Use of Electronic Fare Transaction Data for Corridor Planning

by
Mark E. Hallenbeck
Director

Eric Howard
Graduate Research Assistant

Dmitri Zyuzin
Computer Support Analyst

Ryan Avery
Lecturer

Miki Verma
Research Assistant

Washington State Transportation Center (TRAC)
University of Washington, Box 354802
Seattle, Washington 98105-4631

Washington State Department of Transportation Technical Monitor
Janice Helmann

1. REPORT NO. WA-RD 863.1	2. GOVERNMENT ACCESSION NO.	3. RECIPIENT'S CATALOG NO.	
4. TITLE AND SUBTITLE Use of Electronic Fare Transaction Data for Corridor Planning		5. REPORT DATE June 2017	
7. AUTHOR(S) Mark E. Hallenbeck, Eric Howard, Dmitri Zyuzin, Ryan Avery, Miki Verma		8. PERFORMING ORGANIZATION REPORT NO.	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Washington State Transportation Center (TRAC) University of Washington, Box 354802 University District Building; 1107 NE 45th Street, Suite 535 Seattle, Washington 98105-4631		10. WORK UNIT NO.	
		11. CONTRACT OR GRANT NO. Agreement T1461, Task 15	
12. SPONSORING AGENCY NAME AND ADDRESS Research Office Washington State Department of Transportation Transportation Building, MS 47372 Olympia, Washington 98504-7372 Doug Brodin, Project Manager, 360-705-7972		13. TYPE OF REPORT AND PERIOD COVERED Research Report	
		14. SPONSORING AGENCY CODE	
15. SUPPLEMENTARY NOTES This study was conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.			

16. ABSTRACT

     This report documents the initial phase of a project that developed, tested, and used a methodology and tool set for converting electronic transit agency fare card transaction data, the Puget Sound region's ORCA electronic fare card, into information that describes how customers are using the transit system. The project used two nine week datasets. The first covers February 17, 2015, to April 14, 2015. The second covers from March 26, 2016, to May 27, 2016.

     The report describes the detailed steps necessary to combine ORCA transaction data with vehicle location data and other data sources in order to estimate ridership patterns and other information useful for transit planning and operational needs. Data provided includes origin/destination matrices by day-of-week, and time-of-day, transfer locations, and transfer details such as the distances walked in order to transfer and the time required to make those transfers. The detailed data processing flow is described. Also included is a detailed discussion of data privacy concerns, how those concerns were addressed in this project, and how they might be addressed as part of an ongoing business intelligence system.

     The report also illustrates a wide variety of uses for these types of data.

17. KEY WORDS Electronic transit fare card data, Transit planning data	18. DISTRIBUTION STATEMENT No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22616		
19. SECURITY CLASSIF. (of this report) None	20. SECURITY CLASSIF. (of this page) None	21. NO. OF PAGES	22. PRICE

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Washington State Department of Transportation or Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

TABLE OF CONTENTS

FIGURES

TABLES

EXECUTIVE SUMMARY

INTRODUCTION

This report describes a joint Washington State Department of Transportation (WSDOT)/Sound Transit project that explored the use of fare transaction data from One Regional Fare Card for All (ORCA) to improve corridor and system planning.

Transit agencies across the nation are increasingly using electronic fare media to speed passenger boarding, reduce the cost of fare collection, provide various other benefits to their riders, and support more complex fare transactions. This project demonstrated many ways in which the use of the data collected with these electronic fare media can describe the performance of the transit system, particularly from the viewpoint of the people riding the system. This information can significantly benefit the transportation planning processes for both transit agencies and metropolitan planning organizations (MPOs).

This project obtained and analyzed nine weeks (February 17, 2015 – April 14, 2015) of ORCA transaction data from Sound Transit, with the permission of the ORCA Joint Board. These data were provided under a strict set of privacy and security controls approved by the University of Washington's Institutional Review Board (IRB). Towards the end of the project, an additional nine weeks of data were obtained from the ORCA Joint Board covering the period from March 24, 2016, to May 27, 2016, which was shortly after the UW and Capitol Hill Link Light Rail station opened.

To convert ORCA transaction data into highly useful information for purposes other than revenue collection and financial reporting, additional data sets must be collected. Thus, four additional types of data were collected and used in this project to demonstrate the wide range of planning and operational benefits that can be obtained by using ORCA data. The additional four categories of data collected and used were

- transit agency operations files
- transit service support files
- Commute Trip Reduction files
- non-ORCA/transit files.

Non-ORCA/transit data includes data such as census data and data on the built environment, such as street networks, sidewalk data, and land use data.

ORCA DATA DESCRIPTION AND DATA PROCESSING

The nine-week 2015 ORCA dataset describe over 20,000,000 transit boardings. The 2016 dataset describes over 23,000,000 transit boardings. In both cases, over 28 percent of these boardings are indicated as transfers in the ORCA financial records, where ORCA considers a boarding as a "transfer" whenever that boarding takes place within two hours of when payment was first made. Additional analysis shows that 11 percent of these transfers are being made to and from the same transit route. This indicates that they are actually boardings being made after a short duration activity has been performed. Thus, for planning purposes, they are really the start of a new trip, not a vehicle-to-vehicle transfer. This type of insight is a good example of what can be learned by examining ORCA data.

To help with that analysis process, the project team developed four analytical database files. These are:

- geolocated transactions (where and when ORCA boardings take place)
- transfer details (the locations where riders get off, and then back on transit when transferring)
- stop-to-stop origins and destinations (the base data needed to make an origin/destination matrix, where each trip record includes the number of transfers needed for each trip and the trip duration.)
- no origin/destination found file (which describes those trip boardings for which a destination cannot be determined with confidence.)

To create these four analytical data sets from the initial data supplied by the ORCA Joint Board requires a series of data processing steps. These steps are categorized into five basic steps.

- Preprocessing
- Geolocation processing
- Transaction record clean-up processing
- Origin/destination processing
- Transfer processing.

Each of these tasks is described in detail in the main body of this report.

EXAMPLE ANALYSES

This report includes a number of examples of the types of analyses that can be performed using ORCA, such as:

- geographic location of bus boardings by day of week and time of day
- potential for bias (difference between ORCA boardings and total boardings) in ORCA analyses
- route use profiles (boardings by location and time of day, by type of ORCA user)
- transfer activity (location, time of day, day-of-week, duration of transfer, and route-to-route interactions)
- differences in travel by type of ORCA fare payment type (electronic cash payment versus multiple types of passes),
- analysis of the effectiveness of transit oriented commute trip reduction programs.

For example, Figure Ex-1 presents a map which illustrates the intensity of boardings using ORCA for weekdays. This map could be easily produced for weekends, or adjusted based on observed difference between automatic passenger counts and ORCA boardings. It could also be produced for just boardings made using ORCA Lift (low income) cards, or disability cards. Illustrating the geographic distribution of these important riders.

Figure Ex-1: Map of 2015 ORCA King County Metro Transit Boardings

Figure Ex-2 show how differences in APC and ORCA boarding counts can be used to illustrate those neighborhoods where ORCA is not used heavily. This would be an excellent way

to help direct ORCA marketing campaigns. Future geo-located ORCA transaction data could then be used to evaluate the effectiveness of those campaigns.
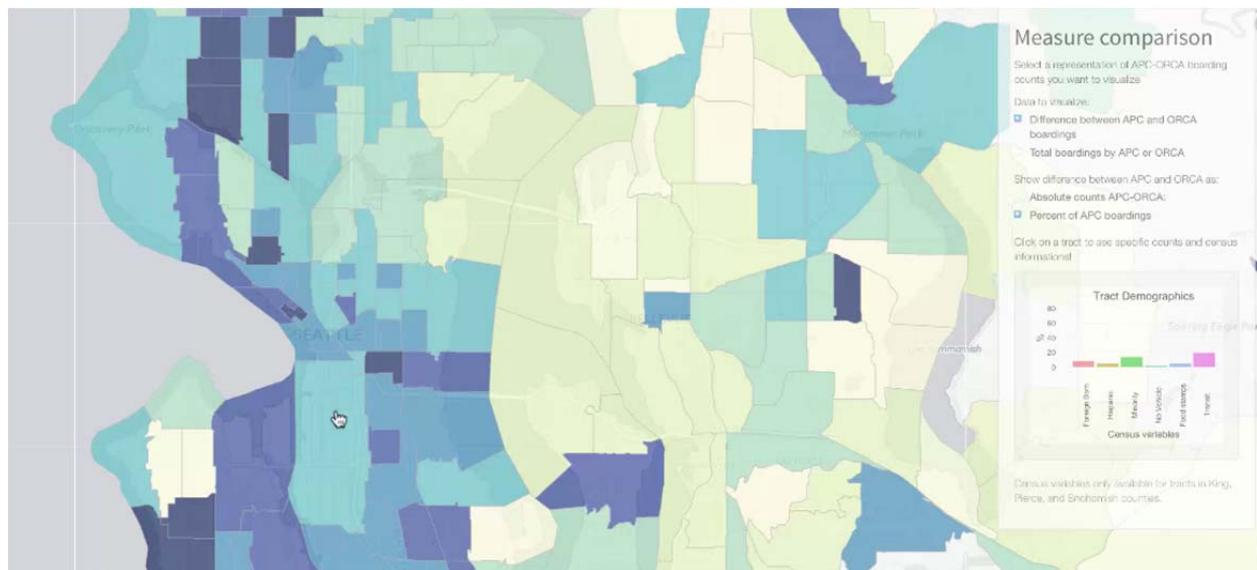
Figure Ex-2: Map of Differences in ORCA and Automated Passenger Count Boardings, by Geographic Zone

Summaries of these types of usage statistics also allow transit use to be compared to the wide array of census variables, allowing the analysis of transit use in neighborhoods of specific socio-economic status.

Similarly, origin/destination matrices aggregated to the zonal level, such as illustrated in Figure Ex-3, which illustrates the zones in the region to which trips starting at SeaTac airport are destined, can be used to help design and refine the transit route structure.
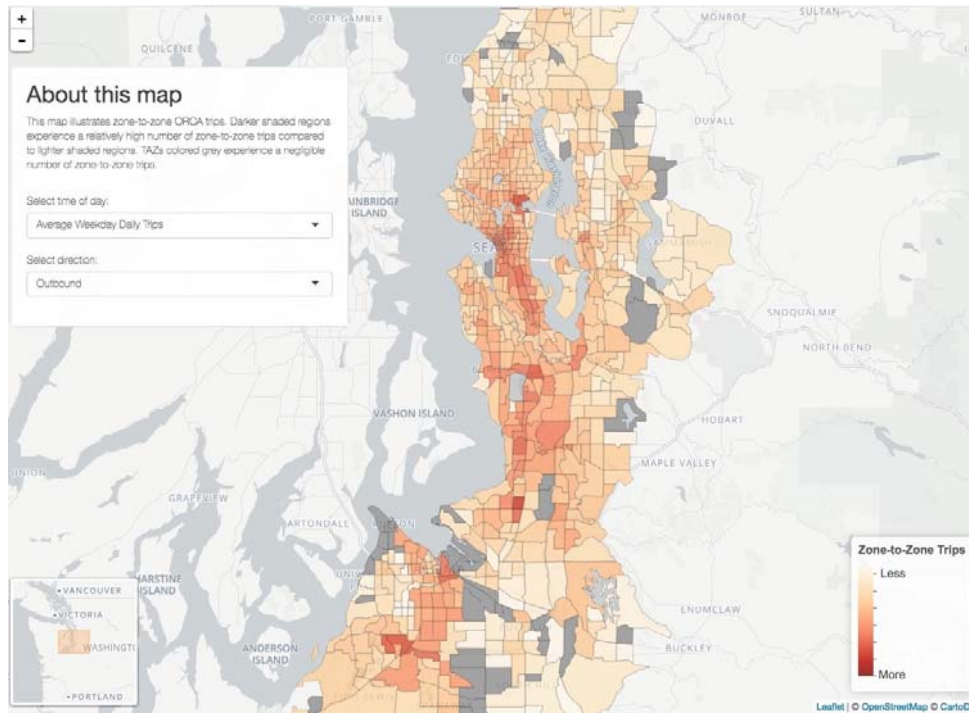
Figure Ex-3: Map of Destinations for Trips Starting at SeaTac Airport

ORCA also allows detailed examination of transfers occurring at rail stations, park and rides, and other major transit centers. For example, Table Ex-1 shows the change in the percentage of trips transferring to Link light rail at the stations open in 2016.

Table Ex-1: Percentages of Link Trips That Involved Transfers from Other Transit Modes

Link Station Name	Transfer Percentage (2016)	Transfer Percentage (2015)	Change 2016-2015
UW	39.2%	N/A	
Capitol Hill	19.8%	N/A	
Westlake	31.9%	38.0%	-6.1%
University Street	27.4%	22.5%	4.9%
Pioneer Square	20.7%	19.1%	1.6%
International District	49.6%	47.8%	1.8%
Stadium	12.4%	17.4%	-4.9%
SODO	39.4%	35.7%	3.7%
Beacon Hill	36.3%	38.2%	-1.8%
Mount Baker	34.3%	35.9%	-1.6%
Columbia City	14.5%	15.7%	-1.2%
Othello	18.2%	18.0%	0.2%
Rainier Beach	36.2%	37.7%	-1.6%
Tukwila Intr. Blvd.	29.1%	28.0%	1.1%
Sea-Tac Airport	27.4%	29.4%	-2.0%

More detailed analysis of transfers at different stations revealed important details of those transfer activities. Table Ex-2 compares the transfer activity at the International District Station in the two nine-week periods in 2015 and 2016. The very large growth in the movement from Sounder to Link is very apparent. Also apparent is the growth in both total ridership caused in part by the opening of the Capitol Hill and UW stadium stations, as well as the growth in streetcar use as a result of the new Broadway streetcar line.

Table Ex-2: Transfer Details at International District Station

Mode Transferring From	Percentage 2016	Number of Trips 2016	Number of Trips 2015
Ferry	0.0%	1	1
Sounder	21.2%	45,672	18,823
Regular Bus	20.1%	43,376	23,317
Streetcar	1.2%	2598	14
BRT	0.1%	193	80
Link	6.9%	14,767	5,033
Demand Responsive	0.2%	410	118
No Transfer	50.4%	108,646	51,694

Of particular interest to this project was the impact of employer based commute trip reduction activities on transit use. Several analyses were undertaken to examine this. Findings from this project show that the price of parking, the level of transit service within close proximity of the employment site, and the built environment around that site are the primary factors that reflect transit use. Employer subsidies are statistically significant, but have modest impact on total transit use. For example, Table Ex-3 shows that the presence of free parking dramatically reduces the percentage of employees that use ORCA at all.

Table Ex-3: Impact of Parking Price on Employee Use of Transit

	Percentage of Employees Who Use the ORCA Card	Standard Deviation of Percentage of Employees	Percentage of CTR Affected Employees Who Use the ORCA Card	Standard Deviation of Percentage of CTR Employees
Pay Parking	55%	31%	80%	102%
Partial Pay Parking	46%	35%	56%	39%
Free Parking	13%	16%	17%	20%

Interestingly, those employees at companies with free parking that do use transit - even occasionally - to commute to work, take transit more often – on average - than employees who have to pay for parking, but use transit at all. It appears that employees that chose to use transit, despite having free parking are more dedicated transit riders.

Where parking is not free, many individuals take transit, but they appear do so reluctantly, likely preferring to drive to work but choosing transit for financial reasons. They therefore periodically take other modes – or telecommute - and therefore use transit slightly less often. This leads to lower transit utilization rates per user, <u>for those people who take transit at all</u>, given free parking. Additional research and analysis is needed to confirm this interpretation of these preliminary results.

Preliminary analysis of other CTR values showed the following results:

- The impact of subsidy value on pass use showed a statistically significant increase in transit use as the transit subsidy value increased. A $10 increase in subsidy resulted in roughly a 3 percent increase in transit use.
- The availability of telecommuting programs and subsidies for other alternative modes reduce transit use.
- The availability of flex time increases transit use.

Analysis of other contributing factors on transit use indicate that the quality of bus service – measured in terms of the number and frequency of buses at stops close to the worksite – is one of the largest predictors of the amount of transit use occurring to a work site. That is, poor transit service will discourage transit use more significantly than an increasing subsidy will stimulate its use.

The complexity of the interactions of the various built environment, transit service, and commute trip subsidy interactions have produced difficulty in the project team's ability to extract reliable statistics from the data. Work in this area continues and will be submitted to WSDOT when it has been completed.

The real benefit of ORCA is that its use unlocks a vast number of analyses. As part of this project, in addition to what is illustrated above, other example analyses performed include:

- creation of new origin/destination matrices by day-of-week, time-of-day, and type of rider,
- transit trip distances and travel times
- information that can be used to compute the outcomes of fare rate changes (e.g., the number of trips fitting specific pricing scenarios),
- estimates of communications requirements for expanded transit revenue system enhancements
- estimates of light rail train and station over-crowding, based on passenger train delays (because some riders cannot board over crowded trains.)
- changes in transit rider behavior based on changes in service levels (e.g., how transit travel changed as a result of the introduction of Link to the University District.)

RECOMMENDATIONS

The project team, including staff from the Washington State Department of Transportation, Sound Transit, the Puget Sound Regional Conference and the University of Washington, recommend that the various ORCA agencies work together to make the use of ORCA data an ongoing activity. When Phase 2 of this project is completed on December 31, 2017, no additional analyses with these data will be performed unless additional funding is identified for those tasks. However, the data will be archived for future work should funding become available. The project team has been working with WSDOT and the various transit agencies to find funding to continue the analyses and to process additional weeks of ORCA transaction data.

Ideally, ORCA data should be routinely processed through the steps described in this document and routinely used for a variety of planning and operational analyses. Both Sound Transit and King County Metro have proposed funding the development of enterprise data systems that would convert the initial code developed under the original ORCA data project to a more stable, sustainable enterprise system meant to routinely perform the required tasks. Several solutions for converting the basic research software into an enterprise system that could be used by all ORCA agencies have been proposed, but no formal decision has been made.

It would be to the advantage of all agencies that employ ORCA that only one such system be built for the region. All ORCA agencies should then have access to those analytical

capabilities. Such a system is well within the technical capabilities of the region. However, before work is completed, the ORCA Joint Board needs to develop policies for the use of the data. This work has been started by Sound Transit and needs to be supported by the ORCA Joint Board.

CHAPTER 1: INTRODUCTION

This report describes a joint Washington State Department of Transportation (WSDOT)/Sound Transit project that explored the use of fare transaction data from One Regional Fare Card for All (ORCA) to improve corridor and system planning.

Transit agencies across the nation are increasingly using electronic fare media to speed passenger boarding, reduce the cost of fare collection, provide various other benefits to their riders, and support more complex fare transactions. This project demonstrated many ways in which the use of the data collected with these electronic fare media can describe the performance of the transit system, particularly from the viewpoint of the people riding the system. This information can significantly benefit the transportation planning processes for both transit agencies and metropolitan planning organizations (MPOs).

Electronic fare transaction data have not been routinely used for the purpose of planning and policy decision making, in large part because

- a number of complex steps are necessary to convert electronic fare transaction data into useful information
- concerns about rider privacy need to be addressed
- these collected data are still relatively new for many transit agencies, and therefore, ways in which these data can be used have not been explored.

This project explored the first two of these issues in detail. This report is meant to help transit agencies realize that the work required to use the electronic fare card data is small in comparison to the benefits that can be gained from their use.

PROJECT INTRODUCTION

WSDOT and its partner agencies are often limited in their ability to plan and implement an integrated, multimodal system by a lack of information about how the current system is used and is performing, particularly from the perspective of users. For example, agencies lack detailed data on how transit services are used given the various demand management strategies employed by companies under the Commute Trip Reduction Act.

This project obtained and analyzed nine weeks (February 17, 2015 – April 14, 2015) of ORCA transaction data from Sound Transit, with the permission of the ORCA Joint Board. These data were provided under a strict set of privacy and security controls approved by the University of Washington's Institutional Review Board (IRB). At the end of the project, an additional nine weeks of data were obtained from the ORCA Joint Board and processed so that they too could be used for analysis. The second nine weeks covered the period between March 24, 2016, and May 27, 2016, which was shortly after the UW and Capitol Hill Link Light Rail station opened.

This project demonstrated how the innovative use of data sources such as ORCA can significantly increase the information available to planners that describe current transit use. For example ORCA data can describe the actual transit rider trip making behavior, including the time required to make trips, the time required to transfer between routes, and the paths taken between specific origins and destinations. In addition, when transit fare card data are combined with other data sources—such as those that describe transit subsidies or the built environment—they allow the construction of analytical models that can be used to better predict future transit ridership, given changes in subsidies or the built environment.

This project was designed to illustrate how the use of ORCA fare card transaction data can improve WSDOT's ability to plan and deliver community-based, multimodal solutions. The project also used ORCA data to help evaluate the effectiveness of WSDOT's transportation demand management activities so that WSDOT and participating employers can make better use of their available Commute Trip Reduction funding.

REPORT ORGANIZATION

This report is divided into five chapters. This first chapter introduces the project. The second chapter describes the data used during the project, including not only the ORCA fare card data but the other data sets that were linked to the ORCA data in order to develop highly useful analytical data sets and/or answer key questions. The third chapter presents and discusses a number of example analyses performed with these ORCA-based analytical data sets. The fourth chapter presents the actual data processing steps required to process and fuse the multiple data sets into key analytical data sets useful for transit and regional planning. The fifth chapter presents conclusions and recommendations for further work.

CHAPTER 2: SUMMARY OF DATA SETS AND DATA PROCESSING

This chapter summarizes the data used for the analyses described in Chapter 3. The chapter provides

- an introduction to the ORCA fare cards and the transaction files that result from their use
- a discussion of other data sets required to take full advantage of the ORCA transaction files, including both transit agency operational data and non-agency data
- a description of the data sets produced from the ORCA transaction data that describe transit rider behavior
- a short discussion of the data processing steps needed to produce the analytical data sets
- a summary of the limitations of the analytical data sets, which need to be considered as caveats when the ORCA data are used
- a short discussion on how rider and business privacy must be addressed when ORCA transaction data are used.

ORCA FARE CARDS AND ORCA TRANSACTION DATA

ORCA fare cards are near-field communication "tap cards[1]" that allow riders to pay for transit rides on nine different transit systems (Community Transit (CT) buses; Everett Transit (ET) buses; King County Metro (KCM) buses and demand responsive paratransit services; the King County water taxi; Kitsap Transit (KT) buses; Pierce Transit (PT) buses; Sound Transit (ST) buses, light rail, and commuter trains; and Washington State Ferries (WSF)). The ORCA system collects, stores, and uses a variety of financial transaction information. These data are used to track all financial aspects of the transit payment system, including when and where revenue is added to specific ORCA fare cards, and when and where those fare cards are used.

This project obtained data only on the use of ORCA fare cards to pay for transit trips. It did not obtain data on how those cards were paid for, or when and how additional value (e.g., a new monthly pass or additional e-purse value) was added to the account represented by that card.

---

[1] To pay for a transit trip, the card holder simply taps the fare card on the reader. The reading device determines the fare card's serial number (CSN) and generates a transaction record that is used to bill the account to which that card is attached. An audible sound, supplemented by a small message on the reader screen, indicates that the fare has been paid. If the read is not successful, a different sound and message indicate that an error occurred so that the card holder can repeat the tap.

The data obtained by the project team were "transaction records," which indicate when transit vehicles were boarded or when an off-board ORCA reader was used to pay for a light rail (LRT), commuter rail (Sounder), or bus rapid transit (BRT) trip.

Nine weeks of ORCA transaction records were initially obtained for the period of February 14, 2015, through April 17, 2015. ORCA transaction records contain the following basic types of data:

- the ORCA card serial number (CSN)
- a business ID that identifies the employer that originally purchased that card to provide it to an employee
- the time and date when the payment occurred (the boarding of a bus, or the time when an off-board reader was tapped for BRT, rail, or ferry modes)
- the device on which the tap occurred (i.e., either the vehicle ID of the bus that was boarded or the device—off-board location—that was tapped)
- the service provider
- the route boarded and direction of travel (if a bus)
- a "trip ID[2]" associated with that route
- the date when the fare payment was made (e.g., e-purse amount, or type of pass used for the payment)
- whether that card provides a specific discount based on its use by a specific passenger type (e.g., youth pass, disabled person pass, senior pass, or Lift-low income pass)
- the number of passengers paid for (when the ORCA card used is an e-purse, a single tap can be used to pay for more than one rider)
- whether the transaction was a transfer, where "transfer" is defined as the payment record having occurred within two hours of the first payment for a trip

---

[2] "Trip ID" is an identifier used to define a specific vehicle journey on a specific route. For example, a given trip ID might be used to reference the third, southbound trip of the day on Route 71. A different trip ID would be used for the fourth southbound taken on Route 71. The Trip ID variable is meant to identify a specific, consistent journey. The same Trip ID is often used for all similar trips for every weekday. These trips would all have the same starting time, and would stop at all of the same bus stops in the same order. So the same Trip ID might be used to identify the third southbound trip of the day for all five weekdays, since that trip starts at the same time on each of those days. A different Trip ID would be needed on Saturday because the third trip would start at a different time of the day.

