NCTCOG

2007 DART Onboard Survey

Database Cleaning Project

Arash Mirzaei, Kathy Yu, Hua Yang
May 2009

2007 DART Onboard Survey

Database Cleaning Project

Table of Contents

Reviewing the Survey data

In the first part of the analysis, the travel model development team reviewed the 2007 DART Onboard Survey Report, the data field definitions and actual data in the database. With these three items, the team looked to get a better understanding of how the data was organized, verify the results in the reports, and understand trends in the data.

Reviewing Data Definitions

Initially, the data field definitions were reviewed to understand how the data was organized, and if the information needed was available. In some cases, more enumerations existed in the database than in the definitions, so revisions were needed. For example, the field of OPURP actually could have values from 1-11, but only values 1-8 were listed in the original definition. In addition, the use of certain flag fields such as RAIL_QC and FLAG were not described anywhere, so additional questions were posted to NuStats to get the latest data definitions.

To understand the determination of the Board and Alight points for the surveyed route, requests were made to NuStats and the GeoStats imputation process and database fields were provided.

Comparing Survey to Database

After understanding the data definitions, the first step of using the database was making sure the values in the report could be reproduced from the database. In order to do this, queries were executed to compare the total number of records, the ridership by day of week, and the ridership for different numbers of transfers. After these initial tests, more in-depth testing could begin.

Understanding User Error Confusion

After reviewing the survey, it was noticed that there were some questions in the survey. To determine if people were consistent in answering questions on the survey, the survey was first reviewed to find redundant questions. Then, the responses to those questions were reviewed to determine if the answers were consistent.

Pinpointing redundant questions

After reviewing the survey, the questions that were

- Question 5 (total buses) and Question 10 (bus sequence)
- Question 4 (transfer from) and Question 10 (bus sequence)
- Question 6 (transfer to) and Question 10 (bus sequence)
- Question 9a/b (first and last rail line/rail station) and Question 10 (bus sequence)

- Question 9a and 9b has a confusing wording, since it asks people to list the first rail station and last rail station as opposed to the boarding and alighting rail station on a single leg of a trip.

Perform checks on inconsistent answers

After reviewing the redundant questions a list of checks was developed to test if inconsistencies exist in the database.

1. Surveyed route should be in sequence in Question 10.
2. Test reasonableness of mode of access in Question 3 and the boarding location.
3. Test reasonableness of walk/wheelchair mode of access in Question 3.
4. If Question 4 is "I did not transfer," the surveyed route should be the first in Question 10 bus sequence.
5. If Question 4 is bus or rail," the first route/rail in sequence in Question 10 should not be the surveyed route, but the surveyed route should be in the sequence.
6. The total bus/lines in Question 5 should match the number of buses and lines in the sequence in Question 10.
7. Check correspondence between transfer route/lines and bus sequence
    - The Transfer From route/line in Question 4 should be listed in the bus sequence in Question 10.
    - The Transfer To route/line in Question 6 should be listed in the bus sequence in Question 10.
8. Test reasonableness of mode of access in Question 8 and the disembarking location.
9. Test reasonableness of walk/wheelchair mode of egress in Question 8.
10. If rail station(s) and associated line(s) are specified in Question 9a/b, they should be in the sequence in Question 10.
11. Check correspondence between rail stations and lines
    - If the first rail station and line is specified in Question 9a, check that the specified station is on the specified line.
    - If the last rail station and line is specified in Question 9b, check that the specified station is on the specified line.
    - If the first or last rail is "Did not/will not use rail on this one way trip," but the corresponding station is specified.
    - If the first or last rail station is undefined.
12. If the mode of access in Question 3 or mode of egress in Question 7 is drive alone, the number of household vehicles in Question 18 should be at least 1.
13. If the total number of bus/lines in Question 5 is 1, then the transfer from in Question 4 should be "I did not transfer."
14. If the total number of bus/lines in Question 5 is 1, then the transfer to Question 6 should be "I did not transfer."

15. If the total number of bus/lines in Question 5 is greater than 1, then the answer to transfer from in Question 4 or the answer to the transfer to Question 6 should not be “I did not transfer.”

Q Check	ID	Wording	Fields	Method	# Records	% Records	Notes
Q3	2	"Test reasonableness of Q3 and Boarding Location"		1. Use distance formula on p. 27. 2. Create histogram of distance for each GETTO			
Q3	3	"Test reasonableness of Walk and Wheelchair in Q3"		1. Use distance formula on p. 27. 2. Create histogram of distance for each GETTO			For Getto = 1, Compared Oxy-bxy vs. blocks and bxy seemed best.
Q3/Q18	12A	"If Q3 is DA, Q18 should be >= 1."	GETTO, HHVEH	(GETTO=6) and HHVEH<>0	4	0.56%	GETTO=6 has 709 records; 4: HHVEH = 0 33/709 = 4.65% = Specified Don't know
Q4/Q10	4	"If Q4 is 'I did not', then Q10 – 1st in sequence should be route.	TRANSFER, ROUTE_MOD, BUS1	TRANSFER=1 and ROUTE_MOD<> BUS1	529	11.99%	TRANSFER =1: 4412
Q4/Q10	5	"If Q4 is 'Bus' or 'Rail', then Q10 – 1st in sequence should not be route, but route should be in the sequence."	TRANSFER, ROUTE_MOD, BUS1234	TRANSFER<>1 and (ROUTE=BUS1 or ROUTE<>BUS234)	420	12.35%	TRANSFER<>1:3401 records + ROUTE=BUS1: ROUTE<>BUS1 but = BUS2or3or4 has 2753 records

Q Check	ID	Wording	Fields	Method	# Records	% Records	Notes
Q4/Q10	7A	"Answer to Q4 should be in sequence in Q10"	TRANSFER, ROUTE1, FINAL, BUS1234	e.g. 1: TRANSFER=2 , ROUTE1<>NULL and ROUTE1= BUS1or2or3 e.g. 2:TRANSFER=3 and BUS1or2or3 = RED LINE e.g. 3: TRANSFER=4 and BUS1or2or3 =BLUE LINE	239	7.03%	TRANSFER=2 and ROUTE1<>NULL has 2097 records; 61 do not have route in bus123 TRANSFER=3 has 711 records; 102 do not have red line in bus1,2,3 TRANSFER=4 has 496 records, 66 do not have blue line in BUS1or2or3 TRANSFER=5 has 97 records; 10 records do not have BUS1or2or3=TRE
Q4/Q5	13	"If Q5 = 1, Q4 = 1."	TOTAL, TRANSFER	TOTAL = 1 AND TRANSFER <> 1	101	4.62%	TOTAL = 1: 2186
Q4/Q6	**17**	**Does Q4 = Q6**	**TRANSFER, FINAL ROUTE1, ROUTE2**	**E.g. 1: TRANSFER = FINAL AND TRANSFER > 2** **E.g. 2: TRANSFER = 2 AND FINAL=2 and ROUTE1=ROUTE2**	**388**	**6.79%**	**For Rail Lines, 1304; matching = 236** **For Rail Lines, 4412; matching = 152**
Q5/Q10	6	"Total in Q5 vs. Sequence in Q10"	TOTAL, BUS1234	e.g.1: TOTAL= 1 and (BUS1 null or BUS2/BUS3/BUS4 not null) e.g. 2: TOTAL= 2 and (BUS1/BUS2 null or BUS3/BUS4 not null)	4	0.05%	TOTAL=1 has 2186 records; 1 has more than BUS1 and BUS1=ROUTE TOTAL=2 has 3120 records; 1 failed TOTAL=3 has 1790 records; 1 failed TOTAL=4 has 717 records; 1 failed
Q5/Q4/Q6	15A	"If Q5 > 1, then Q4 > 1 or Q6 > 1."	TOTAL, TRANSFER, FINAL	TOTAL <> 1 AND TRANSFER = 1 FINAL = 1	453	8.05%	TOTAL <> 1: 5627

