals to zero, but revenue is not zero. In case [A-1], those records

with a ratio greater than the fixed threshold are consequently classified as outliers, while all the records that satisfy the case [A-2] are considered as outliers.

2. *Minimum covariance determinant* (*MCD*) *method* Many approaches have been developed to detect outliers in a dataset and are based on different techniques, including a distance-based approach and density-based approaches (Knorr et al. 2000; Breunig et al. 2000; Peña and Prieto 2001). The distance-based approach assumes that the distribution of samples can be described by the standard sample mean (location) and the covariance matrix (shape) variables. The distance between each sample and the distribution is calculated. A distance value is assigned to each sample. These distance values follow a certain statistical distribution. Using the statistical distribution, the cutoff value of the dataset—working with a user defined significant level—is able to be used to examine whether the value point is an outlier or not.

The method proposed by Hardin and Rocke (2005) is used in this study. This method uses the minimum covariance determinant (MCD) to estimate the location ($\bar{X}^*$) and shape ($\bar{S}^*$) of a dataset, instead of the standard sample mean which is easily affected by the outlier. Following the determination of location and shape, each single data point is numerically assigned with a distance from the MCD to itself. The derived distances follow an F distribution. Equation (1) mathematically defines the relationship of the distances and the F distribution. A user defined significance level and a calculated degree of freedom combine for determining a cutoff value. The outliers may thus be found when the distance of a data point is greater than the cut-off value.

$$\frac{c(m-p+1)}{pm} \times \mathrm{d}^2_{\mathrm{s}*}(X_i,\ \bar{X}^*) \sim F_{p,m-p+1} \tag{1}$$

$$c = \mathrm{E}[\mathrm{s}^*_{jj}]$$

$$m = \frac{2}{CV^2}$$

where, $\bar{X}^*$ is MCD location estimator; s^*_{jj} represent the diagonal elements of $\bar{S}^*$; E[*] is the expectation of *; $\mathrm{d}^2_{\mathrm{s}*}(X_i,\ \bar{X}^*)$ means the distance from X_i to $\bar{X}^*$; p is the number of dimension of this dataset; m is the degree of freedom associated with the F distribution; CV is the estimated coefficient of variation of diagonal element in $\bar{S}^*$

As noted previously, three types of records may exist in the results of outlier detection by applying the MCD method; which are: (B) real outliers, (C) suspect outliers, and valid records of route trips. Various levels of significance are therefore tested, in order to choose an appropriate value for the outlier detection.

3.4.3 Results

The two methods are capable of capturing different types of outliers. The fixed threshold value method is able to detect the (A) ratio outliers with a large gap between revenue and ridership, regardless of whether the values of revenue and ridership fall into a reasonable range. The MCD method is used to capture the (B) real outliers and (C) suspect outliers in which cases either revenue or ridership or both of them are beyond a certain range.

By executing the threshold of the fixed threshold value method on the combinations of revenue and ridership, 209 records are marked as ratio outliers in type [A-1]. The total revenue and ridership of the marked outliers are $2267 (0.02 % of the year's total) and 309 people (less than 0.01 % of the year's total), respectively. In type [A-2], 3854 records are marked as "ratio outliers". The total revenue of these "outliers" is $11,169 (0.09 % of the year's total), and by definition, there is no ridership in these records.

In order to examine the results of type (B) and (C) outlier detection when the significance level is modified, the values from 0.01 to 0.04, at intervals spaced by 0.01, and 0.05 to 0.75, spaced by 0.05, are tested. The results of the examination of significant levels indicate that the significance level of 0.02 gives a good compromise with respect to differentiating the types (B) and (C); while the significance level 0.6 may differentiate between type (C) and valid records. Table 6 summarizes the revenue, ridership, number of records and relevant percentages separated by two types of outliers (B) and (C). Only 15 records are detected as outliers, but the results imply that $1572 and 15,988 people would be overestimated in the year 2011. Further, a total of 0.87 % records in the dataset are found as suspect outliers. The total revenue and ridership impacted by the suspect outliers are relatively large, summarized as $1,252,907 and 2,210,586 people, which contribute 9.95 % of yearly revenue and 8.05 % of yearly ridership, respectively.

