

## **Statistical Evaluation of Transit Onboard Survey Results and Methods**

Arash Mirzaei, Kathleen Yu, Hua Yang, and Behruz Paschai  
Transportation Department, North Central Texas Council of Governments  
616 Six Flags Drive, Arlington TX 76005-5888, USA  
kyu@nctcog.org

### **ABSTRACT**

The statistical accuracy of transit onboard surveys is often defined only through the number of samples required based on daily boardings and the desired confidence interval. It is implicitly assumed that if the target samples are met in data collection, the desired accuracy of the survey is reached. However, the quality of the complete records in the final database has a direct effect on the quality of the survey applications. Without formally defining the accuracy of the final records, it is almost impossible to comfortably rely on the survey data. This project describes methods that were used in evaluation of the Dallas Area Rapid Transit Onboard Survey conducted in Spring 2007.

Through investigational review of the survey instrument and actual responses on the questionnaires, a systematic statistical approach was developed to define the level of confidence that can be expected for major components of the survey including origin and destination locations, transit path, and boarding, alighting and transfer locations. The methods were implemented on 6,447 completed survey records in the DART 2007 Onboard Survey. From this, more than 3,000 survey forms were printed for manual investigation and correction. As a result of statistical tests, a clear and detailed definition of correct and incorrect coding was developed for every field of interest in the database. The result of this effort was a database of corrected records and statistical definitions associated with the evaluated fields.

In addition to the quality control methodology, two supplemental data collections are suggested for future onboard surveys. The first data collection is an effort to adjust the expansion method based on route boarding. It includes recording activities at major park-and-ride and transfer centers by mode of access and egress as a cross sectional check of the expansion. The second data collection is a short interview survey with a small sample to gain an insight into the major biases of the main survey. These biases are introduced by factors such as short trip under-sampling, socio-economic skewing of samples, and the fraction of non-responses. These new surveys were added to the Fort Worth Transportation Authority and Denton County Transportation Authority Onboard Survey which was conducted in Fall 2008.

### **INTRODUCTION**

The Dallas Area Rapid Transit (DART) onboard survey completed in Spring 2007. Subsequently, the final expanded database was delivered to the Travel Model Development group of North Central Texas Council of Governments. The initial review of the database revealed some inconsistencies within the data for each record that triggered the need for a detailed quality control of the database. The respondent's confusion in understanding the context of the questions could have been a possible reason for the observed inconsistencies. As a result of this quality control, a list of suggested improvements to future onboard surveys was

developed. NCTCOG also had the opportunity to implement these changes in the onboard survey conducted in Fall 2008 in the same metroplex with two other transit agencies.

**ANALYZING THE 2007 DART ONBOARD SURVEY DATABASE**

DART is one of three major transit service providers in the Dallas/Fort Worth metroplex. The DART Transit System consists of 111 routes and four modes: commuter rail, light rail, express bus, and local bus. The total weekday boardings for the DART system in 2007 were approximately 218,000. A detailed breakdown of the system modes is shown in **Figure 1**.

**Figure 1. Dallas Area Rapid Transit System Description by Mode**

<b>Transit Mode</b>	<b>Weekday Boardings</b>	<b>Number of Routes</b>
Commuter Rail	9,000	1
Light Rail	62,000	2
Express Bus	6,700	10
Local Bus	141,000	98
<b>Total DART System</b>	<b>218,700</b>	<b>111</b>

***Summary of the 2007 DART Onboard Survey***

DART conducted an onboard survey in Spring 2007. The Request for Proposals (RFP) was posted in December 2006 and the contract was signed in March 2007. With only a short period between signing the contract and the end of data collection, a formal pilot test program was eliminated.

The sample size specified in the data collection plan was based on daily route boardings. They randomly sampled vehicles of each route in each direction for the four time periods: a.m. peak, mid-day, p.m. peak, and off-peak. The sample size was calculated for a confidence level of 95% with a 5% margin of error. One surveyor would record the number of adult boardings onto the vehicle route at each stop; an adult was defined as a person who was at least fifteen years of age. The second surveyor would offer questionnaires to all adults who boarded at the stop and ask them to participate in the survey. In May 2007, the data collection was completed. In November 2008, the final expanded survey database was delivered to North Central Texas Council of Governments (NCTCOG).