Q Check	ID	Wording	Fields	Method	# Records	% Records	Notes
Q6/Q5	14	"If Q5 = 1, Q6 = 1."	TOTAL, FINAL	TOTAL = 1 AND FINAL <> 1	79	3.61%	TOTAL = 1: 2186; 79 failed
Q6/Q10	7B	"Answer to Q6 should be in sequence in Q10"	TRANSFER, ROUTE2 FINAL, BUS1234	FINAL=2, ROUTE2<>NULL and ROUTE2=BUS2or3or4 FINAL =3 and BUS2or3or4 = RED LINE FINAL=4 and BUS2or3or4=BLUE LINE FINAL=5 and BUS2or3or4=TRE	620	18.78%	FINAL=2 and ROUTE2<>NULL has 2122 records; 333 failed FINAL=3 has 653 records; 146 did not have RED Line in bus234 FINAL=4 has 419 records; 114 did not have BLUE line in bus234 FINAL=5 has 108 records; 27 did not have TRE in bus234
Q7/Q18	12	"If Q7 is DA, Q18 should be >= 1."	GETFROM, HHVEH	(GETFROM =6) and HHVEH<>0	6	1.86%	GETFROM =6 has 323 records; 6: HHVEH= 0 8/323 = 2.48% = Specified Don't know
Q8	8	"Test reasonableness of Q8 and Disembarking Location"		1. Use distance formula on p. 27.2. Create histogram of distance for each GETFROM			
Q8	9	"Test reasonableness of Walk and Wheelchair in Q8"		1. Use distance formula on p. 27. 2. Create histogram of distance for each GETFROM			In Distribution Subfolder

Q Check	ID	Wording	Fields	Method	# Records	% Records	Notes
Q9/Q10	10A	"If Q9a has answer, check the rail line in sequence in Q10."	FIRST, RAIL1, BUS1234	FIRST=1 and BUS1or2or3or4=RED LINE	576	7.80%	FIRST=1 has 1889 records, 1712 have Red Line in seq FIRST=2 has 1135 records; 1001 have BLUE Line in seq FIRST=3 has 380 records; 375 have TRE in seq FIRST=4 has 3982 records; 3722 BUS1234 don't have rail line
Q9/Q10	10B	"If Q9b has answer, check the rail line in sequence in Q10."	LAST, RAIL2, BUS1234	e.g. 1: LAST=1 and BUS1or2or3or4=RED LINE e.g. 2: LAST=2 and BUS1or2or3or4=BLUE LINE e.g. 3: LAST=3 and BUS1or2or3or4=TRE e.g. 4: LAST=4 and BUS1234 don't have rail line	286	4.10%	LAST=1 has 1666 records; 1655 have bus1234 = red LAST=2 has 945 records; 933 have bus1234 = blue LAST=3 has 383 records; 380 have bus1234 = tre FIRST=4 has 3982 records; 3722 have bus1234 not rail
Q9	11A	"If Q9a has answers, check station with line."		1. Develop table rail_station of stations/lines 2. e.g. 1: first=1 And rail1_avst=rail_station.station_name And rail_line='RED LINE';	30	0.88%	37/1889 - RED 42/1135 - BLUE 16/380 - TRE 28+25+12 = 65 = 1.95% = No Station
Q9	11B	"If Q9b has answers, check station with line."		1. Develop table rail_station of stations/lines 2. e.g. 1: last=1 And rail2_avst=rail_station.station_name And rail_line='RED LINE';	8	0.27%	45/1666 - RED 21/945 - BLUE 19/383 - TRE 41+21+15 =77=2.57% = No Station

Q Check	ID	Wording	Fields	Method	# Records	% Records	Notes
Q9	11 C	"If Q9a/b has 4 (no rail) how many have station."	FIRST, RAIL1, LAST, RAIL2	(first=4 and rail1 is not null) OR (last=4 and rail2 is not null)	0	0.00%	0/3982 - rail1, rail2
Q9	11 D	"If Q9a/b has 98."	FIRST, LAST,	(first=98) OR (last=98)	898	11.49%	Respondent Error, Q9a/Q9b/Q10 answers do not match.
Q9	16A	Q9a vs. Rail1_XY		e.g. 1: 1st Rail is red and rail1_avst <> bus_on_mod	991	77.18%	1st Rail = Red: 631; 452 don't match 1st Rail = Blue: 317; 253 don't match 1st Rail = Tre: 336; 286 don't match
Q9	16B	Q9b vs. Rail2_XY		e.g. 1: 1st Rail is red and rail2_avst <> bus_on_mod	937	61.04%	Last Rail = Red: 754; 432 don't match Last Rail = Blue: 422; 254 don't match Last Rail = TRE 359; 251 don't match
Q10	1	"Route should be in Sequence in Q10"	ROUTE_ MOD, BUS1234	ROUTE_MOD is not in BUS1234	1	0.01%	All Records: 7813

Check Reasonableness of Mode of Access and Mode of Egress based on distance

To perform the checks testing the reasonableness of access/egress distances, create histograms of the blocks and miles by various modes of accesses so that the trends, maximum and minimum could be reviewed. 88% of the surveys with walk mode of access had a walk of 3 blocks or less, as shown in Exhibit 1. 75% of the surveys with a walk mode of access had a walk distance less than 3 miles as shown in Exhibit 2, but there were also some large values recorded.

