TMIP VISION

TMIP provides technical support and promotes knowledge and information exchange in the transportation planning and modeling community.

DISCLAIMER

The views and opinions expressed during this webinar are those of the presenters and do not represent the official policy or position of FHWA and do not constitute an endorsement, recommendation or specification by FHWA. The webinar is based solely on the professional opinions and experience of the presenters and is made available for information and experience sharing purposes only.

Texas A&M Transportation Institute

TMIP Webinar:
What does (data) integrity and utility mean to you? Painstaking attention to detail comes to mind.

Elaine Murakami & Stacey Bricka

August 21, 2014

Webinar Overview

- Quality Control Resources
- Real-Time QC During Data Collection
- Geocoding / Location Data QC
- Post-Collection Adjustments
- Discussants
- General Q&A

Quality Control Resources

- Travel Survey Manual (Chapters 4, 11, 13)
  - (www.travelsurveymanual.org)
- NCHRP Report 571
  - http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_571.pdf
- Agency Specific Memos
  - http://www.azmag.gov/Documents/TRANS_2012-02-17_2008-National-Household-Travel-Survey-Dataset-for-MAG-Region.pdf

Presentation #1
Real-Time QC During Data Collection

New York Metropolitan

Transportation Council

Sangeeta Bhowmick, NYMTC

Kyeongsu Kim, Louis Berger

Texas A&M Transportation Institute

Presentation #2
Geocoding / Location Data and other QC

Delaware Valley Regional Planning Commission



Christopher Puchalsky, PhD

Benjamin Gruswitz, AICP

Sarah Moran

Presentation #3
Post-Collection Adjustments

Metropolitan Transportation Council

Shimon Israel

Discussants

Oregon Department of Transportation

Christina McDaniel-Wilson

Becky Knudson (Oregon DOT)

NYMTC's Experience in RHTS:

Survey Management by Bi-weekly Report Monitoring and Systematic Monthly Dataset Review

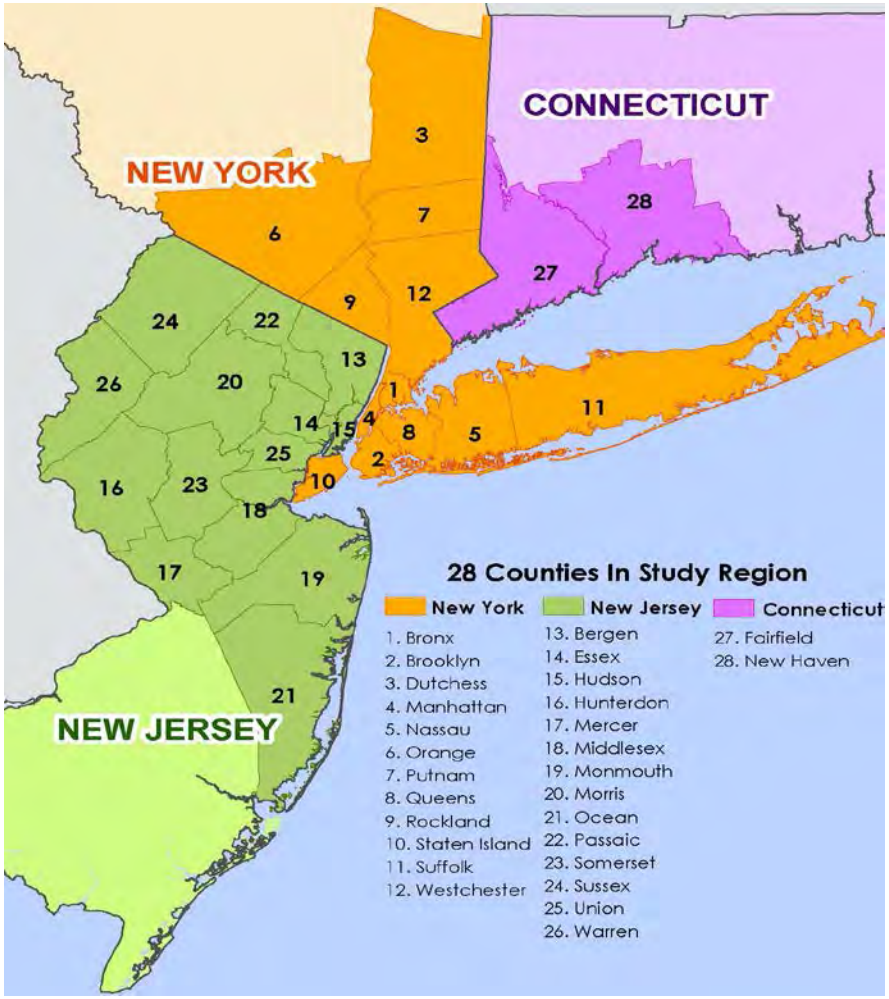
Presented at TMIP Webinar- What does (data) integrity and utility mean to you?
August 21, 2014

Ms. Sangeeta BHOWMICK Acting Director, NYMTC Technical Unit

Mr. KyeongSu KIM Louis Berger, On-site Consultant at NYMTC

2010/2011 Regional Household Travel Survey (RHTS)

- ✓ Jointly sponsored by The New York Metropolitan Transportation Council (NYMTC) & the North Jersey Transportation Planning Authority (NJTPA)



RHTS Study Area

28 Tri-State Counties

- ❑ 12 NY
- ❑ 14 NJ
- ❑ 2 CT

✓ Recruitment - CATI or mail
✓ Retrieval – CATI, mail, or TripBuilder
✓ Available in English, Spanish, Russian and Chinese
✓ GPS subsample: improved accounting for short, non-work walk trips

Available Datasets

- **Household**: 18,965 households (1,104 zero-trip HHs, 5.8%)
- **Person**: 43,558 participants
- **Vehicle**: 29,043 household vehicles
- **Place**: 231,715 unique places
- **Unlinked Trip**: 188,199 unlinked trips or trip-segments.
- **Linked Trip:** 143,925 trips

http://nymtc.org/project/surveys/survey.html

Key Findings

- Slightly more than 82% of all trips in the study area were intra-county, an increase from 78% in the 1997/1998 survey.
- Most intra- and inter- county trips were made by automobile (67% and 95%, respectively), while 66% of travel to Manhattan was often made by rail.
- Manhattan, the other boroughs of New York City, and Hudson County New Jersey had the highest percentages of non-motorized trips within their physical areas (56%, 32% and 31%, respectively).
- Public transit serves 8% of all weekday trips in the region.
- Over 8% of commute trips into Manhattan use some form of public transit.
- 54% of all trips are between home and destinations other than work (e.g., social/recreation, shopping, school, etc.); 23% of trips involve the workplace.
- Work trips in the region normally took between 32 and 35 minutes, with work trips from Manhattan averaging 30 minutes, while work trips from the other NYC boroughs averaged 42 minutes (the high in the region).

Monitoring Bi-weekly progress report

Key QA/QC Tables

- Recruit/Retrieval Productivity
- Special (language) sample
- **Zero-trip monitoring**
- County of household
- Household size and employment status
- Demographics (income, type of phone, language, ethnicity, gender, and age group)

NYMTC's Major Watch List.

1. **Zero-trip monitoring**
2. **Senior sample participation rate**
3. Progress of sample response by counties and bins
   - Minimum of 271 sample for both county level and Census tract-based 21 sampling Bins. (90 Confidence Level & +/-5 % CI)

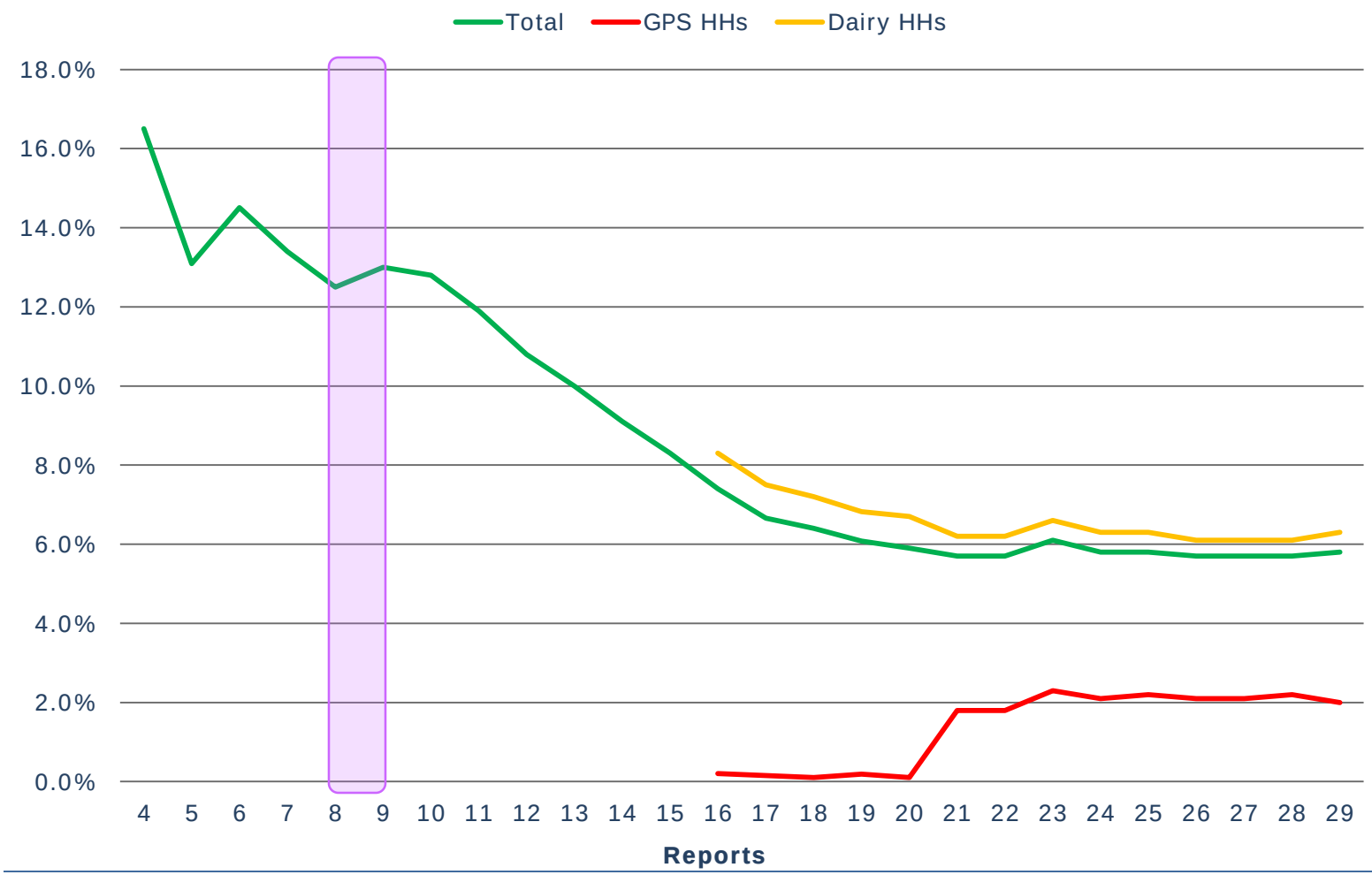
TABLE Zero-Trip Monitoring

At household level	Count	%	GPS	%	Dairy	%
Traveled	17,862	94.2%	1,896	98.%	15,966	93.7%
Did Not Travel	**1,104**	**5.8%**	**39**	**2.%**	**1,065**	**6.3%**
Total	**18,966**	**10%**	**1,935**	**10%**	**17,031**	**10%**