3.4.4 Potential causes

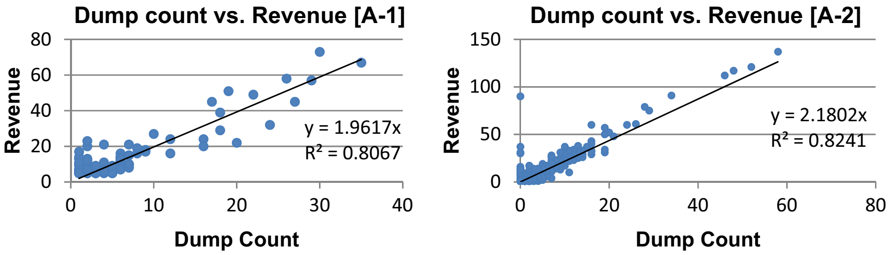
One potential cause of the type (A) may be that the operators forget to push buttons to count the people boarding on buses. A special relationship of revenue and dump count can be found in case of type (A). Figure 2a, b visually show a strong linear relationship between revenue and dump count when type (A) occurs. After applying the linear model to both revenue and dump counts, the resulting linear equations provide the mathematical evidence to demonstrate the strong linear relationship. The slopes of the equations shown in Fig. 2a, b are fairly close to 2 ($2 is Metro's base bus fare). One possible explanation for this phenomenon is that the operators push the dump button, instead of registering riders, to make the cash and coins drop down to the container every time, when people board and put money into the farebox equipment.

In the case of type (B) of outliers, most of the vehicle identifiers are zeros, indicating that the farebox equipment is malfunctioning at that time.

In the case of type (C), the relatively high, but consistent, revenue and ridership of a record may result from multiple trips collapsed together when an operator does not register a new trip on the farebox. This would result in multiple route trips collapsing into a single record, with the revenue and ridership in the record summing up the collapsed trips. The supporting evidence is that the trip numbers are

Table 6 Results of outlier detection by the MCD method

Type of outlier	Revenue ($)	Percentage (%)	Ridership (Person)	Percentage (%)	Number of Records	Percentage (%)
(A) Ratio outlier						
[A-1] (Ridership $\neq$ 0)	2267	0.02	309	0.00	209	0.02
[A-2] (Ridership = 0)	11,169	0.09	0	0.00	3854	0.29
Total	13,436	0.11	309	0	4063	0.31
(B) Real outlier	1572	0.01	15,988	0.06	15	0.00
(C) Suspect outlier	1,252,907	9.95	2,210,586	8.05	11,539	0.87



(a) in case of type [A-1] **(b)** in case of type [A-2]

Fig. 2 The relationship between dump counts and revenue in case of type (A)

inconsistent in case of type (C) or a relief operator failed to log-out the previous run and log-in their own run.

4 Impacts on transit planning

Successful transit planning depends on reliable ridership and revenue information. Three dominant data errors are discussed above, along with their impacts on the estimations of revenue and ridership. Different types of data error may have an impact on these measures at different levels of analysis, from system-level aggregation to trip-level analysis.

Table 7 summarizes the results demonstrated in Sect. 3 and gives an overview of how these data errors could affect the most important estimations in transit planning at system and trip level. The impacts resulting from the data error of simultaneous records perform differently in analyses for these two levels. For example, when taking these simultaneous records into consideration, the values of revenue and ridership might be overestimated in total, while the values might be underestimated for an individual route trip. The impacts of suspect outliers also vary at the system and trip levels. These records may result in overestimating the revenue and ridership of individual route trips. It is found that ridership and revenue could be overestimated up to 8.05 and 9.95 %, respectively, due to the suspect outliers. As mentioned, "zero-reset records" is not considered as a type of data error, because neither revenue nor ridership is positively or negatively offset by the zero-reset records. Nevertheless, "zero-reset records" would still have an impact on transit operations and management reporting, i.e., the mean values of ridership per vehicle trip and revenue per vehicle trip increase by approximately 31.6 and 33.3 % after the "zero-reset records" are removed.