Exhibit 1: Walk Mode of Access Distribution by Block

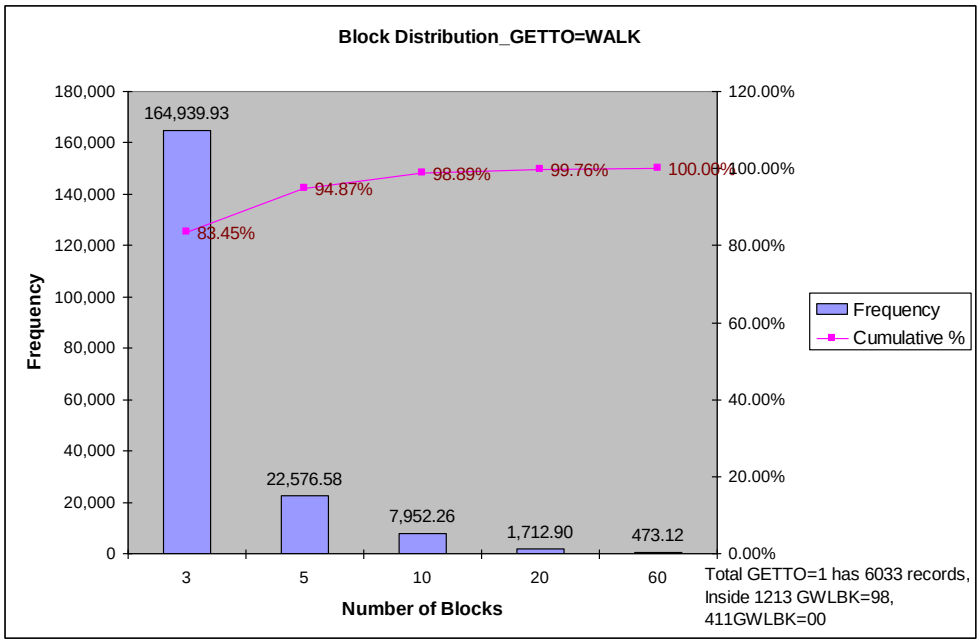
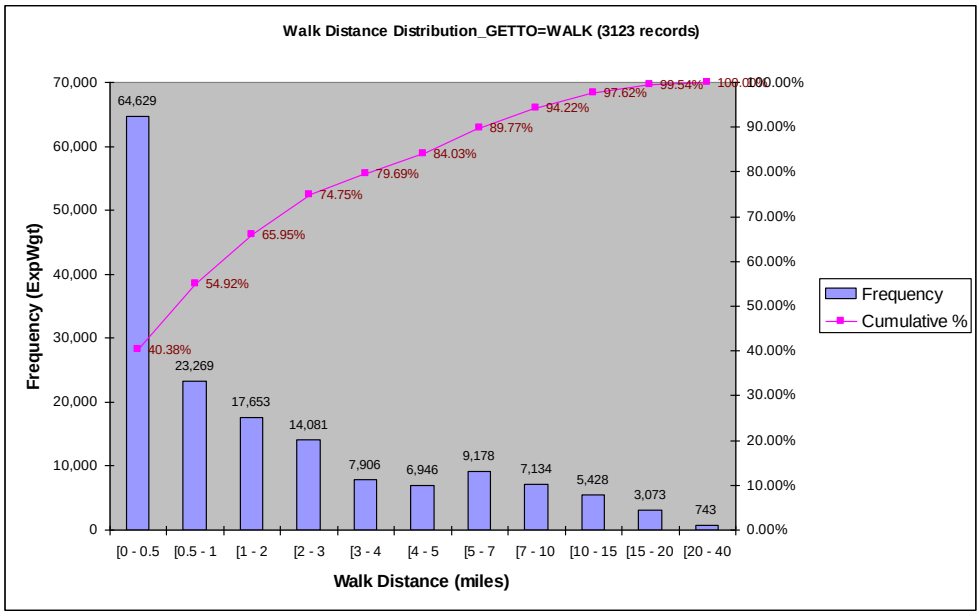


Exhibit 2: Walk Mode of Access Distribution by Miles



The difference between these numbers could be caused by the confusion over the definition of a block, since blocks in different cities or area types could be interpreted differently.

Noticing Trends in Sequence Errors

Some of the trends noticing in sequence errors:

- People described a round-trip and not a one-way trip, so a route/rail was repeated in their sequence.
- People described all possible routes they could take for their trip and not the ones they are specifically taking on this trip.
- People describe the reverse trip than what they are taking.
- People put down origin and destination for roundtrip, but described path for one –way trip. Or vice-versa.
- People are not sure what to include in From/To Modes versus sequence.
- People reversed modes.

Since there were many mistakes that were possible, it was hard to develop one program that could correct all errors that existed. As a result, it was decided that surveys needed to be specifically reviewed to resolve the inconsistencies and determine the intended route sequence.

Database Cleaning – Surveys with Conflicting Answers

It was determined that questions 4, 5, 6, 9a, 9b, and 10 from the survey were repetitive questions and the responses were not always consistent. Using these questions, 2,593 weekday records of the 6,447 were flagged as not having consistent results and requiring a personal review. Surveys were flagged because of the following reasons:

- The route sequence from 4-surveyed route- 6 did not match the first three routes in question 10.
- 9a or 9b mentioned rail, but the sequence did not list it.
- If sequence of routes in question 10 listed rail, but 9a or 9b were not filled in.

Preparations for the Review

Once these surveys were identified as inconsistent, a set of twenty surveys were randomly selected as a test group. A team of three Travel Model Development team members used copies of the actual surveys, the TransCAD roadway and transit networks, DART and Google Map websites, and a DART system map to try to determine the correct sequence for the survey. In doing so, they hoped to determine how long the process would take, understand the difficulties in the process, develop notation for correcting the survey paper copies, determine what tools reviewers would need to correct the survey, and develop instructions for the process, and notation that would be used for correcting the survey.

From this, it was determined that having a map with all of the main survey information on it would allow for faster processing. So, a GISDK macro was written and run to create maps for each survey. Each TransCAD map highlighted the origin, the destination, all rail lines, the survey route, and any route listed by the respondent in questions 4, 6, and 10 for a given survey. Once the instructions and maps were established, an instructional document was created; this document is shown in Appendix A.

To keep the surveys in manageable sets, the surveys were divided up into 25 groups of 90-110 surveys. The surveys were kept in numeric order, so they were easier to keep track of. Also, the surveys were originally conducted in numeric order, so anywhere from two to forty surveys in numeric order might correspond to the same route. The team felt that keeping surveys with the same surveyed route together would help improve efficiency since they would be more comfortable with the path of the route and common transfer routes. Each survey was printed out and placed in their corresponding sets. In addition, the user was provided with a list of the surveyed route for each survey they were given.

Initial Review

A group of ten reviewers were gathered in addition to the Travel Model Development staff. Ten transportation planners were recruited to help with the processing of correcting the surveys. The reviewers attended an orientation session where the instructions included in Appendix A, and the Travel Model Development staff also checked on the progress, and fielding general and specific questions on the process.

Included in these instructions were three notes that could be made on surveys. They were the following:

- If any survey was too difficult to determine a route, then they should be marked with a U and returned to the Travel Model Development staff.
- If any route was listed in the survey but not provided in the survey's TransCAD map, they should be labeled with a NF for Not Found.
- If the reviewer was not confident in their final analysis of a sequence, they should either label the survey with a U or include in the U/Undetermined pile of surveys.
- Surveyors were allowed to use more than 4 lines in their sequence.

After each reviewer completed their first set, randomly chosen surveys from their sets were selected to confirm that the clean-up was reasonable. If there were any problems, the surveys were reviewed again by the Travel Model Development staff and the process was discussed again with the reviewer.

Review of Surveys with Not Found Routes

New maps were created to include routes from a DART route shapefile, a 2007 DART route layer and a 2007 COG layer, so that all coded DART routes were made available to the reviewers. The reviewers were instructed that if a route was not found within these new maps, then the online pamphlets provided by DART or DART route system maps could be used along with the TransCAD survey maps to determine the location of the routes. In some cases, reviewers found a route with transposed numbers fit in a reasonable sequence between the origin and destination, so routes could be found in this manner.