TABLE Demographics by State

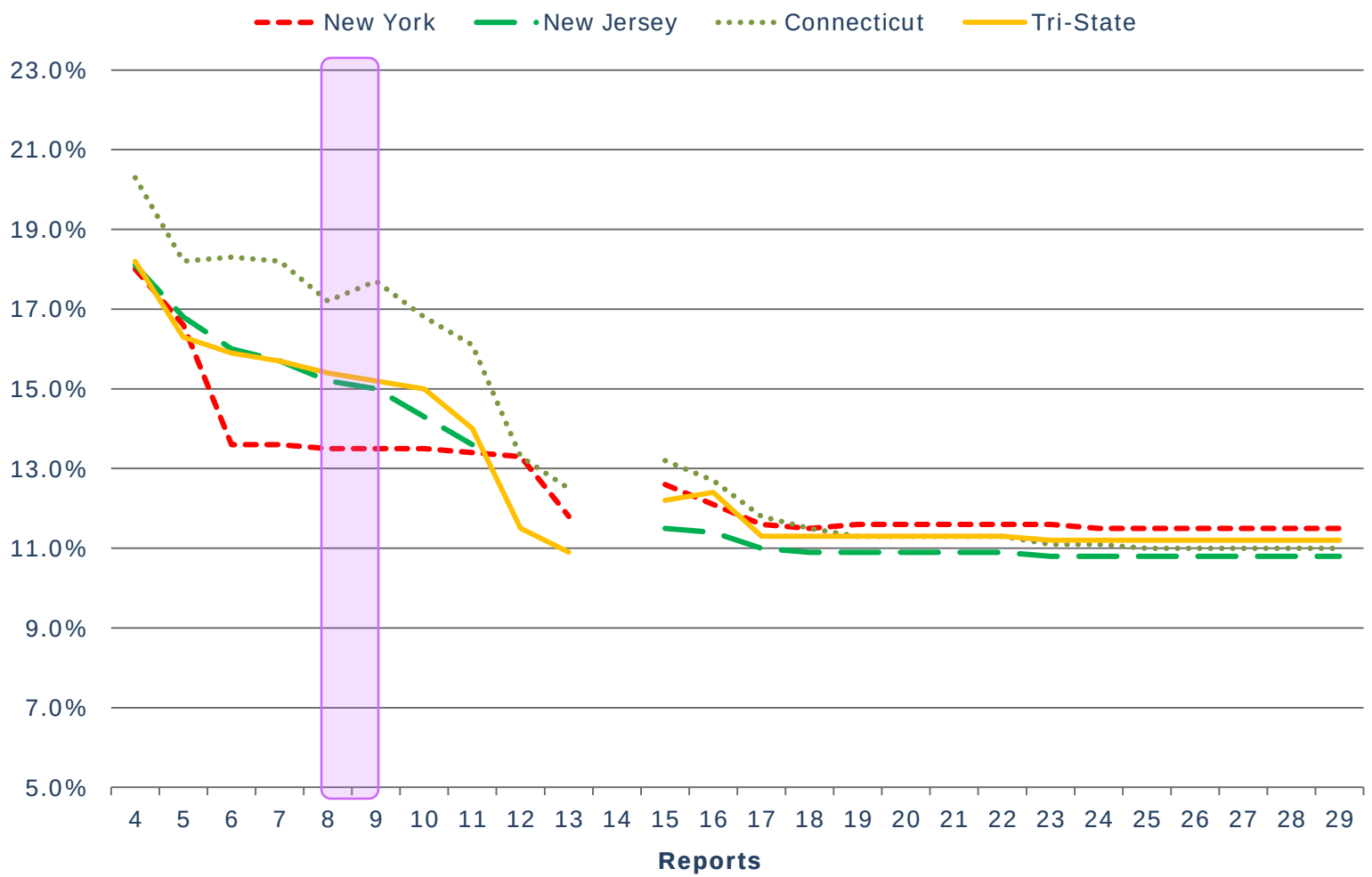
Demographics of Completes	New York			New Jersey			Connecticut			Total		
	REC	RET	ACS	REC	RET	ACS	REC	RET	ACS	REC	RET	ACS
Total Household	17,138	10,129	4,639,082	12,591	7,905	2,465,914	1,425	927	646,087	31,154	18,961	7,751,083
Respondent Age												
Less than 18	21.3%	19.%	23.5%	23.3%	21.3%	23.8%	22.0%	19.8%	24.3%	22.2%	20.0%	23.7%
18-24	7.7%	5.5%	9.6%	6.6%	4.8%	8.8%	6.2%	5.1%	8.9%	7.2%	5.2%	9.3%
25-54	41.1%	40.9%	43.3%	41.1%	40.2%	43.2%	40.8%	40.3%	42.1%	41.1%	40.6%	43.2%
55-64	18.4%	21.9%	10.9%	18.3%	21.5%	11.1%	19.9%	22.4%	11.3%	18.4%	21.7%	11.0%
65+	**11.5%**	**12.7%**	**12.7%**	**10.8%**	**12.1%**	**13.1%**	**11.0%**	**12.4%**	**13.4%**	**11.2%**	**12.4%**	**12.9%**
Don't Know or Refused	3.4%	1.9%	-	3.1%	1.4%	-	2.7%	2.1%	-	3.3%	1.7%	-

WATCH LIST 1:
PERCENTAGE OF ZERO-TRIP HOUSEHOLDS



Note: adding GPS portion was requested from report 16; exclude data in report 1 to 3 (initial periods of data collection)

WATCH LIST 2:
CHANGE IN SENIOR PARTICIPATION RATE (RECRUITMENT)



Note: Exclude data in report 14 due to mismatching total estimates; exclude data in report 1-3 (initial periods of data collection)

WATCH LIST 3:
PROGRESS OF SAMPLE RESPONSE BY COUNTY & BIN

BINS	1	2	3	4	5	6	7	8	9	1	11	12	13	14	15	16	17	18	19	2	21	Total Retrieved	TOTAL GOAL	% Retrieved		CATI/Web Retreived	CATI/Web Goal	% Retrieved		GPS Subsample Retrieved	GPS Retrieval Goal	%
NYMTC County																																
MANHATTAN	0	0	0	0	0	296	0	0	0	696	302	0	19	0	0	0	94	49	0	141	0	1597	**1,511**	*105.7%*		1366	1360	*100.5%*		231	151	*153.0%*
QUEENS	0	0	0	405	0	0	0	0	173	0	0	450	0	0	0	26	51	0	0	0	0	1105	**1,292**	*85.6%*		995	1163	*85.6%*		110	129	*85.3%*
BRONX	0	0	0	0	0	0	434	0	0	0	0	0	110	0	104	0	0	0	112	0	30	790	**1,094**	*72.2%*		691	985	*70.2%*		99	109	*90.8%*
BROOKLYN	0	0	441	0	0	0	0	624	0	0	0	118	0	0	0	8	23	0	11	0	0	1225	**1,323**	*92.6%*		1057	1191	*88.8%*		168	132	*127.3%*
STATEN ISLAND	18	5	0	43	0	0	0	0	0	0	0	0	81	0	0	63	131	60	0	0	0	401	**448**	*89.5%*		366	403	*90.8%*		35	45	*77.8%*
NASSAU	0	0	0	0	265	0	0	0	0	0	0	0	219	0	0	0	337	62	21	19	0	923	**1,062**	*86.9%*		830	956	*86.8%*		93	106	*87.7%*
SUFFOLK	0	0	189	0	0	0	0	0	0	0	0	145	0	0	0	700	6	0	0	19	36	1095	**1,211**	*90.4%*		957	1090	*87.8%*		138	121	*114.0%*
WESTCHESTER	0	87	0	0	0	0	0	0	0	0	215	0	0	182	0	0	139	0	0	0	42	665	**770**	*86.3%*		590	693	*85.1%*		75	77	*97.4%*
ROCKLAND	0	0	24	0	0	0	0	0	0	0	22	0	0	0	118	0	13	8	0	0	87	272	**312**	*87.0%*		241	281	*85.6%*		31	31	*100.0%*
PUTNAM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	226	0	38	264	**271**	*97.4%*		237	244	*97.1%*		27	27	*100.0%*
ORANGE	0	0	0	0	0	0	0	0	145	0	0	0	0	0	0	49	0	61	64	0	0	319	**349**	*91.3%*		288	314	*91.6%*		31	35	*88.6%*
DUTCHESS	0	0	0	0	0	0	0	0	0	0	0	0	0	190	0	0	0	78	0	95	26	389	**458**	*84.9%*		339	412	*82.3%*		50	46	*108.7%*
FAIRFIELD	198	0	0	0	0	0	0	0	0	101	0	0	132	0	0	16	0	0	0	0	0	447	**456**	*97.9%*		396	410	*96.5%*		51	46	*110.9%*
BERGEN	0	0	99	0	0	0	0	0	0	0	0	80	0	0	158	0	388	10	0	121	31	887	**989**	*89.7%*		783	890	*88.0%*		104	99	*105.1%*
PASSAIC	0	0	0	59	0	0	0	0	0	0	0	0	42	0	0	105	127	0	0	0	0	333	**432**	*77.0%*		295	389	*75.7%*		38	43	*88.4%*
HUDSON	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	435	0	242	73	0	45	795	**1,042**	*76.3%*		692	938	*73.8%*		103	104	*99.0%*
ESSEX	0	0	0	0	0	143	22	0	0	0	0	0	0	110	0	0	182	54	0	31	84	626	**758**	*82.6%*		561	682	*82.3%*		65	76	*85.5%*
UNION	0	0	52	0	0	0	0	0	0	0	0	65	0	0	69	0	196	0	0	0	58	440	**548**	*80.3%*		395	493	*80.1%*		45	55	*81.8%*
MORRIS	0	0	107	0	0	0	0	0	0	0	0	71	0	0	124	0	53	48	0	133	0	536	**488**	*109.8%*		482	439	*109.7%*		54	49	*110.2%*
SOMERSET	0	0	0	0	0	109	0	0	0	0	0	0	15	0	0	0	160	52	0	0	0	336	**297**	*113.1%*		304	267	*113.8%*		32	30	*106.7%*
MIDDLESEX	0	0	0	56	0	0	0	0	0	0	0	292	0	0	0	179	81	0	153	0	0	761	**749**	*101.7%*		697	674	*103.5%*		64	75	*85.3%*
MONMOUTH	49	39	0	0	0	0	0	0	0	17	0	0	0	73	0	183	0	0	23	0	235	619	**704**	*87.9%*		546	634	*86.1%*		73	70	*104.3%*
OCEAN	88	0	0	0	0	0	5	0	0	0	0	0	112	0	83	0	0	0	0	217	0	505	**602**	*83.9%*		448	542	*82.6%*		57	60	*95.0%*
HUNTERDON	0	0	0	0	0	0	0	0	0	3	0	0	0	0	66	0	0	71	0	11	179	327	**287**	*113.9%*		298	259	*115.0%*		29	28	*103.6%*
WARREN	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	192	0	0	50	50	292	**271**	*107.7%*		266	244	*109.0%*		26	27	*96.3%*
SUSSEX	0	0	0	0	0	0	0	0	0		0	0	0	149	0	0	0	0	0	0	191	340	**326**	*104.2%*		308	293	*105.0%*		32	33	*97.0%*
NEW HAVEN	0	203	0	0	0	0	0	0	0	72	0	0	0	155	0	0	6	0	0	0	0	436	**467**	*93.3%*		394	420	*93.8%*		42	47	*89.4%*
MERCER	0	0	0	0	0	0	0	0	185	0	0	0	0	0	0	92	0	55	0	0	0	332	**282**	*117.9%*		295	254	*116.3%*		37	28	*132.1%*
TOTAL RETRIEVED	353	334	912	563	265	548	461	624	503	886	539	1221	730	859	722	1856	2179	850	683	837	1132	17057	**18,800**	*90.7%*		15117	16921	*89.3%*		**1,940**	**1,879**	*103.2%*
TOTAL GOAL	**330**	**328**	**934**	**592**	**255**	**577**	**560**	**673**	**573**	**822**	**554**	**1356**	**803**	**1020**	**895**	**1931**	**2450**	**916**	**806**	**1079**	**1347**											
% Retrieved	*106.9%*	*101.8%*	*97.7%*	*95.1%*	*104.0%*	*94.9%*	*82.4%*	*92.8%*	*87.8%*	*107.8%*	*97.3%*	*90.1%*	*90.9%*	*84.3%*	*80.6%*	*96.1%*	*88.9%*	*92.8%*	*84.7%*	*77.6%*	*84.1%*											

1

BINS	1	2	3	4	5	6	7	8	9
NYMTC County									
MANHATTAN	0	0	0	0	0	296	0	0	0
QUEENS	0	0	0	405	0	0	0	0	173
BRONX	0	0	0	0	0	0	434	0	0
BROOKLYN	0	0	441	0	0	0	0	624	0
STATEN ISLAND	18	5	0	43	0	0	0	0	0
NASSAU	0	0	0	0	265	0	0	0	0

2

21	Total Retrieved	TOTAL GOAL	% Retrieved		CATI/Web Retreived	CATI/Web Goal	% Retrieved		GPS Subsample Retrieved	GPS Retrieval Goal	%
0	1597	**1,511**	*105.7%*		1366	1360	*100.5%*		231	151	*153.0%*
0	1105	**1,292**	*85.6%*		995	1163	*85.6%*		110	129	*85.3%*
30	790	**1,094**	*72.2%*		691	985	*70.2%*		99	109	*90.8%*
0	1225	**1,323**	*92.6%*		1057	1191	*88.8%*		168	132	*127.3%*
0	401	**448**	*89.5%*		366	403	*90.8%*		35	45	*77.8%*
0	923	**1,062**	*86.9%*		830	956	*86.8%*		93	106	*87.7%*
36	1095	**1,211**	*90.4%*		957	1090	*87.8%*		138	121	*114.0%*