***Developing a list of the Quality Control Checks***

When the database was received, NCTCOG developed a series of quality control checks to make sure that the database was coded correct from the survey forms. These checks consisted of the following items: evaluation of the geo-coding, identification and correction of inconsistent answers, and assessment of the final quality of the database.

***Evaluation of the Geo-Coding***

The evaluation of the geo-coding involved determining if the origin and destination were geo-coded correctly. The definition of correct geo-coding must consider that small differences are expected when comparing coordinates geo-coded through different tools.

To assess the quality of the geo-coding, the normal approximation of a binomial distribution was used to determine the random sample size for different success rates for a 95% confidence interval. A sample size of 74 (of the 6,447 weekday records in the database) corresponded to a 95% probability of success with 5% margin of error. In order to compare the geo-coding of origins for the sample records, the origin specified on the original survey form was geo-coded by NCTCOG. Then, the straight line distance between the NCTCOG's geo-coded origin and the geo-coded origin in the database was calculated. After reviewing the surveys, it was found that 71 of 74 surveys (95.9%) met the criteria of having the database origin point geo-coded within 0.75 miles of the actual origin location. As a result, NCTCOG was able to assert the following statement: "We are 95% confident that 95% of the origin points are geo-coded within 0.75 miles of the user-specified origin place and address, with a 5% margin of error."

A similar review was conducted on the geo-coding of destinations by evaluating 74 records. It was found that 72 of 74 surveys (97.3%) met the criteria of having the database destination point geo-coded with 0.75 miles of the actual destination location. As a result, NCTCOG was able to assert the following statement: "We are 95% confident that 95% of the origin points are geo-coded within 0.75 miles of the user-specified destination place and address, with a 5% margin of error."

#### ***Identification and Correction of Inconsistent Answers***

The second set of quality control checks involved reviewing responses within a survey for consistency. In this process, it was observed that there were six questions within the survey which were related to determining the respondent's path. These questions asked for the following information:

- the route/line that the respondent transferred from;
- the total number of buses/trains that the respondent will ride on this one-way trip;
- the route/line that the respondent transferred to;
- the first rail line and station that the respondent used on this one-way trip;
- the last rail line and station the respondent alighted on this one-way trip; and
- a list of all bus routes and rail lines in the exact order used on this one-way trip.

Since these questions asked for some redundant information, responses to these questions were checked for consistency. Of the 6,447 weekday surveys, 2,593 were flagged because they possibly contained inconsistent information.

In order to improve the understanding over why inconsistencies occurred, NCTCOG reviewed sample surveys and found some of the reasons for the problems as follows:

- passengers described all possible routes they could take for their trip and not the ones they are specifically taking on this trip;
- passengers described their round trip and not a one-way trip;
- passengers put down origin and destination for their round trip, but described the path for the one-way trip – or vice versa; and
- passengers described their reverse trip rather than what they were taking at the time of the survey.

In order to evaluate the possibility of a reverse trip, a computer program was written to compare whether the origin point was closer to the last route/line in the stated path sequence than to the

first route/line. After this test, an additional 575 weekday surveys were added to the list of flagged surveys. Through this evaluation, it also was determined that there were not many similarities between the errors, so it would not be possible to use database queries to fix the paths. Instead, it was decided to individually review each of the 3,168 surveys to determine whether the stated path was reasonable. The basic rules for this path correction are as follows:

1. Create a map with the origin, destination and all routes reported in the answers to the redundant questions in the surveys.
2. Assume that the origin point, destination point, and surveyed “route” are correct.
3. Review routes listed in the path sequence question and the routes listed in the transfer questions to see if they contain a reasonable path.
4. If needed, use any routes listed on the survey to create a reasonable path from the origin to the destination using the surveyed “route.”
5. For a walk mode of access or egress, confirm that the walk distance is less than or equal to 2.5 miles.
6. If a reasonable path cannot be found, review the origin and destination geo-coding. If the origin or destination needs to be updated, correct them and then use any routes listed on the survey to create a reasonable path for the new origin and new destination using the surveyed “route.”

Following these rules, 164 surveys were removed from the database because there was no path using the surveyed route between the origin and destination or the path included a walk mode of access or egress of greater than 2.5 miles. The other 3,004 surveys were returned to the database with the corrected information on the path, origin and destination.