Once the route was located, the reviewer tried to determine the logical sequence of routes from origin to destination. If a survey could not be determined, the survey was labeled as undetermined and kept with the other surveys classified as undetermined

Review of Undetermined Surveys

After the first pass of all surveys was complete, the surveys classified as undetermined needed to be reviewed again.

To better understand how to streamline this process, the Travel Model Development Team tried to look at undetermined surveys and see if a survey could be successfully found. Through this process, the keys for a successful route of a correct origin, correct destination, and the surveyed route were identified. From this, a set of instructions were developed and is shown in Appendix B.

The instructions noted that the reviewer should first look at the map using the assumption that the geo-coded origin, geo-coded destination, and surveyed route are correct. If no survey could be found from this, the reviewer should consider first geo-coding the origin again and then geo-coding the destination again.

Finally, if a route sequence could still not be determined, then the reviewer should determine if a reasonable sequence could be found without including the surveyed route. If a surveyed route was replaced, then the closest route in sequence to the surveyed route should be chosen as the "new" surveyed route.

If no sequence could be found, then the surveyed was to be ignored and the possible reasons for this are the following:

- No Routes between OD/OD Too Close – There is no route specified on the survey which is between the specified origin and destination.
- O Too Far - Origin is over 2-2.5 miles and the mode of access is walk/wheelchair.
- D Too Far Destination is over 2-2.5 miles and the mode of egress is walk/wheelchair.

The five reviewers and Travel Model Development team involved in this process were taught in a new orientation session where they were given the instructions provided in Appendix B and access to the new maps. When distributing the surveys for this second review, the Travel Model Development team tried to distribute the surveys such that a different reviewer would review the survey than had performed the initial review.

After the reviewers completed the surveys, the Travel Model Development team looked over all surveys which were labeled ignore, and all surveys whose survey route was dropped. They also randomly checked the surveys whose origin and destination had changed.

Review of Missing Surveys

If surveys were missing from the survey DVD, then the information provided from the database was used to determine the sequence and handle any undetermined sequence questions.

Database Entry

The sequences were entered from the sequence notes on the surveys into the spreadsheets. Then, the sequences were loaded from spreadsheets into new fields in the 2007 survey database. During this process, there were some errors encountered loading the database which required updating the sequences in the spreadsheet, or re-reviewing some surveys.

The new fields also include SEQ_REVIEW, SEQ_FLAG, and SEQ_COMMENTS. SEQ_REVIEW had a value of 1 if the survey had been reviewed at all during the database cleaning and 0 otherwise. SEQ_FLAG had a value of 1 if the survey should be ignored; otherwise it was given a value of 0. SEQ_COMMENTS is a field for all surveys with a SEQ_FLAG to include comments on why they should be ignored/left out of the clean database.

After everything was entered, queries were used to test the following:

- Each sequence included the newly defined survey route.
- The total number of routes/lines used was calculated from the new sequence.

Survey Review Summary

2,593 total surveys reviewed.

Database Cleaning – Surveys flagged by Stops Program

Determining Stop Program Logic

A program was developed which found the board and alight points for each route in the sequence, and recorded them into a table. The process began by getting a list of the board and alight points for each route from a transit layer. If the route was a rail line, the list was obtained from a specialized rail file which listed all the rail stations in the system. If the route was a bus line, the route was retrieved from the DART layer; if not found, the route was retrieved from the COG transit layer. From each route or line, the list of stops was obtained.

Using the route stops, the following logic was used to get the actual stops used.
Finding the Origin Board stop:

- If the mode of access is not drive alone or carpool, then find the closest stop to the origin on the first route in the sequence.
- If the mode of access is drive alone or carpool, then first find the closest stop to the origin on the first route in the sequence which is a park and ride. If no park and ride can be found, find the closest stop.

Finding Board and Alight stops on Route/Route transfers:

- Find the alight stop on the one line in sequence and board stop on the next line which minimizes the distance between the stops and minimizes the distance from the board stop of the previous route. In this process, do not allow the alight stop and board stop for the same line to be the same.
- An alternate way to do this which should be considered is find the alight stop on the first line and board stop on the next line which minimizes the distance between the stops and minimizes the sum of the distance from the origin and the distance to the destination. There are cases which get resolved by the current methodology and others which get resolved by the alternate method; it has not been determined which resolved more cases.

Finding the Destination Alight stop:

- If the mode of egress is not drive alone or carpool, then find the closest stop to the destination on the first route in the sequence.
- If the mode of egress is drive alone or carpool, then first find the closest stop to the destination on the last route in the sequence which is a park and ride. If no park and ride can be found, find the closest stop.

Testing Validity of Stops

The stops program was modified to print out warnings/errors in the program. The following items were flagged which caused a survey to be reviewed. This could be triggered for an already reviewed survey, or a survey originally considered logical.

The warnings and errors possible are the following:

- The Destination or Origin was not geo-coded.
- The distance from the origin to the first line in the sequence was greater than the distance from the origin to the last line in the sequence.
- The distance from the destination to the last line in the sequence was greater than the distance from the destination to the first line in the sequence.
- A route in the sequence could not be found.
- Board/Alight point for the same route was the same.

Review of Surveys without a geo-coded Origin or Destination

From the set of surveys which was not originally cleaned, it was found that there were surveys which did not have a geo-coded origin. These surveys were treated as undetermined surveys and reviewed.

There were no surveys without a geo-coded Origin.

Review More Surveys based on Warnings/Errors

Because of the stops program, surveys were flagged because the sequence might be reversed or might not be logical based on stops. Whether they were under they were reviewed originally for their sequence or not, each survey was reviewed again. The same process that was initially used was followed. First, the sequence was reviewed assuming the origin, destination, and surveyed route were correct. If no valid sequence could be found the undetermined procedure was used.

The longitude and latitude of any new origin and destination was based on coordinates found through using Google Maps.

Survey Review Summary

3,168 total surveys reviewed. 164 surveys ignored.

Assigning Weights

Check Mode of Access/Egress Other

If the Mode of Access/Egress is other, it cannot be easily identified as Transit Walk or Transit Drive in the assignment process. To make sure the information in these records was kept, the text include with the Other mode of access and

egress was checked to see if they belong in the existing categories. In cases, where the user specified existing modes in the other field, the mode of access was corrected.

In other cases, the user listed something in the other field such as a bus route or line as a mode, or something which could not be interpret as another mode of access. In order for these surveys to be used, the proportion of each mode of access for weekday surveys was determined. Then, the number of surveys required to maintain those proportions from the surveys with "Other" mode of access was calculated. Finally, using randomly selection, "Other" mode of access surveys were selected and assigned to the mode of access to maintain the calculated distribution. Since it is believe that it is not logical to have a drive alone mode of access and mode of egress, if during the random assignment, drive alone was specified as a mode of access and a mode of egress, then the mode of access was reassigned for that survey.

Similarly, the surveys which had other as the mode of egress were corrected where possible. Then, all remaining surveys with an "Other" mode of egress were randomly assigned a specific mode of egress proportionally to match the weekday mode of egress distribution.