3

HUNTERDON	0	0	0	0	0	0	0	0	0
WARREN	0	0	0	0	0	0	0	0	0
SUSSEX	0	0	0	0	0	0	0	0	0
NEW HAVEN	0	203	0	0	0	0	0	0	0
MERCER	0	0	0	0	0	0	0	0	185
TOTAL RETRIEVED	353	334	912	563	265	548	461	624	503
TOTAL GOAL	**330**	**328**	**934**	**592**	**255**	**577**	**560**	**673**	**573**
% Retrieved	*106.9%*	*101.8%*	*97.7%*	*95.1%*	*104.0%*	*94.9%*	*82.4%*	*92.8%*	*87.8%*

Initial Routine Checks

- Ordinary Routines
  - Data interaction b/w HH, Person, Vehicle, and Place files
  - loop trips, OD match and auto driver-passenger combination
  - Geographies and others

Systematic Routine Checks

- Speed (by auto, non-motorized)
- Long distance trips

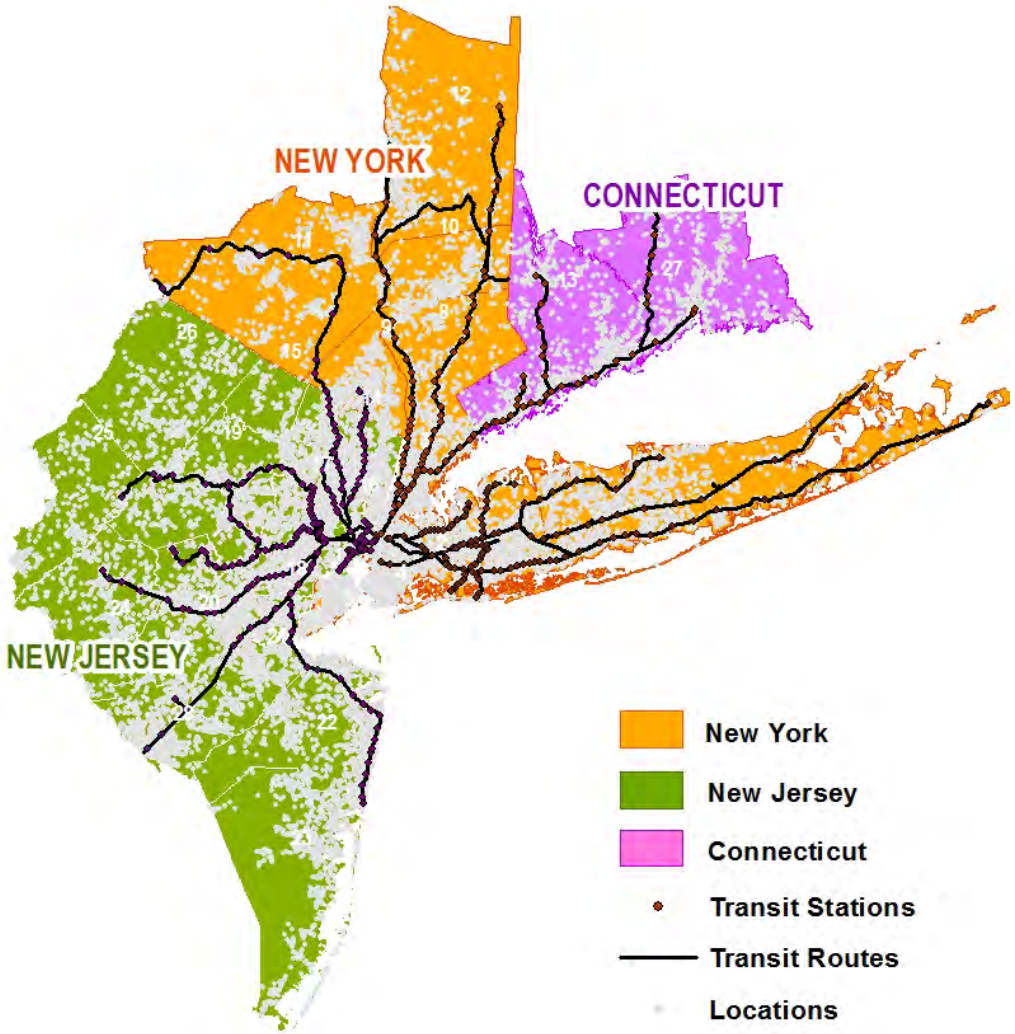
Systematic Routine Checks

1. **Transit Trips**
   - **Checked the availability of transit service between OD (from an unlinked trip*) in a municipality level**
     - prepared a database (look-up table) that has information of commuter rail stations with municipality FIPS codes (tagged in ArcGIS)
     - displayed trip ODs using x,y coordinates in ArcGIS; then tagged municipality FIPS codes
     - exported the attribute table; developed scripts for validity check in STATA (merge)
     - flag if transit is not available or wrong operator or service info

       turns out that many had (1) missing station access/egress trip segment from/to home, workplace, or other places, and(2) invalid or missing service operator information; corrected them accordingly.
2. **Public transportation in general** (including buses, subway, light rail, ferry and etc.)
   - **Reviewed trip OD, operator, service route, fare, pass type and unit and others**
     - flag if irrational information (i.e., missing data, incorrect operator, service information, and missed intermodal transfers); corrected them accordingly

** A place file was delivered as a part of an interim dataset. This unlinked trip (trip segment) data format, having OD information in a same row (line), was recreated to facilitate data checks.*

Checked
the availability of transit service



Other Systematic Routine Checks

- Linked Trip*
  - Checked if it is correctly converted from an unlinked trip (originally from a place) file.
    - Travel time,
    - Travel distance,
    - Travel mode,
    - Trip purpose,
    - Different levels of geography,
    - Labels.

** combination of unlinked trips (at least more than 1 unlinked trips)*

Thank You

Questions?

Cleaning Household Surveys

What to Do with Trip Purpose: "Relocating Chipmunks" and Other Fun Tips

August 21, 2014

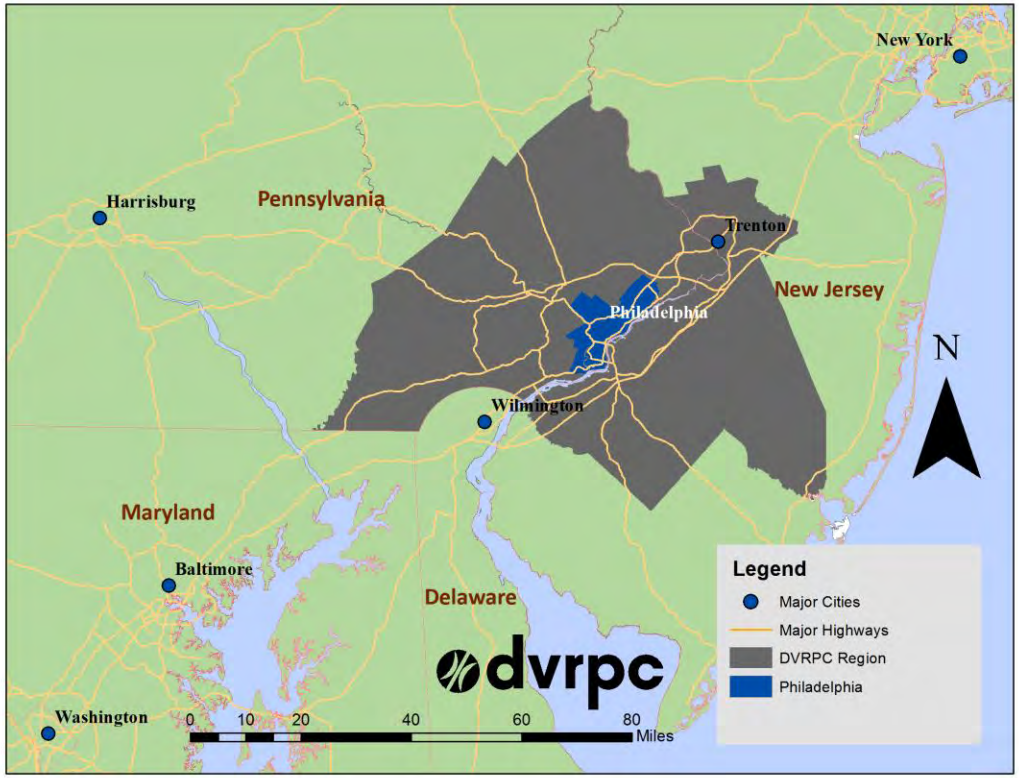
Chris Puchalsky, PhD
Associate Director - Systems Planning

Ben Gruswitz, AICP
Transportation Planner - Office of Modeling & Analysis

Sarah Moran
Intern - Office of Modeling & Analysis

WHAT'S DVRPC?

- Metropolitan Planning Organization (MPO)
  - 2 States
  - 9 Counties
  - 351 Municipalities
  - 5.6 Million Population
  - 3,800 sq. miles
  - ~115 employees
- Activities –
  - Long Range Plan (LRP)
  - Transportation Improvement Program (TIP)
  - Wide range of planning and technical support for regional partners



OUTLINE

- Background on Survey
- Setting yourself up for quality
- What's in those fields?
  - Recoding "other" responses
  - Inputs and outputs for geocoding
- Our "pick a winner" geocode method
- The search for the missing trip

Background on Survey

SURVEY BACKGROUND

- 1-day paper diary survey
- 10,000 households goal, 9,384 actual complete good surveys (almost 900,000 HHs contacted)
- 3 day GPS sub-sample (500 HH goal, 380 actual)
- Address based sampling frame
- 12 month roughly equal sample, weekdays
- State, area-type, HH-size x income, and HH-size x auto ownership as control variables
- Diary data retrieved by either phone, web, or mail

Setting yourself up for quality

TMIP Webinar

QUALITY CONTROL FROM DAY ONE

- Pilot study
- Periodic meetings with contractor (weekly at first, less often later)
- Weekly monitoring reports
- Several preliminary data deliveries
- 3rd party QA/QC contractor (left in middle of project)

IN THE BEGINNING…

- Contractual Data Quality
  - Clear definitions of what constitutes “quality” data (or, as clear as can be)
    - 100% response rates on “key” questions
    - e.g. # of people in HH, mode of trip, work status
    - 90% response on difficulty question – HH income
    - 95% response on all others
    - Geocoding quality standards
    - e.g. 95% of regional locations geocode-able to parcel level
  - As with all contracts, exact wording is important
    - Envision your query

TRANSLATING QUALITY STANDARDS

- Example
  1. Contractual language

     "All modes of transportation identified"
  2. Translate to technical speak

     "MODE is not null and is >= 1 and <= 30. This should be 100% complete—no large household exemptions."
  3. Construct and run query

```
INSERT INTO checking (Description, Result, QueryName,
  [Percent], IsPass, Limit )
SELECT "Is there a Mode? (non-exempt)" AS Epxr2,
  COUNT(*) AS Expr1, "qry_2_w_MODE_check" AS Expr3,
  COUNT(*)/getTableSize('h_complete_trips_nogo_exempt')
  AS Expr4, IIf([Expr4]>=[Limit],"Pass","Fail") AS
  Expr5, 1 AS Limit
FROM 4_Trip
WHERE (({4_Trip].MODE >= 1 AND [4_Trip].MODE <= 30) OR
  [4_Trip].TRIPNUM = 97 OR [4_Trip].PExempt = 1 OR
  [4_Trip].NOGO IS NOT NULL) AND [4_Trip].complete = 1;
```

  4. Review summary table of all queries

dvrpc

HOUSEHOLD TRAVEL SURVEY

QUERY SUMMARY

Result	Data Table	Description	Delivery Count	Delivery Percent	Contract Standard
Pass	Household	Are there at least 9,500 complete HHs?	9,502	100.02%	100%
Pass	Household	Are all HH locations geocoded?	9,502	100.02%	100%
Pass	Household	How many HHs provided their income?	8,856	91.98%	90%
Pass	Person	Do # of people in each complete HH = # of complete person records for that HHID?	9,502	100.00%	100%
Pass	Person	Do all complete HHIDs in Household and Person tables match?	9,502	100.00%	100%
Pass	Person	Are household worker counts less than HH size?	9,502	100.00%	100%
Passable	Person	Are all people over 15 employed or unemployed?	17,962	99.82%	100%
Fail	Person	How many work locations of employed people at fixed work sites geocoded?	8,557	93.86%	95%
Passable	Person	Are all people a either a student or non-student?	21,266	99.89%	100%
Pass	Person	Do all students have a school type?	4,228	100.00%	100%
Pass	Person	How many students have coordinates for their school?	4,037	95.48%	95%
Pass	Vehicle	Do # of vehicles in each complete HH = # of complete vehicle records for that HH?	9,502	100.00%	100%
Pass	Vehicle	Do all HHIDs in Vehicle table match Household table HHIDs?	9,502	100.00%	100%
Pass	Trip	Do # of trips in complete households = # of complete trip records?	9,502	100.00%	100%
Pass	Trip	Do all complete HHIDs in Household and Trip tables match?	9,502	100.00%	100%
Fail	Trip	How many HHs have trips with coordinates or have a household size exemption?	5,719	60.19%	100%
Fail	Trip	How many trips are geocoded among non-exempt households?	59,892	93.17%	100%
Fail	Trip	Is there a mode for each trip? (non-exempt)	64,684	94.79%	100%

Pass Passable Fail

What's in those fields?