- if the trip was defined as a transfer, then the mode of the vehicle taken just before the current boarding, the service provider of that mode, and the route (when it was a bus) previously taken.

Note that the location of the transit vehicle boarding activity is not directly available as part of the initial ORCA transaction record. Additional data processing is necessary to determine the physical location where that tap occurred by matching the time, date, and "device" information (vehicle ID or off-board reader ID) with either agency vehicle location records or a table that indicates the physical location of the off-board readers. As a result, additional data are needed for fusion with these records.

OTHER DATA SETS REQUIRED

In addition to the base ORCA transaction files, four types of data were used to demonstrate the wide range of planning and operational benefits that can be obtained by using ORCA data. Those four categories of data were

- transit agency operations files
- transit service support files
- Commute Trip Reduction files
- non-ORCA files.

Each of these categories is discussed briefly below.

Transit Agency Operations Files

To geolocate the ORCA payments that occurred on buses, the ORCA transaction data had to be fused with transit agency bus location data. Transit (automatic) vehicle location (AVL) data were provided by each of the participating agencies (CT, KCM, and PT, with ST data provided by those three operating agencies). For off-board ORCA payments, the locations of all fixed readers for Link Light rail, Sounder heavy rail, and both BRT (King County's Rapid Ride and Community Transit's SWIFT) services were obtained. The off-board reader table also listed the stop IDs found in the AVL data associated with those stops, and the BRT routes that were served by those locations. Note that some of the fixed reader locations changed between 2015 and 2016, as did the routes that served some specific locations.

Transit AVL typically reports location on a periodic basis (roughly every minute or two), plus whenever a bus opens its doors or turns on or off its engine. The AVL files provided by the agencies were in a format that also gave the project team access to Automatic Passenger Counter (APC) data. APC data describe how many people board or exit a bus at each stop. All PT and CT buses are equipped with APCs, although there are times when the APCs are not working but data are still collected by the AVL system. For King County Metro, about 30 percent of all bus journeys also have APCs. The remaining 70 percent of trips do not have APCs, but AVL data are collected in the same format. Sound Transit buses are operated, under contract, by CT, PT, and KCM. Not all Sound Transit buses operated by CT, PT, or KCM have APCs. However, even when a vehicle does not carry an APC, location data are provided.

Separate data files were provided by each of the service agencies, and these data files use slightly different file formats, variable names, and variable formats. (For example, bus direction may be reported as "North" or "NORTH.") While the AVL files from all three agencies (PT, KCM, and CT) included both latitude and longitude (X/Y) coordinate locations for their buses, the three agencies use different geographic projection systems for those coordinates. Consequently, before they could be used within a single analysis framework, these coordinate systems had be converted to a single projection system. The KCM AVL X/Y coordinates were converted from projection 3690 to projection 4326. The CT projection was converted from 2926 to 4326. The PT data used the 4326 system and did not need to be converted.

Sound Transit data were provided by the individual agency under contract for a given Sound Transit service. Those data arrived in the X/Y coordinate system used by the respective agency. Therefore, it was necessary to identify, for each Sound Transit route, which county transit agency operated that service for Sound Transit, and then apply the appropriate coordinate system translation for that specific agency to all data points for that route.

Everett Transit provided access to its vehicle location files for 2016 (but not 2015), but limitations in project resources did not allow their use in this project. Vehicle location data for Kitsap Transit were not available. Washington State Ferries' fixed reader location data were also not used. Actual streetcar location data were not used in this project, although streetcar boarding locations in downtown Seattle were included.

Transit Service Support Files

A number of other data files were obtained from the transit agencies to assist in the processing of the ORCA data. These files included the following:

- route, schedule, and stop order files (obtained as **GTFS files**), which describe all of the stops and the order of those stops for all transit routes operated by day of the week
- the **Sound Transit Route Service Provider Table**, which links specific Sound Transit bus routes to the agency that operates those routes under contract for Sound Transit.

The route, schedule, and stop order files obtained from the transit agencies were used to ensure that analyses were clear regarding the order of stops made during each transit route and how those stops changed over the course of the day, as the AVL data present for any given bus journey might not include all bus stop IDs for that route. This could occur when the bus did not stop at a given bus stop during a given trip. It could also occur when the AVL system on-board the vehicle failed at some point during a trip.

The table linking Sound Transit routes to the transit agency that operated those routes was needed to link the Sound Transit services to the correct AVL file when vehicle position look-ups were performed. This table is different for the 2015 data than the 2016 data, as Sound Transit routes changed between 2015 and 2016.

Commute Trip Reduction (CTR) Files

The state's Commute Trip Reduction law requires that companies with work sites to/from which more than 100 workers arrive or depart during peak periods must have a commute trip reduction program designed to limit the number of vehicle trips arriving at and departing from that location during peak periods. Each company must submit a biennial survey that describes their commute trip reduction activities. The project team obtained a spreadsheet summarizing each company's plan, based on that survey, from WSDOT. Among other data, the spreadsheet described

- the number of workers at each work site
- the number of workers at each site who were subject to the CTR requirements
- the location (street address) of each work site
- the types of travel subsidies provided by the employer, if any (e.g., transit pass or bike use subsidies)
- the cost of on-site and nearby parking

- parking subsidies paid by the employer
- other commute trip reduction activities undertaken by the employer (e.g., vanpool and carpool matching or preferential parking programs)
- the availability of flex time.

Many of the companies affected by CTR purchase transit passes for their employees. Some companies subsidize those passes. Other companies give them away free, and still others sell them at cost to their employees. The project team obtained from Sound Transit a list of all ORCA cards that have been purchased by companies to provide to their employees. This table indicated which ORCA card numbers were associated with particular companies. This table allowed the subsidy information contained in the CTR employer spreadsheet to be associated with individual ORCA cards.

Files Describing Non-CTR Employer Supplied ORCA Cards

Not all employers that purchase ORCA cards for their employees are required to fill out the CTR survey. However, when they sign the contract to purchase ORCA cards, employers indicate whether a specific subsidy will be provided to their employees. The project team worked with King County Metro to extract from the contract information these subsidy data and the specific ORCA card IDs to which each of those subsidies applied. These data were also formatted so that the subsidy information could also be linked to individual ORCA cards.

Non-ORCA Files

Finally, the project team gathered a large number of other data sets that were needed to support the desired planning analyses. These files included
- built environment data in GIS files
- census population and employment data in GIS files.

The built environment data included street, bike, and sidewalk networks, parking costs, the densities and land uses associated with specific parcels in the region, and traffic analysis zone boundaries. The census data were taken from the most recent five-year totals from the American Community Survey. Those data included residential and employment density by geographic

area, along with race/ethnicity, household income distributions, household size, the number of children in the household, and education level distributions for each census block.

ANALYTICAL DATA SETS PRODUCED AND USED

The data processing system invented for this project creates four basic data files that can then be used in a variety of analyses. The four analytical files are as follows:

- a geo-located Transactions file (where and when each boarding occurred)
- an origin/destination file that describes all of the trips for which destinations can be identified
- a no-destination file that contains boarding records for which no destinations can be identified
- a transfer details file that describes the transfers reported in the ORCA transaction records (exit stop, time, and route as well as boarding stop, time, and route).

Three of these files may then be used for a variety of analytical purposes. The fourth (the no-destination found file) is intended primarily for the quality assurance process. For many analyses, these main analytical files will be linked to the GIS information described previously. Each of these four analytical files is described below. A summary of the data processing steps required to obtain these files is provided in the next major section.

Geo-Located Transactions File

This file contains all of the ORCA boarding transactions. For this project that comprised data from the nine-week study periods, with separate files maintained for the 2015 data and the 2016 data. Two separate versions of these files are also created. One contains hashed card identifiers and links to the CTR subsidy tables. The second removes all identifiers. The file with no identifiers is designed to be shared with other data users while meeting all privacy requirements. The file that contains hashed identifiers can only be used by individuals who have met security and privacy rules and regulations for access to sensitive data concerning specific individuals (PII). While the hashed ID's make it impossible to directly identify an ORCA card, studies have shown that even hashed identifiers can be linked to outside information sources in order to re-identify individuals. Consequently, those data cannot be released. However, the

identifiers are necessary if users want to perform analyses about the frequency of system use by specific (if anonymous) individuals. That is, if it is important for analysts to know how often individuals make specific trips, then it is necessary to use the hashed IDs to separate trips made by different individuals.

The geo-located Transactions file contains the following data items:

CSN (hashed card number)	Route_id_from
BID (hashed business identification number)	Route_id_to
	Service_id_from
Date	Service_id_to
Time	Vehicle_id
Mode_id_from	Trip_id_to
Mode_id_to	Trip_start
Num_passengers	Stop_id
Passenger_type	Stop_error
Product_id	Stop_location_x
Direction_from	Stop_location_y
Direction_to	Multi-use_pass

The term "to" in a variable name refers to the route, mode, or service being boarded at that time. The term "from" refers to the previous route, mode, or service provider for the immediate previous transit boarding. These variables contain data only when the current boarding is considered a transfer by ORCA.[3] If the boarding is not considered a transfer, "from" variables are left blank.

Stop_ID is the number associated with a specific transit stop. Stop_IDs are not unique when examined across transit agencies. For example, both CT and PT may use the same Stop_ID, but that number will refer to totally different geographic locations. For example, both may have a Stop_ID = 100, but the CT stop will be located in Snohomish County and the PT stop will be located in Pierce County. In addition, Stop_IDs are associated with specific locations where buses stop, so a major transit center typically has multiple Stop_IDs. For example, the Bellevue Transit Center has more than ten Stop_IDs.

The "direction" variables in ORCA are set to either "inbound" or "outbound." This is different than direction values used by the transit agencies, which use cardinal directions (i.e.,

---

[3] ORCA considers a boarding a transfer when it occurs within two hours of a previous fare payment boarding via ORCA.

north, south, east, west). A specific table is created to convert ORCA's inbound/outbound directional values to the cardinal directions used by the transit agencies.

For ORCA transaction records that were made off-board, the original ORCA record does not contain data for several of these variables, including Route_to, Direction_to, Vehicle_ID, and Trip_id_to. As part of the data processing flow, these values are added to the geo-located record whenever possible; however, there are times when it is not possible to assign those values, in which case the process treats them as missing variables.

Finally, the geo-located Transaction file does contain the ORCA transactions for Kitsap Transit, Everett Transit, Washington State Ferries, and for Demand Responsive rides. However, those transactions are not geo-located. Therefore, the Stop_location_x and Stop_location_y variables for those transactions are also missing values. There are also instances in which data errors in either the ORCA or AVL records prevent the project team's software from identifying specific stop locations. The Stop_location_x and Stop_location_y values for those records are also treated as missing values.

Stop_error is a code used within the geo-location process. A -1 value indicates that this record has not been processed. A Stop_error = 4 means that the required Vehicle ID cannot be found in the AVL database, and therefore bus location cannot be determined. Stop_Error = 5 means that the service provider was not PT, ST, CT, or KCM, and therefore no AVL data are currently available. If the time difference between the boarding time and the closest AVL record for that vehicle ID is greater than 5 minutes, then the Stop_error is set to 1. If the AVL record look-up returns a value of "ERR" (meaning that the AVL record exists within the time limit but that the GPS system on the bus did not function correctly and no Stop_id was assigned for that record), then Stop_error is set to 2.

The Product_id variable describes the mechanism used to pay for a given trip. This numerical code (for which a decoding table exists) indicates if the ORCA pass was used as an electronic purse or one of many types of monthly passes.

Origin/Destination Files

These files contain all trips for which a likely destination can be identified. The process by which likely destinations are identified is summarized later in this chapter. Both trip "origins" and "destinations" are Stop_ids in these files. That is, the origin is not a geographic area or

specific land parcel but a physical transit stop that was the initial boarding location of a trip. Likewise, the destination is the stop at which a rider is estimated to have left the transit vehicle, not a land parcel or larger geographic area.

If the traveler transferred between vehicles as part of a trip, the origin/destination process incorporates that transfer into the O/D estimation process. The number of transfers required to make any trip is recorded on the O/D record. Only one O/D record is produced per trip, no matter how many transfers were required. The origin for a trip is the stop at which the rider first boarded the first vehicle ridden on that trip. The destination is the estimated exit stop from the last transit vehicle ridden as part of that trip.

At this stage in the development of the ORCA data analysis process, "transfers" are taken directly from the ORCA transaction records. This definition (any boardings made within two hours of the first trip boarding) can result in multiple stops within a trip chain[4] being considered transfers. While work has been started on further refining the origin/destination process to separate trip chains into separate origins and destinations, that functionality is not yet available.

The transfer points and transfer activities for each ORCA defined "trip" are placed in the "Transfer" file described in the Transfer subsection below.

As with the geo-located Transactions file, separate origin/destination files were maintained for the nine-week 2015 and 2016 periods. In addition, the process creates two versions of each file. One set of files contains hashed ORCA card and business identifiers. The other set contains no identifiers. If the transit agencies recreate these data sets, they too will need specific data access controls that limit access to records that contain even hashed ORCA card or business IDs.

The origin/destination files contain the following data items:

---

[4] A trip chain is a transportation planning term that refers to a single set of destinations being visited in short order. A classic example of a trip chain is to stop at the grocery store on the way home from work. The chain consists of a trip from work to the store, and from the store to home. In the ORCA transaction records, this is treated as one trip, with a transfer taking place at bus stops near the store if the time between the initial boarding (near the work place) and the boarding near the store occur within two hours of each other.

CSN (hashed card number)
BID (hashed business identification
     number)
Trip_start_date
Trip_start_time
Origin_mode
Destination_mode
Number_passengers
Passenger_type
Product_id
First_direction
First_route_taken
First_service_provider
Origin
Origin_service_agency
Origin_lat

Origin_lon
Destination
Destination_service_agency
Destination_lat
Destination_lon
Destination_time
Last_route_taken
Last_direction
Last_service_provider
Last_mode
Number_of_transfers
Last_vehicle_id
Last_trip_id
Transfer_required
Trip_start
Trip_duration

In these files, "Origin" is the Stop_id for the initial boarding to start the trip. "Destination" is the Stop_id for the stop at which the rider is estimated to have exited a transit vehicle at the end of his/her trip.

The other variables in the origin/destination file that that contain "origin" in their name (e.g., Origin_mode) are taken from the first (non-transfer) boarding observed. The variables with "destination" in their name are associated with the last segment of that trip. If it is a no-transfer trip, then the origin value and destination values are associated with the same vehicle, route, and journey. The destination times are extracted from the AVL data files for bus trips. They are taken from "tap off" times for rail trips. (Note that rail trip destinations can be estimated without a "tap off," but no data on travel time can be inferred unless a transfer occurs at the destination end of the rail trip.)

All latitude and longitude values in these files are in the 4326 coordinate system.

No Origin/Destination File

This file contains the subset of transaction records for which a destination cannot be identified. It includes all boarding transactions that are not part of an identified trip. Transfer boardings that occurred in the middle of a trip for which a destination is identified are <u>not</u> included in this file. Those transfers <u>are</u> included in the Transfer file. The file contains the same variables as those found in the geo-located Transactions file.

Transfer Details File

This set of two files (for this project, there were two each for 2015 and 2016, one with identifiers and one without identifiers) contains variables that describe each transfer. All transfers are included in this file, although not all transfers contain detailed information about the transfer activity. If a given trip (origin to destination) includes three transfers, then in the O/D file described above there is just one origin/destination record, but in the transfer file there are three separate records, one for each of the transfers that were part of that trip.

Each transfer record describes the two routes that were involved in that transfer, the stops used (exit and boarding stops), their locations, and the time required to make the transfer. For buses, the exit time comes from the AVL databases. After the Stop_id on the previous route that was closest to the transfer (new) boarding stop has been identified, a "reverse look-up" occurs in the AVL databases. That look-up identifies when the previously used vehicle stopped at the estimated exit Stop_id (the closest stop to the transfer boarding location). That stop time is assumed to be when the rider exited the first bus. The next boarding time is the time when s/he completed the transfer. The transfer duration is the second time minus the first time. (Note that time must be date inclusive in order to handle transfers that occurred around midnight.)

For trips involving rail modes, the "exit time" is the "tap off" time, and the "boarding time" is the "tap on" time. Note that these assumptions will understate the actual transfer times. For off-board BRT trips, the system determines the "boarding time" by examining the time when the first BRT bus stopped at the off-board location <u>after</u> the off-board tap occurred. (For example, the rider tapped off-board at 10:05:30, but the next BRT bus did not arrive at that stop until 10:08:25. The rider is assumed to have boarded at 10:08:25, even though s/he tapped almost three minutes earlier.)

The transfer file contains the following variables:

CSN (hashed card number)	First_route_taken
BID (hashed business identification number)	First_service_provider
	First_trip_id
Transfer_boarding_date	Exit_stop_id
Transfer_boarding_time	Exit_time
Origin_mode	Exit_lat
Destination_mode	Exit_lon
Number_passengers	Transfer_boarding_stop_id
Passenger_type	Transfer_stop_lat
Product_id	Transfer_stop_lon

Transfer_route_taken   Transfer_to_vehicle_id
Transfer_direction   Last_trip_id
Transfer_newservice_provider
Number_of_transfers

BASIC DATA PROCESSING DESCRIPTION

This subsection summarizes the data processing steps for geo-locating the ORCA transaction records. A more detailed discussion of the processing steps is included in Chapter 4. The steps required to create the four desired analytical data sets from the initial data supplied by the ORCA agencies are organized into the following tasks:

- Preprocessing

- Geolocation processing

- Transaction record clean-up processing

- Origin/destination processing

- Transfer processing.

Each of these tasks is described below.

Preprocessing

A number of preprocessing steps are required. These steps standardize the geographic references and create a number of look-up tables that greatly speed the complex data processing. The following pre-processing tasks are performed:

- Import and standardize AVL files,

- Create stop location table

- Update off-board stop location table

- Create (or update) the quarter-mile look-up table

- Create subsidy table

- Link the subsidy table to ORCA cards (CSNs)

- Hash the CSNs and Business IDs in the subsidy table, maintaining the link between the subsidy table and the hashed CSNs

- Preprocess date and time values in the transaction data

- Remove duplicate boarding records.

Each of these tasks is described below.

Import and Standardize AVL Data

The vehicle location data files supplied by the three participating transit agencies (KCM, PT, and CT) arrive in different formats and contain slightly different contents. In addition, the three agencies use different latitude/longitude projection systems. This initial data preparation task reads in these files and creates the databases and look-up queries necessary to create analytical files.

Update Off-Board Stop Location Table

A key issue when individual bus boardings are geographically located is, "*Where are the stops actually located*?" The AVL files that describe actual bus journeys contain that information, but the X and Y values reported in the AVL files can be different each time a bus reports its location at a stop because of the error inherent in GPS reports. To provide a single definition of where each individual transit stop is located, and to ensure that all transit stops are present for an analysis (because each new transit schedule can add, remove, or move stops to account for changing ridership patterns and new services), this pre-processor step creates a single table that lists all transit stops and their locations.

To perform this task, the pre-processor removes extreme X/Y coordinates for each stop and service provider from the AVL files (this accounts for bad GPS data) and then finds the median values for the X and Y coordinates from the remaining coordinates. Stop locations for off-board payment modes (BRT, LRT, and Commuter rail) are obtained from tables supplied by the agencies. These locations must then be linked to the bus Stop_ids for King County Metro and Community Transit BRT stops that are used by the AVL system. This is because the off-board Stop_ids use ID values that are different than those reported within the AVL system.

Create Quarter-Mile Look-up Table

A "quarter-mile look up table" is then created. This table is used to determine where an individual departs a bus, both in the transfer analysis and to estimate destinations. The quarter-mile look-up table is produced by a series of GIS queries. A row in the table exists for each transit stop (Stop_id) active during a study period. (Note that a new table will need to be produced for every new schedule shake up. Schedule shake ups typically occur three times per year.) Each row in the table indicates which routes pass within one-quarter mile of that Stop_id,

and which Stop_id on each of those routes is the closest stop to that specific location for that route. A table is generated and used to perform the ¼-mile look-up because GIS queries are relatively slow in terms of computer processing time, whereas database look-up tables are fast. Creation of a look-up table speeds the creation of the O/D and Transfer files.

Create Commute Trip Reduction Subsidy Table

One area of interest in this project is the effect of employer subsidies on transit use. To support those kinds of analyses, the subsidy information obtained from CTR employers, as well as the transit pass subsidy information obtained from KCM ORCA Business Passport contract documents, is summarized in a subsidy table. A row exists in the table for each employer. The employer identity is then linked to specific ORCA card serial numbers (CSNs) and then the subsidy table rows are linked to individual CSNs.

Hash ORCA Card Numbers and Business IDs

Once the subsidy values have been linked to the CSNs, the CSNs are hashed with a two-factor hashing routine. Business IDs are also hashed. However, the (hashed business ID based) subsidy table rows remain linked to the individual (hashed) CSNs. This allows analysts to understand the subsidies associated with each ORCA card, but will not allow the direct association of an employer with a specific card.

Preprocess Date and Time Values in the Transaction Data

The next pre-processing step is to load the transaction data, provided to the project team as CSV-formatted text files, into a database. When the data are loaded into the project's MySQL database, the separate Date and Time fields found in the CSV file must be transformed into a single Datetime field. Simultaneously, three additional variables are added to the new database formed from the transaction records. These (initially blank) variables are Stop_id, Stop_location (X, Y coordinates), and Stop_error. The Stop_error variable is then set equal to -1 to indicate that the record has yet to be subjected to geo-location processing.

Remove Duplicate Boarding Records

The final preprocessing step removes duplicate boarding records that occur when an ORCA card is used both as a pass and as an e-purse to pay for a single boarding transaction. This occurs when the value of the transaction is greater than the value of the pass stored on the ORCA

card, and that ORCA card has an electronic purse cash value. In such an instance, two records are created in the ORCA transactions database. These two records are identical except for the "Product_id" variable. On one record this variable indicates the pass being used, and on the other record, this variable indicates that an e-purse was used. In this step, the e-purse record must be deleted. A new variable, "Multi-use_pass," is created and set to "1" in the remaining record when the e-purse record is deleted. (This variable is set to "zero" in transaction records when duplicate records are not present.)

Geolocation Processing

Geolocation processing is performed in two separate tasks. The first is for payments made on-board buses. These are identified in the ORCA records as having Mode_to variables equal to 128. The second is for payments made at fixed readers. This latter task includes all Link and Sounder payments, as well as all SWIFT (CT BRT) and a large percentage of Rapid Ride (KCM BRT) payments.

The location of a bus when a payment was made on the bus is determined by using data found in the transaction records to identify the specific bus being boarded and then looking up that bus's location in the AVL files for the operating agency. This task is complicated by a number of important details. The first complication is the fact that the time reported on the ORCA transaction is not directly found in the AVL file. Time differences between the two files occur for two reasons. First, the clocks used by ORCA fare collection readers are not synchronized to clocks used by the AVL system. Second, the time recorded by ORCA is when the tap occurs, whereas the time recorded by the AVL system is when the bus opens its doors. Thus, the "match" between the ORCA payment and the AVL record is inexact. Therefore, the process looks up the AVL record for *that bus with the closest time to the ORCA record.* A constraint of 5 minutes is placed on that search. (That is, if the bus being examined does not have a stop recorded within 5 minutes of the ORCA transaction record, that ORCA boarding cannot be geolocated.)

The second major constraint is that errors occur in both the AVL system and in the ORCA fare collection system. Common errors include missing values in the ORCA record (e.g., no vehicle ID present in the transaction record), invalid GPS points in the AVL record, and

missing AVL records (e.g., the AVL system failed to report a given bus's location for some period).