In the clean database, there still exists one survey with a drive alone mode of access and a drive alone mode of egress; this was specified by the user and not through the random weighting process.

Data re-expanded

Expand the data using the original data for the expansion factors with the following notes:

- If a surveyed route was changed, remove the particular "completed review" from the previous surveyed route, and set the survey's Response Factor to 1.

Process

1. Create new fields in the busstopfile
2. Use the data in cog_final_dataset to see which survey routes have changed. (Did this also include which surveys have been ignored?). Use this to update the completes for each boarding point of each trip.
3. Recalculate the completes, bus response factor, and completes after applying bus response factor.
4. Calculate the total number of adult_boardings for each trip and the total number of completes for each trip, and use this to determine the trip level weight.
5. Calculate the response factor as trip weight * bus stop response factor.

6. Total Trips and Sample trips for each RTDD remain the same. Totals Trips divided by sample trips defines the Vehicle Factor.
7. The Boarding factor is calculated by multiplying the Response factor by the vehicle Factor. Although the Linked Trip factor is used in the 2007 Dart Onboard Survey reports, COG believes this should not be used in the expansion weighting.
8. Use the boarding factors for route and day of week to weight the ridership.
9. Calculate the Expansion Factor as Population Average Daily Ridership / Ridership Weighted by Boarding Factors.
10. The final expansion weight is calculated as the product of the Response Factor and the Expansion factor.

Production/Attraction Matrix

The production/attraction matrix will need to be developed from the DART Onboard Survey database. In order to do this, the trip purpose needs to be assigned to each record based on whether the origin or destination is home and/or work. Based on the trip purpose and the mode of access/egress, the origin and destination ends must be classified as the production and attraction end of each trip. In the case of a Non-Home based trip (NHB) with a drive alone, carpool, or pick up/drop off mode of access or egress, the production and attraction can also be determined by labeling the end with the drive mode as the production. The SQL associated with this process is provided in Exhibit 3.

Exhibit 3: Steps to Create Production Attraction Matrix

Index	Step	Logic	SQL
1	Create Trip_Purpose		
1A	Set Trip_Purpose = HBW	orig is work and dest is home or orig is home and dest is work	UPDATE cog_final_dataset SET trip_purpose = "HBW" WHERE (opurp=1 And dpurp=4) Or (opurp=4 And dpurp=1);
1B	Set Trip_Purpose = HNW	orig is not work and dest is home or orig is home and dest is not work	UPDATE cog_final_dataset SET trip_purpose = "HNW" WHERE (opurp=4 And dpurp<>1) Or (opurp<>1 And dpurp=4);
1C	Set Trip_Purpose = NHB	orig is not home and dest is not home	UPDATE cog_final_dataset SET trip_purpose = "NHB" WHERE opurp<>4 And dpurp<>4;
2	Create Production/Attraction		
2A	Set P=Orig, A=Dest	orig is home or trip purpose is NHB	UPDATE cog_final_dataset SET production = 'ORIGIN', attraction = 'DESTINATION', p_mode = new_getto, a_mode = new_getfro, p_xcoord = new_ox, p_ycoord = new_oy, p_tsz = new_o_tsz, a_xcoord = new_dx, a_ycoord = new_dy, a_tsz = new_d_tsz WHERE new_opurp=4 Or trip_purpose='NHB';
2B	Set P=Dest, A=Orig	dest is home	UPDATE cog_final_dataset SET production = 'DESTINATION', attraction = 'ORIGIN', p_mode = new_getfro, a_mode = new_getto, p_xcoord = new_dx, p_ycoord = new_dy, p_tsz = new_d_tsz, a_xcoord = new_ox, a_ycoord = new_oy, a_tsz = new_o_tsz, WHERE new_dpurp=4;
3	Determine P/A Mode		
3A	Initialize PAIGNORED	Initialize PAIGNORED	UPDATE_PAIGNORED update cog_final_dataset set PA_IGNORED = 'YES' where P_MODE = 4 or P_MODE = 97 or P_MODE = 99

Index	Step	Logic	SQL
3B	Ignore Bicycle, Other, DK/RF Modes	ignore mode = bicycle, other, dk/rf	UPDATE_PAIGNORED update cog_final_dataset set PA_IGNORED = 'YES' where P_MODE = 4 or P_MODE = 97 or P_MODE = 99
4	Create PA Matrix		
4A	Create matrix where group by p_tsz and a_tsz and aggregate by 1 or more fields		
4B	Load data into matrix and define all tsz as indices by adding 4874 rows where production is an unused tsz, and another 4874 rows where attraction is an unused tsz.		

Confidence in the Data

After cleaning the database, a subset of the data was reviewed to produce statements of our confidence in the data.

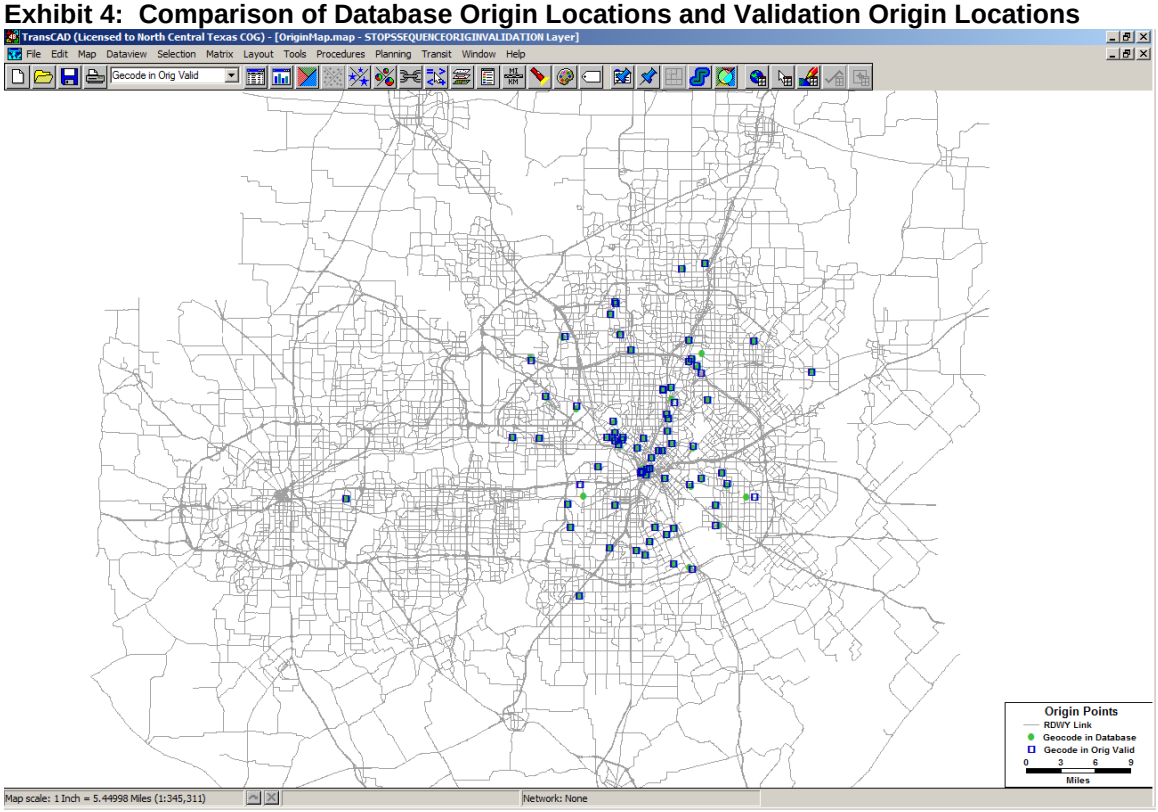
Confidence in Origin and Destination Geo-coding

We are 95% confident that 95% +/- 5% of the origin points are geo-coded within 0.75 miles of the user-specified origin place and address. We are 95% confident that 95% +/- 5% of the destination points are geo-coded within 0.75 miles of the user-specified destination place and address.