Recoding "other" responses

WHAT'S AN "OTHER" RESPONSE?

Work Status Codes & Categories
1. Retired
2. Disabled/on disability status
3. Homemaker
4. Unemployed but looking for work
5. Unemployed and not looking for work
6. Student
7. Volunteer
97. Other
98. Don't know
99. Refused

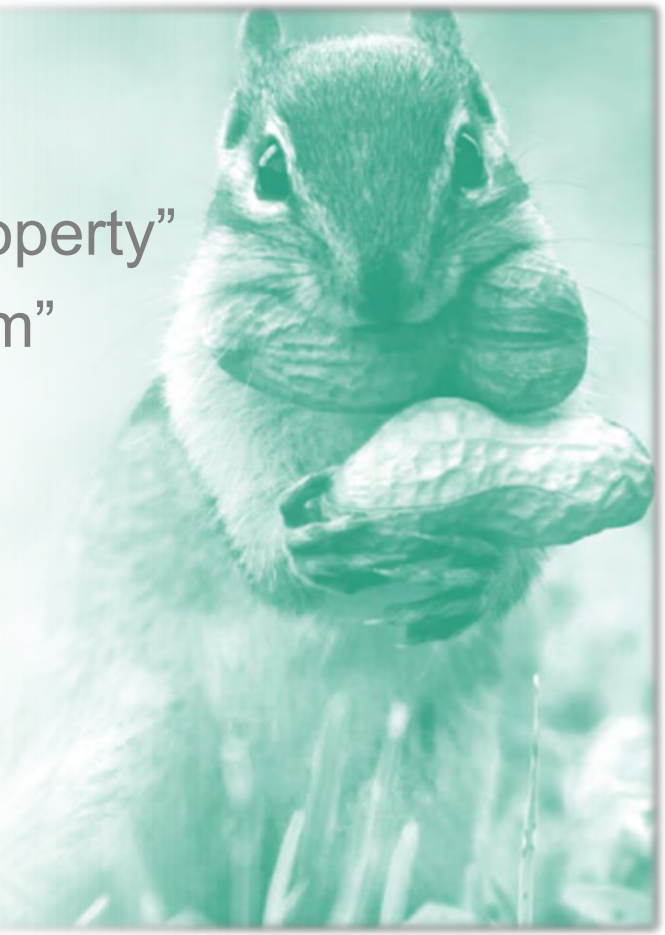
Person ID	Work Status	Work Status - Other
UUUUU	6	<null>
WWWWW	3	<null>
XXXXX	97	freelancer
YYYYY	1	<null>
ZZZZZ	1	<null>

WHAT'S AN "OTHER" RESPONSE?

- "Other" Fields
  - Description of Residence
  - Race
  - Method of Transit Fare Payment
  - Work Status
  - Occupation
  - Level of School
  - Preschool Type
  - Highest Level of Education
  - Mode of Transportation to Work
  - Vehicle Make
  - Vehicle Body Type
  - Vehicle Ownership Status
  - Origin Activity
  - Toll Road Used
  - Toll Bridge Used
  - Parking Cost Unit

NOTABLE "OTHER" ACTIVITY RESPONSES

- Detailed responses
  - "Relocating Chipmunks"
  - "Attending to bee hive on property"
  - "Enjoyed soft serve ice cream"
  - "Dumpster diving"
- Privacy protection
  - "Not your business"
  - "Rather not say"
- TMI
  - "@!#$%^&"



http://www.mrwallpaper.com/wallpapers/Chipmunk-Peanuts.jpg

RECODE: VEHICLE OWNERSHIP

Vehicle Ownership Codes & Categories
1. Owned by household member
2. Leased by household member
3. Owned or leased by employer
4. Owned or leased by person not living in household
97. Other (specify)
98. Don't know
99. Refused

"Other" Responses

"Owned and borrowed against"

"Making payments to own"

"Under finance"

"Borrowed"

"Loaned"

"Company Vehicle"

"Company Car"

"Work Vehicle"

1. Owned by household member

4. Owned or leased by person not living in household

3. Owned or leased by employer

RECODE: LEVEL OF SCHOOL ATTENDING

Level of School Codes & Categories
1. Daycare
2. Nursery school/Preschool
3. Kindergarten to Grade 8
4. Grade 9-12
5. Technical/Vocational school
6. 2 year college
7. 4 year college or university
8. Graduate/Professional
97. Other (specify)
98. Don't know
99. Refused

"Other" Responses

"For retirees"

"Doctoral"

"Life long learning"

"Real Estate School"

"Adult Education"

"Nursing refresher class"

"Over 60 program"

"Craft classes"

8. Graduate / Professional

Not a Student

ALSO FOUND IN "OTHER" FIELDS

Mode of Transportation to Work
- "this diary is blank"
- "this whole diary was not received, disregard"
- "this section is blank, disregard"

Why No Trips Made
- "blank diary"
- "did not receive diary"
- "missing"
- "does not exist"
- "not a household member"
- "respondent did not live in HH on travel date"
- "John moved"

Origin Activity
- "this whole diary is empty"
- "this diary was missing, pushing the case through"

- Had to either
  - adjust household size
  - throughout household and reweight

What's in those fields?

Inputs and outputs for geocoding

TMIP Webinar

GEOCODE QUALITY

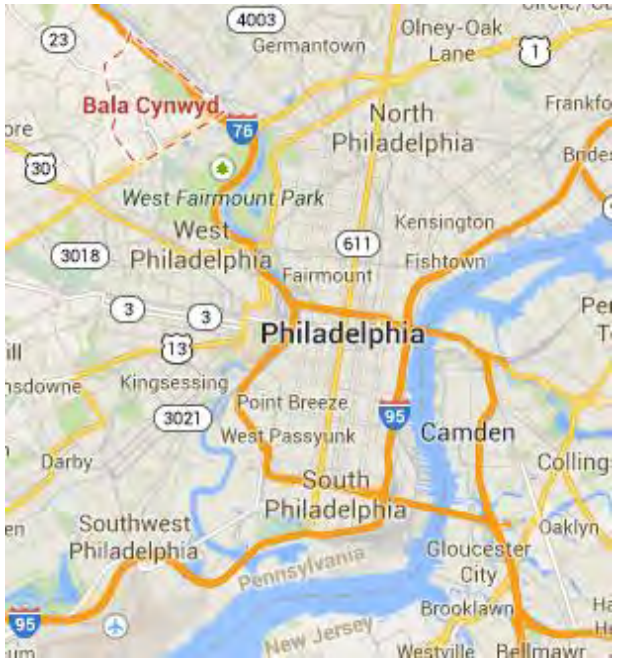
- Database queries
  - Can tell you whether a latitude or longitude field is blank
  - Can't tell you when they're wrong
- Short story of geocoding experience
  - Contractor tasked with geocoding – used esri product
  - We received diary data (place name, address, etc.) and coordinates
  - We looked under the hood and started questioning results
  - Contractor said esri product not what it used to be

GEOCODERS AREN'T ALWAYS THE PROBLEM

- Good address inputs are critical for accurate geocoding
  - "rns" (rather not say) is the airport code for Rennes-Saint-Jacques Airport in France
  - "dk" (don't know) is the country code for Denmark

GEOCODING GREATER PHILADELPHIA

Which is correct?



Baa Cynwyd
Bala Cwynd
Bala Cyawyd
Bala Cyn Wyd
Bala Cynwd
Bala Cynwood
Bala Cynwy
Bala Cynwyd
Bala Cynwyo
Bala Lynwyd

Balacynwid
Balacynwld
Balacynwyd
Balacywyd
Balan Cynwyd
Balla Cynwid
Balla Cynwyd
Ballard Cynwood
Bela Cynwyd

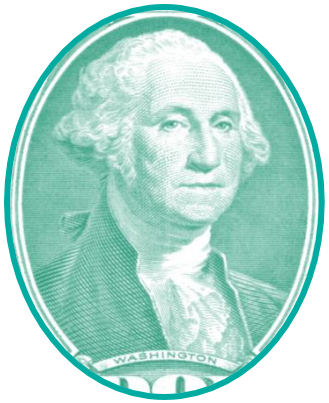
GEOCODING GREATER PHILADELPHIA

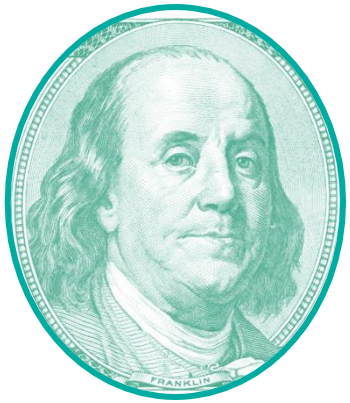
Other strange names in Pennsylvania

- Conshohocken
  - Cons Hockey, Conschahawken, Conshohocker
- Quakertown
  - Quacker Town, Wuakertown
- And don’t confuse
  - Wyalusing, Wycombe, Wyncote, Wyndmoor, Wynnewood, Wyoming, and Wyomissing

GEOCODING GREATER PHILADELPHIA

- Townships and Boroughs don't always have original names
  - Municipalities named after:
    - Washington
      - 22 in Pennsylvania
      - 5 in New Jersey
    - Franklin
      - 18 in Pennsylvania
      - 5 in New Jersey
  - Need a zip code or at least a unique street address





SO PHILLY'S WEIRD, SO WHAT?

- You too might have some geographic peculiarities
- Locations might be best handled by locals
- What if you can't do it yourself?
  - Have contractor provide
    - Detailed procedures on cleaning inputs
    - Geocoder's match score, geocoder used, flags for locations they overwrite
    - Original and cleaned address as well as geocoder's matched address (output)
  - You'll want to
    - Spatial join to find out if it matches input's state, city, zip code, etc.
    - Spot check a sampling of coordinates
    - Investigate identical coordinates

Our "pick a winner" geocode method

GOOGLE/BING AUTOMATED PROCESS

Welcome to GeoPy's docu

geopy.readthedocs.org/en/1.1.0/

GeoPy

Search docs

- Welcome to GeoPy's documentation!
- Geocoders
- Calculating Distance
- Data
- Exceptions
- Logging
- Indices and search

Read the Docs v: 1.1.0

Docs » Welcome to GeoPy's documentation! Edit on GitHub

Welcome to GeoPy's documentation!

geopy is a Python 2 and 3 client for several popular geocoding web services.

geopy makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders and other data sources.