Overall, based on the features of farebox data errors and the potential causes, the revenue in total would likely be overestimated. The analysis results listed in Table 7 indicate that the data errors identified in this study will either have no impact or will overestimate the revenue.

In order to investigate the impacts on trip level analysis, transit agencies often consider the frequently-posed questions:

Table 7 Overview of revenue and ridership estimations

Types of data errors	Mainly caused by	System-level		Trip-level	
		Revenue	Ridership	Revenue	Ridership
Potential duplicate record	Equipment	Overestimation (up to 0.01 %)	Overestimation (up to 0.01 %)	Overestimation	Overestimation
Simultaneous record	Operators	Overestimation/non[a] (up to 0.02 %)	Overestimation/non[a] (up to 0.08 %)	Non[a]/ underestimation	Non[a]/ underestimation
Outliers					
Ratio outlier	Operators	Non[a] (up to 0.11 %)	Underestimation (up to 0.01 %)	Non[a]	Underestimation
Real outlier	Equipment	Overestimation (up to 0.01 %)	Overestimation (0.003 %)	Overestimation	Overestimation
Suspect outlier	Operators	Non[a]/overestimation	Non[a] overestimation	Non[a] (up to 9.95 %)	Non[a] (up to 8.05 %)

[a] Non means this type of data errors may have no impact on estimation

1. How many scheduled trips were served on Route #65?
2. How many trips were run by Bus #2501?
3. What is the pattern of revenue and ridership on Route #65?
4. Are the buses on Route #65 at capacity in the morning peak hour?
5. What is the total revenue and ridership on Route #65?

These questions could not be perfectly addressed using the raw farebox records. Specifically, the number of trips running on Route #65 would be overestimated if the zero records are kept, because zero records would create fake trips. Duplicated records, aside from overestimating the number of trips, would also change the pattern of revenue and ridership on the route. The number of trips run by each bus would be valuable information for use in maintaining buses, but all the data errors discussed above make this count difficult to obtain reliably from the farebox data. While suspect outliers would not negatively affect the total estimates of revenue and ridership, the collapse of trip data that they represent makes it difficult to answer questions about ridership per trip versus capacity. Finally, even the basic planning question of estimation of total ridership and revenue by route is challenged by potential overestimation or underestimation at the trip level.

5 Conclusions

System design, policy development and budgeting decisions at all levels of transit management heavily rely on a constant stream of transit operations data. Electronic farebox equipment, widely used across the nation, is one of the major transit data sources. This study used the farebox data provided by Bi-State Development Agency dba Metro Transit. Since the quality of operation management data greatly affects the transit operation report generation and decision making, a systematic data-driven approach to processing ridership and revenue data was developed to examine the farebox data.

This study identified three generalized types of data error: potential duplicate records, simultaneous records and outliers. Several methods were developed to evaluate quantitatively the error impacts on trip and system level transit management analysis. According to the results, operators' behavior may not account for the errors in estimations of system-level total revenue. This conclusion is based on two conditions:

- Cash and coins are successfully recognized by the full-functioning farebox equipment.
- The large revenue values are caused by trips collapsed together and are consistent with the associated ridership values, with a ratio that falls into a reasonable range on the basis of operational experiences.

The estimation of total revenue is therefore trustworthy, assuming that the farebox equipment counts money correctly. However, the collection of ridership information relies heavily on the operators' responsibility and reliability. The ridership could easily be underestimated or overestimated if the operators do not operate the farebox equipment correctly. Therefore, potential future work could include a thorough investigation of the causes of the data errors due to human factors. The drivers' behaviors may highly relate to the drivers' training, driver rest time between two trips, etc. A comprehensive survey should be conducted to further investigate the relationship between drivers' behaviors and data input accuracy. The survey results could serve as a guidance to reduce the impact of the human factors on data quality. In addition to human factor investigation, another future work could be focused on identifying the potential data error issues caused by farebox software and hardware. Therefore, a data quality control computer program should be developed to automatically identify the issues of software and hardware, cleaning the erroneous data, and generate daily farebox health reports.