After cleaning the database, comparisons were made between the successes of the original path questions compared to the cleaned path records. The transfer questions are the two question sin the survey which asked for the “transfer from route” and “transfer to route.” The question asking for the list of all routes and lines used on this one-way trip was labeled as the sequence question. The table in **Figure 2** describes how often the sequence question and transfer questions corresponded to the final corrected path for all weekday surveys in the database. This table shows that the sequence question was more frequently closer to the corrected path than the transfer questions were; as a result, the sequence question may be preferred to the transfer question for obtaining the path sequence in future onboard surveys.

**Figure 2. Percentage of Surveys where the Transfer Questions or Sequence Question corresponded to the Corrected Path**

<b>Path</b>	<b>Transfer Questions</b>	<b>Sequence Question</b>
No Transfer	86%	98%
1 Transfer	73%	73%
2 Transfers	26%	59%
3 Transfers		24%

Path	Transfer Questions	Sequence Question
Overall	60%	72%

***Assessment of the Final Quality of the Database***

As the final step of cleaning the paths, each survey was reviewed to determine if the path sequence in the database between the origin and destination was reasonable. It was found that 71 of the 74 randomly selected records (95.9%) had a correct path. Therefore, NCTCOG was able to assert that “We are 95% confident that 95% of the path sequences are correct, with 5% margin of error. A path sequence is correct when the respondent-identified sequence from its origin to its destination is feasible.”

After cleaning the path sequence, NCTCOG ran a program to find the boarding and alighting stops for each route/line in the final path sequence for each survey. The boarding and alighting stops were then reviewed to determine the reasonableness of each stop within the one-way trip. As a result, NCTCOG was able to assert that “We are 95% confident that 95% of the geo-coding of boarding and alighting locations of all records are correct, with a 5% margin of error. Geo-coding of boarding and alighting locations of a record is considered correct if a visual inspection indicates that (1) geo-coding of both the first boarding stop and last alighting stop are considered reasonable and (2) geo-coding of at most one middle stop is unreasonable.”

***Suggested Improvements for Future Onboard Surveys***

Based on this experience with cleaning the database, there are some recommendations that were developed for planning onboard surveys. First, pilot tests should be conducted in a timely manner in order to allow for a complete evaluation of possible survey instruments. The list of questions should be reviewed so that redundant questions are eliminated to minimize respondent confusion and to improve the response rate. In addition, the questionnaire should only contain necessary questions since people will be more able to complete a shorter questionnaire than a longer one. Another important recommendation from this work is the quality of the data can be quantified using a statistical evaluation of the fields in the database with the actual surveys. Stating the required confidence expected within the surveyor contract may improve the quality of the resulting onboard survey database.

**PLANNING AND MANAGING THE 2008 FWTA AND DCTA ONBOARD SURVEY**

In 2008, NCTCOG prepared to manage an onboard survey for the Fort Worth Transportation Authority (the T) and the Denton County Transportation Authority (DCTA). The T and DCTA consist of nine express bus routes representing a total of 1,126 in weekday boardings in 2007 and 53 local bus routes representing a total of 32,882 weekday boardings in 2007. The T and DCTA transit systems are shown in

**Figure 3** with modes highlighted by color.

**Figure 3. Transit System of the Fort Worth Transportation Authority and Denton County Transportation Authority**



Since NCTCOG was managing the T and DCTA Onboard Survey, it was possible to implement the suggested recommendations developed from the previous work with the 2007 DART Onboard Survey database. The improvements included the following: clarification of the expected quality in the surveyor contract, reduction of the number of questions, clarification of the one-way trip, improvement of the survey instrument through pilot tests, and improvement of the sample expansion plan with supplemental data.

#### ***Clarification of the Expected Quality in the Surveyor Contract***

In order to improve the quality of the final database, NCTCOG included the quality expected within the RFP and the final contract. The definition of a complete survey was stated using the wording given below. This also stated the expected confidence interval and margin of error for complete survey records.