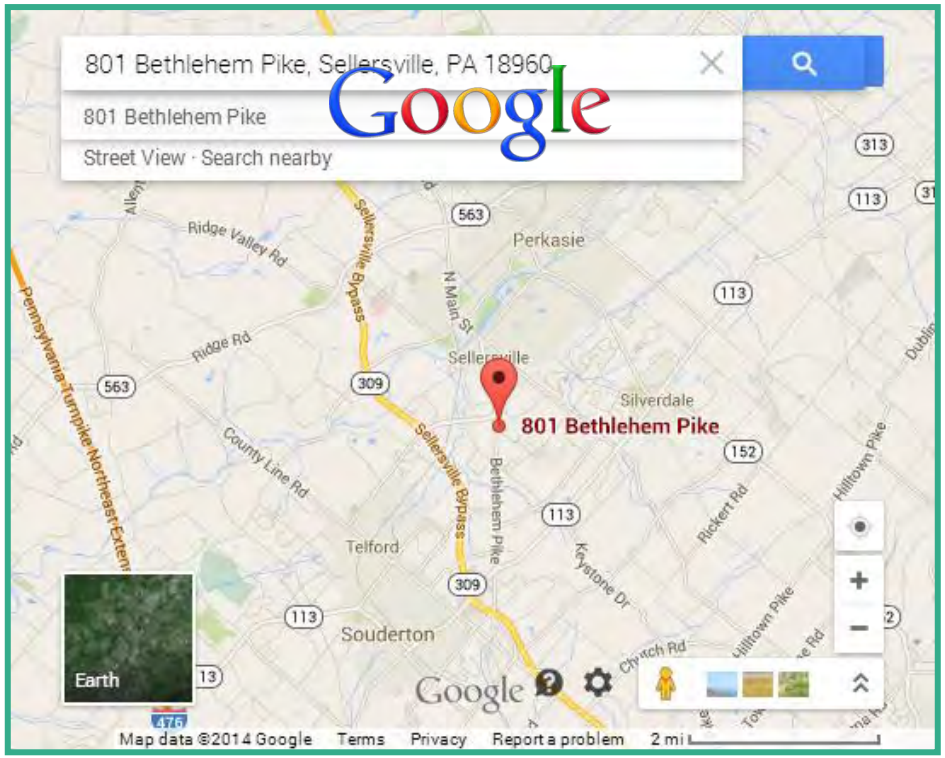
Geocoders

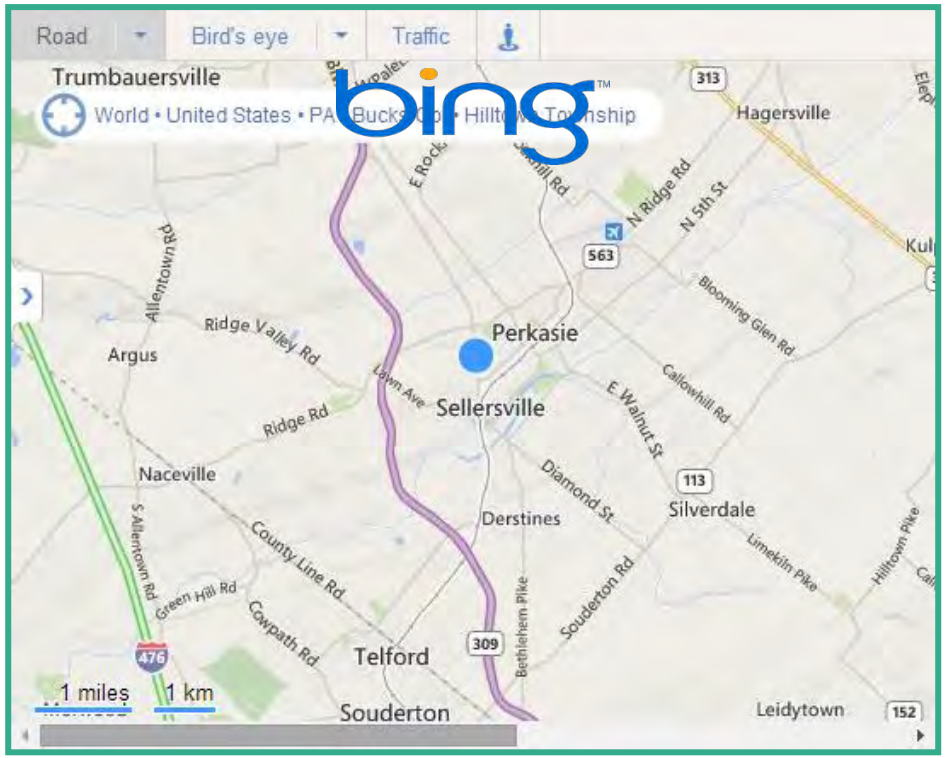
Each geolocation service you might use, such as Google Maps, Bing Maps, or Yahoo BOSS, has its own class in `geopy.geocoders` abstracting the service's API. Geocoders each define at least a `geocode` method, for resolving a location from a string, and may define a `reverse` method, which resolves a pair of coordinates to an address. Each Geocoder accepts any credentials or settings needed to interact with its service, e.g., an API key or locale, during its initialization.

To geolocate a query to an address and coordinates:

GOOGLE/BING AUTOMATED PROCESS

- Example: 801 Bethlehem Pike, Sellersville, PA 18960 (place name = “AT Subaru”)

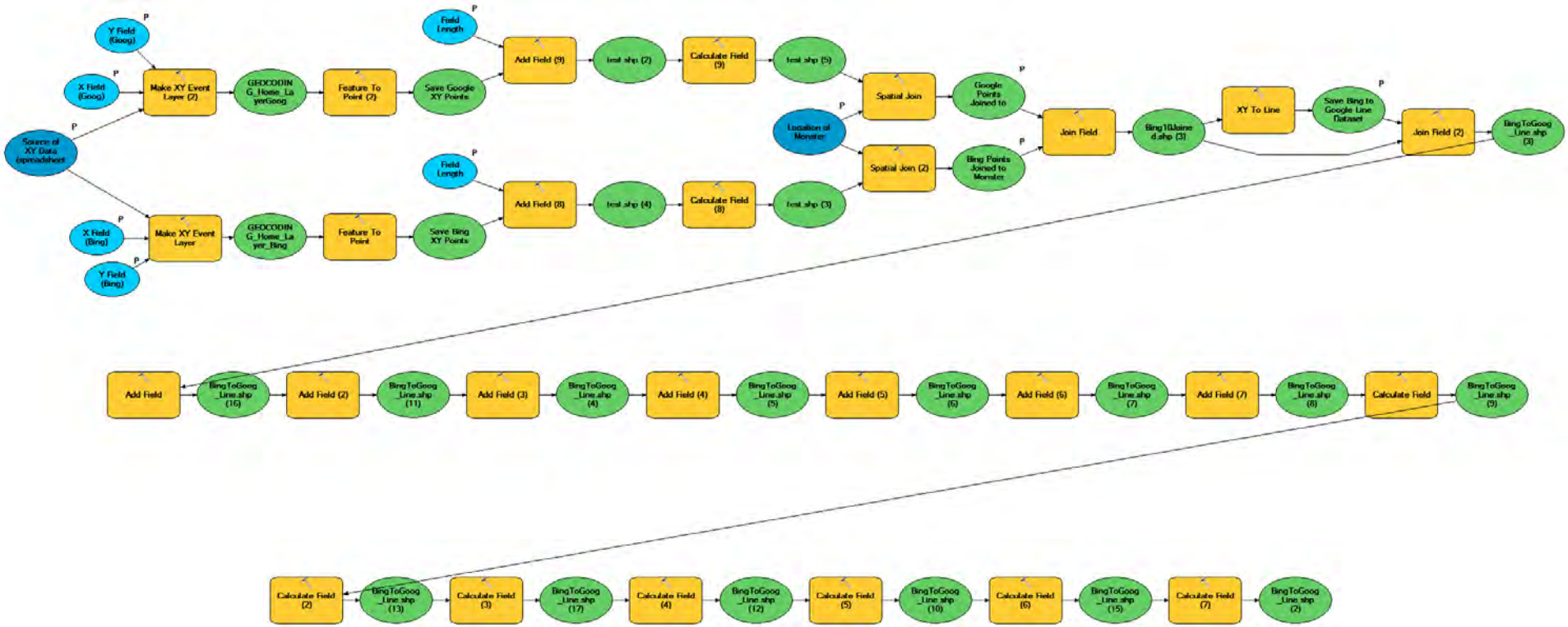




PUTTING DATA INTO BINS

Location Type	Address	Place Name	Cross Streets	Other	Total
Home	9,628	-	-	-	**9,628**
Work	4,799	1,865	182	-	**6,846**
School	-	-	-	3,514	**3,514**
Other (work in progress)	17,853	13,167	-	-	**31,020**
Total	**32,280**	**15,032**	**182**	**3,514**	**51,008**

ARCGIS MODEL

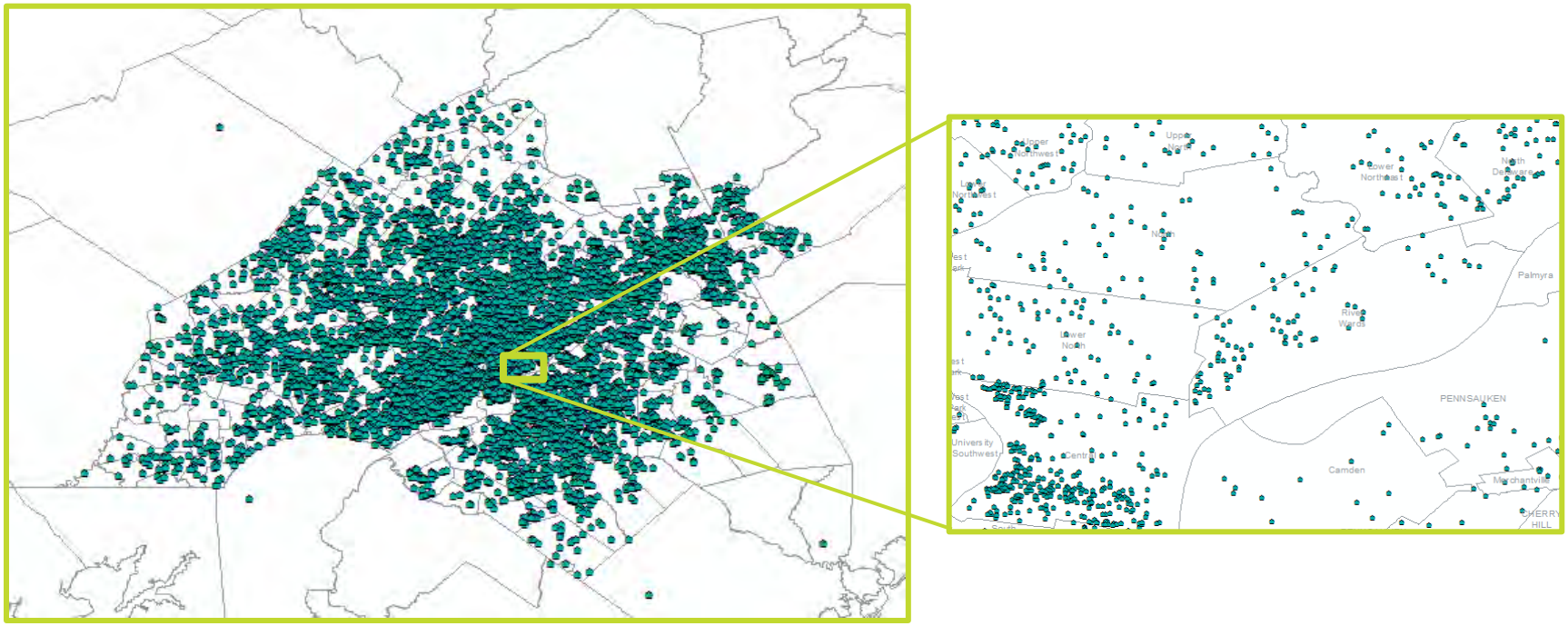


COMPILED "THE MONSTER"

- Regional boundaries combined into single file
  - State
  - Region
  - County
  - Planning Districts
  - Municipality
  - TAZ
  - Census Block

COMPARING RESULTS

- Home locations where Bing & Google geographies matched to Census Block level

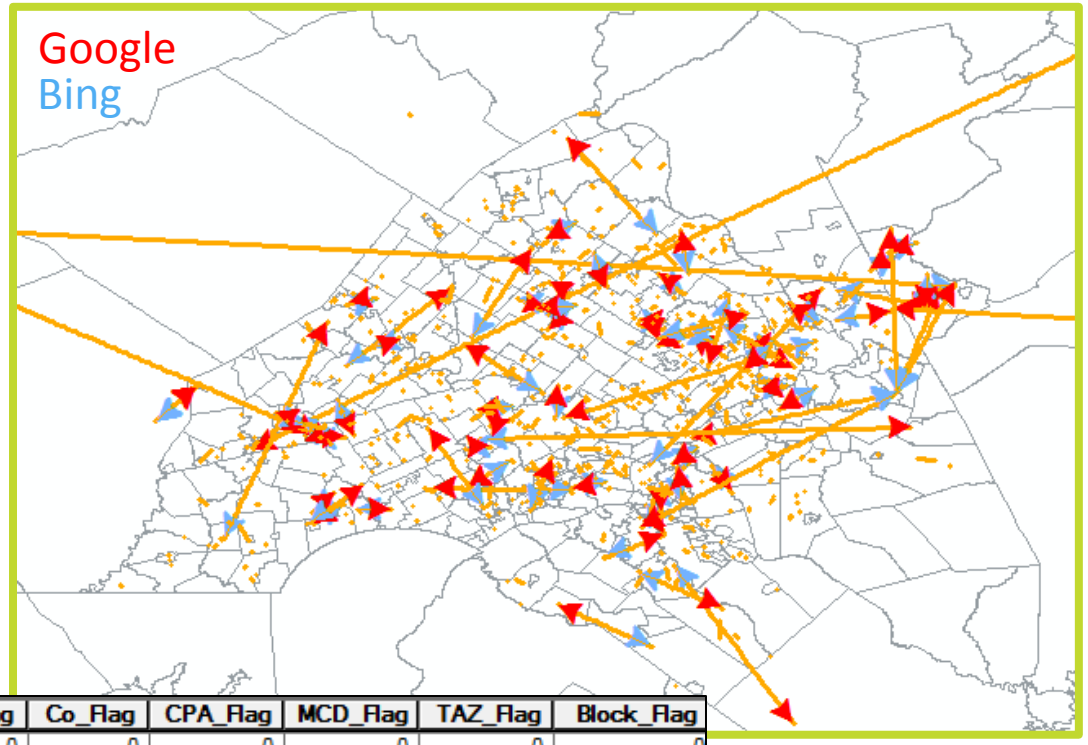


MODEL RESULTS

Result	Home	Work	Other	Total
Matched to Census Block	8,611 *89.4%*	3,407 *71.0%*	13,167 *73.8%*	**25,185** ***78.0%***
Mismatched	1,017	1,392	4,686	**7,095**
Total Records	**9,628**	**4,799**	**17,853**	**32,280**