These confidence statements were developed after reviewing 75 randomly-picked surveys. The review of each survey consisted of looking at the type of place, place name, address and cross streets listed in the original response to Question 2. All of this information was entered into Google Maps to find the location. The exact location names, address, and/or cross streets and longitude and latitude were recorded into the origin validation spreadsheet. The full spreadsheet was then loaded into TransCAD, and the Locate by Address feature was used to get the Longitude and Latitude for each exact address. If TransCAD could locate the address, the TransCAD longitude and latitude was used as the validation point; if an address was not located in TransCAD, the Google Maps longitude and latitude was used at the validation point. The TSZ of each validation point was also found.

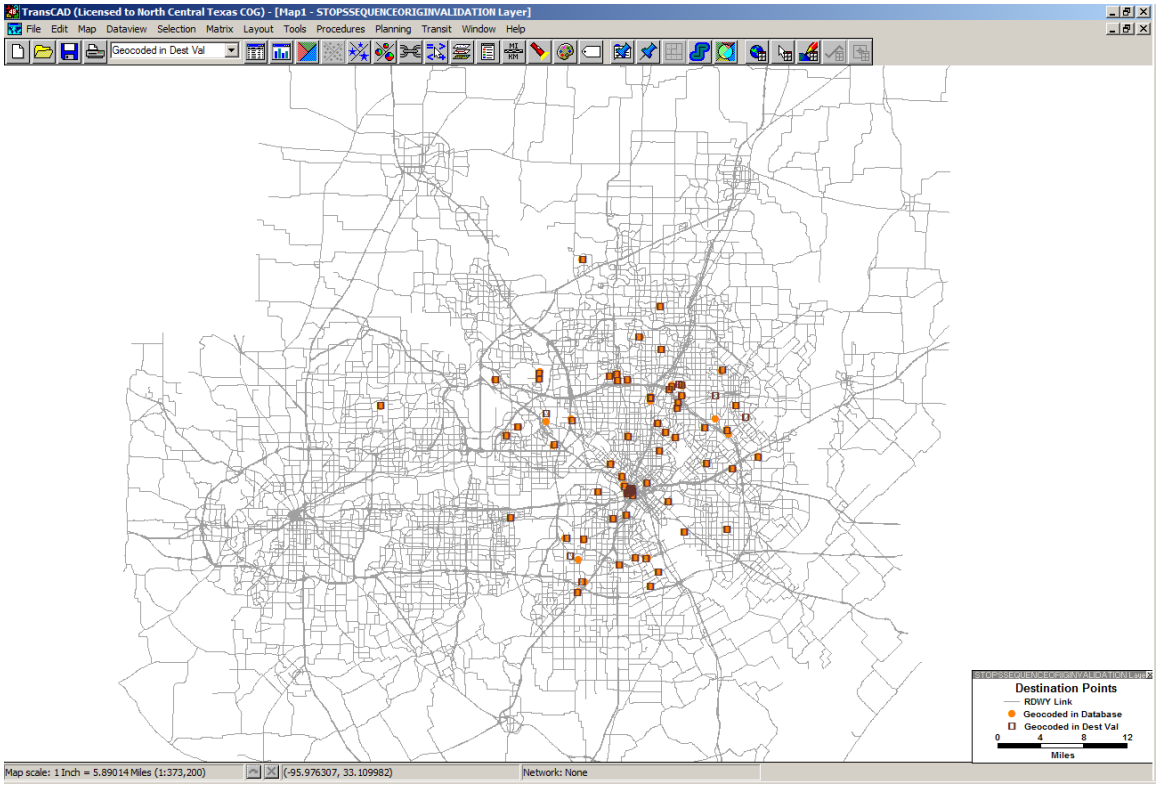
When this was complete, the distance between the database geo-coded origin points and the validation geo-coded origin points was compared. It was calculated that 72 of 75 of the validation geo-coded points were within 0.75 mile of the database geo-coded points. The comparison between the two sets of points is shown in Exhibit 4.

Exhibit 4: Comparison of Database Origin Locations and Validation Origin Locations



This same process was done with the original response to Question 8 and compared to the database geo-coded destination points. In this case, it was calculated that 72 of the 75 validation geo-coded destination points were within 0.75 miles of the database geo-coded destination points. Because of these tests, the origin and destination confidence statements were developed. The comparison between the two sets of points is shown in Exhibit 5.

Exhibit 5: Comparison of Database Destination Locations and Validation Destination Locations



Confidence in Route Sequences

We are 95% confident that 95% of the path sequences are correct. A path sequence is correct when the respondent-identified sequence from its origin to its destination is feasible.

This confidence statement was developed after projecting a 95% confidence and then reviewing 74 randomly-picked surveys. The review of each survey consisted of looking a map showing the origin, destination, each route or line listed in the sequence, and all board and alight points for each route. Each board and alight point was reviewed for reasonableness; reasonableness was recorded as 1 for reasonable and 0 for not reasonable. After these tests, the stops confidence statements were developed.

Confidence in Stops

We are 95% confident that 95% +/- 5% of the geo-coding of boarding and alighting locations of all records are correct. Geo-coding of boarding and alighting locations of a record is considered correct 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.

This confidence statement was developed after projecting a 95% confidence and then reviewing 75 randomly-picked surveys. The review of each survey consisted of looking a map showing the origin, destination, each route or line listed in the sequence, and all board and alight points for each route. Each board and alight point was reviewed for reasonableness; reasonableness was recorded as 1 for reasonable and 0 for not reasonable. After these tests, the stops confidence statements were developed.

Appendix A
DART Onboard Survey Instructions

A DART Onboard Survey was conducted in 2007. The Travel Model Development Group is trying to use the questionnaire to reconstruct the path of each respondent. The survey contains some redundant questions when it asks for the "transfer from" route/line (Q4), "transfer to" route/line (Q6), 1st rail line/station (Q9a), last rail line/station (Q9b), and the full sequence of routes/lines (Q10). When comparing the answers to these questions, 2593 questionnaires were noted to have some consistency problem(s) between the responses to these questions. The purpose of this project is to check which of the paths they supplied in their questionnaires are reasonable, and correct the questionnaire.

Example Correspondence of Question Responses:
For example, the respondent is currently riding on route 582.