If no AVL record with a valid location is found to match an ORCA transaction record within 5 minutes of the ORCA payment time, then that transaction record is not geolocated, and the appropriate Stop_error is assigned.

Once all on-board ORCA records have been processed, the off-board BRT transactions (Mode_to = 250) can be processed. Most (but not always all) off-board payments are already included in the Transactions file. This process requires two passes through the database, with the transactions in the database being processed by CSN, by date, by time. The first pass through the off-board BRT transactions assigns the Route_to, Direction_to, Stop_id, and Stop_location (X/Y) variables to all off-board transactions found in the database that occurred at BRT stops served by one and only one BRT route. This comprises the majority of the stop locations. Once a route and direction have been assigned, the AVL file is then searched for that date, time, Stop_id, direction, and route to determine which actual vehicle (Vehicle_id) operating that route, in that direction, that stopped at that location closest in time to, but after, that transaction time. This physical bus is assumed to be the bus boarded by that individual. The Vehicle_id variable for that transaction record is then updated with all of those values, and a record indicator is placed in the record indicating which record in the AVL file is associated with that boarding. (Note: this same approach could be used to assign streetcar boardings to specific vehicles, but the project team does not have a streetcar AVL file.)

Unfortunately, a number of transit stops in downtown Seattle serve multiple Rapid Ride lines (the C, D, and E lines). Trips boarding at those locations cannot be directly assigned to routes and directions because the rider may have gotten on any one of the three lines. During the first pass through the database, the Route_to variable for these records is set to "CDE." To determine which of the routes, and thus vehicles, should be assigned to each boarding requires a second pass through the database.

On this second pass through the database, the record immediately following the CDE record is examined. If that record contains the same CSN as the CDE record, then a quarter-mile look-up[5] is performed around the boarding location for that record (i.e., the record immediately

---

[5] See the origin/destination subsection for a thorough discussion of the quarter-mile look-up process and the reasoning behind that process.

19

following the CDE record). If one and only one BRT route is within that quarter-mile look-up OR if that record describes a boarding of an identified BRT route, then that route (either the BRT route being boarded or the one BRT route operating within a quarter mile of the new boarding location) is the BRT route assumed to be boarded during the CDE transaction.

By using that new route assignment, the system can now look up in the AVL files which vehicle, operating on that route in that direction, passed by the off-board location for the CDE record closest to—but after—the CDE payment tap recorded at that off-board location. The CDE record in the Transactions file is then updated to include the newly identified Route_to and Vehicle_id information, and a pointer is created which indicates which AVL record is associated with that boarding.

If there is no "following record" for that CSN, or if the following record does not occur at a transit stop within one-quarter mile of one of the C, D, or E Rapid Ride routes, then the route boarded cannot be identified. When this occurs, Route_to is left equal to CDE, and Vehicle_id is left as a missing value.

Sounder and Link off-board locations are processed next. In theory, all Sounder and Link boardings should be found in the Link and Sounder transactions files provided by the ORCA system. However, in practice, the project team found that some Sounder and Link trips not present in the original ORCA transactions file were found in the separate ORCA rail payment files. These ORCA payment files also include both the tap on and tap off information needed to determine the cost of the distance-based rail fares.

Thus, for the rail files, the separate Link and Sounder rail payment files are transformed so that a single record exists for each rail trip. The station at which the rail trip started is selected as the boarding location. The record in the ORCA transactions file that matches the date, time, and CSN for this record is then identified. If no record is found, a new record is inserted with the appropriate CSN, date, time, and location information. The Route_to and Vehicle_id variables in the geolocated transaction are then set to LINK or SOUNDER, and the Stop_id, Stop_location, and Stop_error variables are updated.

Transaction Record Clean-Up Processing

Once geolocation processing has been completed, one additional data processing step is necessary. Once the off-board payment transactions have been processed, transfer records

occurring <u>from</u> those trips need their "*_from" variables (e.g., "Route_from") updated. To do this, transaction records are processed in CSN, date, time order. Whenever a record contains a "Mode_from" value equal to 250 (BRT), 251 (Link), or 9 (Sounder), the variable Route_from must be updated to the value found in the previous record's Route_to variable.

Origin/Destination Processing

Once all records that can be geo-located have been processed, it is possible to estimate the destinations of many trips, as well as compute the details (distance traveled and time required) of many transfers. The origin/destination and transfer processing procedures are simultaneous but produce separate files. This is because determining a rider's likely destination requires following that rider's trip across transfers. Therefore, the transfer detail file is created at the same time that destinations are determined. These procedures are discussed briefly below, with more programmatic details provided in Chapter 4.

Each non-transfer ORCA transaction is the origin of a trip.

The assumption behind the estimation of the likely destination of that trip is that, for many trips, the origin point for that rider's next trip can serve as a good estimator of the destination of that rider's previous trip. This basic assumption is true when a traveler either 1) uses transit for consecutive trips, or 2) any intervening trips made by that transit rider take the rider back to the destination of his previous transit trip.

This assumption is not time dependent. For example, I can take the bus to work (trip #1) on Friday morning, and then take it back home in the afternoon (trip #2). My after-work trip's origin serves as my morning trip's destination. But when I take transit back to work on Monday morning (trip #3), my Monday morning trip origin still serves as a good destination estimator of my Friday afternoon trip home, even though it occurs two and a half days later, and several intervening weekend trips have taken place. While this process will create errors for specific trips, the project team believes that the basic process captures the majority of movements.

Consequently, the basic origin/destination processing follows specific ORCA CSNs through space and time. The basic rule is that if Trip number (N+1) for ORCA card X starts within one-quarter mile of the route last used by card X (Route Y) during trip N, then the destination of Trip N is the bus stop on Route Y that is closest to the boarding location for Trip N+1. (Note that the "one-quarter mile" rule can be changed. For a recent request by ST, a 1/3-

mile look up table was created and used in place of the ¼-mile table for all origin/destination and transfer procedures. The use of the 1/3-mile table added several hundred thousand extra trips to the origin/destination database. No significant research has been done yet comparing these different files.)

In order to determine the last route used (i.e., Route Y), the origin/destination process must incorporate transfers into Trip N. That is, if Trip N includes one or more transfers, then Route Y changes to a new route with each transfer boarding. The route boarded as a result of the last transfer boarding of Trip N is used when the quarter-mile look up is performed around the stop at which Trip N+1 starts in order to find the destination of Trip N.

If Trip N+1 starts at a location that is not close to Route Y, then we know that the transit rider used a non-transit mode of transport to move to a different part of the region before boarding transit for Trip N+1. We then state that we cannot estimate a destination for Trip N.

For Link and Sounder trips that do not involve a transfer to another vehicle after the rider exits the train, the "tap off" record indicates the destination (stop_id) for that trip. However, if the train rider transfers to a bus, the process described above must be used. If the train rider forgets to tap off and does not transfer, then the process used to identify bus destinations is used to find the destination of the train trip.

Finally, because the destination determination process requires looking at the next ORCA boarding, by definition the last record for each ORCA card has no destination. For our nine-week data sets, this meant that a destination cannot be identified for roughly 600,000 out of 20,000,000 boarding records (~3 percent) simply because there is no "next boarding" that can be used to estimate that location unless that last trip was a Sounder or Link train for which the rider tapped off.

The origin/destination process is described in more detail in Chapter 4.

Transfer Processing

As noted previously, transfer processing is performed in the same software step as the origin/destination process. Thus, in the transfer process, ORCA records are processed in date/time order, by ORCA CSN.

When an ORCA record (record N+1) indicates that a transfer occurred, the previous record (record N) describes the specific vehicle, route, and trip that was boarded before the transfer boarding was made.

The transfer detail process starts with the stop location where boarding N+1 occurred. The program then looks around that stop to determine the stop closest to where Route (N) passes. If that stop is farther away than one-quarter mile, the program assumes that some activity occurred between the first transit ride (N) and the new transit ride (N+1). No transfer details are computed for this case.

If the look-up finds one or more stops that are within one-quarter mile of boarding stop (N+1) serving Route (N), then the closest stop serving Route N to the new boarding stop (N+1) is assigned as the Exit_stop for boarding N. The AVL file for the service provider that operates Route (N) is then queried for the vehicle_ID on record N, while it is on Route (N), on the specific bus journey (trip of the day) being served when the boarding N occurred, to determine from the AVL file when that specific bus stopped at Exit_stop (N). Exit_stop (N) and the time vehicle_ID (N) stopped at Exit_stop (N) are then placed in the transfer details file, along with the time when the rider boarded the vehicle, as found on record (N+1). The latitude and longitude for Exit_stop (N) are also extracted and placed in the transfer details file, along with the route previously used.

This information allows the computation of the transfer duration, which is defined as the time difference between the new boarding and the time when the previously used vehicle stopped at the Exit_stop. A GIS query can also be run to determine the physical distance traveled between the Exit_stop and the new transfer boarding location. If a more precise distance is required, this GIS query can be made via a street or sidewalk network to determine actual walk distance, rather than the airline distance. (Note that the quarter-mile look up function uses airline distance, not network distance.)

If an error in the AVL file occurs, either for the current transfer boarding or during the look-up for the time when the previous transit vehicle reached the exit stop, an error code is placed in the appropriate transfer detail file variable.

Finally, if Link or Sounder is used as the transit vehicle transferred either to or from, then the off-board taps are used to define boarding time and exit time. So "exit time" is set equal to the "tap off" time stamp. This does create some additional error in the computation of actual

"transfer time duration" since the off-board tap occurs an unknown time after the passenger has left the train, and the time required to exit a station – and possibly tap off – can vary considerably from station to station, especially given the distance a rider must walk to exit the station at some underground stations (e.g., Beacon Hill, UW) versus the distance traveled at other at ground stations (e.g., Othello, Columbia City.)

If no "tap off" occurs, then an error is set in the transfer details file. Note that the use of off-board taps under-estimates transfer times both to and from the rail modes. This is because the boarding tap can occur before the train arrival. The exit tap occurs not only after the rider has left the vehicle but often after the rider has walked some distance and is in the process of exiting the station.

The transfer detail process is described in more detail in Chapter 4.

CAVEATS ON THE USE OF THE ANALYTICAL DATA SETS

Analysis of the ORCA data has shown that a number of artifacts in the data require data users to be careful. Further research and data processing are needed in the following areas.

Impact of Financial Transfers on the Transfer Details and Origin/Destination Files

One of the most important data artifacts is the fact that ORCA defines a "transfer" as any boarding that occurs within 2 hours of the initial payment for a trip. The 2-hour window was developed to allow for a long regional trip and still allow enough time for the rider to transfer to a bus on a route with a long headway. For example, the rider arrives and taps his/her ORCA pass at the Lakewood train station, 5 minutes before boarding the train to Seattle (a 70-minute trip). S/he wishes to transfer to the 522 to Woodinville. This requires a walk from King Street Station to 4th Ave S. and S. Jackson St, which takes 4 or 5 minutes. If s/he misses one 522 bus, s/he will have to wait up to 30 minutes until the next bus. In that case, s/he might not board the transfer bus to until as many as 110 minutes have passed since s/he tapped on in Lakewood. Similar scenarios are the reason for the 2-hour transfer window.

However, this same 2-hour transfer grace period can allow riders to make one or more short duration trips, all based on a single payment. In the ORCA records, three short trips can easily be marked as "transfers." For example, an individual who works in Pioneer Square could board a bus in the bus tunnel at lunch, get off at Westlake, eat lunch, get back on a bus in the

tunnel, and return to work, all within an hour. If s/he met a work colleague at the office after lunch, and then left for a business meeting via transit, s/he could easily have two transfer boardings (three total boardings) within a 2-hour period.

From a transit planning perspective, we would call this activity three trips. ORCA records this as one trip, with a transfer at Westlake, and then a second transfer near the office in Pioneer Square. The project team has adopted the term "financial transfer" for these transfers. They are a "transfer" only in terms of how they are being treated by ORCA's financial system. They are not transfers as defined by transit agency operational rules or as modeled in transportation planning analyses.

At this time, the origin/destination process has developed only a partial solution for these transfers. An analysis of transfer records shows that around 11 percent of all transfers take place to and from the same bus route. For example, the rider gets off a Route 40, and then some time later (but in less than 2 hours) boards a different Route 40 bus. If the rider boards the second Route 40 bus going in the same direction as the first Route 40, then the project team calls this a "pit stop" financial transfer. If the rider boards the second Route 40 bus going in the opposite direction of the first boarding, then it is defined as an "end point" transfer. We understand that these are "transfers" only in financial terms – not a practical travel necessity in order for a rider to reach their next activity – because the rider gains no travel benefit from getting off one bus only to get on a second bus for the exact same route. Thus, the fact that they re-board the same route is an excellent indicator that the traveler performed some activity (other than ride transit) in between the first and second boardings. Thus, from a transportation planning perspective, these need to be treated as two separate transit trips, not as one trip with a transfer in the middle of that trip.

Financial transfers (both pit stop and end point transfers) that involve just one route can be readily identified. The latest version of the data processing software, removes these transfers from the Transfer detail file, treats the Route re-boarding as a new trip, and follows the previously described process to identify the likely exit stop for the initial trip. The origin/destination file is then updated by adding that trip.

However, algorithms are still being developed and tested to detect other obvious financial transfers, when that transfer activity does not include same route – to – same route transfer activity. This means that the Transfer Details file includes a number of transfer times that also

include time spent at some activity (e.g., shopping, lunch, or a business meeting). At the same time, these short trips are not found in the current origin/destination file.

A third set of financial transfers can also be identified whenever a transfer requires a very long distance between the first and second routes. Unfortunately, the destination of the first trip cannot be identified for these financial transfers. The quarter-mile look-up process is just a first cut at determining a valid transfer distance. Work has been proposed (but not yet performed) that would examine the effect on the transfer process if transfer activities were allowed to occur one-third or one-half mile away. But a number of financial transfers occur between routes or modes that are separated by multiple miles. For example, in the 2015 data set, some individuals "transferred" from Washington State Ferries to Community Transit buses at the Ash Way park and ride. The closest WSF terminal is in Edmonds, 7.6 miles away. The Mukilteo Ferry is 8.8 miles away. It can be stated with confidence that any transfer between a WSF ferry and a bus at Ash Way included an additional mode of travel between the ferry and the bus. These transfers should not be included in an operational analysis of transfer activity at Ash Way. Similar long distance transfers occur periodically at other locations. For example, in 2016, 28 people transferred from the Sounder rail line to Link at the Tukwila International Boulevard Station. The closest Sounder Station is 2.7 miles way.

Impacts of the Origin/Destination File Assumptions

With the exception of single-mode rail trips during which entrance and exit ORCA taps occur, the ORCA system does not directly determine trip destinations. The assumptions used in this project to estimate the destinations of transit trips are essentially the same assumptions used by other researchers around the world when their electronic transit data records lack off-board taps.

The basic assumption in this process (that the origin of the second trip serves as a good estimate of the destination of the first trip) is reasonable, if imperfect. This assumption causes unknown error. Even when the basic assumption is correct (the rider got off in downtown), the actual exit stop could easily be close to the selected stop but not that specific stop. This is particularly true in dense areas where stops are closely spaced. In a downtown area, a rider might easily get off one stop earlier or later, simply to visit a preferred coffee shop in the morning, rather than the stop that is closest to his office.

The project team believes that the selected approach captures the general origin/destination pattern of trips well, but we have no corroborating data to confirm or refute that belief.

A good follow-on project would collect actual origin/destination stop information for specific ORCA cards to determine the accuracy of these assumptions. Until that follow-on work occurs, the current origin/destination files should be used with caution.

Errors in the Initial Stop Location Estimate

The first major task in using the ORCA data set is to geolocate the ORCA bus boardings. That task is central to the vast majority of the innovative ORCA analyses demonstrated by this project. For bus transit, the geolocation task uses the time when each ORCA tap occurs and the vehicle, route, and schedule information incorporated in that record. These data are then linked via those variables to the transit agency AVL files.

The first limitation in this task is that the time when an ORCA card is tapped does not directly correlate to any time reported by the bus AVL system. The ORCA time is recorded when the ORCA card is tapped on the reader. The bus AVL system records its time when the bus door opens. These times are different. If a large number of people board the bus, or if a wheelchair must be loaded or unloaded, then a significant time difference can accumulate between the time when the door opens and the time when an ORCA tap occurs.

This time difference is further complicated by the fact that the ORCA hardware clocks and AVL hardware clocks are not synchronized. Therefore, in addition to the differences between when the two systems report an activity, potential biases between the two systems can differ from bus to bus and from day to day.

This is resolved in the geolocation assignment process by picking stop locations from the AVL system that are closest to the time when the ORCA tap occurs. However, some bus stop assignments may be off by at least one stop along the route.

Data quality reviews of this process also revealed a number of errors in the original data files provided by the transit agencies that further degrade these computations for a limited number of transactions. For example, there are times when the ORCA and AVL records disagree on what work a specific bus is doing at a specific point in time (e.g., the ORCA record says that the bus is a Route 40 northbound at 8:05 AM, whereas the AVL file says that the bus is a Route

40 southbound at 8:05 AM). The geolocation software attempts to reconcile these data discrepancies whenever possible. However, it may sometimes choose the wrong solution.

When these software solutions are in error, that error cascades down the ORCA data processing chain. For example, for the above example, if the ORCA processing system picks the Route 40 northbound, but the rider was actually boarding the Route 40 southbound, and the next boarding is a transfer boarding, then the software may not find the correct exit stop, and it definitely will not select the correct exit stop time. This typically results in a transfer detail record that has irrational transfer time. It can also cause a failure to find a reasonable transfer exit stop, and thus no transfer details for that trip.

There are also times when the AVL system has errors, for example, a bus does not report its location for several hours or days. When this occurs, the software cannot assign a stop location to the ORCA boardings that occurred while the AVL system was malfunctioning.

All of these types of errors, while fairly infrequent, create errors in the ORCA data processing that become especially apparent in the transfer details and origin/destination results.

Consequently, care must be taken before the transfer and origin/destination records are used. Illogical records must be removed before these data sets are summarized.

Bias in the ORCA Data

Not all users of the transit system have ORCA cards. ORCA use varies among transit agencies and specific transit user population groups. For example, the 2016 Transit Integration Report to the Puget Sound Regional Council reported that whereas 84 percent of Community Transit trips are paid for with ORCA, only 64 percent of King County Metro bus trips were paid for with ORCA, and only 51 percent of Pierce Transit trips were paid for with ORCA.

Commuters who work for employers that subsidize transit use are heavy users of ORCA. Tourists are much less likely to use ORCA.

Consequently, it is important that when planners use ORCA data to develop transit service improvements, they understand that the analyses have inherent biases. The relative size of those biases can be computed, but estimating transit use that does not involve ORCA is more difficult.

This project has developed an initial mechanism for measuring the differences between ORCA and cash payment trips. It is a table that compares – when and only when APC data have

been collected by the transit agency – APC boardings against boardings recorded by ORCA. This "bias measurement" file is included in the project database. However, additional work is needed to operationalize the understanding of those differences. Some of this work is planned for summer 2017 as part of the digital Sciences for Social Good program.

DEALING WITH DATA PRIVACY

This project deals with data that contain private information. While the project team has no access to information that directly identifies the name of any individual, and the identifiers in the records have been hashed for further privacy, research has shown that records that indicate travel by time and location can be easily fused with other data sets to re-identify individuals.

As a result, the project team set up, and had approved by the UW Institutional Review Board (IRB), a specific data handling protocol. That protocol placed the data in a highly secure data storage location and restricted access to those data to only individuals who had both signed a non-disclosure agreement and been trained in the allowable uses of the data. The IRB agreement limits the analyses that can be done. No data that contain individually identifiable trips can be released to the public. No analyses performed will report on the activities of employees of a specific employer.

One key outcome of the data use constraints is that all analytical requests must currently be passed through the project team. This constrains the use of the existing data sets to the UW team members.

The project team is currently working with another UW research team to develop a data privacy and use policy for Sound Transit and the ORCA Joint Board. The intent is to help the ORCA agencies develop, adopt, and deploy a data use policy that that will allow the transit agencies to gain the type of insight and benefits from ORCA that are demonstrated in this report while also ensuring that the privacy of all individuals is maintained. The combination of an adopted data use policy and the data processing procedures described in this document will allow the transit agencies to build a robust system that securely performs the types of analyses illustrated in this report.

However, until that additional work has been conducted, the continued performance of the types of analyses discussed in this report is constrained to the resources associated with the UW project team.

CHAPTER 3: EXAMPLE ORCA ANALYSES

This chapter presents some of the example analyses performed as part of this project. They were designed to illustrate how ORCA data can be used to better inform transportation agency decision making in a variety of areas. The following analyses were performed:

- geographic location of bus boardings
- potential for bias in ORCA analyses
- route use profiles
- transfer activity
- differences in travel by type of ORCA product
- analysis of commute trip reduction program effectiveness
- miscellaneous other analyses.

Each of these types of analyses is described below. What is particularly powerful about the use of ORCA data is that each of these types of analyses can be performed by type of transit rider. That is, all of these analyses can be performed not only for all riders who use ORCA, but special groups of riders, such as low income riders, or only riders who use senior discounts. This will allow agencies to examine the use of transit by type of rider.

GEOGRAPHIC LOCATIONS OF BUS BOARDINGS

The first set of analyses simply describes where ORCA transit boarding activities occurred. Figure 3-1 is an example map of where ORCA boarding payments were made in 2015. By selecting different transit agencies, modes, or routes, it is possible to quickly illustrate where and when people who use ORCA boarded vehicles. These maps can be drawn quickly by simply loading the geographically located ORCA boardings into a GIS. In addition, it is possible to use simple database tools to assign day of week and time of day (e.g., AM peak, PM peak) variables to the boarding records, which allows quick illustration of geographic ridership patterns by time and date. Finally, as will be illustrated later in this chapter, it is possible to examine geographic use patterns—and changes in those use patterns—by important rider sub-groups, such as low income riders.

Figure 3-1: Map of 2015 ORCA King County Metro Transit Boardings

To a certain extent, this capability is a duplication of the ridership data that can be obtained from the automatic passenger counters (APCs) used on some buses. However, many transit agencies do not equip all of their buses with passenger counters. For example, only about 30 percent of King County Metro's bus fleet is equipped with APCs. On the other hand, electronic fare collection data are available on all vehicles and therefore provide an better source of transit ridership information across time periods, routes, and vehicle trips.

Notable for analysis, however, is that not all transit payments are made with ORCA. Therefore, the ORCA data produce their own biases in the ridership statistics produced with ORCA, as does the use of APC data. Consequently, this project demonstrated tools to help data users understand where that bias occurs and whether its impact.

POTENTIAL FOR BIAS IN ORCA ANALYSES

One of the biggest concerns when ORCA data are used is that not all riders use electronic fare cards. There may be significant differences between how users of electronic fare media use the system and how people who pay with cash use the system. Consequently, if system usage

analysis is performed with electronic fare card data, it is possible that significant travel movements will be under-represented.

This task within the project determined how to inform ORCA data users where there are differences between ORCA measured ridership and APC measured ridership so that the biases between the two data sets can be taken into account. As part of this task, it must be acknowledged that geolocation of the ORCA data is not perfect. Because ORCA does not directly include geolocation information in the transaction record, the location of most bus boardings must be computed by linking the ORCA transaction time with the location of the bus being boarded as that location is reported by the transit agencies' fleet AVL systems. Reducing the precision of this task are the fact that the clocks used by the ORCA and AVL systems are not synchronized and the fact that the AVL and APC systems contain a modest number of errors caused by various equipment failures. As a result, there can be errors in the ORCA geolocation estimates, and there can be errors in the current APC-based estimates.

To determine the bias in ORCA data, we first assume that the APC counts are correct measures of actual passenger boarding volumes when those data are present. The developed software then compares recorded ORCA boarding volumes with APC boarding volumes and reports the differences.

First, ORCA data were aggregated to provide a count of the number of ORCA payments made at each bus stop for each bus journey[6] for each day. This gives an ORCA estimate of the number of passengers boarding at that stop on that bus and paying with ORCA. This number is then compared with the APC boarding statistic for that same bus journey. The first complication is that not all buses are equipped with APC equipment, and therefore not all ORCA boarding counts can be compared to APC data. These comparisons can only be performed when APC data are available in the AVL data file.