*At the end of the data collection, the consultants should show that they are 95% confident that 90 to 100% of the collected surveys are useable. A filled questionnaire is considered useable if the respondent properly answered all of the following questions: origin, destination, purposes, path questions that identify the routes that the users take, mode of access and egress, time of survey, and home address.*

Assuring confidence that the surveys are coded correctly was done by defining the expected quality of the coding of the data from the surveys in the RFP. The document stated this requirement using the following wording:

*The consultants should show they are 95% confident that 90 to 100% of location data and all other collected information in the survey instrument are coded correctly. This quality will be tested by random sampling of the final coded records. A record is considered incomplete if there is missing or wrong information in any of the following fields: origin, destination, purpose fields, path fields that identify the routes that the users take, mode of access and egress, time of survey, and home address.*

In addition to incorporating these new contract requirements, NCTCOG followed up with the consultant to make sure these requirements were met during the course of the survey.

### ***Reduction of the Number of Questions***

Among ways to improve the survey questionnaire are minimizing the number of questions and removing confusion in the questions. This will increase the passengers' ability to complete the questionnaire while on the bus. Because of this, NCTCOG created a question list using the 2007 DART Onboard Survey questionnaire as their source. Then they met with representatives of the transit agencies involved to determine which questions were needed for their work. In the end, six questions were removed from the question list; these six questions asked the following: the main reason for taking the survey route, the perceived length of their trip, the frequency they use transit, whether they were weekday or weekend users, what mode would they use if the bus was not available, and the number of adults in their household. In addition, the number of questions which asked about the path was limited to the following two questions: state the total number of buses/trains that the respondent will and list all bus routes and rail lines in the exact order used on this one-way trip.

### ***Clarification of the One-Way Trip***

In the 2007 DART Onboard Survey, there was confusion about what a one-way trip was since respondents listed one-way trips, reverse trips, and round trips in different portions of the survey. So, in addition to reducing the number of questions about the path sequence, NCTCOG staff was of the opinion that it was important to re-enforce the idea of a one-way trip. This was accomplished by using graphical images within the survey form to describe a one-way trip, advertising the one-way trip and the survey, and establishing consistency in the use of the phrase "one-way trip" in the survey instrument.

The 2007 DART Onboard Survey instrument included only one graphic diagram of a one-way trip. However, when preparing the survey instrument for the 2008 survey, NCTCOG included an additional graphical image in the design. The two graphical images would display the use of different trip purposes and use of different modes of access and egress. It was thought that this would help the riders to realize that they did not have to fit their trip into a particular mold. An example of the image used is shown in **Figure 4**.

**Figure 4: Graphical Image of One-Way Trip used in the 2008 T and DCTA Survey Instrument**



**Note: Your trip may be different from our examples.**

This image was then included in advertising campaigns that appeared on the buses one week before the surveys were distributed. On the T, rail hangers containing the graphic images and information about the survey were created. On the DCTA buses, signs were placed on the inside of the buses. Each advertising material appeared in both English and Spanish.

Finally, on the actual survey instrument, the questions were reviewed to find that the wording of the "one-way trip" was not consistent. Once the phrase "This One-Way Trip" was finalized as

the preferred wording, all questions were reviewed until it was confirmed that both the wording and the font of the phrase were consistent throughout the questionnaire.

### Review the Survey Instrument through Pilot Tests

Another way in which NCTCOG attempted to improve response rate was to include a pilot test in the contract. By improving how the survey instrument asked the information about the path, it was believed that the origin, destination, and path sequence information might be filled out more correctly. The proposed format of the new questionnaire design was a graphical design which would provide a visual way of viewing the path sequence. The final version of the graphical survey instrument is shown in **Figure 5**.

**Figure 5: Graphical Design of the Questionnaire**

Two types of pilot tests were conducted to test the graphical design against the traditional design of the survey instrument. The first pilot test was an in-field test of the questionnaire where both survey instruments were offered alternately to passengers on a bus on one day. The second pilot test was a set of cognitive interviews. Twenty-eight transit riders were recruited and interviewed for one hour. In each interview, the rider was asked to fill out one survey instrument, review the other instrument, describe how they answered the questions, and offer their perspective on the pros and cons of each instrument.