Acknowledgments The authors would like to thank Metro Transit, St. Louis for data support.

References

Breunig MM, Kriegel HP, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM Sigmod international conference on management of data, pp 93–104 **(Association for Computing Machinery)**

Corey J, Lao Y, Wu Y, Wang Y (2012) Detection and correction of inductive loop detector sensitivity errors by using Gaussian Mixture Models. Transp Res Rec J Transp Res Board 2256(2011):120–129

Furth PG (1995) Integrating electronic fareboxes with other on-board equipment (pp. 1–30). http://onlinepubs.trb.org/onlinepubs/archive/studies/idea/finalreports/transit/Transit4_Final_Report.pdf. Accessed 10 Sep 2014

Furth PG (1996) Integration of fareboxes with other. Transp Res Rec J Transp Res Board 1557:21–27. doi:10.3141/1557-04 **(Public Transportation 1996: Bus, Rural and Intercity, and Paratransit)**

Greater Cleveland Regional Transit Authority (2005) How to Use RTA's Fareboxes Fares Greater Cleveland Regional Transit Authority. http://www.riderta.com/fares/farebox.asp

Hardin J, Rocke DM (2005) The distribution of robust distances. J Comput Gr Stat 14(4):928–946

Hofmann M, O'Mahony M (2005) Transfer journey identification and analyses from electronic fare collection data. In: Proceedings of the 2005 IEEE Intelligent Transportation Systems, pp 34–39

Knorr EM, Ng RT, Tucakov V (2000) Distance-based outliers: algorithms and applications. VLDB J Int J Very Large Data Bases 8(3–4):237–253

Lee H, Coifman B (2012) Identifying and correcting pulse-breakup errors from freeway loop detectors. Transp Res Rec J Transp Res Board 2256(2011):68–78

Lee S, Hickman M (2014) Trip purpose inference using automated fare collection data. Public Transp 6:1–20. doi:10.1007/s12469-013-0077-5

Metro Transit-St. Louis (2010) Fare chart. http://www.metrostlouis.org/FaresPasses/FareChart.aspx. Accessed 10 September 2014

Lu A, Reddy A (2012) An algorithm to measure daily bus passenger miles using electronic farebox data for national transit database (NTD) section 15 reporting. Transp Res Rec J Transp Res Board 2216(2011):19–32

Ma X, Wu YJ, Wang Y, Chen F, Liu J (2013) Mining smartcard data for transit riders' travel patterns. Transp Res Part C Emerg Technol 36:1–12. doi:10.1016/j.trc.2013.07.010

Moore, G. R. (2002). Transit ridership efficiency as a function of fares. J Public Transp 5(1)

Navick DS, Furth PG (2007) Estimating passenger miles, origin-destination patterns, and loads with location-stamped farebox data. Transp Res Rec J Transp Res Board 1799(2002):107–113

O'Malley M, Brown AG, Swaim-Staley BK, Mobley DB (2011) Farebox recovery attainment and operational requirements. Hanover, Maryland. http://dlslibrary.state.md.us/publications/Exec/MDOT/MTA/TR7-208(b)_2011.pdf

Pelletier M-P, Trépanier M, Morency C (2011) Smart card data use in public transit: a literature review. Transp Res Part C Emerg Technol 19(4):557–568. doi:10.1016/j.trc.2010.12.003

Peña D, Prieto FJ (2001) Multivariate outlier detection and robust covariance matrix estimation. Technometrics 43(3):286–300 **(American Statistical Association and American Society for Quality)**

Robinson S, Narayanan B, Toh N, Pereira F (2014) Methods for pre-processing smartcard data to improve data quality. Transp Res Part C 49:43–58. doi:10.1016/j.trc.2014.10.006

Wong ASF (1997) The role of new technology in improving data collection for public transportation. Thesis, Massachusetts Institute of Technology

Wu Y, Zhang G, Wang Y (2010) Volume data correction for single-channel advance loop detectors at signalized intersections. Transp Res Rec J Transp Res Board 2160(2010):128–139