When both APC and ORCA data are present, a simple subtraction tells the degree to which ORCA under-counts APC-measured ridership at a given stop for a given bus journey. ORCA should be a subset of APC counts. (That is, the APC should count everyone boarding the bus, and ORCA taps should be a subset of those boarding activities.) However, as a result of errors in the APC counting process and the ORCA geolocation process, it is possible for ORCA to register more passengers than the APC system. For example, discussions with King County

---

[6] A bus journey is defined as a single, directional trip on a defined route on a specific day.

Metro staff about differences in APC and ORCA counts have determined that a significant number of stop locations are incorrectly mapped within the KCM APC GIS system. This causes APC boarding data to be incorrectly assigned to other stops, since the X/Y being reported for some boardings is not associated with a specific bus stop ID..

The data file that results from the comparison of totaled ORCA and APC data can be aggregated to show which stops, routes, or geographic areas have ridership that is well represented by ORCA data, and which stops, routes, and geographic areas have relatively low levels of ORCA use.

One problem with these summaries is that APC data availability is not a random sample of routes or bus journeys. Some bus routes are more heavily sampled with APCs than others. This can create bias in the aggregated results. At this time, the project team has not attempted to address this issue. Instead, the project team has created a series of summaries as an informational output about ORCA.

There are multiple ways to examine the differences between ORCA boarding estimates and APC-based estimates. Figure 3-2 presents a simple map display that indicates, at the bus stop level, where ORCA data differ from APC counts. These summary statistics can be given as absolute values, or as percentage values. When absolute values are used, high-use stops tend to be highlighted as places where large errors occur. When percentage values are used, low volume transit stops (where small absolute errors can show up as large percentage errors) often stand out as being the locations of large errors.
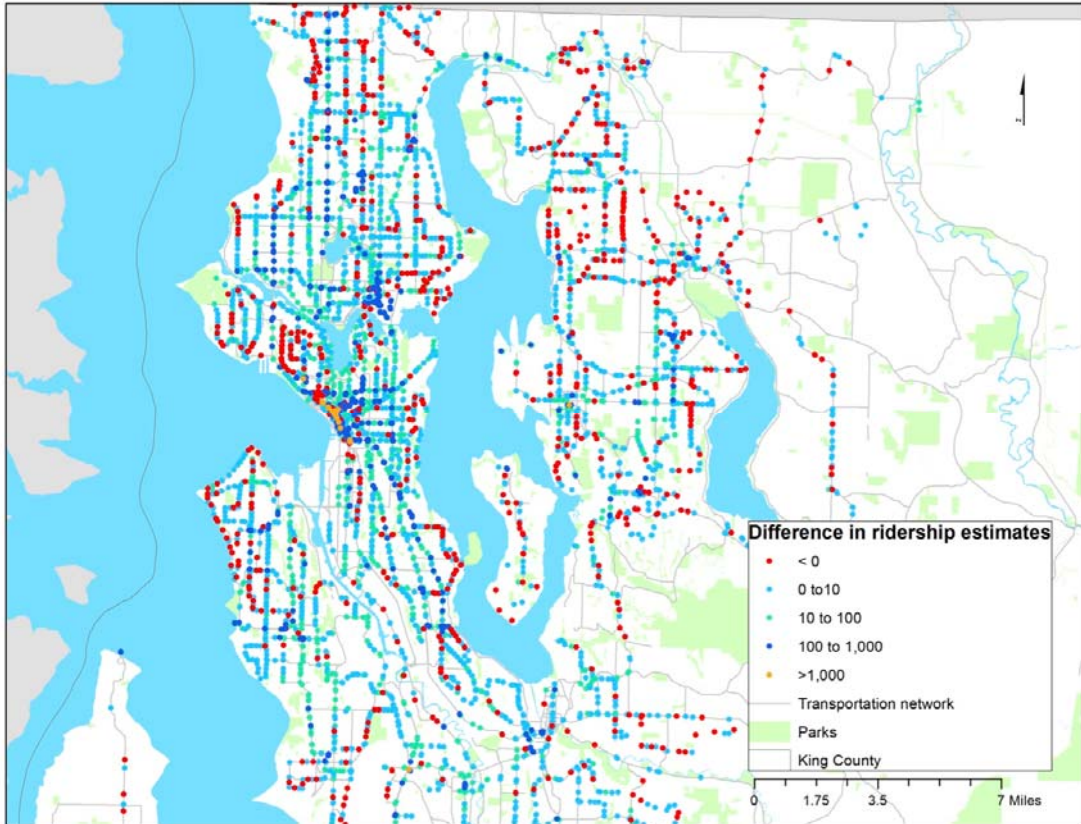
Figure 3-2: Map of Differences in ORCA and Passenger Count Boardings

Because we are interested in not just whether a specific stop or route has a low (or high) ORCA utilization rate, the project team also developed a technique to illustrate the ORCA/APC bias by geographic area. Figure 3-3 illustrates this concept. In the figure, percentage bias differences between aggregated ORCA and APC counts are shown by zone. The darker the coloring of the zone, the greater the difference in measured boardings between the APC and ORCA data sets. For example, a dark blue zone indicates where relatively low levels of ORCA use occur. In this example application, census data could then be extracted for each zone to help a user determine the socio-ethnic make-up of that zone or to extract other census details such as car ownership.
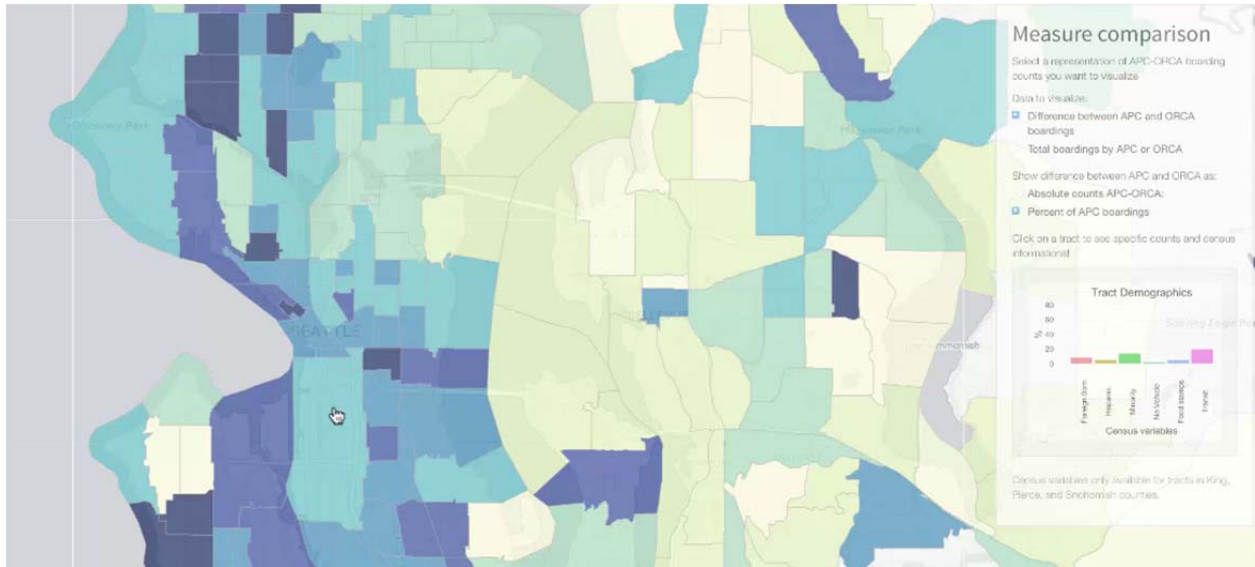
Figure 3-3: Map of Differences in ORCA and Automated Passenger Count Boardings, by Geographic Zone

This visualization has also helped the project team understand failures in the data processing system developed and applied. For example, the Figure 3-3 graphic is from an early version of the software processing and illustrates that the early versions of the software did not correctly compare off-board paying BRT riders with the APC counts.

To produce Figure 3-3, the project team also developed mechanisms for converting stop-based ORCA statistics with geography-based zone systems. Geographic zones are the most common aggregation level used in most travel modeling. They are used for origin/destination matrices, to store all kinds of aggregate census statistics, and for many other functions.

Unfortunately, the geographic data aggregation process is not as simple as it might seem. The problem is that ORCA data are geolocated to transit (e.g., bus) stops, and those bus stops are the geolocations that serve to locate transit boardings, transfer locations, bias estimates, and origin/destination tables. Bus stops are physically located on the side of the road, so that stops in opposite directions are located on opposite sides of the road, and the center of the road is typically used as the border between geographic zones. This creates some undesirable artifacts of a simple GIS-informed aggregation, which are illustrated in Figure 3-4.
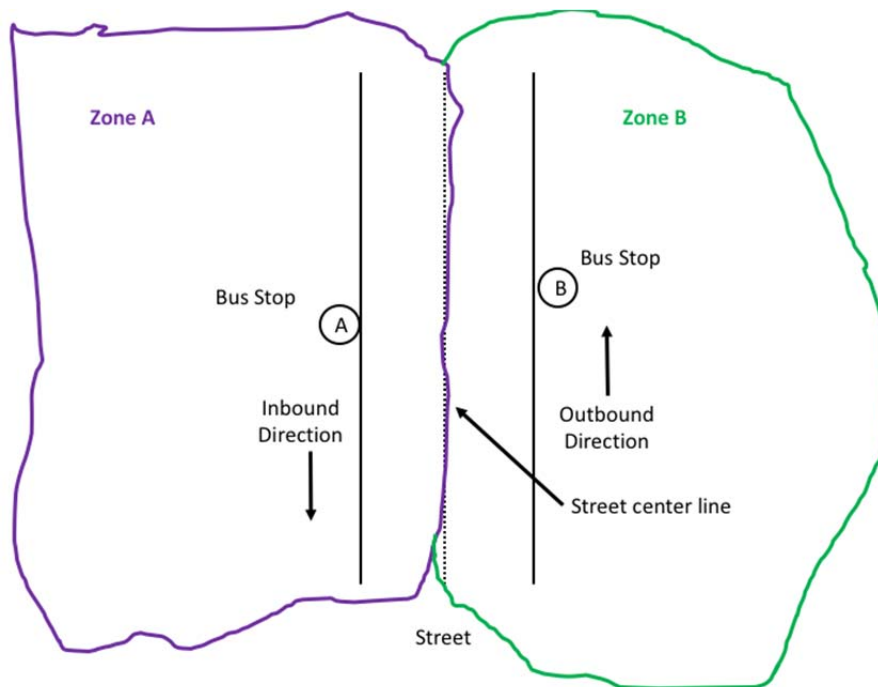
Figure 3-4: Example of Problems in Assigning Bus Stops to Zones

Figure 3-4 shows that logical pairs of bus stops are often split into two different zones, simply because the GIS boundary for zones is typically the centerline of arterials. When this occurs, the inbound and outbound bus stops at a given location are assigned to different zones if a simple GIS query is used to assign each stop to the zone in which it is physically located. In the case illustrated in Figure 3-4, if a person got on the bus at Stop A in the morning and exited the bus at Stop B in the evening, the origin of his morning commute trip would be assigned to Zone A, but the destination of his evening commute trip would be assigned to Zone B. This would create significant errors in the origin/destination tables generated from the ORCA data, especially if those tables were used to calibrate or validate the forecasting models used by the Puget Sound Regional Council (PSRC) or the transit agencies.

To account for this problem, the project team created rasters[7] in the GIS. The bias measures for all of the stops within each raster were then aggregated. The raster summaries were then allocated to each of the zones in which they fell. So a raster square that fell entirely within a census tract would be fully allocated to that tract. A raster that was split between two census tracts would be allocated between those two zones on the basis of the relative amount of the

---

[7] A raster is a uniform square shape overlaying the GIS map.

raster in each zone. The "relative amount" of the raster can be computed on the basis of geographic area, population, employment, or some combination of those and other census variables.

This project used a simple version of this approach, based on population and employment taken from census block-level information. Additional research and analysis should be performed with different aggregation metrics before the ORCA data are directly used.

ROUTE USE PROFILES

One simple use of ORCA data is the ability to examine where and when people board buses. Because ORCA tracks different types of passengers (regular fare adults, seniors, youth, disabled, and low income), it is possible to examine how the travel patterns differ for these types of riders.

Figure 3-5 illustrates one way to look at how people use ORCA. This summarizes boardings by time of day for a specific route. This small application, built by summer 2016 Digital Science for Social Good students, allows the user to select specific providers, routes, days of the week, travel directions, and types of users. Boardings on each route are displayed by time of day. Figure 3-6 illustrates the different ridership patterns on one route for different categories of riders. For example, it is possible to see a significant afternoon spike in youth boardings early in the afternoon when school lets out, but before the start of the afternoon commute period. Interestingly, on this specific route (the 271 serves the University of Washington as well as parts of Bellevue), no morning peak is observed.

Similar graphics can be used to examine differences in boarding patterns at specific stops as well as for individual routes. It is also possible to use these graphics to watch the growth in low income (Lift) cards over time. By selecting consecutive weeks in the 2015 data set as the date range in this graphic, and selecting only low income riders, it is possible to observe the growth in ORCA Lift usage as the program rolled out in the Spring of 2015.
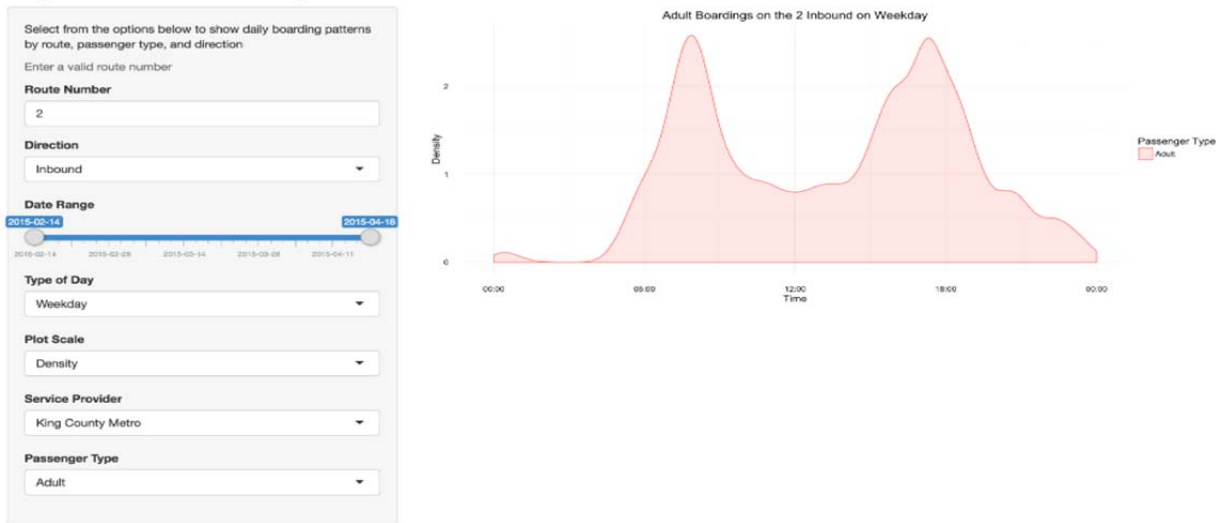
Figure 3-5: Example of a Route Profile
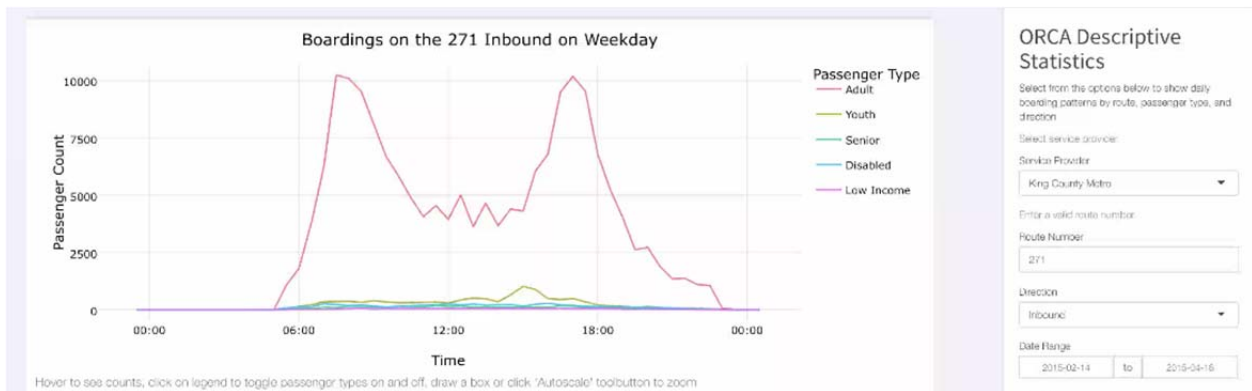
Figure 3-6: A Route Profile Showing Different Boarding Patterns by Passenger Type

TRANSFER ACTIVITY

Many of the most immediately useful analyses that were conducted with the ORCA data involved an analysis of transfer activity. The only difficulty, as noted in Chapter 2, is that ORCA measures <u>financial</u> transfers, which are not always direct transfers between transit vehicles. That is, in the ORCA system a "transfer" is defined as any boarding that occurs within 2 hours of a fare payment. This allows transfers between poorly coordinated services and also allows long-line haul trips to occur before the transfer occurs. (For example, a traveler boards a bus in Pierce County and rides that bus to Seattle on a congested I-5, and then needs to wait in downtown

Seattle for 20 minutes to transfer to the bus that will take her to work, having missed her normal transfer.)

However, that same 2-hour rule can cause short trips to be marked as "transfers" in the ORCA records. For example, a traveler gets on a bus at the south end of downtown Seattle and rides the bus to the Pike Place Market over lunch. S/he shops at the market and then returns on a southbound bus to his/her starting location.

Whereas in transit planning terms that activity is considered two trips, ORCA's record keeping considers them to be one trip with a transfer.

Roughly 5,900,000 of the approximately 21,000,000 ORCA transaction records for 2015 were marked as transfers (28.2 percent). Of those transfers, riders re-boarded the same bus route that they originally boarded 637,000 times. Therefore, at a minimum, just under 11 percent of boardings marked by ORCA as transfers were not actually operational transfers but rather short trips that were only transfers from a financial perspective. Additional "financial transfers" will occur between different routes. A technique to automatically identify these short trips has not yet been perfected.

In the 2016 data, transferring to and from the same route grew slightly to just over 11 percent of all ORCA boardings marked as transfers.

Analysis of transfer activity is highly beneficial to the transit agencies. For example, transit agencies expect that many riders using Link light rail will arrive at the Link station via other transit modes. As more Link stations open, ORCA data will allow agencies to monitor the effectiveness of different transit feeder services they create.

To demonstrate this functionality, several different transfer analyses were performed. Table 3-1 shows a comparison of the number of ORCA Link trips that transferred to Link from other modes in 2015 and 2016.

More detailed analysis of transfers at different stations revealed important details of those transfer activities. Table 3-2 examines transfers at the International District Station for nine weeks in both 2015 and 2016. The very large movement from the Sounder Train to Link, which takes Sounder passengers to the downtown core, is very apparent. Also apparent is the growth in both total ridership caused in part by the opening of the Capitol Hill and UW stadium stations, as well as the growth in streetcar use as a result of the new Broadway streetcar line.

Table 3-1: Percentages of Link Trips That Involved Transfers from Other Transit Modes

Link Station Name	Transfer Percentage (2016)	Transfer Percentage (2015)	Change 2016-2015
UW	39.2%	N/A	
Capitol Hill	19.8%	N/A	
Westlake	31.9%	38.0%	-6.1%
University Street	27.4%	22.5%	4.9%
Pioneer Square	20.7%	19.1%	1.6%
International District	49.6%	47.8%	1.8%
Stadium	12.4%	17.4%	-4.9%
SODO	39.4%	35.7%	3.7%
Beacon Hill	36.3%	38.2%	-1.8%
Mount Baker	34.3%	35.9%	-1.6%
Columbia City	14.5%	15.7%	-1.2%
Othello	18.2%	18.0%	0.2%
Rainier Beach	36.2%	37.7%	-1.6%
Tukwila Intr. Blvd.	29.1%	28.0%	1.1%
Sea-Tac Airport	27.4%	29.4%	-2.0%

Table 3-2: Transfer Details at International District Station

Mode Transferring From	Percentage 2016	Number of Trips 2016	Number of Trips 2015
Ferry	0.0%	1	1
Sounder	21.2%	45,672	18,823
Regular Bus	20.1%	43,376	23,317
Streetcar	1.2%	2598	14
BRT	0.1%	193	80
Link	6.9%	14,767	5,033
Demand Responsive	0.2%	410	118
No Transfer	50.4%	108,646	51,694

These same types of transfer analyses can be done at major bus transfer points or between specific transit routes. For example, an analysis was performed of transfer activity at the Ash Way park and ride. It found that 37.5 percent of ORCA users boarding buses at Ash Way were transferring to buses from other buses. However, if those numbers are disaggregated by time of day, it can be seen that transfer activity changed dramatically throughout the day (see Table 3-3).

Table 3-3: Transfer Details by Time of Day at Ash Way Park and Ride

Time of Day	Percentage of Transit Boarding Activity Coming from Another Bus
Peak (6 - 8 AM	16.6%
Midday (8 AM – 4 PM	50.9%
PM Peak (4 PM – 7 PM	82.5%
Night (7 PM – 6 AM	19.0%

Table 3-3 also illustrates the importance of understanding that ORCA data describe the transit boarding activity, not total transit activity. This is shown in the very large percentage of transfers that occurred in the evening peak period. Actually, the vast majority of people who use Ash Way in the PM peak period arrive on buses, disembark from those buses, get into their cars, and leave the park and ride. So they do not show up in the ORCA data. The PM statistic in Table 3-3 shows that the majority of people boarding buses in the evening commute period came from other buses. That is different than the total number of people using the park and ride.

Another important transfer analysis examines which routes actually interact with each other at specific locations such as Ash Way. This tells transit agencies which route schedules must be coordinated in order to facilitate these movements. Table 3-4 below shows a small excerpt of the route-to-route cross-tabulation that describes the transfer activity at Ash Way.

Table 3-4: Route to Route Transfer Details at the Ash Way Park and Ride

Previous Route	Route Being Boarded								
	112	115	116	119	201	511	512	532	810
2	0	0	1	0	0	0	1	0	0
5	0	0	0	0	0	0	2	0	0
101	0	2	1	1	2	1	5	0	1
105	1	0	0	0	0	0	2	0	0
106	0	0	0	0	0	1	0	0	0
112	8	114	96	149	31	63	384	180	77
113	0	3	1	0	0	0	6	3	0
115	151	51	67	265	381	240	958	564	20
116	67	90	41	351	605	236	1107	727	79
119	102	459	248	36	54	202	788	469	67
201	204	489	897	115	27	200	395	200	3
202	155	674	520	124	26	101	383	299	4
410	14	64	46	18	10	0	3	13	0
413	6	213	178	83	143	18	20	130	22

In Table 3-4, columns are the routes operating at Ash Way. Each row of the table then describes the route from which an ORCA card user transfers from. These cross-tabulations also

illustrate the occurrence of "financial transfers[8]" within the data set. They are very apparent in the first two rows. Routes 2 and 5 do not actually serve Ash Way. Therefore, these boardings are marked as transfers for financial purposes, not direct transfers from one bus to another. These riders have conducted some other activity in between riding the Route 2 or Route 5, then traveled to Ash Way, and then boarded a bus.

Cross-tabulations like this table can be easily developed by stop and by time of day in order to determine where and when passengers transfer between routes. High levels of interaction between routes would be good indicators that the schedules for those routes should be coordinated to make those transfers as seamless as possible.

The effectiveness of those transfers can then be examined by using the transfer details data set. The transfer details data set allows the computation of transfer travel time statistics between routes at a stop. That query can be restricted to specific route pairings and stops (e.g., from the route 532 to route 271 at the Bellevue Transit Center), days of the week, or time periods. Such analyses will allow the transit agencies to not only determine which routes are heavily used by riders to transfer between routes, but also how efficiently riders can transfer between those routes, by time of day or day of week.

Whether specific stops or route-to-route interactions are analyzed, there are multiple ways to look at those interactions. Figure 3-7 shows a frequency histogram diagram of transfer times between routes. This figure covers all times and days of the week. It also covers the interactions among all routes. Because this particular stop is downtown, it illustrates the long tail of transfer travel times that are most likely associated with financial transfers.