The results of the in-field test found that the completion rate of the traditional and graphical designs were 69% and 50%, respectively. The results of the cognitive interviews were not conclusive, since there was a mixed response to both questionnaires. Results of the cognitive interview suggested some small additions to the traditional survey instrument which were implemented to produce the final questionnaire design shown in

**Figure 6.**

Figure 6: Final Questionnaire Design

**2. How did you GET FROM THAT PLACE (START) to the VERY FIRST BUS on THIS ONE-WAY TRIP? (Please fill in one bubble only)**  
**NO TRANSFER INFORMATION PLEASE**

Walked  Bicycled  Other (specify): \_\_\_\_\_  
(transferring from another bus is not a valid response)  
 Wheelchair  Carooled and parked:  → Parking location? \_\_\_\_\_  
Parking Lot Name/Cross Street  
 Dropped off  Drove alone and parked:  → Parking location? \_\_\_\_\_  
Parking Lot Name/Cross Street

**4. What TYPE OF PLACE are you GOING TO NOW? (END of THIS ONE-WAY TRIP) (Please fill in one bubble only)**

Work  Shopping  Restaurant  Other (specify): \_\_\_\_\_  
 College/University (student only)  Social/Recreational  Other (specify): \_\_\_\_\_  
 School (K-12) (student only)  Medical appointment/Hospital visit  
 My Home → If you gave your Home address in Question 1 → Go to Question 5

**a. What is the NAME of the PLACE, BUSINESS OR BUILDING you are GOING TO NOW?** Example: (S)M(I)T(I)H (M)O(S)T(I)E(T)A(L)

Name of Place \_\_\_\_\_

**b. What is the ADDRESS? (Provide the NEAREST CROSS STREETS if you don't know the EXACT ADDRESS.)** Example: L(1)R(1)O(S) (1)S(1) M(A)J(I)N(1)S(T)

Address \_\_\_\_\_

Cross Street #1 \_\_\_\_\_ Cross Street #2 \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

**5. How will you GET FROM the VERY LAST BUS to the END of THIS ONE-WAY TRIP? (Please fill in one bubble only)**  
**NO TRANSFER INFORMATION PLEASE**

Walk  Bicycle  Other (specify): \_\_\_\_\_  
(transferring from another bus is not a valid response)  
 Wheelchair  Carpool and park:  → Parking location? \_\_\_\_\_  
Parking Lot Name/Cross Street  
 Get picked up  Drove alone and park:  → Parking location? \_\_\_\_\_  
Parking Lot Name/Cross Street

**6a. Including this bus, how many TOTAL BUSES AND TRAINS will you use to make THIS ONE-WAY TRIP?**

One, only this bus → Go to Question 7  Two  Three  Four

**6b. List the BUS ROUTES or RAIL LINES in the exact order you use them in THIS ONE-WAY TRIP.**

FIRST - bus route/rail line I took → SECOND - bus route/rail line I took or will take → THIRD - bus route/rail line I took or will take → FOURTH - bus route/rail line I took or will take

**7. Where will you get off THIS BUS? (Name of the place and nearest cross streets)**

Name of Place (including Transit Center or Park & Ride Lot) \_\_\_\_\_

Cross Street #1 \_\_\_\_\_ Cross Street #2 \_\_\_\_\_

**8. How did you PAY for THIS ONE-WAY TRIP? (Please fill in one bubble only)**

One-way Ticket  Monthly Pass  None - I rode within the Downtown Free Zone only  
 Day Pass  Annual Pass (E-pass)  Other Pass (MITS + 1, 9th grade pass, TCU Student Pass)  
 Weekly Pass

**9. Was your FARE ...**

Local (Regular fare)  High School Student with School ID  Disabled with ID or Medicare Card  
 Premium Fare (Express, TRE Dallas-zone, DART System and Arlington Park+Ride lots)  Over 65 with ID

**10. HOW MANY MINUTES did you wait for THIS BUS you are on now?**

Less than 5 minutes  11 - 15 minutes  21 - 25 minutes  More than 30 minutes  
 6 - 10 minutes  16 - 20 minutes  26 - 30 minutes

**11. How many registered CARS, TRUCKS, or MOTORCYCLES are available to your household?**

None  1  2  3  4 or more

**12. Including yourself, how many TOTAL PEOPLE live in your household?**

1  2  3  4  5  6 or more

**13. Do you have a VALID DRIVER'S LICENSE?**

Yes  No

**14. Are you... (fill in the bubble that best describes you)**

Employed (full-time or part-time)  Unemployed  Homemaker  Retired  Student

**15. What is your AGE?**

15 - 24  25 - 34  35 - 44  45 - 54  55 - 64  65 + years of age

**16. What is your ETHNICITY? (fill in the bubble that best describes you)**

Asian  Hispanic  White  
 Black/African American  Native American  Other (specify): \_\_\_\_\_