- The sequence the user wrote (Q10) is 161, RED LINE, 582.
- The answer to where they transferred from (Q4) should be RED LINE (Light Rail).
- The answer to the total number of routes/lines (Q5) on this trip should be 3.
- The answer to where they transferred to (Q6) should be "I will not transfer."
- Question 9a and 9b should answer "Red Line (light rail)."

Logic: Each questionnaire has one map which it corresponds to. This map will list the origin and destination of the one-way trip surveyed and all routes listed by the passenger in his/her questionnaire. By checking the answers (routes sequence, transfer, etc.) in the questionnaire with the map, you can decide whether the path in the questionnaire can be achieved reasonably.

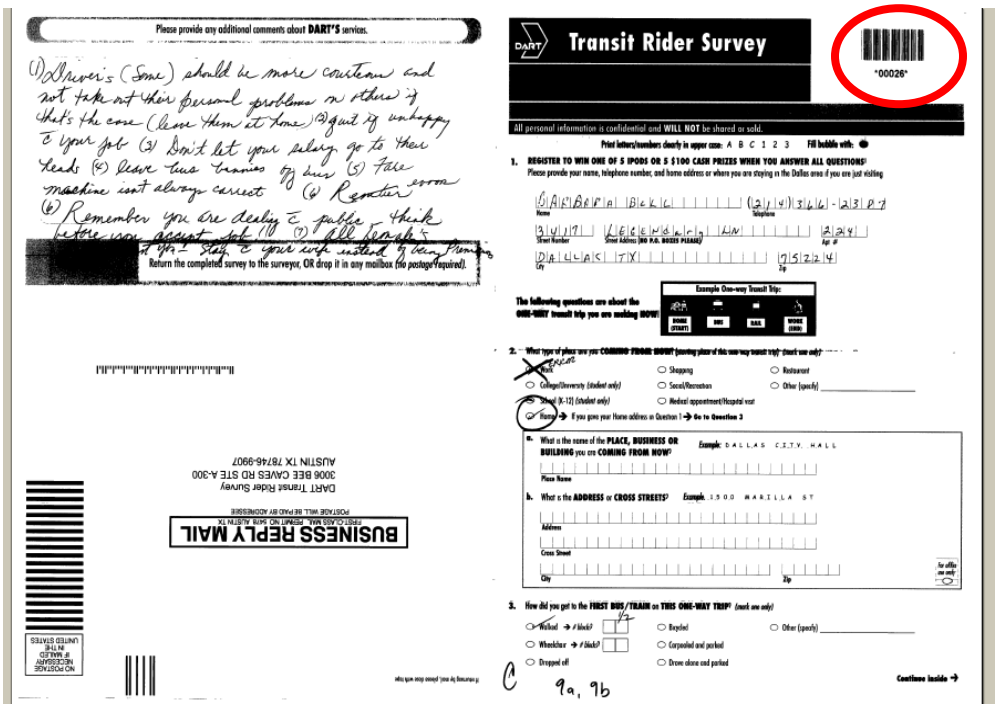
Tools:
You will be provided with

- A blank 2007 DART Onboard Survey for reference.
- A set of approximately 100 surveys. Each survey is 1 double-sided sheet.
- A spreadsheet table with lists the Sample Number, Route, and answers to Q4, Q6, Q9a, Q9b, and Q10. This list contains all surveys in your stack. The fields listed are SAMPN (questionnaire Number); Route is where the survey was done. Fields "total", "transfer_from," "transfer_to," "9a_first_rail," and "9a_last_rail" are correspond to question 5, 4, 6, 9a, and 9b in questionnaire. Fields "bus1/2/3/4" corresponds to question 10.
- A colored marker

Method:

1. Copy I:\Model_Development\External_Projects\DART_OnboardSurvey_2007\GeoCode into your local drive C:\Temp.

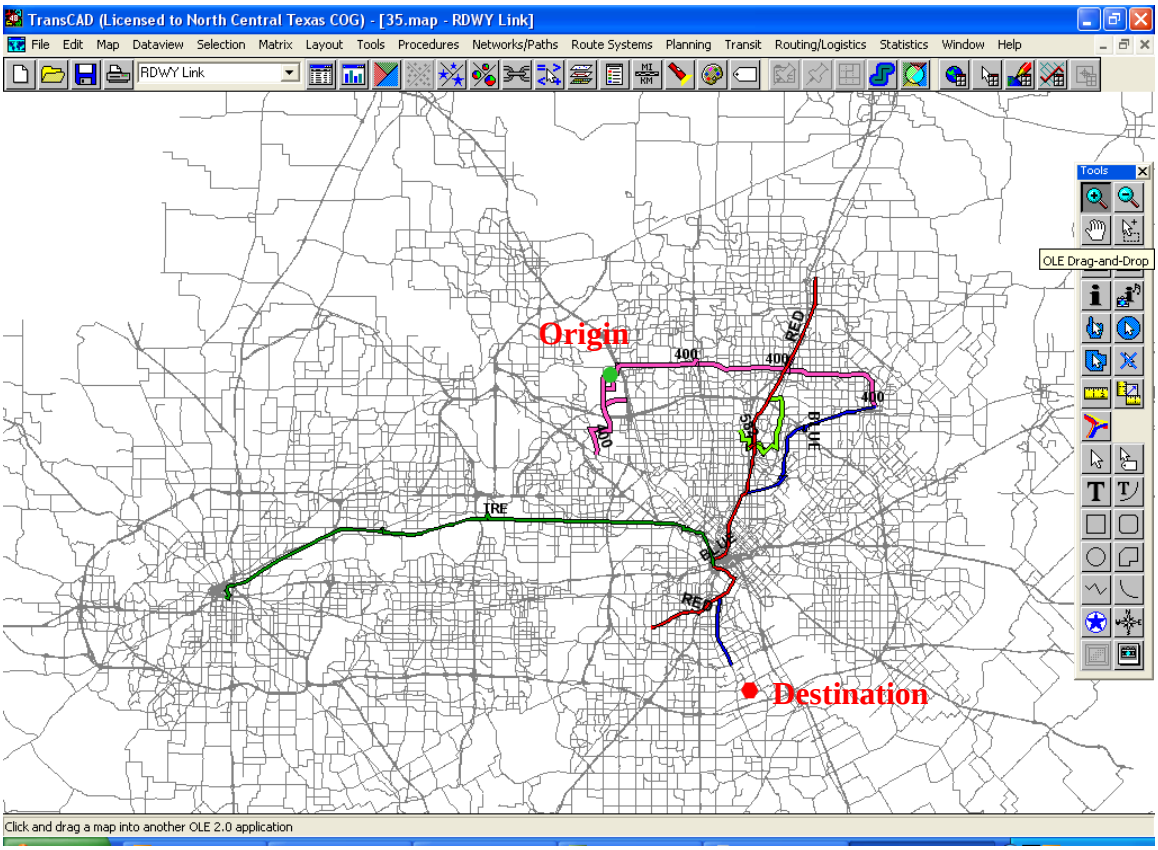
2. Pick up a questionnaire, and locate the questionnaire ID (SAMPN). It is listed in the upper right corner of the front page of the questionnaire (as shown in the image below). In this image, the questionnaire is number 00026.



3. Each questionnaire has a map associated with it. The map name is the same as the questionnaire ID (SAMPN). For example, the map corresponding to the survey above would be called 26.map.