---

[8] See the initial discussion at the top of this section.
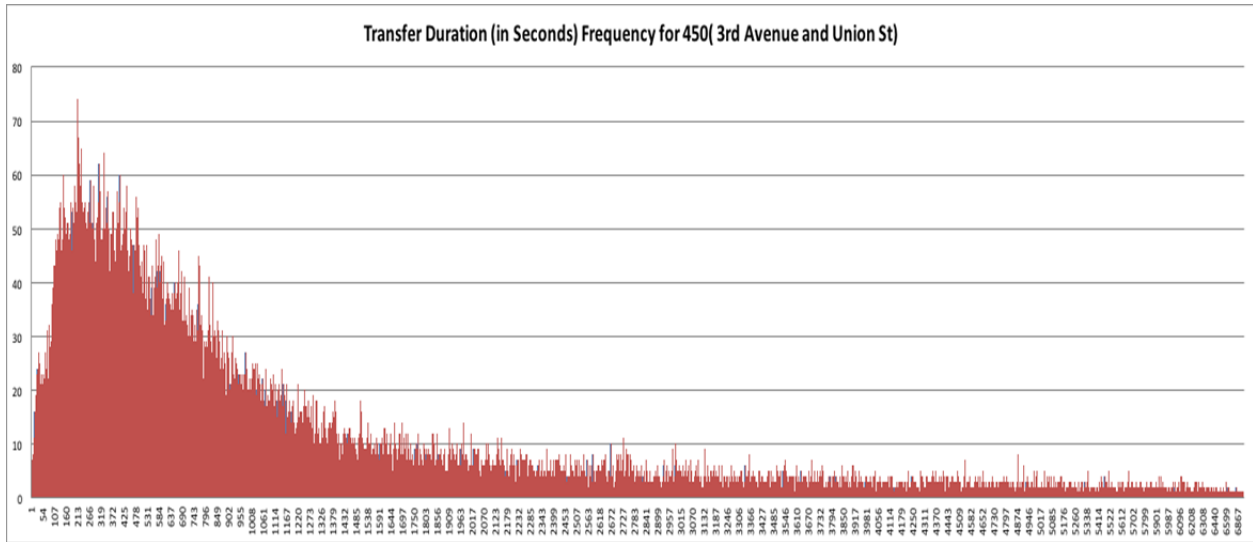
Figure 3-7: Example of Transfer Time Distributions at Different Locations

Additional work is needed to examine these statistics, but the potential to learn more about transfer activity is significant.

DIFFERENCES IN TRAVEL BY TYPE OF ORCA PRODUCT

Another set of powerful and informational analyses look at how different groups of people use their ORCA cards. There are two important categorizations of people within ORCA. The first is the type of passenger. ORCA tracks five different types of cards:

- Adults (the ORCA default)
- Youth (mostly student passes)
- Senior citizens
- Disabled
- Low income (Lift).

ORCA also tracks a variety of product types. That is, an ORCA can be used as an electronic purse, paying for each trip on a cash basis, or it can also be used as one of many different types of monthly passes. These passes have specific value, ranging from 50 cents to the cost of any transit trip. They can allow use of specific transit vehicles, such as the state ferry system, or a specific transit system, or they can allow travel across some combination of transit services. One of the most common pass types is the Business Passport, which allows the use of all transit

services in the region (although some business passports allow use of ferries, while others do not).

There are 46 defined ORCA product codes. The e-purse is one code. The majority of the other codes are some type of monthly pass along with a small number of day-passes and multi-trip passes (e.g., WSF sells an ORCA product that is good for 10-rides by a passenger on Washington State Ferries. The availability of these pass type indicators allow examination of how riders choose to pay for rides, by type of passenger, by geographic location, and by time of day.

The ORCA data set also connects employer subsidy information to the ORCA CSNs. This allows analysis of the effects of employer commute trip reduction programs on trip making (see below) because a large part of travel behavior is based on the type of pass purchased or used.

An example of the analyses that are possible with ORCA product types is illustrated in Figure 3-8. This figure shows a frequency distribution that illustrates how often each ORCA card (CSN) was used in the 2015 nine-week data set. The Y-axis gives the fraction of all cards that were used to board a transit vehicle (not including transfers) during the nine-week period. (Note that the Y-axis is not uniform across the graphs that illustrate the different major pass types.)
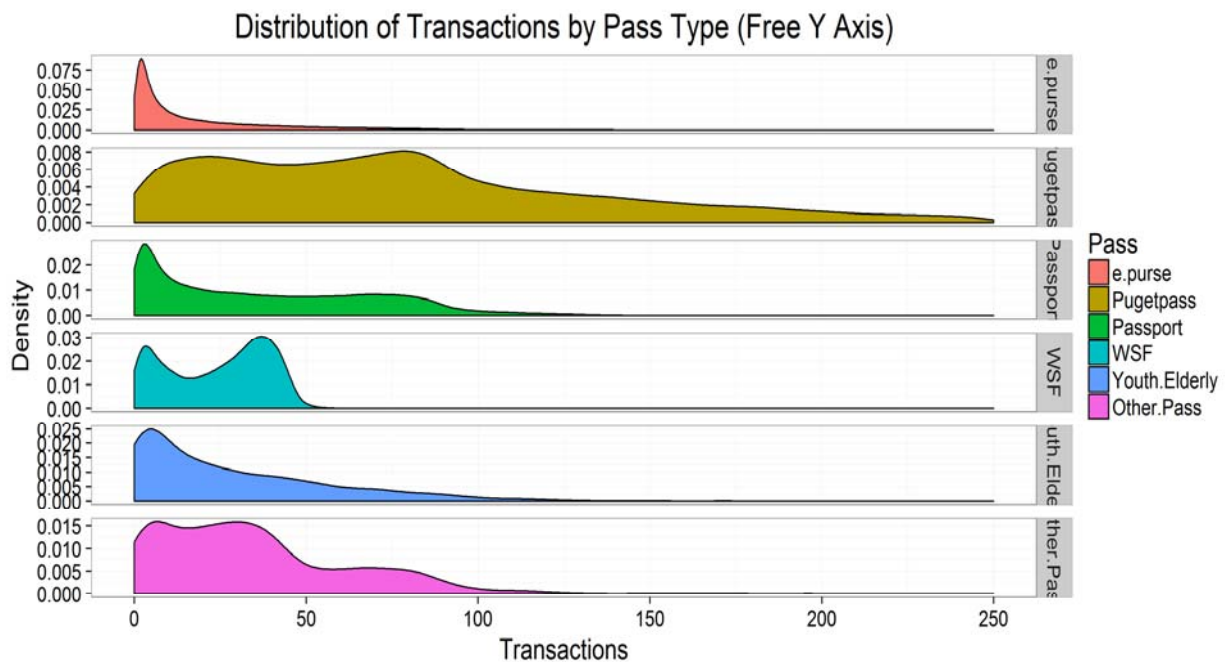
Figure 3-8: Frequency of ORCA Use by Type of ORCA Product

Figure 3-8 shows that holders of e-purse ORCA passes tend to use transit fairly infrequently. The frequency distribution of e-purse users peaks around three trips. At the other extreme is the PugetPass category, whose utilization peaks around 90 transactions each month. This would represent slightly more than one commute trip (one trip inbound and a second trip outbound) every weekday during the nine-week period. (Note that the PugetPass category includes a variety of pass values. A monthly Puget Pass can be obtained with a per trip value of as little as $0.50 or as much as a $5.50.) This pass also has a large contingent of users that use their pass more than 100 times each month. A large number of the PugetPasses are purchased directly by riders and are not subsidized by employers. Anyone choosing to buy a monthly pass, would choose to pay for that pass only if s/he were going to use it enough to pay for the value of the monthly pass.

Conversely, the Business Passport is the most common employer provided and subsidized pass. Many of these passes are used less often than twice every business day. In fact, the peak utilization is found around 4 to 5 trips per month. But unlike the e-Purse, a large fraction of these passes is used more than 50 times each month. The cost to the employee is still low (the pass is sometimes free) relative to the cost of paying cash for the transit trips they make. In addition, many company-subsidized passes are paid for by automatic payroll deduction. This means that an employee's pass is automatically renewed each month, and so some months may have less use than a direct economic analysis might predict. Thus unlike the self-purchased PugetPass, because the Business Passport is often heavily subsidized and purchased annually, it still makes economic sense for an employee to have and use such a pass, even if they use the pass infrequently.

The project team has done relatively little work to date on analysis by passenger type, although this type of analysis also has the ability to provide considerable insight into transit use. In the 2015 data set, the ORCA Lift (low income subsidy) program was just starting. Consequently, only 278,000 trips were taken by riders using Lift. In 2016 that number grew to 1,190,000 trips during the nine-week study period. In 2015, the growth in the Lift program's use was actually visible in the data set, as Lift use grew from week to week. In fact, the low income program appears to be at least partly responsible for a large portion of the increase in ORCA use between the 2015 and 2016 data sets. Lift was responsible for 48 percent of the increase, and

Adult riders were responsible for 58 percent of the increase. Youth and disabled pass use actually declined from 2015 to 2016. (See Table 3-5.) Note that because the project team has no data on what individuals use which ORCA cards, it is not possible for the project team to determine if new ORCA Lift users previously used other types of ORCA passes, in particular discounted passes for seniors or those who have disabilities.

Table 3-5: Changes in Boardings by Passenger Type 2015 to 2016

Type of Passenger	2016 Boardings	2015 Boardings	Difference	Percentage Difference
Adult	18,168,100	16,937,400	1,230,715	58.3%
Youth	1,253,500	1,276,600	(23,016)	-1.1%
Senior	887,700	811,100	76,582	3.6%
Disabled	1,565,300	1,648,100	(82,780)	-3.9%
Low Income	1,187,500	278,200	909,388	43.1%
Total	23,062,100	20,951,400	2,110,889	

Additional analysis looking into the use of both product type variables and passenger type variables must be performed. In particular, the simple analyses shown above do not account for the fact that many ORCA passes carry both a pass and an e-purse value. The e-purse value is used to pay for a portion of a trip when the pass value does not cover the entire trip. This occurs commonly on the Sounder train. For Sounder, a Puget Pass may cover the majority, but not all, of the trip cost. Frequent Sounder riders that have these types of passes then routinely also place cash value on their ORCA card so that a single exit tap will pay for the trip.

ANALYSIS OF COMMUTE TRIP REDUCTION PROGRAM EFFECTIVENESS

One of the specific interests of this project was to examine the impacts of employer commute trip reduction programs on travel behavior. That analysis has not yet been completed, but preliminary results are available. The complexity of the ORCA data sets meant that this effort took longer than anticipated. That work is still continuing as part of two student efforts and will be delivered later in 2017. Preliminary results from the project are discussed below.

It is not surprising that employee transit use is a function of many factors, including the quality of transit service between the employee's residence and the work site, the size of the employer transit subsidy, the cost of parking to the employee, other commute alternatives and subsidies provided by the employer, and the built environment characteristics of both the origin and destination ends of the commute trip.

Of perhaps most significance is the availability of free parking at the work site. Using CTR survey data that describe whether a company's work sites have free parking, paid parking, or a combination of the two (at some worksites, some employees have free parking, while others do not, and for some companies, some worksites have free parking, while other work sites do not) an analysis of the percentage of employees that used their ORCA card for transit at all was performed across these three parking categories. The analysis showed that providing employees with free parking was an extremely powerful variable describing whether employees took transit during the nine-week study period. (See Table 3-6) Table 3-6 shows that the presence of free parking dramatically reduces the percentage of employees that use ORCA at all.

Table 3-6: Impact of Parking Price on Employee Use of Transit

	Percentage of Employees Who Use the ORCA Card	Standard Deviation of Percentage of Employees	Percentage of CTR Affected Employees Who Use the ORCA Card	Standard Deviation of Percentage of CTR Employees
Pay Parking	55%	31%	80%	102%
Partial Pay Parking	46%	35%	56%	39%
Free Parking	13%	16%	17%	20%

Interestingly, those employees at companies with free parking that do use transit - even occasionally - to commute to work, take transit more often – on average - than employees who have to pay for parking, but use transit at all. It appears that employees that chose to use transit, despite having free parking are more dedicated transit riders.

Given the effect of free parking at reducing overall transit use, analyses of travel behavior associated with different pass types were undertaken. These showed that users who had a PugetPass used their ORCA card almost 3 times as much as users who had an e-purse. ORCA cards with a Business Passport were used just less often than ORCA PugetPasses, but were used almost twice as often as e-purse cards associated with businesses.

Where parking is not free, many individuals take transit, but they appear do so reluctantly, likely preferring to drive to work but choosing transit for financial reasons. They therefore periodically take other modes – or telecommute - and therefore use transit slightly less often. This leads to lower transit utilization rates per user, <u>for those people who take transit at all</u>, given free parking. (That is, a person who takes transit to work, despite having access to free parking at that site, tends to use transit more often than a person who has to pay for parking but gets a

subsidized transit pass from his employer.) Additional research and analysis is needed to confirm this interpretation of these preliminary results.

Preliminary analysis of other CTR values showed the following results:

• The impact of subsidy value on pass use showed a statistically significant increase in transit use as the transit subsidy value increased. A $10 increase in subsidy resulted in a 3 percent increase in transit use.
• The availability of telecommuting programs and subsidies for other alternative modes reduce transit use.
• The availability of flex time increases transit use.

Analysis of other contributing factors on transit use indicate that the quality of bus service – measured in terms of the number and frequency of buses at stops close to the worksite – is one of the largest predictors of the amount of transit use occurring to a work site. That is, poor transit service will discourage transit use more significantly than an increasing subsidy will stimulate its use.

The complexity of the interactions of the various built environment, transit service, and commute trip subsidy interactions have produced difficulty in the project team's ability to extract reliable statistics from the data. Work in this area continues and will be submitted to WSDOT when it has been completed.

MISCELLANEOUS OTHER ANALYSES

A large number of other analyses have been requested by various members of the Technical Advisory Committee for this project, city and transit agency staff, and even members of the public. Some of these analyses have been performed; others are simply not possible given the limited resources available to the project. Each of the sections below describes specific analyses that have been requested and can be examined by using the ORCA data.

Origin/Destination File

One of the desired analyses is to compare the origin/destination table obtained from the ORCA data with the transit origin/destination tables that PSRC and the various transit agencies use. The ORCA table serves as an excellent validation tool for the regional transit modeling effort.

Under a separate funding agreement with Sound Transit and their contractor (WSP/Parsons Brinckerhoff), an origin/destination table based on a 1/3-mile look-up table was

produced, quality assurance checked, and given to ST so that it could be used in the development of their latest update to their ridership forecasting model.

Percentage of KCM Trips That Cross Zone Boundaries

King County Metro requested data early in the study that would summarize the origin/destination patterns for trips crossing zone boundaries. Crossing zone boundaries affects the cost of a trip. It was a simple matter to overlay the KCM zone boundaries and summarize the number of trips occurring entirely within one fare zone and the number that used more than two fare zones.

Ability to Model Revenue Given Origin/Destination and Transfer Activity

Also early in the study both Sound Transit and King County Metro staff asked whether the ORCA data could be used to model new fare structures. The answer to that question is "yes," with the exception that the ORCA data only provide a good measure of current ridership patterns. They do not describe the changes to that behavior given changes in fare structure.

The results of the fare analysis might also be affected by the bias inherent in the ORCA data. That is, if the people not using ORCA travel in significantly differently patterns than the people who pay with ORCA, that bias will carry over to the revenue estimated by the analysis.

However, the ORCA data sets, if used carefully, and with specific attention to where significant bias does and does not exist, would be very useful for informing a study about how new fare structures might affect revenue. The origin/destination data set would provide an excellent view of the lengths of trips occurring and the transfers taking place as part of those trips. The Transactions file describes the interaction between different agencies (where and when they occur), which is an important part of any new fare structure. The bias calculations by route and stop, combined with total boardings by stop and route, would also be valuable in understanding actual ridership patterns and thus the total impact of changes in fare structure on total revenue.

KCM Boarding Locations by Time of Day

Early in the study, King County Metro asked for a geolocated data set that indicated when and where ORCA taps on KCM-operated vehicles were occurring. The project team gave KCM a data table that listed date, time, X, and Y for all KCM boardings.

This data set was used as input to the agency's ongoing work to size upgrades to its fleet communications system. That is, the current ORCA boardings are a good starting point for estimating the communications load required, geographically and temporally, if real-time communications are needed to support Next Generation ORCA fare payments.

Transit Trip Length

Multiple agencies asked the project team to measure transit travel lengths. The origin/destination table allows the direct calculation of on-board travel time, but data quality issues related to the destination stop on many trips have made the research team reluctant to produce a number that describes the distribution of travel times.

However, the real interest of these requests has been the actual distance traveled. This request can be performed—subject to the same limitations about the quality of the destination estimation—but it requires a very significant GIS process to determine the network distance between transit stops. This work has not yet been done, as it is well beyond the scope of this project. However, it is possible to estimate average trip distances for origin/destination pairs of interest.

Similarly, it is possible to determine the number of transfers required to go from various origins to various destinations. For example, it would be possible to perform an analysis which compared the number of transfers required for trips from lower income neighborhoods (or neighborhoods with high minority populations, or high levels of Lift use), and compare those trips to those from higher income neighborhoods. These analyses have not yet been performed.

Overcrowding of Link Trains

Sound Transit asked whether the ORCA data could be used to indicate when Link trains were overcrowded and passengers were consequently left behind on the station platform. A preliminary analysis suggests that it is possible to perform such an analysis.

Work performed with subway fare collection data in Washington, DC, has indicated that it is possible to identify station overcrowding by observing bi-modal and tri-modal distributions in passenger travel times between stations. A similar technique was used on a subset of Link data, and the results were proved promising.

Limited project resource prohibited a complete test of this technique.

Change in Behavior in the U-District

With the opening of the Link light rail station in the University of Washington, a number of major transit service changes took place. For example, KCM Route 72 was eliminated, Routes 71 and 73 changed considerably in frequency and destination, and Route 74 was re-routed to skip the University District. In their place, was new light rail service to the southeastern corner of the University District.

Both the transit agencies and the University are very interested in how travel to the University District has changed as a result of these service changes. Table 3-2, presented earlier, shows that transfers between Sounder and Link increased by more than a factor of 2 between 2015 and 2016. How many of those trips on Link are being made to the University District? Is it an increase in travel to the University District via transit from Pierce County? If not, how did those trips previously get to the University? Did they take University bound buses from Pierce County, or did they take more frequently running bus services to downtown Seattle and then transfer to one of the many bus routes from downtown to the U-District?

Agencies are also interested in travel patterns associated with transfers occurring at the University of Washington Link station. There is interest in knowing the duration of transfers, as well as whether transfers are evenly balanced (i.e., are people willing to walk up the hill?)

Finally, there is considerable interest in understanding the level of intra-University District travel. How many trips are being made via transit within the University District, as opposed to walking?

CHAPTER 4: DATA PROCESSING FLOW

This chapter describes the data processing flow that was followed to create the analytical data sets used in the ORCA data analyses described previously. If ORCA data are included as part of a business intelligence (BI) system for the region, similar data processing will be necessary. However, that BI system may start with somewhat different data sets than were provided to this project team. As a result, the initial preprocessing steps will likely be somewhat different than those described below.

The project team also identified a number of additional data processing steps that could further improve the accuracy and completeness of the resulting analytical data sets. Although the budget and schedule for this project did not allow the project team to develop and implement the software necessary to add those improvements, they could be included at a future date, and they would change the data flow from that described in this chapter.

The data processing flow as currently implemented is described below.

THE INPUT DATA SETS

The input data sets drive the data processing flow. If a transit agency starts from a different set of initial input files, then the data processing flow will need to change accordingly. The input files are separated into five categories:

- ORCA transactions files
- transit agency operations files
- transit service support files
- commute trip reduction files
- non-ORCA files.

Each of these categories of files is described below.

ORCA Transactions Files

The initial ORCA transactions files were obtained from Sound Transit, which downloaded them as "reports" from the ORCA system run by Vix, the company that supplies the ORCA data processing functionality. The primary data files supplied to the project team were produced from the ORCA Boardings report. This provides the following types of data within each record:

- the card number (csn) that performed the transaction
- the card reader on which the ORCA card was tapped (the bus ID or off-board device)
- the route boarded, direction of travel, and service provider for that route
- the date and time at which the payment was made
- the type of fare transaction (e-purse or type of pass used for the payment)
- whether the transaction was a transfer and if it was, the route and service provider from which that transfer was made
- whether a discount was applied to that payment based on the specific characteristics of use (e.g., youth pass, disabled person, etc.)
- the number of passengers paid for (when the ORCA card was used as an e-purse)
- the service being used (the route boarded and the service provider)
- whether the transaction was part of a "paired" transaction, meaning that both a pass value and an e-purse value were taken from that card during that transaction, and that two records (one for the pass, and one for the e-purse) reside in the database.

While all ORCA bus boardings are included in the initial transactions file, the off-board transactions do not contain all of the data needed to identify the buses that were boarded. As a result, two additional files were provided for each nine-week period. One file was for King County's Rapid Ride off-board taps, and one was for Community Transit's SWIFT BRT off-board taps.

The off-board BRT ORCA files were provided as CSV files and are described as "Station Boarding Detail (FP Universe)" files. They contain a location code that uniquely identifies each specific stop location. The meta-data for these off-board ORCA taps uses a different code than is used in other processes to identify bus stops. Consequently, these off-board stop locations are associated to the stop_id variables used in the transit agency AVL files by means of a look-up table. (See the Transit Agency Support Files subsection below.) The look-up table describes the specific geographic location (X/Y) of that stop, as well as the stop_id used by the AVL file for that location.

In addition to the code for the location of the off-board BRT tap, the off-board BRT Station Boarding File, like the on-board Boardings Report, also describes the number of passengers paid for and the types of passengers (adult, disabled, etc.) boarding.

While the initial data table produced by the ORCA Boardings Report contains some rail trips, for some reason it does not include all rail ORCA trips or taps. To obtain those, Sound Transit provided a third set of files. Those files were specific to ORCA-based rail system use. Separate files were sent for Sounder and Link. The Link and Sounder files are identical in format and arrived as Excel worksheets. They reflect the "multi-tap" nature of the distance-based rail transactions. The spreadsheet has one row for each "tap" recorded. For any given rail trip, that can mean between one and three records.[9] The first record is for the boarding tap. The second record is for the exit tap—if the rider remembered to tap on exit. A third record is present when the rider's ORCA card contained a pass value, but the pass value was insufficient to pay for the entire trip. The third record then shows the ORCA e-purse payment required to cover the remaining cost of the trip. Each of the records contains the following basic data items:

- the card number (csn) that performed the transaction
- the origin station and destination station (when known) for the trip (these are present both as numerical values and as a text name for each station)
- the device ID on which the ORCA card was tapped
- the number of passengers paid for
- the date and time at which the payment was made (note that the time of the exit tap is the same for the e-purse and pass payments when a third record is required)
- the type of fare transaction (e-purse or type of pass used for the payment)
- whether the transaction was a transfer and if it was, the service provider from which that transfer was made
- whether a discount was applied to that payment based on the specific characteristics of use (e.g., youth pass, disabled person, etc.).

For records that describe the boarding tap, the "destination location" variable is filled with the value "4294967295," which indicates "unknown." If an exit tap occurred, then the code for the destination station is placed in this field.

---

[9] Note that in cases where a "same-route to same-route transfer" occurs on Link, more than three records can be present for a single "trip." After the same route to same-route transfers are accounted for, no more than three records should be present for the rail portion of a trip.

Transit Agency Operations Files

With the exception of the off-board reader locations, the ORCA transaction data do not include geographic location information. To geolocate ORCA payments on buses, the ORCA transaction data must be fused with transit agency bus location data. Transit vehicle location (AVL) data were provided by each of the participating agencies.

In addition, this project was very interested in the question of bias in the ORCA data. That is, not all riders use ORCA to pay for their transit use. Therefore, before ORCA data are used as a descriptor of all transit use, it is important to understand the differences in ORCA use and total use of the transit system. To help with this bias analysis step, the participating transit agencies also included their detailed automatic passenger count (APC) data. These data describe how many boardings and alightings occurred at each stop, for each bus equipped with a functioning APC.

Separate files are required from each of the service agencies. Data were obtained from three of the five bus agencies: King County Metro, Community Transit, and Pierce Transit. Sound Transit bus data were also provided by these three agencies. Everett Transit did provide access to its vehicle location files for 2016, but project resources did not allow their use in this project. Kitsap Transit vehicle location data were not available. Washington State payment location data were not used. Streetcar location data were also not used, although streetcar boarding locations (off-board taps) were included.