COMPARING RESULTS

- ArcGIS Model Output
  - "XY to Line" tool gives point to point distance
  - Model created fields to flag records where Google/Bing geographies disagree



St_Flag	Reg_Flag	Co_Flag	CPA_Flag	MCD_Flag	TAZ_Flag	Block_Flag
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	1	0	1	1
0	0	0	0	0	0	0
0	0	0	0	0	0	1
0	0	0	0	1	1	1
0	0	0	0	0	0	0

GEOGRAPHY MISMATCHES

- Home Addresses
  - Google/Bing disagreed on 1,017 of 9,628 home locations

Boundary Level	Number of Mismatches	Percent of Mismatches
State	195	19.2%
Region	204	20.1%
County	224	22.0%
Planning District	285	28.0%
MCD	342	33.6%
TAZ	453	44.5%
Block	1,017	100.0%

GEOGRAPHY MISMATCHES

- Work Addresses
  - Google/Bing disagreed on 1,392 of 4,799 work locations

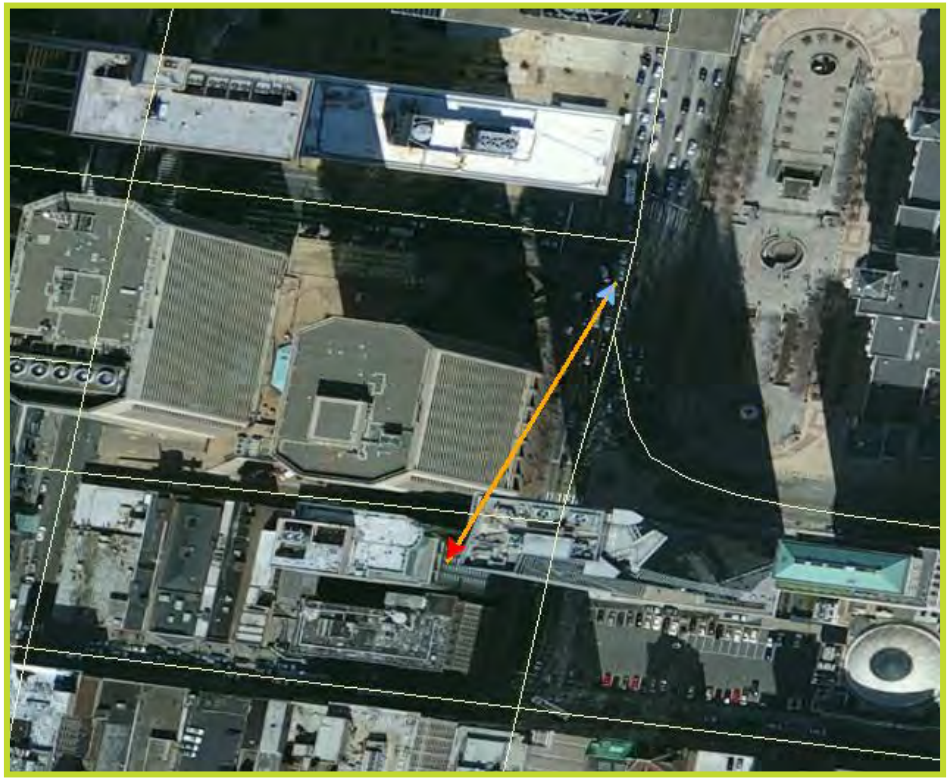
	Number of Mismatches	Percent of Mismatches
State	85	6.1%
Region	116	8.3%
County	183	13.1%
Planning District	268	19.3%
MCD	354	25.4%
TAZ	749	53.8%
Block	1,392	100.0%

GEOGRAPHY MISMATCHES

- Other Trip (Non-Home/School/Work) Addresses
  - Google/Bing disagreed on 4,686 of 17,853 other locations

Boundary Level	Number of Mismatches	Percent of Mismatches
State	175	3.7%
Region	214	4.6%
County	382	8.2%
Planning District	723	15.4%
MCD	1,099	23.5%
TAZ	2,405	51.3%
Block	4,686	100.0%

Easily resolved mismatch - Bing in ROW

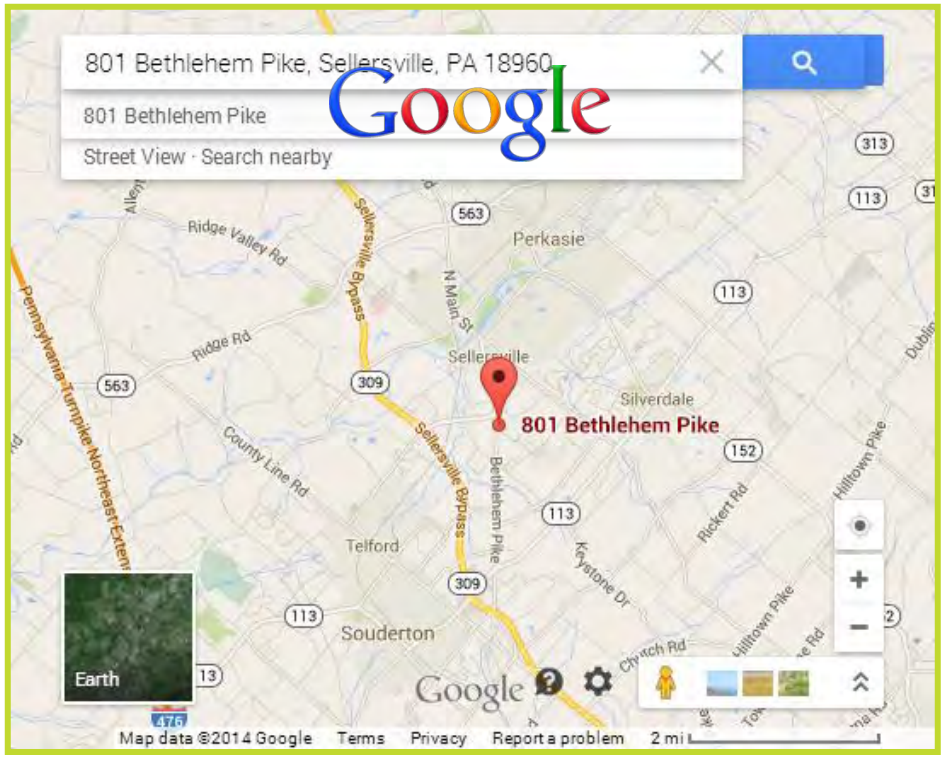


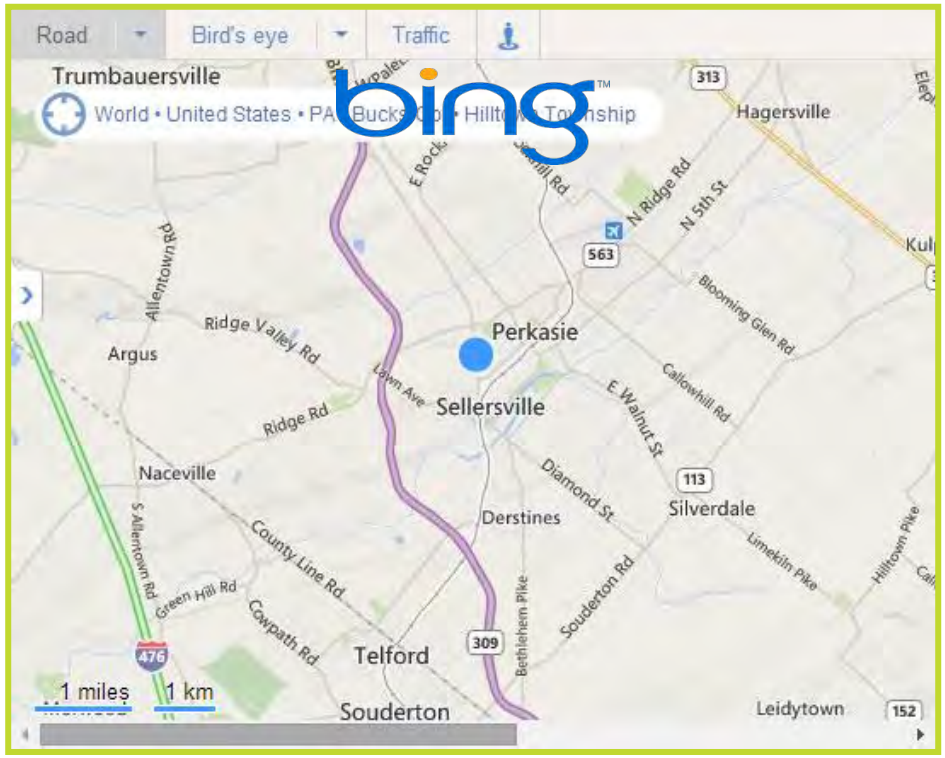
PICK A WINNER

Hard to resolve mismatch - takes some research

- Now what?
  - 801 Bethlehem Pike, Sellersville, PA 18960
  - Place Name = “AT Subaru”

PICK A WINNER





PICK A WINNER

PICK A WINNER

Winner	Home	Work
Bing chosen	52.7%	9.4%
Google chosen	46.5%	87.8%
Manually Overwritten	0.1%	2.8%

The search for the missing trip

TMIP Webinar

NOT-STRAIGHT-FORWARD CHECKS

- Partial list of what constitutes a complete trip record for a person from the contract:
  - Other trip elements complete or inferable in a systematic manner
  - Passes edit check for missed reporting of trips or activities
  - All trips and activities pass logical sequencing check
- But, how do we know if we have a missing trip?

CHECKING FOR MISSING TRIPS

- How do we know if we have a missing trip?
  1. Use some common sense. e.g. – person leaves work at 4 PM, stops at convenience store, makes no other trips. Is it likely that they stayed there till 3 AM the next day?
  2. Establish some common-sense checking rules and automate
  3. Evaluate rules vs. GPS sub-sample households
  4. Do spot checks.

SELECTED IMPUTATION FOR MISSING TRIPS

- Logical – Adjacent trip start/ends at home?

  Yes? → send them home

- Feasibility – Given the last destination arrival and duration, is it possible to travel home?

  No? → keep record as is, stop imputing

- Person
  - Total hours worked > 6
  - Typically works the following day
  - Special 'fatiguing' activities (e.g. medical, major shopping)

  If so, then flag for manual review

OH, & I'M NOT EVEN GOING TO MENTION…

- Reweighting
- Tour identification & classification
- GPS errors
- Formatting
- Misaligned fields

ADDITIONAL RESOURCES

- GeoPy geocoding APIs
  - http://geopy.readthedocs.org/
- Google Refine
  - https://code.google.com/p/google-refine/
- TMIP HTS resources
  - http://www.travelsurveymanual.org/
- MAG HTS Report: Appendix A
  - http://www.azmag.gov/Documents/TRANS_2012-02-17_2008-National-Household-Travel-Survey-Dataset-for-MAG-Region.pdf

Chris Puchalsky, PhD
Associate Director - Systems Planning
cpuchalsky@dvrpc.org

Ben Gruswitz, AICP
Transportation Planner - Office of Modeling & Analysis
bgruswitz@dvrpc.org

Sarah Moran
Intern - Office of Modeling & Analysis
smoran@dvrpc.org

Special thanks…
Will Tsay
Office of Modeling & Analysis

Kim Korejko
Office of Geographic Information Systems

dvrpc

HOUSEHOLD TRAVEL SURVEY

THANK YOU

Sample Weighting and Expansion

California Household Travel Survey (CHTS) 2012/13: Statewide Sample

TMIP Webinar – QC in Travel Surveys
August 21, 2014

Shimon Israel
Metropolitan Transportation Commission
sisrael@mtc.ca.gov

Presentation Overview

1. CHTS background
2. Sample weighting and expansion
3. IPF or “raking” of data
4. Initial weighting scheme of CHTS data
5. MTC’s approach to re-weighting data

CHTS 2012/13

- Data collected Feb. 2012 – Jan. 2013
- 42,500 sample HHs statewide
- Collaborative effort
- Address-based recruitment
- One-day activity diary survey
- Vehicle, OBD, wearable GPS components
- Supplemental sample purchase

Sample Weighting and Expansion

- "Naive" weight for CHTS : 12.6M / 42,500 = 296
- Weighting corrects for geographic and demographic biases.
- Expansion factors up to aggregate demographic and travel characteristics.