All map files are located in I:\Model_Development\External_Projects\DART_OnboardSurvey_2007\GeoCode\Maps. In the map, the following items are shown

- The origin – shown as a green dot.
- The destination – shown as a red hexagon (looks like a dot)
- All routes mentioned in questionnaire will show up and have bus route/line labels (TRROUTE) on them in the map.
- *ALL RAIL LINES (TRE, RED Line and BLUE Line) WILL ALWAYS SHOW UP* no matter whether they are mentioned in questionnaire or not.



4. Use the table provide to you to find the questionnaire number in the field SAMPN. Find the corresponding **"route" number** for that SAMPN; this is the route on which the respondent was surveyed. Note: Since this is the one route to be verified as on the respondent's path, it must be in the final sequence.

sampn	route	total	transfer_from	transfer_to	9a_first_rail	9a_last_rail	bu
26	582	3	RED LINE		N/A	N/A	161
27	582	4	1	583	RED LINE	RED LINE	1
28	582	3	RED LINE		Error	RED LINE	350
32	582	3	RED LINE		RED LINE	RED LINE	11
33	582	2			RED LINE	RED LINE	RED LIN
34	582	2			RED LINE	RED LINE	RED LIN
35	582	3	RED LINE		RED LINE	Error	400
40	582	3	RED LINE	BLUE LINE	N/A	N/A	582
43	582	3	RED LINE	RED LINE	RED LINE	TRE	TRE

5. You will be doing your corrections on the side of the survey listing questions 4-26. At the top of this side, write the route number with your marker and circle it. At the left up corner of this side, put your initial on there.

6. Review the origin, route, and destination on the map. If the route does not appear on the map, go to step 9. Note: The user may have walked/biked/driven from the origin to the first bus/line, and may have walked/biked/driven from the last bus/line to the destination. Alternately, they could have taken a bus which is not included in this survey or our network (FWTA buses, airport shuttles, company shuttles, DART On Call).

7. Consider the possible sequences that the respondent has provided to see which route sequence makes sense as how to get from the origin to the and destination. Please use the following tools.

    - The route they were surveyed on must be in the sequence.
    - The routes/lines listed in question 10.
    - The route that they transferred from to get to the current route is listed in question 4.
    - The route that they transferred to after the current is listed in question 6.
    - The rail they said they used would be listed in Question 9a and 9b.
    - Note: It is not uncommon for people to have reversed the route or provided a round-trip sequence instead of a one-way sequence. In the end, only a one-way trip must be described.
    - Question 2 and 8 give the type of Origin and Destination (Home-Work, etc.). Question 3 and 7 give the mode of access for Origin and Destination (Walk, Drive alone, etc). This information can help you to figure out if they have put down the reverse path.

NOTE: There are some routes that cannot be displayed on the map, because they are not in our TransCAD Network, not in the DART System, or are non-specific. For the purposes of this study, please flag these surveys and skip to step 9 without determining the sequence.

386	**FTW BUS (all buses listed as FTW)**
631	**LOCAL**
866	**TROLLEY**
DART ON CALL	**TI SHUTTLE**
DFW BUS	**TRANSIT BUS**

8. If you are able to determine a path the origin to the destination, then you must then correct and check off the answers to the questions.

    - Question 10 - Write the correct **sequence** in question 10 in the 4 rectangles provided. Circle the route/line in that sequence which corresponds to the route that the user was surveyed on (determined in step 4).

- Question 4 – Correct, then check off the **"transfer from"** route line. This is the route/line that the respondent transferred from to get to the current "route."
- Question 5 – Correct, then check off the **number** of bus/lines in the sequence.
- Question 6 – Correct, then check off the **"transfer to"** route/line. This is the route/line that the respondent transferred to after getting of the current "route."
- Question 9a – Correct, then check off the line listed as the **FIRST** rail line. If the rail line did not correspond, place an x over the rail station listed. Otherwise, leave do nothing to the rail station.
- Question 9b – Correct, then check off the line listed as the **LAST** rail line. If the rail line did not correspond, place an x over the rail station listed. Otherwise, leave do nothing to the rail station.

9. If you are unable to determine a sequence, you must flag a survey with one of the following marks in the upper right hand corner.
    - U – UNDETERMINED – You were unable to determine the route sequence.
    - NF – Not Found – You were unable to find one of the routes listed in the sequence.

10. Repeat steps 2-9 for the next questionnaire.

11. Return the set of questionnaires to Kathy Yu and Hua Yang in piles:

    - Completed Surveys
    - Undetermined/Not Found Surveys

12. If your time permits, you will be provided a new set of surveys to be reviewed.

Appendix B

DART Survey – Processing Undetermined Routes

Preparation:

1. Copy the folder I:\Model_Development\External_Projects\DART_OnboardSurvey_2007\GeoCode\DartSurveyReview to c:\Temp.
2. Locate maps in the following folder: I:\Model_Development\External_Projects\DART_OnboardSurvey_2007\GeoCode\MapsDART
3. To find the route associated with a survey, use the following document: I:\Model_Development\External_Projects\DART_OnboardSurvey_2007\GeoCode\ Survey_Route.xls

For each survey, do the following:

1. Retrieve surveyed route from Survey_Route.xls
2. Try to determine the sequence between the origin and destination using the route.
    a. If possible, correct sequence and include your initials in upper left corner. You are done with this survey.
    b. If not possible, continue to step 3.
3. Check if origin is correct on map with response in Q2, and no other point is valid.
    a. If it needs to be corrected, write correct address in Q2 and write "**Fix O**" in upper right corner. Continue to Step 4.
    b. If O does not exist and cannot be found, write **"Ignore – O Invalid"** on upper right hand corner, and include your initials in upper left corner. You are done with this survey.
    c. If O is correct, put a check mark next to Q2.
4. Check if destination is correct on map with response in Q8, and no other point is valid.
    a. If it needs to be corrected, write correct address in Q8, and write "**Fix D**" in upper right corner. Continue to step 5.
    b. If D does not exist and cannot be found, write **"Ignore – D Invalid"** on upper left hand corner, and include your initials in upper left corner. You are done with this survey.
    c. If D is correct, put a check mark next to Q8.
5. See if route can be incorporated into a sequence between Origin and Destination for Q10, and consider mode of access responses (Q3/Q7) in reaching the Origin and Destination points.
    a. If a sequence can be created with the route,
        i. Enter the sequence in Q10,
        ii. Correct answers in questions 4, 5, 6, and 9
        iii. Cross off the markings on the upper right corner.
        iv. Include your initials in the upper left corner.

b. If a sequence can be created only without the surveyed route,
    i. Enter the sequence in Q10,
    ii. Select the closest route to be the new “surveyed route” and write its name with a circle in the center-top of the survey form.
    iii. Correct answers in questions 4, 5, 6, and 9
    iv. Cross off the U markings on the upper right corner.
    v. Write “**Drop Survey Route**” and **#** of Survey route in upper right hand corner.
    vi. Include your initials in the upper left corner.
c. If a sequence cannot be created
    i. Write **“Ignore – No Routes between OD”**
    ii. Include your initials in the upper left corner.