Pierce Transit, Community Transit, and King County Metro conveniently provided their APC and AVL data in a single set of files. The combined AVL and APC data describe when each bus arrived at each bus stop. When an APC is present on a bus, the numbers of people who boarded and alighted at that location are provided on the record. The files from all three transit agencies also contain additional GPS data points that indicate where buses were at specific points in time between bus stops.

For Pierce Transit and Community Transit, all buses are equipped with APCs, although at times the APCs are not working but data are provided by the AVL system. In addition, not all Sound Transit buses operated by PT have APCs. The locations of those vehicles are provided in the same format as the vehicles with APCs. For King County Metro, about 30 percent of bus journeys include APC data. The remaining 70 percent of trips do not include APC data, but

AVL data are present. The AVL reports location (X/Y, and stop_id) and a time stamp whenever the vehicle door is opened.

While the AVL files from all three agencies (PT, KCM, and CT) provide X/Y locations, each of the three agencies uses a different geographic projection system, and each uses a different file format, different variable names, and different meta-data.

The differences in coordinate projection systems is the most problematic of these issues. Before they are used, the KCM AVL X/Y coordinates are converted from projection 3690 to projection 4326. The CT projection is converted from 2926 to 4326. The PT data use the 4326 and do not need to be converted. For Sound Transit bus routes, a table is maintained that indicates which routes are operated by which transit agency (PT, CT, or KCM) and thus which AVL file must be queried and which coordinate system must be used.

Transit Service Support Files

A number of other data files were obtained from the transit agencies to assist in the processing of the ORCA data. These files included the following:

- a table (the Off-Board Reader Table) listing the off-board reader IDs, the stop IDs at which each reader was located, and the routes that were served by those locations
- route, schedule, and stop order files (obtained as GTFS files) that describe all of the stops and their order for all transit routes operated by day of the week
- a table (Sounder Route Service Provider Table) that links specific Sound Transit routes to the agency that operates those routes under contract.

The off-board reader files are used to identify the routes boarded when ORCA taps are made off the bus for the Rapid Ride and Swift BRT lines. Note that this list changes over time and must be updated periodically, and the appropriate list must be applied to any given set of transaction records. (That is, if the 2015 table is used for processing 2016 ORCA taps, errors will occur.)

The route, schedule, and stop order files obtained from the transit agencies are used to ensure that the analysis is clear on the order of stops made for each transit route, as well as how those stops change over the course of the day, as the supplied AVL data may not include all bus stop IDs for a given bus journey. This occurs most commonly when the bus does not stop at a stop location, but it can also occur when the on-board AVL system fails during a trip.

Commute Trip Reduction Files

State law (CTR regulation) requires that companies with work sites at which more than 100 workers arrive during the peak period must have a commute trip reduction program designed to limit the number of vehicle trips arriving or departing during the peak period. Each company is required to submit a survey that describes that plan every two years. A spreadsheet describing those plans was obtained from the Washington State Department of Transportation. Among other data, the survey describes

- the number of workers at each work site
- the number of workers at each site that are subject to the CTR requirements
- the location of each worksite
- the types of travel subsidies, if any, provided by the employer (e.g., transit pass or bike use subsidies)
- the cost of on-site and nearby parking
- parking subsidies paid by the employer
- other commute trip reduction activities undertaken by the employer (e.g., vanpool and carpool matching or preferential parking programs)
- the availability of flex time.

Many CTR-affected companies purchase transit passes for their employees. A significant portion of those companies subsidize those passes. Other companies give them away free, and still others sell them at cost to their employees. The project team obtained from Sound Transit a list of all ORCA cards that have been purchased by companies to provide to their employees. This table indicates which ORCA card numbers are associated with which companies. This table allows the subsidy information contained in the CTR employer spreadsheet to be associated with individual ORCA cards being used.

Non-ORCA Files

Not all employers that purchase ORCA cards for their employees are required to fill out the CTR survey. However, when they sign the contract to purchase ORCA cards, employers must indicate whether a specific subsidy will be provided to their employees. The project team worked with King County Metro to extract this subsidy data from the contract information along

with the specific ORCA card IDs to which each of those subsidies applied. These data were also formatted so that this subsidy information could be linked to individual ORCA cards.

Finally, the project team gathered a large number of other data sets that were needed to support the desired analyses. These files included:

- build environment data in GIS files
- census population and employment data in GIS files.

The built environment data included data such as street, bike, and sidewalk networks; parking cost; the density and land use associated with specific parcels in the region; and traffic analysis zone boundaries. The census data were taken from the most recent 5-year totals from the American Community Survey. Those data included residential and employment density by geographic area, along with race/ethnicity, household income distributions, household size, the number of children in the household, and education level distributions for each census block.

OUTPUT DATA SETS TO BE CREATED

There are four output data sets:

- geolocated Transactions file
- transfer details file.
- origin/destination file
- no origin/destination found file.

These four files are described below.

Geo-Located Transactions File

This file contains all of the ORCA boarding transactions. The geo-located Transactions file contains the data items shown in Table 4-1.

Table 4-1: Geo-Located Transactions File Description

Variable Name	Variable Type	Description
ID	int(10) UN AI PK	A record identifier
csn	varchar(32)	Hashed ORCA Card CSN
bid	varchar(32)	Hashed ORCA business ID
date	date	Date of boarding
time	time	Time when boarding tap occurred
mode_id_from	int(10) UN	Previous mode taken if a transfer record (see App. B)
mode_id_to	int(10) UN	Mode being boarded (see App. B)
num_passengers	smallint(5) UN	Number of passengers for which payment is made
passenger_type	int(10) UN	Type of passenger (e.g., adult, senior, disabled, see App. B)
product_id	int(10) UN	The type of pass (or e-purse) used for payment
direction_from	varchar(16)	The direction (inbound/outbound) the route is going for the previous route taken, if a transfer, blank otherwise
direction_to	varchar(16)	The direction (inbound/outbound) the route being boarded is currently going
route_id_from	varchar(16)	The route previously taken if a transfer, blank otherwise
route_id_to	varchar(16)	The route being boarded
service_id_from	int(10) UN	The providing agency of the route previously taken if a transfer record (see App. B), blank otherwise
service_id_to	int(10) UN	The providing agency of the route being boarded (see App. B)
vehicle_id	varchar(16)	The identifying number of the vehicle boarded (blank if a train)
trip_id_to [10]	varchar(16)	An identifier used by the transit agencies to identify a specific bus trip pattern (e.g., "the xth trip of the day on Route Y on weekdays")
trip_start	smallint(5) UN	The time when that day's trip started
stop_id	varchar(16)	The estimated transit stop at which this boarding took place
stop_location	point	Latitude and longitude of the stop_id location
stop_error	smallint(6)	A value that indicates whether problems were encountered when attempting to identify the location at which the boarding occurred. (See App. B)
avl_id	varchar(12)	A pointer that identifies the specific AVL record that contains the time and trip when the bus picks up this rider.
purse_added		A value of 1 indicates that an e-purse was also used, in addition to the shown monthly pass to pay for this trip.

Transfer Details File

This file contains all ORCA-identified transfers. The file contains each boarding time and location, as well as the stop and time the rider is estimated to have exited the previous transit vehicle. See Table 4-2. <u>A record is present for all transfers</u>. Records for transfers for which exit stops cannot be identified are present, but values are missing for those details that cannot be estimated.

---

[10] Trip_id is a value that describes a specific route a bus will operate on, with a given start time and given set of time points. The same trip_id can be used for all weekday trips on a given route, but only if the schedule is the same for all five days of the week. For example, a trip_id value will be used for the first trip of the day southbound on Route 71. A different Trip_id value will be used for the second southbound trip of the day. If the first southbound 71 starts at the same time every day and stops at all of the same bus stops, then the same trip_id will be used for all five days of the week. If on Friday the bus follows a slightly different route, then a different trip_id will be used for Friday.

Exceptions to this are "same-route to same-route" transfers found in the geolocated Transactions file. (For example, where a rider exits on Route 71 and then re-boards a bus on Route 71.) Because there are no good reasons to transfer to the same previously boarded route, these transfers are assumed to be "financial transfers" and are removed from the Transfer Details file. (The estimated transfer exit_stop is identified as a destination for the first trip, and an origin/destination record is created to reflect that trip. The boarding location previously considered a transfer boarding location is then considered the start of the next trip.)

Table 4-2: Transfer Details File Description

Variable Name	Variable Type	Description
id	int(11) AI PK	A record identifier
csn	varchar(255)	Hashed ORCA Card CSN
bid	varchar(255)	Hashed ORCA business ID
transfer_boarding_date	date	Date of boarding
transfer_boarding_time	time	Time when boarding tap occurred
origin_mode	int(11)	Mode of previous vehicle
destination_mode	int(11)	Mode being boarded (see App. B)
num_passengers	int(11)	Number of passengers for which payment is made
passenger_type	int(11)	Type of passenger (e.g., adult, senior, disabled, see App. B)
product_id	int(11)	The type of pass (or e-purse) used for payment
first_direction	varchar(255)	The direction (inbound/outbound) the route is going for the route used prior to the transfer boarding
first_route_taken	varchar(255)	The route taken before the transfer
first_service_provider	int(11)	The service agency for the route used before the transfer boarding
first_trip_id	varchar(255)	The identifier of the bus journey for the route used before the transfer boarding
exit_stop_id	varchar(255)	The transit stop where the rider is estimated to have exited the vehicle they used before making this transfer boarding
exit_time	varchar(255)	The time at which the rider is estimated to have left their previous transit vehicle
exit_lat	varchar(255)	The latitude of the exit stop
exit_lon	varchar(255)	The longitude of the exit stop
transfer_boarding_stop_id	varchar(255)	The transit stop at which the rider is boarding a transit vehicle as part of a transfer
transfer_stop_lat	varchar(255)	The latitude of the boarding stop
transfer_stop_lon	varchar(255)	The longitude of the boarding stop
transfer_route_taken	varchar(12)	The route number that the rider is boarding
transfer_direction	varchar(255)	The direction (inbound/outbound if rail, cardinal direction if a bus) of the route the rider is boarding
transfer_newservice_provider	varchar(12)	The service agency for the bus boarded as part of the transfer
number_of_transfers	int(11)	The number of transfers occurring on this trip up to this point (counting this transfer)
transfer_to_vehicle_id	varchar(255)	The vehicle_id of the vehicle the rider is boarding
last_trip_id	varchar(12)	The trip_id of the vehicle being boarded, representing the route and bus journey now being used
transfer_duration	time	The time duration of the transfer. (Transfer_boarding_time – exit_time)

Origin/Destination File

This file contains the origins and destinations of all trips for which those destinations have been identified, given specific assumptions described in this text. In general, a destination for a trip is assumed to be located at the stop that is closest to the next transit boarding on that route made with that ORCA card. This is only assumed to be true when the next transit boarding is located within a defined distance of the route for the previous trip. (This is explained in detail in the Origin/Destination Processing section later in this chapter.) Somewhat different origin/destination files result from different distances allowed in this destination process. Also, not all trips have a destination that can be identified/estimated. The variables found in this file are described in Table 4-3.

No Destination Found File

This file is structured the same as the origin/destination file described above. It contains all trips for which a destination cannot be identified. That generally means that the starting location for a following trip does not fall within 1/4 mile or 1/3 mile of the last route used for this trip. It can also mean that a transfer was not found to be feasible (i.e., a transfer boarding did not occur within 1/4 mile or 1/3 mile of the previous route taken), as this prevents the determination of the trip path.

This file is used mostly for forensic analysis.

Table 4-3: Origin/Destination File Description

Variable Name	Variable Type	Description
id	int(11) AI PK	A record identifier
csn	varchar(255)	Hashed ORCA Card CSN
bid	varchar(255)	Hashed ORCA business ID
trip_start_date	date	Date when trip started
trip_start_time	time	Time when first boarding tap occurred
origin_mode	varchar(12)	Mode of the first transit vehicle taken as part of this trip
destination_mode	varchar(12)	Mode of the last transit vehicle taken as part of this trip
num_passengers	int(11)	Number of passengers paid for that are making this trip
passenger_type	int(11)	Type of passenger (e.g., adult, senior, disabled, see App. B)
product_id	int(11)	The type of pass (or e-purse) used for payment
first_direction	varchar(12)	The direction (inbound/outbound) of the first route taken for this trip
first_route_taken	varchar(64)	The first route boarded for this trip
first_service_provider	int(11)	The agency providing the service on the first route taken
origin	varchar(255)	The stop_id where the first boarding for the trip took place
origin_service_agency	varchar(12)	The agency providing the service on the first vehicle taken
origin_lat	varchar(255)	The latitude of the stop_id at which the first boarding occurred
origin_lon	varchar(255)	The longitude of the stop_id at which the first boarding occurred
destination	varchar(255)	The estimated stop_id where the rider exited the last transit vehicle before going to their trip destination
destination_service_agency	varchar(12)	The agency providing the service used for the last segment of this trip
destination_lat	varchar(255)	The latitude of the stop_id estimated as the destination for this trip
destination_lon	varchar(255)	The longitude of the stop_id estimated as the destination for this trip
destination_time	varchar(32)	The time taken from the AVL, which indicates when the rider debarked from their last vehicle. (or the tap-off time from a train)
last_route_taken	varchar(32)	The last route taken before reaching the rider's destination
last_direction	varchar(12)	The direction (inbound/outbound) of the last route taken
last_service_provider	varchar(32)	The agency operating the service used to reach the destination
last_mode	varchar(32)	The mode of the last transit vehicle used to reach the destination
number_of_transfers	int(11)	The number of transfers required to reach the destination
last_vehicle_id	varchar(255)	The vehicle_id of the last vehicle taken for this trip (blank if a train was the last mode taken)
last_trip_id	varchar(32)	The trip_id for the vehicle_id used for the last leg of the trip used to reach this destination
transfer_required	varchar(32)	A value of 1 indicates if a transfer was required for this trip
trip_start	varchar(32)	The time of day when this trip started
trip_duration	time	Duration is computed as (destination_time – trip_start_time)

INTRODUCTION TO DATA PROCESSING

The data processing steps required to create the four desired analytical data sets from the initial data supplied by the ORCA agencies are broken into the following five tasks:

- Preprocessing
- Geolocation processing
- Transaction record clean-up processing
- Origin/destination processing
- Transfer processing.

Note that while the origin/destination and transfer processing are separate functions, they are performed simultaneously. Each of these tasks is described below.

Preprocessing

In order to fuse the input data sets together, geolocate the transactions, and then create the origin/destination and transfer files, a number of preprocessing steps must first be performed. These steps standardize the geographic references and create a number of look-up tables that greatly speed the complex data processing. The following are the pre-processing tasks:

- Import and standardize the AVL files

- Create a stop location table

- Create a table that correlates the ORCA transaction record's directional variable (i.e., inbound or outbound) with the cardinal directions used by the transit agency's directional variable (i.e., north/south/east/west)

- Create (or update) the quarter-mile look-up table

- Update the off-board stop location table

- Preprocess ORCA transactions data and reformat date and time variables

- Create a subsidy table

- Link the subsidy table to ORCA cards (CSNs)

- Hash the CSNs and Business IDs in the subsidy table, maintaining the link between the subsidy table and the hashed CSNs

- Remove duplicate boarding records.

These tasks are described below. Schema for each of the data sets are presented in Appendix C.

<u>Import and Standardize AVL Files</u>

The vehicle location data files supplied by the three participating transit agencies (KCM, PT, and CT) arrive in different formats and with modestly different contents. In addition, the three agencies use different latitude/longitude projection systems. This task loads the AVL data into three different database tables and converts some of the variables provided by the agencies into a consistent SQL data format. The separate database tables are required because of differences in the structure of the three data files. These difference not only lead to the need for slightly different database tables but also to the need for different data processing steps and for the geolocation process to convert the X/Y coordinates to a consistent projection system during data processing.

Note that the 2015 CT AVL file did not contain the trip_id variable, which is used to help identify which trip a given bus is operating on a specific route. This changed how the CT geocoding process had to be performed for 2015. The trip_id variable was added to the 2016 CT AVL file.

The preprocessor task is simple. There are three separate routines, one for each transit agency.

For the <u>KCM AVL file</u>, the data supplied are directly imported and one new variable field is added, called **`stop_datetime`**. It is constructed by combining the **`OPD_DATE`** and **`ACT_STOP_TM`** fields found in the KCM AVL file. This field is used later as part of the trip and transfer duration computations.

For the <u>Community Transit AVL</u> import, three minor reformatting issues need to be addressed, and then one variable is created. The tasks are as follows:

- **`Date`** is scanned into the SQL **`date`** format from: **`'%m/%d/%Y'`** format.
- **`ArrivalTime`** is scanned into the SQL **`time`** format from the seconds format.
- **`DepartureTime`** is scanned into the SQL **`time`** format from the seconds format.

A **`stop_datetime`** variable is computed from the CT Date and ArrivalTime variables.

For <u>Pierce Transit AVL,</u> three minor reformatting steps are required:

- **`TransitDateValue`** is scanned into the SQL **`date`** format from **`'%m/%d/%Y'`** format.
- **`apc_date_time`** is scanned into the SQL **`datetime`** format from **`'%m/%d/%Y %h:%i:%s%p'`** format.
- **`StartTripTime_Trip`** is scanned into the SQL **`time`** from **`'%k:%i'`** format.

The PT AVL data file includes the location as a point value in the **`GPS_LOCATION`** variable.

Create the Stop Location Table

A key issue when individual bus boardings are geographically located is where the stops are actually located. The AVL files that describe actual bus journeys contain that information, but the X and Y values reported in the AVL files tend to jump around because of the error inherent in GPS reports. To provide a single definition of where bus stops are located, the following steps are performed:

- Remove all X/Y locations that are "zero" (an indicator that the GPS did not report correctly).
- Compute the mean and standard deviation of all remaining X/Y values for each stop location.
- Remove any X/Y values that are outside of 3 standard deviations from that mean value.

The median value of all remaining X/Y coordinates is then taken to indicate the location of each stop. This is converted into a table.

It is important to remember that the same stop ID can be used by different service providers to represent different physical locations. For example, Stop_ID = 1 could mean a location in Snohomish County for Community Transit, but a location in Pierce County for Pierce Transit. Therefore, when stop_ids are referenced both the stop_id value and the service provider code must be provided. This is the only way to ensure that the stop_id identified is the specific location desired.

Create Table Mapping Agency Route Cardinal Directions to ORCA Inbound/ Outbound Directions

The ORCA transaction records have a directional variable that is either "Inbound" or "Outbound." None of the transit agencies use this coding. Instead they use cardinal directions (North/South/East/West) to indicate the directions for routes. In addition, different agencies use slightly different versions of these terms in their data streams (e.g., NORTH versus North.) In this process, the data processing software uses the agency supplied AVL data to normalize all the directional variables (ie. E/East/EAST) to uniform East, West, North and South, and then create a table which maps these values to the ORCA inbound/outbound direction coding.

Create (or Update) the Quarter-Mile Look-up Table

The project team determined that to perform a transfer analysis and to find the likely destination of a trip, it is important to know which bus routes pass close by individual bus stops. For example, we know that an individual boards a Route 71 northbound bus, and then transfers to the Route 44 westbound at bus stop ID = 3405. We want to be able to look up which bus stops the Route 71 uses that are close to bus stop ID = 3405 because the stop served by Route 71 northbound that is closest to the bus stop ID = 3405 is likely the stop at which that rider exited the Route 71 bus to make that transfer.

This look-up process can be done as a series of GIS queries, but those queries are slow and CPU intensive. To dramatically speed up the data processing, the project team initially developed a look-up table that indicates for each bus stop what other bus stops are within a 1/4 mile (airline distance) of that stop, which routes serve those other stops, and the airline distance to those stops. We also determined that the transfer boarding stop and route need to be included into the quarter-mile look-up for itself. That is, individuals may very well exit at Stop ID = 3405 and then board another bus at Stop ID = 3405. It is also possible for those individuals to get back on the same route, at the same stop where they got off earlier.

The process needs to be performed for each new set of schedules (a shake-up typically happens three times per year), as the locations of bus stops can appear, disappear, and move as part of each new set of schedules. In addition, routes change, stopping at different stops, even though an old stop on that route still exists and is served by other bus routes. Finally, the table needs to be sensitive to differences in schedule on different days of the week or even times of the day. (For example, a stop might be used by Route 71 on weekdays but not on weekends.)

Once constructed, the quarter-mile look-up table defines which routes, and which stops on those routes, can serve as exit locations from a transit vehicle either as part of a transfer or as the trip destination.

The initial 1/4-mile distance was selected because that distance is a common value used in the transit literature as a distance people will typically walk to transit service. However, the project team has always assumed that other distances would also be tested as part of the ORCA system. A 1/3-mile table was built in April 2017 at the request of Sound Transit and was used as part of the origin/destination data processing performed for Sound Transit's ridership forecasting

model update. Both (1/4-mile and 1/3-mile) transfer and origin/destination files are stored in the project database.

The project team believes that a 1/2–mile table should also be constructed and tested at some point, but this has not yet been done.

Note that the differences between the transfer details and origin/destination tables when they are created with the 1/4- and 1/3-mile tables are relatively small. This is because all transfers and destinations found in the tables based on the quarter-mile table will also be found in the tables built with the 1/3-mile table—since the same stops are always "closest" to a given boarding location. The differences in outcome between the two tables is caused only by the fact that some exit stops and destinations can be successfully "found" with the 1/3-mile table that cannot be "found" with the 1/4-mile table. Whether these stops are correctly identified is a topic that needs further study. In the most recent PSRC travel survey—which tracked volunteers using smartphone-based travel diaries—may provide data that can serve as ground truth for testing these different distance assumptions.

Update the Off-board Stop Locations Table

Whenever a shake-up or new set of transit schedules is implemented, the new AVL data file can be used to identify all new stop IDs and their locations. However, the AVL file does not reflect the location codes used by the off-board ORCA card readers that Link, Sounder, and the streetcar use. Therefore, for every new set of schedules, it is imperative that the off-board locations table also be updated. This update should include new off-board card reader locations and changes in specific BRT stop locations. The off-board locations must then be linked to the stop_ids used in the bus AVL files.

If the wrong look-up table is used, none of the boardings at those new locations will be assigned to a stop.

Preprocess ORCA Transaction Data and Reformat Date and Time Variables

Before the ORCA transaction data can be used, they must first be loaded into a database. MySQL was used for this project, although in the future Postgres will likely be the preferred database. During this upload, several variables in the ORCA transaction record need to be formatted as they are read into a database. The `date` and `time` columns in the database are generated from the `Txn_Date` field found in the transaction csv file data provided by ORCA

and formatted into an SQL `datetime` format. Five columns are added to the end of the table (`stop_id, stop_location, stop_error, avl_id, purse_added`) to be filled in as part of the geolocation processing and later steps.

Table 4-4 shows the final table structure used for the Transactions file. Note that **bold** indicates an indexed column.

Table 4-4 ORCA Transactions Table Structure

ID	int(10) UN AI PK
csn	varchar(32)
bid	varchar(32)
date	date
time	time
mode_id_from	int(10) UN
mode_id_to	int(10) UN
num_passengers	smallint(5) UN
passenger_type	int(10) UN
product_id	int(10) UN
direction_from	varchar(16)
direction_to	varchar(16)
route_id_from	varchar(16)
route_id_to	varchar(16)
service_id_from	int(10) UN
service_id_to	int(10) UN
vehicle_id	varchar(16)
trip_id_to	varchar(16)
trip_start	smallint(5) UN
stop_id	varchar(16)
stop_location	point
stop_error	smallint(6)
avl_id	varchar(12)
purse_added	smallint(6)

Create a Subsidy Table

This preprocessing task is needed only if ORCA card subsidy data should be added to the analyses. That should not necessary for many transit agency queries, but it is necessary if the analyses include the effects of CTR programs on travel behavior.

Unfortunately, CTR survey questions are answered differently by each employer. For example, some employers report their transit subsidies as dollar values. Others report transit subsidies as percentages of a transit pass value. This project created a simplified subsidy table

that was based on a combination of King County Metro's published rates for the Business Passport program and a computed estimate of the monthly dollar value of each company's reported CTR transit pass subsidy. A single record must be created for each employer, which is identified by its ORCA business ID.