IPF or “Raking”

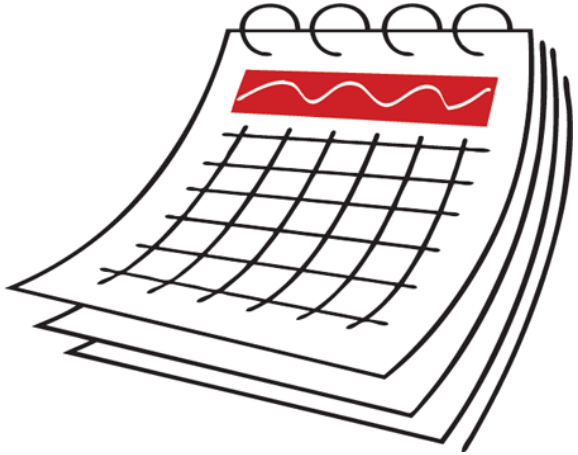
- Balances different marginal control totals
- Typically uses census data
- Best fit for the raked variables
- Balanced representation of population totals
- Automated script routine

Initial Weighting Scheme for CHTS HHs

- Statewide HHs by size
- Statewide HHs by income
- Statewide HHs by workers
- Statewide HHs by vehicles
- County of residence

Four Sets of MTC Weights

- Combined Sample / “Average Daily”
- Weekday Sample (n=30,216)
- Saturday Sample (n=5,979)
- Sunday Sample (n=6,236)



Raking Models - MTC: Combined & Weekday Samples

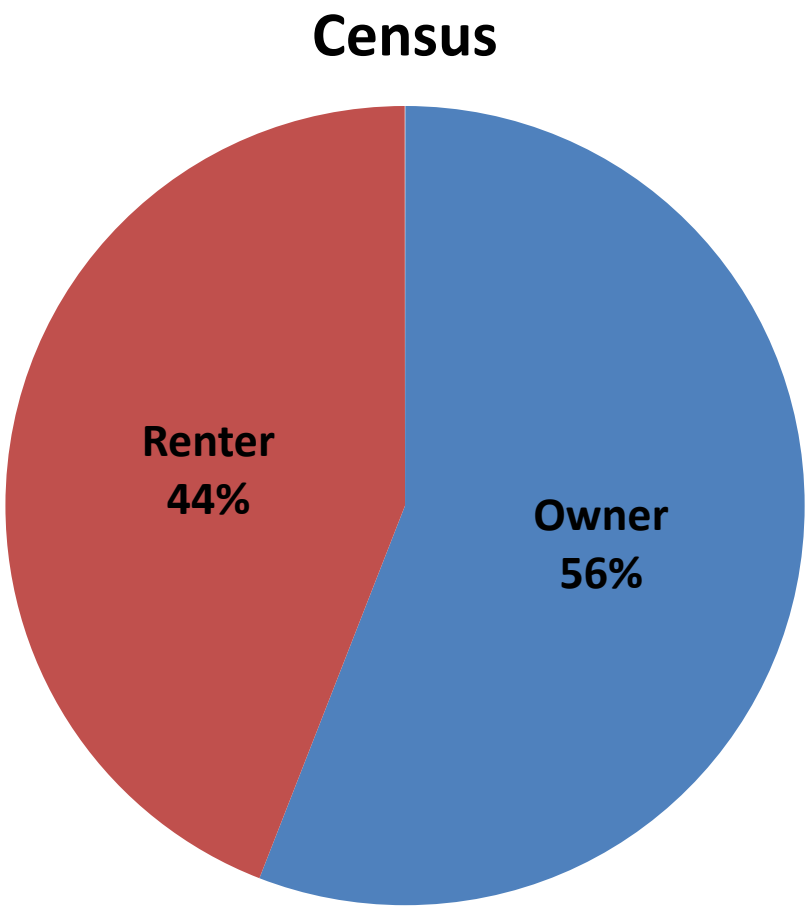
- County (58) by Tenure (2)
- County (58) by Age of Householder (5)
- County (58) by Minority Status (2)
- County (58) by Vehicles (4)
- Super-County (41) by Workers (4)
- County (58) by Household Size (5)

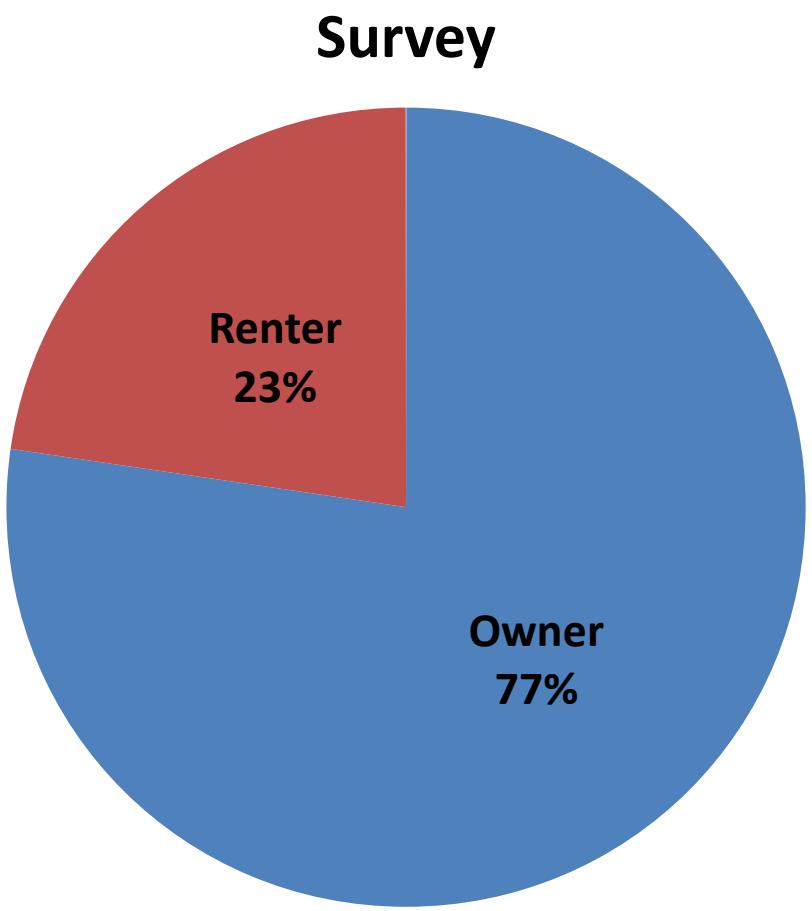

Del Norte / Siskyou / Modoc / Lassen
Trinity / Tehama / Glenn / Colusa
Plumas / Sierra / Nevada
Mendocino / Lake
Yuba / Sutter
"Sierra Nevada"
Monterey / San Benito
Nevada
Utah
Arizona
"Super-Counties"
California Household Travel Survey 2012-13
Scale:
1 in = 95 miles

Raking Models - MTC: Saturday and Sunday Samples

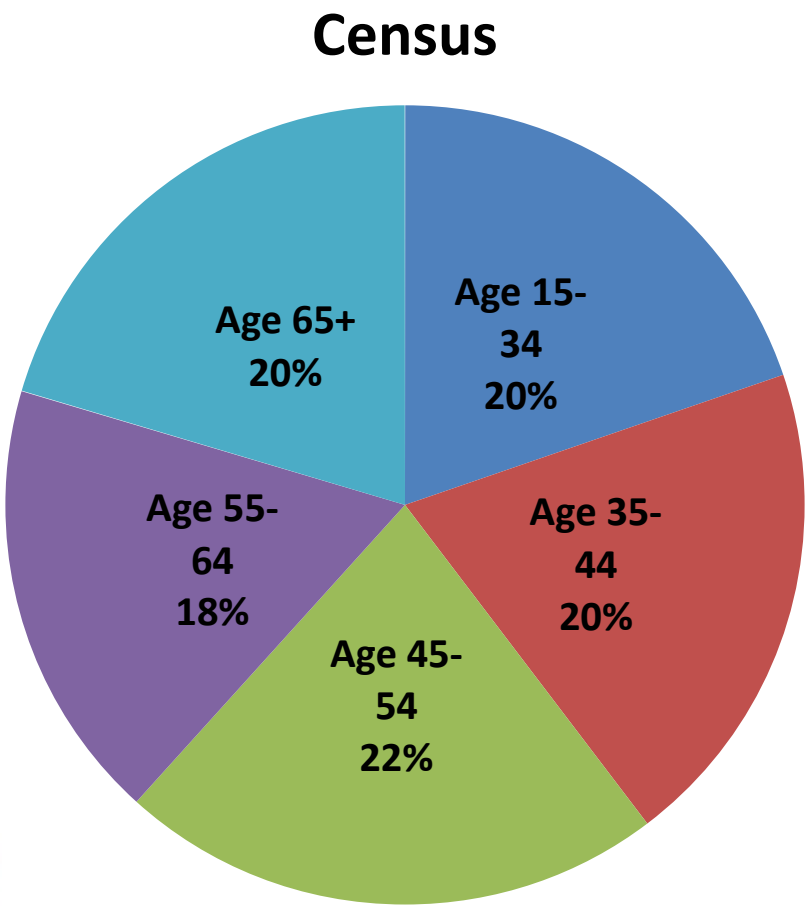
- County (58) by Tenure (2)
- Super-County (41) by Age of HHlder (5)
- County (58) by Minority Status (2)
- Super-County (41) by Vehicles (4)
- Super-County (41) by Workers (4)
- Super-County (41) by Household Size (5)