**17. What was your estimated TOTAL HOUSEHOLD INCOME in 2007 before taxes?**

Less than \$10,000  \$15,000 - \$24,999  \$35,000 - \$49,999  \$75,000 or more  
 \$10,000 - \$14,999  \$25,000 - \$34,999  \$50,000 - \$74,999

250174-DL-0-2102000&e Please continue on the back →

**Improvement of Sample Expansion Plan with Supplemental Surveys**

In order to improve the survey expansion, a boarding and alighting study and a non-response survey were included in the contract. When expanding the final database, the expansion includes consideration of the number of complete questionnaires and boardings at the corresponding stop on the surveyed trip; the number of trips surveyed and total trips of the same route, time of day, direction, and day of the week; and the total ridership of the surveyed route<sup>1</sup>. After this expansion, the ridership of the surveyed route is matched, but there are no checks to see if the boardings at each stop on the route also match. Automatic People Counter (APC) counts were not available on the surveyed routes, and boarding counts would only be available on surveyed routes. In the boarding and alighting study, surveyors recorded the total ons and offs by mode of access and egress for 14 park-and-ride and transfer stations in the T and DCTA transit systems. These counts will be used as a cross-sectional check of the expansion. An Iterative Proportional Fitting (IPF) method will be used to develop new expansion factors for the records that match the counts at the stations.

As another way to improve the understanding of the response, a non-response survey was also included in the contract. The non-response survey would help measure major biases of the onboard survey due to suspected sources of non-response bias. These sources are considered to be short trips, ethnicity, age, and income. The survey was conducted at the same time as the

<sup>1</sup>NuStats, “Dallas Area Rapid Transit (DART): 2007 Transit Rider Study – Final Report,” prepared for Dallas Area Rapid Transit, November 2007. pp. 57-59.

onboard survey; if a person refused the onboard survey questionnaire or did not complete the form on the bus, they were then asked to participate in a short interview. The interview form is shown in **Figure 7**. Interview forms were recorded in the database if at least one of the five questions were answered. The response rate for the interview form is shown in **Figure 8**. Using the result of this survey can help adjust the expansion factor to account for non-response and decrease the expansion biases.

**Figure 7: Non-Response Interview Form**

**NCTCOG (The T) On-Board Non-Response Survey**

Assignment #: \_\_\_\_\_ Trip #: \_\_\_\_\_ Route #: \_\_\_\_\_

**1) Reason for not taking/completing a survey:**

- Never participate
- Too many questions
- Not interested/Don't Care
- Conditions on bus not suitable
- No time to complete it on this trip / trip too short
- Other (specify): \_\_\_\_\_

**2) How many minutes will you be traveling on THIS BUS for THIS TRIP?**

- 5 or less
- 6-10
- 11-15
- 16-20
- 21-25
- 26-30
- More than 30

**3) Age**

- 15 - 24
- 25 - 34
- 35 - 44
- 44 - 54
- 55-64
- 65+

**4) Ethnicity**

- White
- African American
- Hispanic
- Asian
- Native American
- Other (specify): \_\_\_\_\_

**5) HH Income 2007**

- <10K
- 10K–14.9K
- 15K–24.9K
- 25K–34.9K
- 35K-49.9K
- 50K–75.9K
- 75K+

**Figure 8: Response Rate to Non-Response Interview Questions**

Survey Question	# Responded	% Response
Reason for refusal	1,424	98.27%
Minutes traveled	903	62.32%
Age	1,332	91.93%
Ethnicity	1,366	94.27%
Income	500	34.51%

**CONCLUSION**

Through the work on the onboard surveys in the Dallas/Fort Worth modeling area, a list of suggested improvements for future onboard surveys was developed. The addition of expected quality in the scope of work will help with evaluating the quality of the final database. Onboard survey projects should try to refine the survey instrument and test innovative ways of gathering the information to improve the response rate and the quality of the survey. Using tools such as boarding and alighting studies and non-response bias surveys, surveyors can better perform the expansion and understand the major biases of the onboard survey. NCTCOG modeling staff is of the opinion that creation of robust analytical models directly depends on the quality of the collected data and the understanding of its inherent limitations.