A simplified subsidy table was also created by the project team from the CTR and King County contract data. This simple subsidy table contains two variables. The first subsidy is a binary value for which 1 = the employer provided a subsidy for that pass . The second variable, SubsidySize, is a categorical variable that places the subsidy into one of four categories:

0   no subsidy
1   < 50 percent subsidy of the pass
2   > 50 percent subsidy of the pass, but some cost of the pass to the user
3   100 percent subsidy of the pass by the employer.

The subsidy data also must be linked to the individual ORCA card CSNs. This simply requires the CSN to business ID linkage table provided by the transit agencies. The result is a list of ORCA CSNs with a link to specific records in the subsidy table, where that linkage is the ORCA Business ID for that record. Note that subsidies will change for some employers as often as every two years and that new ORCA cards are continuously being introduced into the system.

Hash the CSNs and Business IDs

Once the subsidy table has been created and linked to the CSNs used in the ORCA transactions files, both the ORCA CSN values and the business IDs in the subsidy table must be hashed to help protect the privacy rights of both individuals and companies.

This is done by first creating a hash for each CSN. For this project, the MySQL function OLD_PASSWORD is used to generate a 16-byte hash string for each CSN and business ID. Note that the link between specific CSNs and business subsidy records must be maintained. The hash is applied to all transaction records, replacing the original CSN with the hashed ID during the import into the database.

A similar hash for the business ID is then created and applied to both the subsidy table and the business ID field in the transaction records. Again, make sure that the linkage between the subsidy table and transaction table is maintained.

For this project, the hash keys were then destroyed. However, if this overall analysis process will be performed continuously, then a new process will be necessary to either keep the

hash key or periodically change the hash key. Some continuity in hash keys over time will be necessary to perform the destination processing. (That is, the destination estimation process will require tracking specific CSNs over multiple days. If the hash key changes on a daily basis, this will not be possible unless more than one key is maintained for at least some period of time.)

Remove Duplicate Boarding Records

The final preprocessing step removes duplicate boarding records in the Transactions database that occur when an ORCA card is used both as a pass and as an e-purse to pay for a single boarding transaction. This happens when the value of the transaction is greater than the value of the pass stored on the ORCA card, and that ORCA card also has an electronic purse cash value. The two records are identical except for the Product_id variable. On one record this variable indicates the pass being used, and on the other record, this variable indicates that an e-purse was used. In this step, the e-purse record must be deleted. The variable purse_added is set to "1" in the remaining record when the e-purse record is deleted. (This variable is set to "zero" in transaction records when duplicate records are not present.)

Geolocation Processing

Once the pre-processing described above has been performed, then the available ORCA transactions must be geolocated. With the exception of the train and off-board BRT transactions, ORCA transaction data are not geolocated. Furthermore, while location is present for the BRT data, route and vehicle_id are not. The geolocation task for on-board taps requires linking the ORCA transaction record to the appropriate AVL record for the boarded bus and extracting the location of that bus when that transaction occurs. For off-board ORCA transactions, the geolocation process involves determining which specific vehicle was boarded, given a payment at a specific location and time.

On-Board Bus Payment Geolocation Processing

The initial task is to geolocate bus boardings where payment occurred on the bus. This task is performed in the following steps:

- Sort ORCA transactions by date and time.
- Set the value of `stop_error in orca-transactions` table to -1 for all records. (This is a flag to indicate records that have not been processed yet.)

- For only those transaction records in which `Mode_to` = 128 (bus, with on-board payment) perform the following tasks:
  - Use the **Service_id_to** variable to identify the AVL file to use for geolocation look-up. If the service provider is Sound Transit, then also use the **Route_to** variable and the ST Route linkage table to determine the correct AVL file to use for the look-up.
  - Use the variables `date, time,` and `vehicle_id` and `trip_id` to perform a look-up in the AVL records for that vehicle ID, operating that trip_id, on that date, and then pick the location record that has the smallest absolute time difference between the AVL recorded date and time and the date and time from the ORCA transaction for that record's `vehicle_id and trip_id`.

If the trip_id variable is not found, then the route_to variable and time are used to identify the appropriate AVL record for determining the correct boarding location. The stop_id and its associated X and Y values for that record are then extracted from that AVL record and used to update the ORCA transactions record. The record ID for the AVL file in which that stop location record was found is also extracted. The variable **avl_id** in the Transactions file is set equal to the AVL record ID. This speeds up later data processing by allowing later look-up procedures to directly identify the record used to geolocate this transaction. This in turn simplifies the task of finding the exit stop and the time when that vehicle stopped at that exit location, which is needed for the transfer detail and trip duration computations.

Several additional errors can occur in this process

- If the time difference between the boarding time and the closest AVL record for that vehicle ID is greater than 5 minutes, then the stop_error is set to 1. (A time differential greater than 5 minutes indicates errors in the AVL data collection process often related to either a failed GPS device or failed communications on the bus.)
- If the AVL record look-up returns a value of "ERR," then stop_error is set to 2. (The code "ERR" means that the AVL record exists within the time limit, but the AVL system did not function correctly and there is no stop_id assigned for that point in time.)
- If no matching vehicle_id is present in the AVL database, then the stop_error is set to 4.
- If further re-processing using different lookup parameters is attempted, but still results in AVL record not existing in the database, stop_error is set to 5.

- If the service provider is not KCM, ST, PT, or CT, the stop_error remains as -1 (e.g., record not geolocated).

The outcome of this process is a set of geolocated ORCA Transactions files, with stop_error codes added when data are not found in the AVL records and when conflicts occur between the data found in the ORCA records and data found in the AVL records.

Off-Board Bus Payment (BRT) Geolocation Processing

The next step in the geolocation process is to identify the stop_id where off-board BRT payments[11] were made, the route that was boarded, and which actual vehicle_id was boarded after the off-board payment was made. BRT payments that are made on the vehicle will have been geolocated in the previous processing step. Off-board payments are present in the Transactions file but have not yet been fully processed.

The off-board payments are also present in separate files. These off-board payment files are used for the initial off-board data geolocation processing because the off-board files contain the long device ID value needed to identify the off-board stop locations, whereas the ORCA transaction records do not. Data in the off-board file are processed and then used to update the record already in the Transactions file. If the off-board file contains a record that is not found in the Transactions file, then a new record is inserted by using the information in the off-board file.

The steps for using the off-board payments (BRT) file are as follows:
- Read the value for Origin Location.
- Look up that value in the Off-Board Reader Table and extract the stop_id, stop_location_x, stop_location_y, and route and direction values for that location.
- Use the Txn Date value (a date and time value) to select the AVL file for that service provider (KCM or CT) and perform a look-up to identify the first vehicle that passes that stop_id, for that route and direction, <u>after</u> that transaction time. The first vehicle on that route that passes that stop_id is the vehicle assumed to be boarded.
- Store the vehicle_id, the stop_id, stop_location_x, stop_location_y, and the route values with that transaction value, to be used to update the ORCA transactions file variables.

---

[11] Note that for Community Transit's SWIFT BRT routes, only off-board payment is possible. For King County Metro's Rapid Ride BRT, it is possible to pay either on the bus or off-board.

There is one significant limitation in this process. A number of off-board bus stops in downtown Seattle are served by more than one BRT route. Therefore, it is not possible to directly assign the off-board payments to a route at this point, and it is also not possible to assign the boarding to a vehicle, since a vehicle look-up cannot be completed without route information. Additional processing is necessary to resolve this issue. Therefore, when the look-up of the Stop Location Table indicates that the stop in question is served by more than one route, then stop_id, stop_location_x, stop_location_y are correctly assigned for the boarding location but Route is set to "CDE." (Several downtown BRT stops are served by the C, D, and E BRT lines. Several other downtown stops serve just the C and D lines.)

Once all off-board BRT boardings have been processed, the entire new off-board BRT payment data table is used to update the main ORCA Transactions database.

The first step in that process is to sort the Transactions database by CSN, then date and time. Records are matched between the revised off-board payment file and the main Transactions database by using CSN, date, and time. The Route and Vehicle_id variables taken from the BRT off-board file are used to update the Route_to variable in the record being updated. Note that vehicle_id cannot be updated in the CDE records, since we do not know either the BRT route or vehicle.

In addition, whenever a record is updated, the NEXT record in the Transactions file (that is sorted by CSN, date, and time) is examined. IF that next record has the same CSN AND it is a transfer from BRT (Mode_from = 250), THEN the Route_from variable in that second record is also updated by setting Route_from = Route_to taken from the previous record's Route_to variable. This update places the correct BRT route designation in the next record if, and only if, that next record describes a transfer from the BRT line.

Once this update process has been completed, the entire Transactions database must be reprocessed to finish the off-board geolocation update process. As before, the Transactions file must be processed in CSN order, by date and time. In this case, each record is searched to determine whether any records contain Route_to = CDE. This locates records for which correct BRT route has not yet been identified and the correct vehicle ID added to the transaction record.

To identify the correct BRT route, the next transaction record must be examined for that CSN. (If the CDE record is the last record in the database for that CSN, then the route or vehicle cannot be identified. Leave those variables blank.) Where the next record does contain the same

73

CSN as the CDE record, perform a 1/4-mile look-up[12] around the boarding location for that next record. If one and only one BRT line is within that 1/4-mile look-up <u>OR</u> if that record describes a boarding of an identified BRT line, then that line (either because the BRT line or the one BRT line operating within 1/4-mile of the new boarding location) is assumed to be boarded during the CDE transaction.

By using that new route assignment, it is now possible to look up in the AVL files which vehicle, operating on that newly assigned route, passed by the off-board location for the CDE record closest to—but after—the CDE payment tap recorded at that off-board location. The CDE record in the Transactions file is then updated to include the newly identified Route_to and Vehicle_id information. If next record in the database is a transfer record, then the Route_from variable of that second record is also updated to reflect the route from which that transfer occurred. Also updated at this time is the avl_id record that indicates the AVL record from which the vehicle_id variable has been extracted.

The database search then continues to find the next CDE record that must be updated.

Note that this process does not allow a route to be assigned to some ORCA transaction records. For those records, also cannot be assigned a vehicle_id, and consequently, those records cannot be used in some transfer analysis computations.

Sounder and Link Geolocation Processing

The Sounder and Link files supplied by ORCA contain geolocation references for the origin and destination points for the majority of trips taken by train. The only time that the beginning and ending of the rail portion of an individual's trip cannot be directly identified is when a boarding tap occurred, but an exit tap did not occur. (Note that the rail trip may be only a portion of the total transit trip, as transfers may be involved on either end of the rail portion of that trip.)

Most, but not all, Sounder and Link trips are found in the ORCA Transactions database. However, all Sounder and Link trips are found in the Sounder and Link files. Like the off-board BRT files, these rail transaction files are therefore used as the starting point for the geolocation effort.

---

[12] See the origin/destination subsection for a thorough discussion of the 1/4-mile look-up process and the reasoning behind that process.

Link and Sounder files are processed in exactly the same way but are processed in separate steps because they are held in separate database tables. That is because they were transmitted to the project team by Sound Transit as separate files.

The first step in the process is to create a single record for each train trip, regardless of how many taps were made, as this record needs to be included in the Transactions file. One, two, or three ORCA taps can have been made for any given trip. One tap occurs if the rider taps on to start the trip but forgets to tap off. Two taps occur if the rider taps on and off. Three taps appear in the rail file if the ORCA card contains a pass value, and that value was insufficient to pay for that trip. In that case, the e-purse value on the card shows up as a third tap record. Note that because of the definition of a transfer within the ORCA financial system, it is possible for more than three ORCA taps on Link or Sounder to be associated with a given "trip." This occurs when the second rail trip starts within 2 hours of the first ORCA tap. This registers as a "same-route to same-route transfer." The project team breaks these transfers into two separate trips, and then no more than three taps/records should exist for the rail portion of any trip.

At the same time that the single boarding record is being developed to add to the Transactions database, a rail origin/destination file is created that can be used later in the data processing effort. To perform the above tasks, the rail transaction records are sorted first by CSN and then by transaction date and time.

After the file has been sorted, if more than one consecutive record for that CSN and date have the same origin station, those records are grouped together and treated as a single trip. From that group of records, the record with the earliest time stamp is treated as the origin tap. If more than one record have the same time stamp, one record will be a pass exit, and the other will be an e-purse exit. A new variable "purse_added" is added to the end of the record and set to "1" when this occurs. The record with the latest time stamp is treated as the destination tap; again, there may be two records with the same late time stamp. If the purse-added flag has not been set, do so.

If at least two consecutive records are present with the same Origin Location, and the second (or third) record has a non-zero (or non-blank) value in the Destination Location variable, then an off-tap was performed.

If two consecutive records have the same Origin Location and the second record does not have a valid Destination Location (i.e., Destination Location = 4294967295 for both records),

then those two records represent two different trips, and the first trip does not have a defined destination. (The next—or third—record will then determine whether the second trip has a destination.)

On the basis of these details, a "rail only" origin/destination record can be created in which one record defines each specific rail trip. The rail file is processed by CSN, in date time order, as follows:

(a) Each time a new CSN is observed, start a new trip record.

For that trip record, the "boarding time" is the Txn Date for that record.

If the next record has that same Origin, and the Destination Location of the second record is not equal to 4294967295, then use that Destination Location as the destination of the rail trip, and use the Txn Date for that second record as the exit time for the rail portion of that trip.

If the next (third) record has the same CSN and Origin Location and Destination Location of the second record, then it is the second exit tap (either epurse or pass) for that trip. When this occurs, set the purse_added flag = 1.

If the next (second) record has a different Origin Location value, then an exit tap was not performed.

Then the first trip record is completed, with no known rail destination.

Use the second record (the new trip) as the first record, and return to the top of this data processing sequence (step (a) above).

If the next record to be processed has a different CSN than the first record,

Then the first trip record is completed, with no known rail destination.

Use the second record (the new trip) as the first record, and return to the top of this data processing sequence (step (a) above).

Note that following these procedures will result in separate rail origin/destination tables for Link and Sounder. These tables are used both to update the Transactions file and in the origin/destination processing.

The rail origin/destination file completed by processing both the Link and Sounder rail files is then used to update the Transactions file. This update process searches the existing Transactions file to find matches between the CSN, Date, and Time on a transaction record with the CSN, boarding (origin Txn) date and time, respectively, of a rail trip. That transaction record

is then updated by setting Route_to = "Link" or "Sounder" (depending on the mode is being processed). Vehicle_id is also set to "Link" or "Sounder." Next, a look-up is performed for the Link or Sounder Directional table to determine the direction of travel on the rail system. This allows the updating of the "direction_to" variable. Purse_added is also updated where required. The transaction record containing that trip is then updated. If no record with that CSN, date, and time is present, then a new transaction record is created.

Transactions Record Clean-Up Processing

Once geolocation processing has been completed and off-board payment transactions have been processed, one additional data processing step is necessary: for the Route_from and Direction_from variables, transfer records that result <u>from</u> trips paid for off-board must be updated. To do this, transaction records are processed in CSN, date, and time order. Whenever a record contains a "Mode_from" value equal to 250 (BRT), 251 (Link), or 9 (Sounder), the variables Route_from and Direction_from in those records must be updated to equal the value found in the previous record's Route_to and Direction_to variables.

Origin/Destination Processing

One problem with trying to use the ORCA transaction data for service and transit operations planning is that ORCA is a "tap-on only" system for buses. That is, patrons pay to ride buses by tapping their ORCA cards when they board the bus, but they do not tap when they exit the bus. Therefore, ORCA has a direct record of when they board a bus but no record of when they exit. This is unlike transit trips on either of the rail modes, for which payment of the rail trip is distance based, and the rider is supposed to tap both at the beginning and at the end of a trip.

Identifying the Destination of a Trip for Buses

If one simple assumption is made, then a large fraction of the transit destinations can be estimated by using the available ORCA data. The premise behind the approach is that, for many trips, the origin point for a trip can serve as a good estimator of the destination of that rider's previous trip. This basic assumption is true when a traveler either 1) uses transit for consecutive trips, or 2) any intervening trips made by that transit rider return the traveler to the destination from his/her previous transit trip.

One good thing about this assumption is that it is not time dependent. For example, I can take the bus to work (trip #1) on Friday morning, and then take it back home in the afternoon (trip #2). But when I take it back to work on Monday morning (trip #3), my Monday morning trip origin still serves as a good estimator of my Friday afternoon trip home, even though it occurs two and a half days later, and several intervening weekend trips have occurred.

Now this assumption is not fool proof, and no detailed work has yet been performed to date to test the accuracy of the assumption. But for this project, it is a reasonable assumption. Logic also suggests that a large percentage of trip destinations can be correctly attributed by using this assumption, and that the errors that result from the use of this assumption are unlikely to significantly bias trip pattern estimates.

To operationalize this assumption, the following logic was developed. The simplest version involves trips that do not include transfers. For this condition,

- Identify the initial boarding stop (Stop1), route (Route1), vehicle (Vehicle1), and direction of travel (Direction1) for trip #1.
- Identify the initial boarding stop for the next trip taken by that traveler (Stop2).
- Look to determine whether Route1, going in Direction1, passes within 1/4 mile[13] of Stop2.
- If Route1 does pass within 1/4 mile of Stop2, then select the closest bus stop to Stop2 that is served by Route1 when it is traveling in Direction1. This stop (ExitStop1) becomes the destination of trip #1.
- The time the rider arrives at this exit stop can then be determined by looking in the transit agency's AVL file when Vehicle1 stops at ExitStop1.
- If Route1 does not pass within 1/4 mile of Stop2, then no destination can be determined for trip #1.

This process continues for a specific ORCA CSN (i.e., comparing Route2 against Stop3, and Route3 against Stop4, etc.) until all that ORCA card's boarding transactions have been processed.

Figure 4-1 illustrates this process. The example rider boards vehicle ID 2198, operating as Route 63, southbound, at point X1. The next time s/he boards a bus, s/he gets on a Route 70

---

[13] Different distances may also be used in this look-up process. The project team recommends that future work pursue sensitivity tests examining the use of 1/4-, 1/3-, and 1/2-mile distances. The project was not able to perform these tests because of time and resource constraints.

bus, northbound at point X2a. The 1/4-mile look-up finds two bus stops served by the southbound Route 63 that fall within the 1/4-mile distance. Stop Y is the closest of these two points to the second boarding location. It is the selected destination for the first trip.
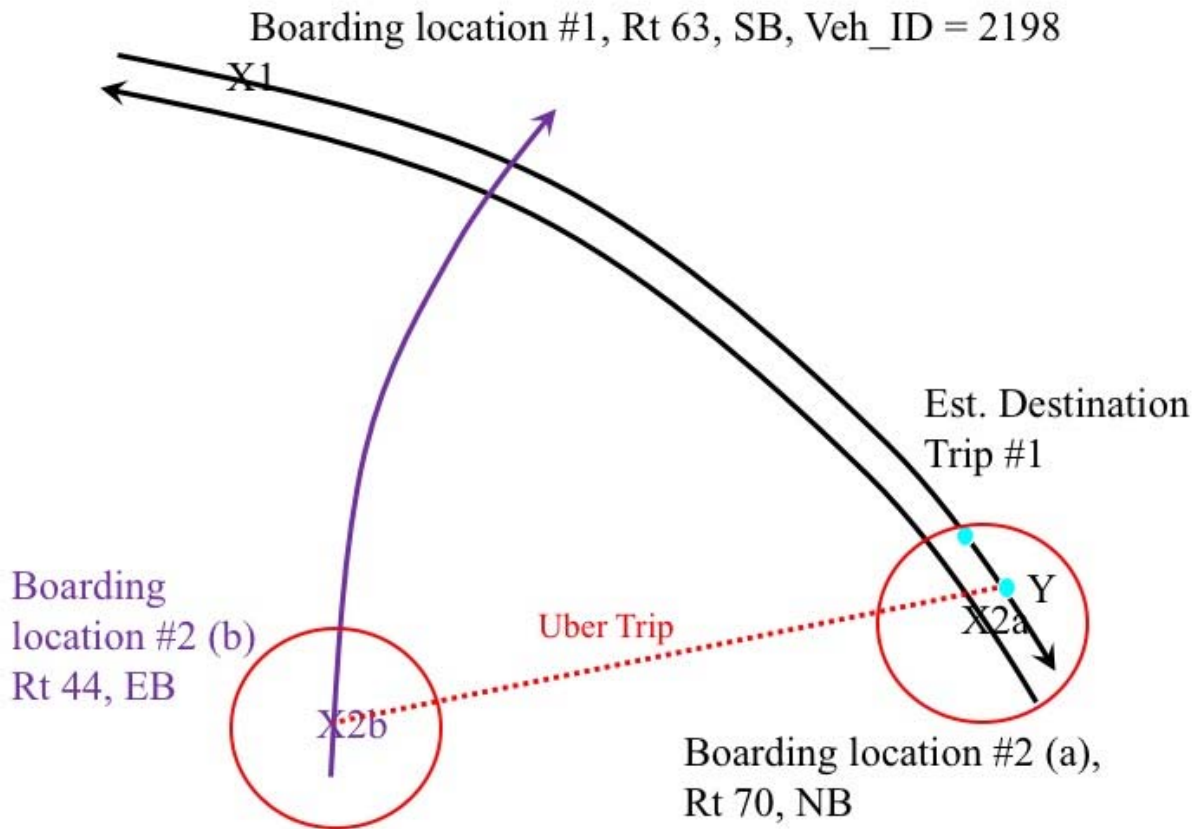
Figure 4-1: Illustration of How Destinations Are Determined

If instead of boarding the bus at point X2a, the rider takes an Uber to a different part of the city for a meeting, and then boards a Route 44 bus, eastbound at stop X2b, then the 1/4-mile look-up function will fail to find a bus stop on the Route 63 that is within 1/4 mile of the bus boarding location. If that happens, no destination will be predicted for the first trip.

This example shows that the origin/destination pairs for routine transit riders are likely to be well represented by this process. This is true even for trip chains. Where this process breaks down is when intervening travel activity occurs between transit trips and the traveler does not return to his/her original transit destination before taking the next trip. In most cases, this will

result in a "not possible to determine destination" outcome. In some cases, it will provide a false report of a destination.

Also clear from looking at Figure 4-1 is that the selection of the destination bus stop is not very exact. This is particularly true in places such as dense activity centers, where trip chaining often includes walking, or where other factors (hills, better sidewalks, coffee shop locations on the way to work) may influence the selection of the exit stop more than just that stop's distance to the next bus stop taken. In these cases, a rider may exit from a bus to perform some activity and then walk to the ultimate destination before boarding a bus to travel to another activity. In these cases, the destinations are identified as occurring close to the end of the (walking) trip chain, as opposed to where the rider actually exited the bus. This is an acceptable limitation in the process, given limitations in the data set and the relatively modest importance of those errors in the overall transit planning effort.

A major complication in programming this approach to identifying transit destinations is the fact that many trips involve a transfer. When a transfer occurs, additional steps are required in the software algorithm. These additional steps are as follows:

1) If the boarding transaction for trip#2 is marked as a transfer (the Mode_from variable in the transaction record is not NULL, which means this is not a trip but a transfer, and thus this is really "record#2, not trip#2), then

2) Reset Route#1 equal to Route#2, Direction1 to Direction2, and Vehicle1 to Vehicle2 (the origin stop is still Stop1).

3) Read transaction (boarding) record #3 to ensure that record #3 is for the same CSN (i.e., CSN2 – CSN3). If so continue; if not, then no destination can be estimated for this trip.

4) If CSN2 = CSN3 and Mode_from is NULL (record #3 is not a transfer), then

    4a Identify the boarding stop (Stop3) and location from record #3.

    4b Look to see whether Route1 (now set equal to Route2), going in Direction1 (now equal to Direction2), passes within 1/4 mile[14] of Stop3.

---

[14] Different distances may also be used in this look-up process. The project team recommends that future work pursue sensitivity tests examining the use of 1/4-, 1/3-, and 1/2-mile distances. The project was not able to perform these tests because of time and resource constraints.

4c If Route1 does pass within 1/4 mile of Stop3, then select the closest Stop3 that is served by that route while it travels in direction2. This stop (ExitStop2) becomes the destination of the trip, including the transfer.

4d The time the rider arrives at this exit stop can then be determined by looking in the transit agency's AVL file when Vehicle2 stopped at ExitStop2.

4e If Route1 (now defined as being equal to the value in Route2) does not pass within 1/4 mile of Stop2, then no destination can be determined for trip #1.

4 cont.) If CSN2 = CSN3, and Mode_from contains a valid entry, then record #3 <u>is</u> a transfer. In this case, return to step 2 and use the values from this third record to replace the values in Route#1, and Direction1, and Vehicle1.

Continue with steps 3 and 4 until either a destination is identified or it is determined that a destination cannot be identified for this trip. Multiple transfers are allowed as part of this process.

<u>Estimating a Destination if a Trip Includes Train Travel</u>

An exception to the above procedure occurs if the mode_to variable is either 251 (Link) or 9 (Sounder). In this case, a specific subroutine looks up the specific boarding transaction in the Link or Sounder origin/destination table that is built as part of the off-board geolocation processing. If the origin/destination record for that trip has a defined destination, then that destination is set equal to the destination of the rail trip.

If the next record in the Transactions file is not a transfer record for that same CSN, then the trip destination is set equal to the rail destination, and that next transaction record is used to start a new trip.

However, if that next record <u>is</u> a transfer record, then the transfer details process (see below) is performed by using the destination train station as the known exit stop and the "tap off" time as the exit time from the previous transit vehicle. (Note that the train (Link or Sounder) station does need to fall within 1/4 mile of the transfer boarding stop. If this is not true, the entire trip is placed in the "no destination found category.") The destination for the trip is then determined by whether that bus trip has an identifiable destination.

After a Destination Has Been Identified

Once a trip destination has been identified, then an origin/destination record is created. The origin/destination record contains the CSN, the first transit stop used during a trip, the time when the payment was made at that stop (the bus boarding time if payment occurred on-board a bus, or the time an off-board reader was tapped for off-board BRT and rail trips), the estimated exit stop for that trip, and the time the last transit vehicle used during that trip is estimated to have stopped at that location. When a rail vehicle is the last transit vehicle used, the exit time is the time when the "tap off" occurred.

Transfer Processing

Roughly 25 percent of all transaction records indicate that the boarding is a transfer boarding. When a transfer occurs, the goals of the transfer processing step are to 1) determine the details of the transfer—including where the rider left the first vehicle, the distance s/he walked to the transfer boarding, and the total amount of time between leaving the first vehicle and boarding the second vehicle. In many cases, these details can be determined. In other cases, the "transfer" is a financial transfer, and no logical exit point from the first transit vehicle can be determined. In still other cases, the data indicate that a financial transfer occurred, but the exit stop can be identified. This allows those "transfers" to be converted into destinations for the previous trip, and the "transfer boarding" to be used as the origin of a new trip.

The following steps are taken as transfer records are processed. Details of these processes are described as part of the origin/destination processing above.

Successful Transfers

If the Route being boarded (Route2) contains a valid Route_from variable, and that route is different from the route identified in the Route_to variable (Route1), then the 1/4-mile check is performed around the new boarding location (Stop2) to find the exit stop just before the transfer boarding.

If that look-up succeeds (Route1 serves a stop that is close to Stop2), then the closest Route1 stop to Stop2 is chosen as the exit_stop for the transfer. The Lat/Lon coordinates for that stop are then taken from the 1/4-mile look-up table. The AVL record ID from the previous transaction record (Trip1) is then used to find the correct AVL record for the bus onto which the rider previously boarded. The time that bus, on that bus journey, stopped at the exit_stop is then

extracted from the AVL file. The variables exit_stop_id, exit_lon, exit_lat, and exit_time are then updated in the Transfer Details file.

Failed Transfer Exit Stop Look-Ups

If Route1 is not equal to Route2 and the 1/4-mile look-up fails (Route1 is not close to Stop2), then it is assumed that at least one trip occurred between the first and second transit trips. Therefore 1) this is a financial transfer, not an operational transfer, and 2) no destination can be found for Trip1. This is illustrated in Figure 4-1, where the Transfer Boarding stop is at the point labels X2(b). In this case in Figure 4-1, the rider has taken an Uber trip between exiting the bus taken for Trip 1 and boarding the bus for transit Trip #2.

When this occurs, Trip1 is placed in the No_Destination file, and this Transfer is placed in the Transfer Details file with the variables exit_stop_id, exit_lon, exit_lat, and exit_time all set to = N/A.

Handling Same-Route to Same-Route Transfers

If the Route being boarded (i.e., the Route_to value in Record2) is the same as the route being transferred from (i.e., the Route_from value in Record2, which should be equal to the Route_to value in Record1), then an activity is assumed to have taken place between the first boarding and the second boarding (i.e., this person performed a "trip chain" or "tour"). This is because an individual rarely gets off a bus in order to get on a later bus serving that same route just to improve their trip.[15] Consequently, if Route2 = Route1, then boarding Stop2 is the identifier of the logical approximate destination for Trip1. This "transfer" is therefore not a transfer and is not placed in the Transfer Details file. Instead, the origin/destination processing described above is performed,[16] and the Transfer Boarding becomes the origin point for Trip#2.

---

[15] An exception to this statement might be when one bus is so late that the bus behind it has caught up. The first bus is therefore full, and the second bus is empty. Some passengers might get off the first bus in order to board the second bus to gain a seat. These instances are extremely rare, so we make the logical assumption that re-boarding the same route previously used is the sign of a trip chain.

[16] Thus, a 1/4-mile look-up is performed around Stop2. This is allowed to return boarding stop2 as the destination of Trip1. Where the direction of travel for Trip1 equals the direction for Trip2, it is logical that the rider would exit and re-board at the same location. This is what is termed a "pit stop": the rider gets off at a bus stop, does an activity, and then re-boards that same route, going in the same direction, at the same stop. Alternatively, the rider may be involved in an "end point" transfer: that is, the rider does an activity and returns to their initial starting point via the same route. In this case Exit Stop1 is essentially across the street from Boarding Stop2, and the rider is likely returning to the origin for Trip1 after having completed an activity near Exit stop1 and Boarding stop2.

In both cases, once the same-route to same-route transfer is converted into an origin/destination pair, the next record (Record3) for this ORCA card number is then examined, proceeding either to the transfer or origin/destination subroutines based on whether Record3 is or is not a transfer.

Rail Transfers

If Trip1 occurs on a rail vehicle (Sounder or Link), and a transfer is indicated for that CSN in the next transaction record, then a transfer is assumed, even if there is an exit tap.

If the rider taps out, the validity of the transfer boarding (that is, whether it's a "valid transfer" or a financial transfer) is verified by checking the location of the transfer boarding (Stop2) against the exit station for the rail portion of the trip by using the rail origin/destination file. If the new boarding is not within 1/4 mile of the exit rail station, then the transfer is assumed to be a financial transfer and the trip destination is assumed to be the rail station exit point.

If the transfer is valid, then the time when the "exit tap" occurred is used as the time when the exit occurred from the rail trip and is entered into the transfer details file. The Latitude and Longitude values are those associated with the rail station. Note that because many Link stations have multiple exit points, this assumption slightly limits the actual size of the transfer area around that station.

If the rider does not "tap out," then the 1/4-mile computation takes into consideration the entire rail line both to determine the validity of the transfer and to identify the rail station that served as the exit point from the rail portion of the trip. If the new boarding is not within 1/4 mile of the exit rail station, then the transfer is assumed to be a financial transfer and the trip destination is assumed to be the rail station exit point.

Even if the transfer is valid, the time when the rider exited the rail station cannot be determined. Therefore, while the station can be located, the transfer duration cannot be determined because the time of the "exit tap" is not known. The Latitude and Longitude values in the Transfer Details file are those associated with the rail station, but the exit_time variable is left blank, and the duration variable is set to "not applicable."

Off-Board BRT Transfers

While BRT payments can also be made off-board, the bus that a patron subsequently boards is determined for most boardings as part of the geolocation processing steps describe above. When the transfer process is performed, trips that include BRT are processed in the same way as all other bus trips because it is possible to look up when the specific bus ridden passed by an identified stop. The one exception to this rule is boardings that are designated as CDE and for which the specific route taken (and thus bus boarded) cannot be identified. In those cases, as with rail trips that are not tapped out, the transfer location can be identified, but the transfer duration cannot be computed.

.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

CONCLUSIONS

This project has succeeded in developing a prototype data set that uses ORCA transactions to answer key policy and planning questions for the region's MPOs and transit agencies. The data set fuses ORCA fare card transaction data with automated vehicle transaction data from three transit agencies: King County Metro, Pierce Transit, and Community Transit. The fused analytical data sets describe the origins and destinations of over 11,500,000 transit trips during nine weeks in 2016. This data set allows planners for WSDOT, PSRC, and the regional transit agencies to examine trip making behavior.

The data summarized and described in this report were supplied by Sound Transit to the UW with the express permission of the ORCA Joint Board. The UW is not permitted to supply the transit agencies with the geolocated data as part of the Human Subjects agreements signed with the UW Institutional Review Board at the time the project was initiated. The UW is allowed to supply summaries of those data to the transit agencies and other interested parties, as long as no personally identifiable data are released. These summaries were developed manually by using a variety of statistical and database tools (e.g., R, SQL, Postgres, Python). A number of specific analyses were performed as part of this project to illustrate how data from electronic fare media—when combined with other available data—can be used to answer a wide range of questions. Some of the work performed included the following:

- Maps of the geographic location of bus boardings with ORCA were constructed. Different maps can be produced for total boardings versus boardings by specific passenger types, including seniors, low income users, youth, or disability pass holders.
- Maps were produced that examined the differences between ORCA boardings and total boardings by stop or census tract.
- Tools were built that describe how boardings differ by time of day or day of week, on each route, by type of rider.
- Analyses examined how the frequency of transit use differs between riders with monthly passes and riders paying on an individual trip basis (e.g., e-purse).

- Preliminary results were produced that illustrate the effectiveness of employer transit subsidies on the use of transit to work, given different built environment factors (e.g., the availability of free or paid parking, the amount/quality of transit service available).
- Origin/destination matrices were developed for the region for use in improved regional model calibration and transit service planning.

The project demonstrated to a variety of regional stakeholders a proof of concept for how electronic fare card data might be used to study a variety of transit operations and planning activities. The project team also developed, refined, and documented the steps needed to systematically process these data to allow public agencies to use them in a variety of ways that both protect privacy and prove useful to transportation planners across the region.

RECOMMENDATIONS

The project team, including staff from the Washington State Department of Transportation, Sound Transit, the Puget Sound Regional Conference and the University of Washington, recommend that the various ORCA agencies work together to make the use of ORCA data an ongoing activity. When Phase 2 of this project is completed on December 31, 2017, no additional analyses with these data will be performed unless additional funding is identified for those tasks. However, the data will be archived for future work should funding become available. The project team has been working with WSDOT and the various transit agencies to find funding to continue the analyses and to process additional weeks of ORCA transaction data.

Ideally, ORCA data should be routinely processed through the steps described in this document and routinely used for a variety of planning and operational analyses. Both Sound Transit and King County Metro have proposed funding the development of enterprise data systems that would convert the initial code developed under the original ORCA data project to a more stable, sustainable enterprise system meant to routinely perform the required tasks. Several solutions for converting the basic research software into an enterprise system that could be used by all ORCA agencies have been proposed, but no formal decision has been made.

It would be to the advantage of all agencies that employ ORCA that only one such system be built for the region. All ORCA agencies should then have access to those analytical

capabilities. Such a system is well within the technical capabilities of the region. However, before work is completed, the ORCA Joint Board needs to develop policies for the use of the data. This work has been started by Sound Transit and needs to be supported by the ORCA Joint Board.

APPENDIX A
ACRONYMS AND DEFINITIONS

AVL   Automatic Vehicle Location system (Each transit vehicle carries a GPS device that allows the vehicle's location to be determined and its future location predicted. AVL systems produce a historical file that can be used to identify where a vehicle was located at a specific date and time.

APC   Automatic Passenger Counter system (Some, but not all transit vehicles count the number of people getting on and off the vehicle. All PT and CT buses have these automatic passenger counters. 15% of KCM vehicles have APCs. ST buses operated by CT have APCs. A subset of the ST fleet operated by PT or KCM have APCs. Where not all vehicles have APCs, ridership estimates are based on non-random samples of the ridership counts that are collected by the vehicles that have APCs.

BID   Business ID, the unique number that identifies a specific company that purchases ORCA cards in order to provide them to their employeees

Block the set of trips that a specific transit vehicle performs during a given day. That block may require more than one operator, and may involve more than one route. For example, the block for the day for bus #1207 may first include Trip #1 on SB Route 71, then become trip #8 on SB Route #132, then become trip #14 on NB Route #132, and then trip #23 on NB Route 73.

CSN   Card Serial Number, the unique number that identifies a specific ORCA card

CT   Community Transit, the transit authority which operates in most of Snohomish County (everywhere but in the city of Everett.)

CTR   Commute Trip Reduction, used both to reference a Washington State law, and to reference the activities performed by employers to comply with that law.

KCM King County Metro transit, the agency that operates transit services in King County

Off-board fare collection – Refers to the fact that fare payment for BRT, Light Rail, Sounder, and Ferry services is, or can be, paid on the side of the road (for BRT) or at a station (for the rail and ferry services.) In some cases, BRT payments can also be made when boarding the bus. With off-board fare payment, the actual payment does not occur when the vehicle arrives, thus the time associated with the payment is not equivalent to the time when the patron boards the transit vehicle.

ORCA One Regional Card for All – the electronic fare card used by nine transit agencies in the Puget Sound Region

PT   Pierce Transit, the transit authority which operates transit services in Pierce County.

ST   Sound Transit, the regional transit agency serving Pierce, King, and Snohomish Counties.

Stop_ID   Each bus stop has a unique identifier. Unfortunately, each transit agency may use the same number, but refer to a different physical location. That is, KCM Stop_ID = 5 and CT Stop_ID = 5 may reference two completely different locations, one in King County and the other in Snohomish County. Therefore, identification of a stop_id typically requires the use of both the actual stop_id and the service provider.

Trip   A piece of work for a transit vehicle that starts at the beginning of a route and ends at the terminus for that route. Trips are directional and numbered from the first to last trip for that route and direction for each day. (So Trip #4 for Southbound Route 71, would be the southbound direction's fourth trip of the day for Route 71.)

APPENDIX B
TRANSACTION RECORD VARIABLE CODES

Mode

Variable Code	Code Value
8	Ferryboat
9	Commuter Rail
128	Bus (on-board payment)
249	Streetcar
250	Bus (off-board payment, always BRT)
251	Light Rail
254	Demand Response

Service Provider

Variable Code	Code Value
2	Community Transit
3	Everett Transit
4	King County Metro
5	Kitsap Transit
6	Pierce Transit
7	Sound Transit
8	Washington State Ferries

Passenger Type

Variable Code	Code Value
1	Adult
2	Youth
3	Senior
4	Disabled
5	Low Income

Stop_Error

Variable Code	Code Value
-1	Transaction not processed (typically means that there is not a corresponding AVL record for this Transaction.)
1	No AVL time found within five minutes of the ORCA boarding time
2	An AVL record exists, but no valid GPS location is associated with that time (AVL GPS = ERR)
4	Vehicle_ID not found
5	Service provider has no AVL data

SubsidySize

Variable Code	Code Value
0	None
1	< 50 percent
2	>50 percent but less than 100 percent
3	Free pass

Product or Pass Type

Variable Code	Product or Pass Type
1	Purse
11	PugetPass $0.50
12	PugetPass $0.75
13	PugetPass $1.00
14	PugetPass $1.25
15	PugetPass $1.50
16	PugetPass $1.75
17	PugetPass $2.00
18	PugetPass $2.25
19	PugetPass $2.50
20	PugetPass $2.75
21	PugetPass $3.00
22	PugetPass $3.25
23	PugetPass $3.50
24	PugetPass $3.75
25	PugetPass $4.00
26	PugetPass $4.25
27	PugetPass $4.50
28	PugetPass $4.75

29	PugetPass $5.00
30	PugetPass $5.25
31	PugetPass $5.50
52	All Day Pass
66	Metro Monthly Access Pass
71	Metro Monthly Adult Pass Offpeak
72	Metro Monthly Adult Pass Peak
100	KCM Ten Ride Pass
131	Everett Transit Monthly Reduced Fare Pass
151	Kitsap Transit Full Fare Pass
152	Kitsap Transit Reduced Fare Pass
154	Kitsap Transit Worker/Driver Full Fare Pass
201	WSF Mukilteo-Clinton Monthly Pass
202	WSF Vashon Island Monthly Pass
203	WSF Fauntleroy-Southworth Monthly Pass
204	WSF Central Sound Monthly Pass
231	WSF Mukilteo-Clinton Passenger 10-Ride
232	WSF Vashon Island Passenger 10-Ride
233	WSF Fauntleroy-Southworth Passenger 10-Ride
234	WSF Central Sound Passenger 10-Ride
235	WSF Anacortes/Lopez 10-Ride
352	PT Adult All-Day Pass
353	PT Youth/Senior/Disabled All-Day Pass
402	Regional Bus and Train Passport
411	Community Transit Business Passport
412	Everett Transit Business Passport
413	Kitsap Transit Business Passport
414	King County Metro Business Passport
415	Pierce Transit Business Passport
416	Sound Transit Business Passport

APPENDIX C
DATABASE SCHEMA

Table: orca-transactions

ID	int(10) UN AI PK
csn	varchar(32)
bid	varchar(32)
date	date
time	time
mode_id_from	int(10) UN
mode_id_to	int(10) UN
num_passengers	smallint(5) UN
passenger_type	int(10) UN
product_id	int(10) UN
direction_from	varchar(16)
direction_to	varchar(16)
route_id_from	varchar(16)
route_id_to	varchar(16)
service_id_from	int(10) UN
service_id_to	int(10) UN
vehicle_id	varchar(16)
trip_id_to	varchar(16)
trip_start	smallint(5) UN
stop_id	varchar(16)
stop_location	point
stop_error	smallint(6)
avl_id	varchar(12)
purse_added	smallint(6)

Table: csns

ID	int(10) UN AI PK
csn	varchar(45)

Table: stops_qtr_mile_directional

ID	int(10) UN AI PK
stop1_agency	varchar(6)
stop1_stopid	bigint(20)
stop1_routeid	varchar(16)
stop1_direction	varchar(16)
stop2_agency	varchar(6)
stop2_stopid	bigint(20)
stop2_routeid	varchar(16)
stop2_direction	varchar(16)
distance	double

Table: qtr_mile_stops

ID	int(10) UN AI PK
stop1_agency	varchar(6)
stop1_stopid	bigint(20)
stop1_longitude	double
stop1_latitude	double
stop1_routeid	varchar(16)
stop1_direction	varchar(16)

Table: subsidy

ID	int(11) AI PK
InstitutionID	varchar(45)
CompName	varchar(200)
TransitPass	int(11)
Subsidy	int(11)
SubsidySize	int(11)

Table: orca-institution

ID	int(10) UN AI PK
csn	varchar(32)
institution	varchar(32)

Table: kcm-avl

Column	Type
ID	int(10) UN AI
DAYCODE	mediumint(8) UN
TRIP_ID	int(10) UN
PATTERN_ID	int(10) UN
PATTERN_QUALITY	tinyint(3) UN
BLK	mediumint(8) UN
RTE	smallint(5) UN
DIR	char(2)
SCH_ST_MIN	smallint(5) UN
OPD_DATE	date
VEHICLE_ID	mediumint(8) UN
STOP_ID	mediumint(8) UN
STOP_SEQ	tinyint(3) UN
STOP_NAME	varchar(256)
SCH_STOP_SEC	mediumint(8) UN
ACT_STOP_ARR	mediumint(8) UN
SCH_STOP_TM	time
ACT_STOP_TM	time
DWELL_SEC	mediumint(8) UN
DOORS_OPEN	enum('Y','N')
DOOR_OPEN_SEC	smallint(5) UN
APC_VEH	enum('Y','N')
ONS	tinyint(3) UN
OFFS	tinyint(3) UN
LOAD	tinyint(3) UN
POINT_ROLE	varchar(8)
GPS_LONGITUDE	float
GPS_LATITUDE	float
stop_datetime	datetime
direction_to	varchar(12)
longitude	double
latitude	double

Table: ct-avl

Column	Type
ID	int(10) UN AI PK
VehicleID	mediumint(8) UN
StopID	smallint(5) UN
Date	date
ArrivalTime	time
DepartureTime	time
Route	smallint(5) UN
Direction	varchar(12)
ONs	smallint(5) UN
OFFs	smallint(5) UN
BalancedONs	smallint(5) UN
BalancedOFFs	smallint(5) UN
Load	smallint(5) UN
GPS_LOCATION	point
stop_datetime	datetime
longitude	double
latitude	double
trip_id	varchar(25)

Table: pt-avl

Column	Type
ID	int(10) UN AI PK
current_route_id	smallint(5) UN
TransitDateValue	date
apc_date_time	datetime
block_id	smallint(5) UN
direction_description	varchar(12)
StartTripTime_Trip	time
ons	smallint(5) UN
offs	smallint(5) UN
GPS_LOCATION	point
bs_sname	smallint(5) UN
bs_lname	varchar(256)
Month	smallint(6)
ext_trip_id	mediumint(8) UN
Duplicate	varchar(10)
description	varchar(12)
sched_version	smallint(5) UN
TripCount	smallint(5) UN
import_error	smallint(5) UN
import_trip_error	smallint(5) UN
TripStartHour	smallint(5) UN

Table: orca-link / orca-sounder

Column	Type
ID	int(10) UN AI PK
business_date	date
csn	varchar(32)
origin	varchar(20)
destination	varchar(20)
passengers	int(10) UN
passenger_type	varchar(32)
product_desc	varchar(64)
settlement_date	date
txn_date_time	datetime
txn_desc	varchar(128)
txn_value	double

Table: link-od / sounder-od

Column	Type
ID	int(10) UN AI PK
csn	varchar(32)
business_date	date
origin	varchar(20)
destination	varchar(20)
origin_datetime	datetime
destination_datetime	datetime
direction	varchar(45)

Table: subsidy

Column	Type
ID	int(11) AI PK
InstitutionID	varchar(45)
CompName	varchar(200)
TransitPass	int(11)
Subsidy	int(11)
SubsidySize	int(11)

Table: orca-service-key

Column	Type
ID	int(10) UN AI PK
service	varchar(32)

Table: **orca-mode-key**			Table: **orca-passenger-type-key**			Table: **orca-product-key**	
ID	int(10) UN AI PK		**ID**	int(10) UN AI PK		**ID**	int(10) UN AI PK
mode	varchar(32)		**passenger-type**	varchar(32)		**product**	varchar(64)

Output Tables

Table: **od-trip** / **od-trip-nodestination**			Table: **od-transfer**	
id	int(11) AI PK		**id**	int(11) AI PK
csn	varchar(255)		csn	varchar(255)
bid	varchar(255)		bid	varchar(255)
trip_start_date	date		transfer_boarding_date	date
trip_start_time	time		transfer_boarding_time	time
origin_mode	varchar(12)		origin_mode	int(11)
destination_mode	varchar(12)		destination_mode	int(11)
num_passengers	int(11)		num_passengers	int(11)
passenger_type	int(11)		passenger_type	int(11)
product_id	int(11)		product_id	int(11)
first_direction	varchar(12)		first_direction	varchar(255)
first_route_taken	varchar(64)		first_route_taken	varchar(255)
first_service_provider	int(11)		first_service_provider	int(11)
origin	varchar(255)		first_trip_id	varchar(255)
origin_service_agency	varchar(12)		exit_stop_id	varchar(255)
origin_lat	varchar(255)		exit_time	varchar(255)
origin_lon	varchar(255)		exit_lat	varchar(255)
destination	varchar(255)		exit_lon	varchar(255)
destination_service_agency	varchar(12)		transfer_boarding_stop_id	varchar(255)
destination_lat	varchar(255)		transfer_stop_lat	varchar(255)
destination_lon	varchar(255)		transfer_stop_lon	varchar(255)
destination_time	varchar(32)		transfer_route_taken	varchar(12)
last_route_taken	varchar(32)		transfer_direction	varchar(255)
last_direction	varchar(12)		transfer_newservice_provider	varchar(12)
last_service_provider	varchar(32)		number_of_transfers	int(11)
last_mode	varchar(32)		transfer_to_vehicle_id	varchar(255)
number_of_transfers	int(11)		last_trip_id	varchar(12)
last_vehicle_id	varchar(255)		transfer_duration	time
last_trip_id	varchar(32)			
transfer_required	varchar(32)			
trip_start	varchar(32)			
trip_duration	time			