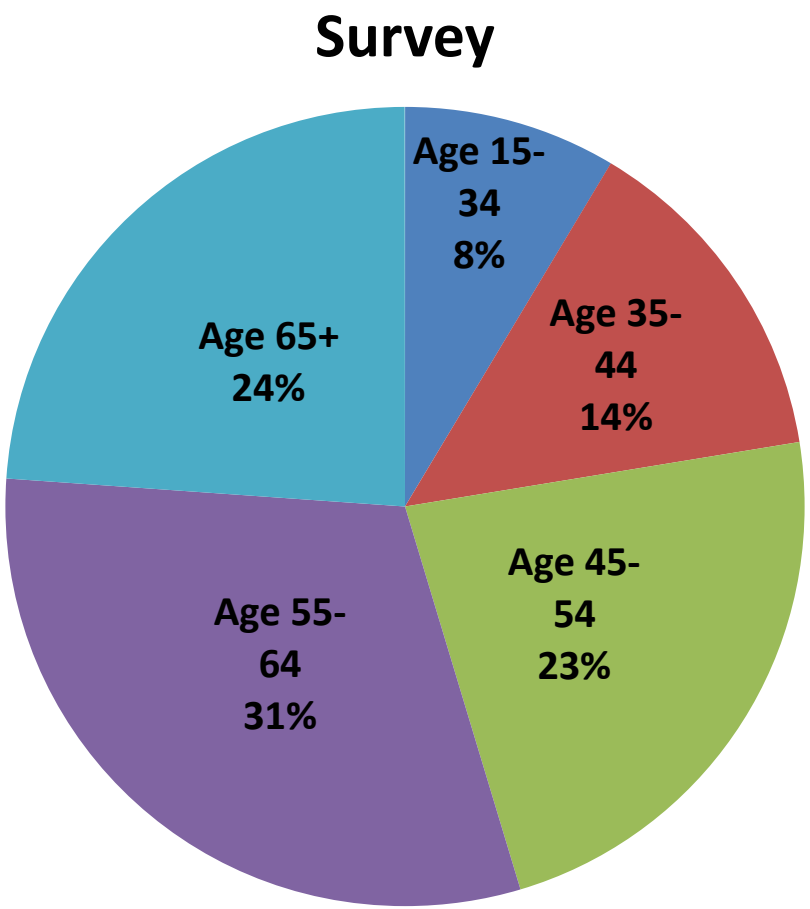
Tenure



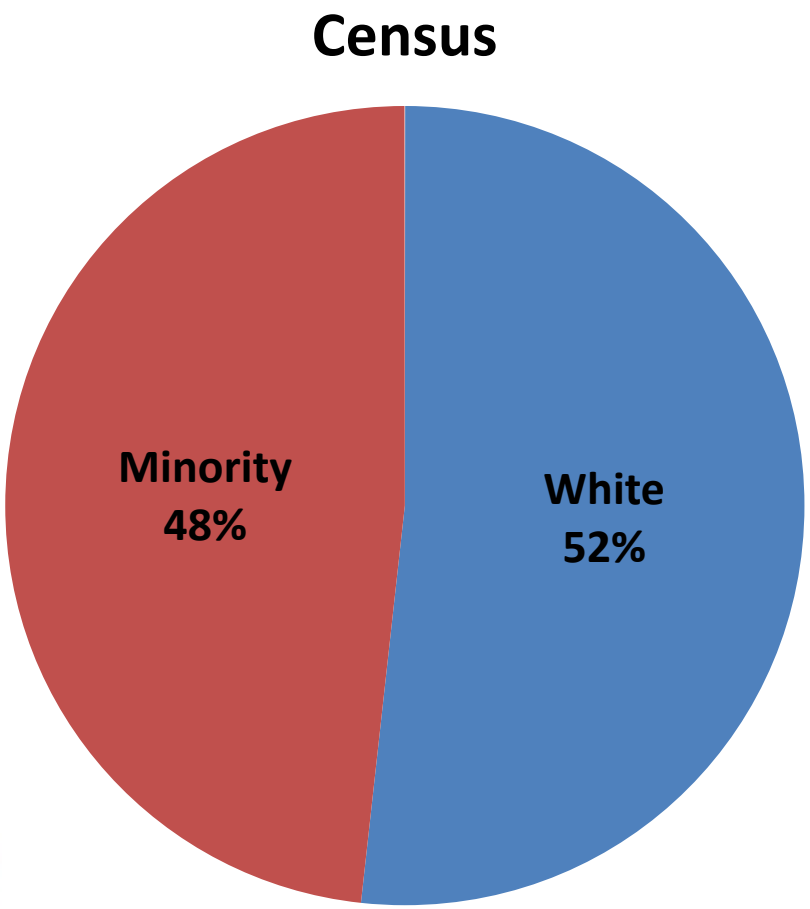


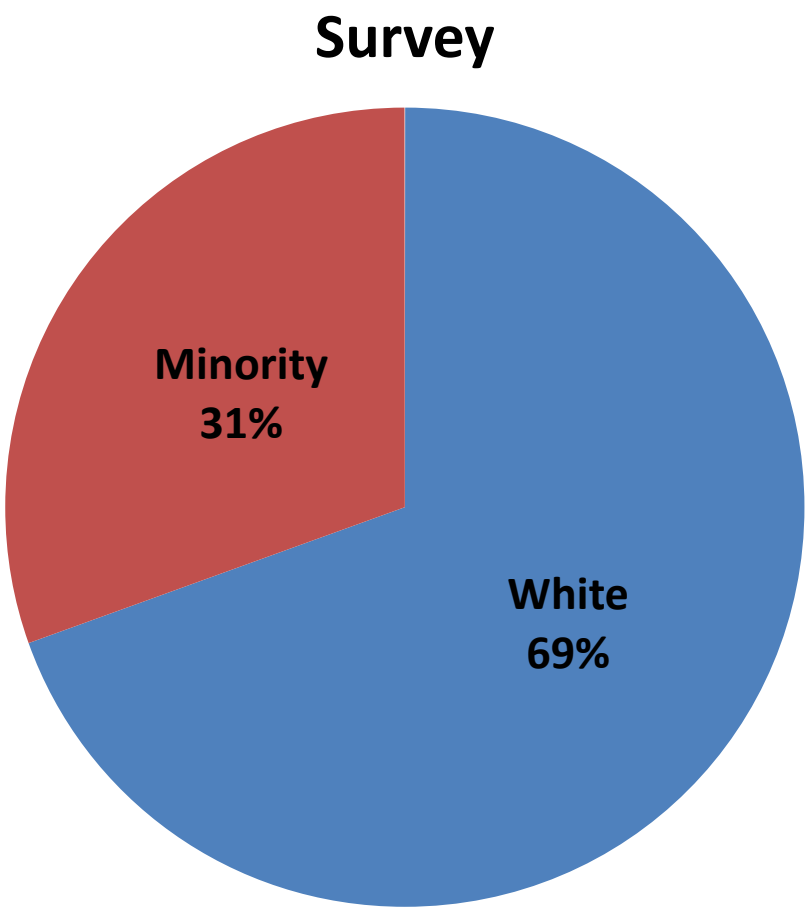
Age of Householder



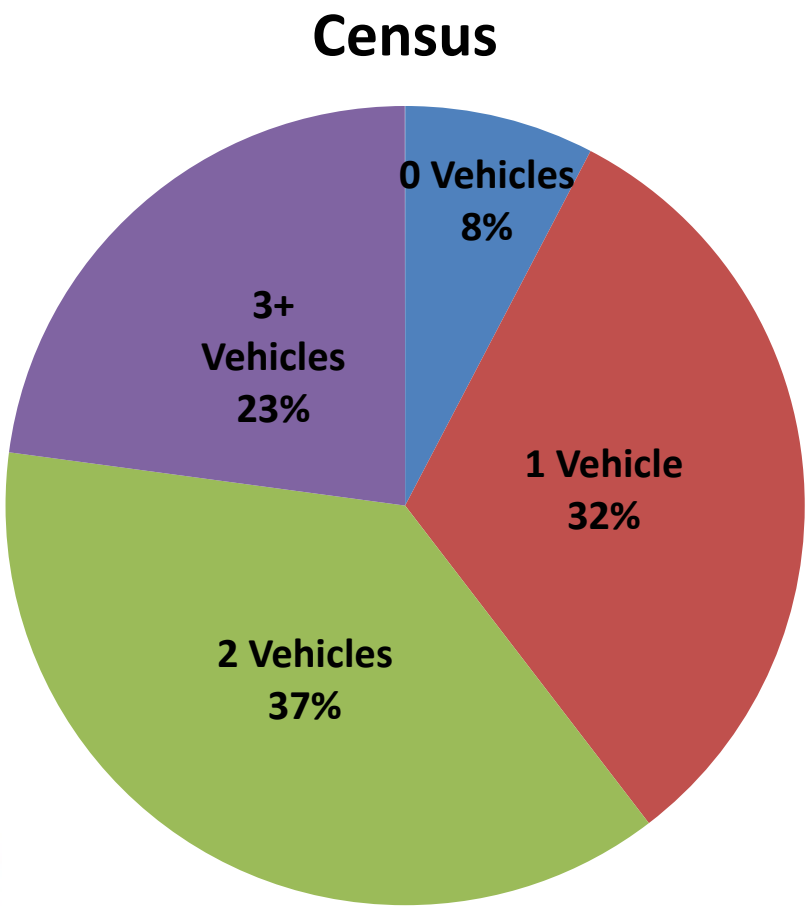


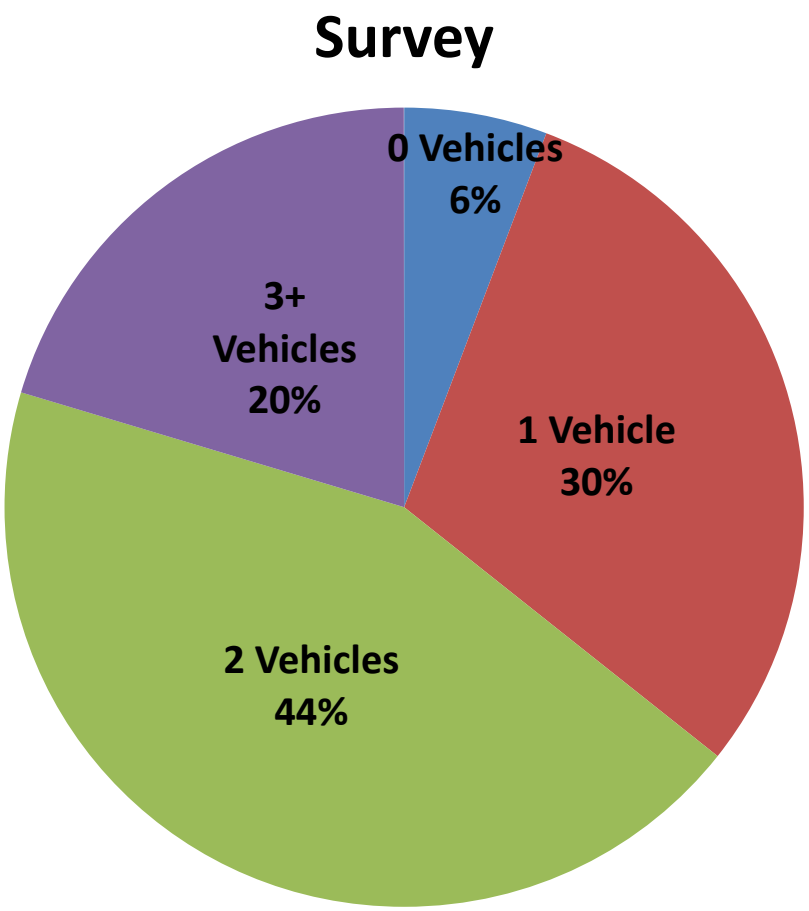
Race of Householder



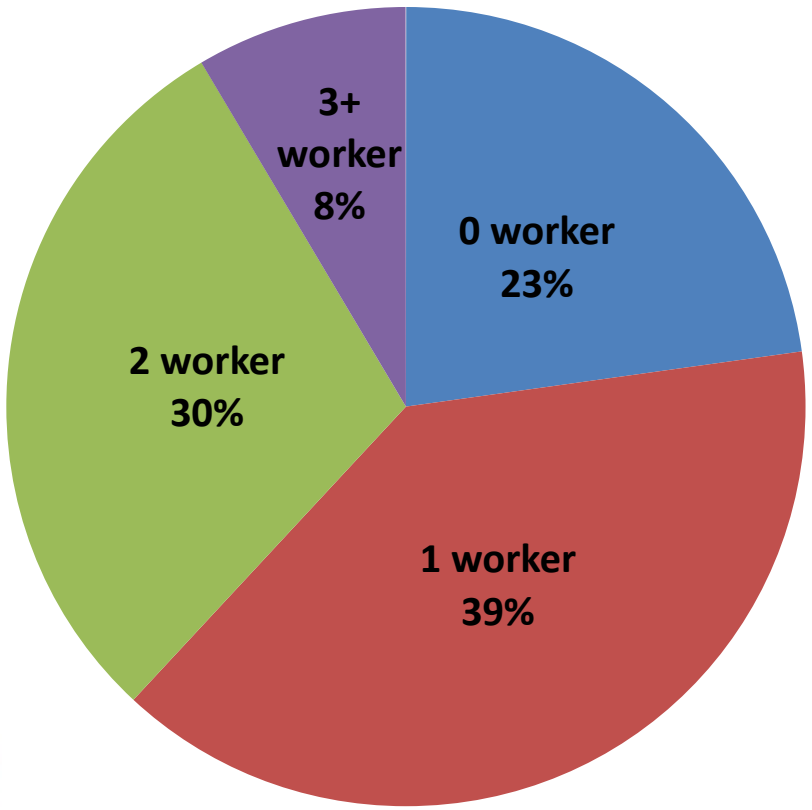


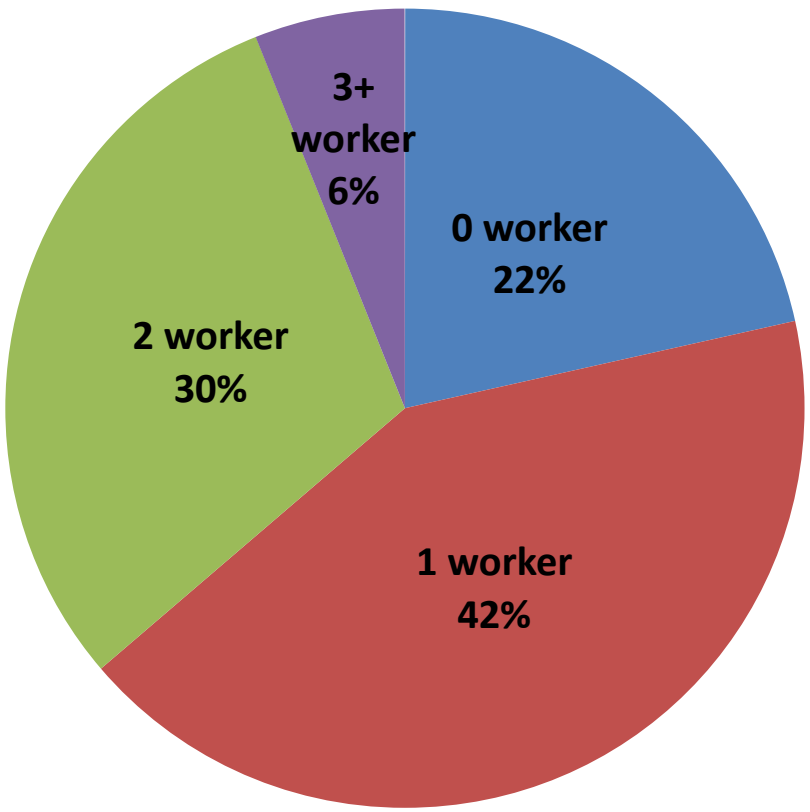
Vehicles in Household



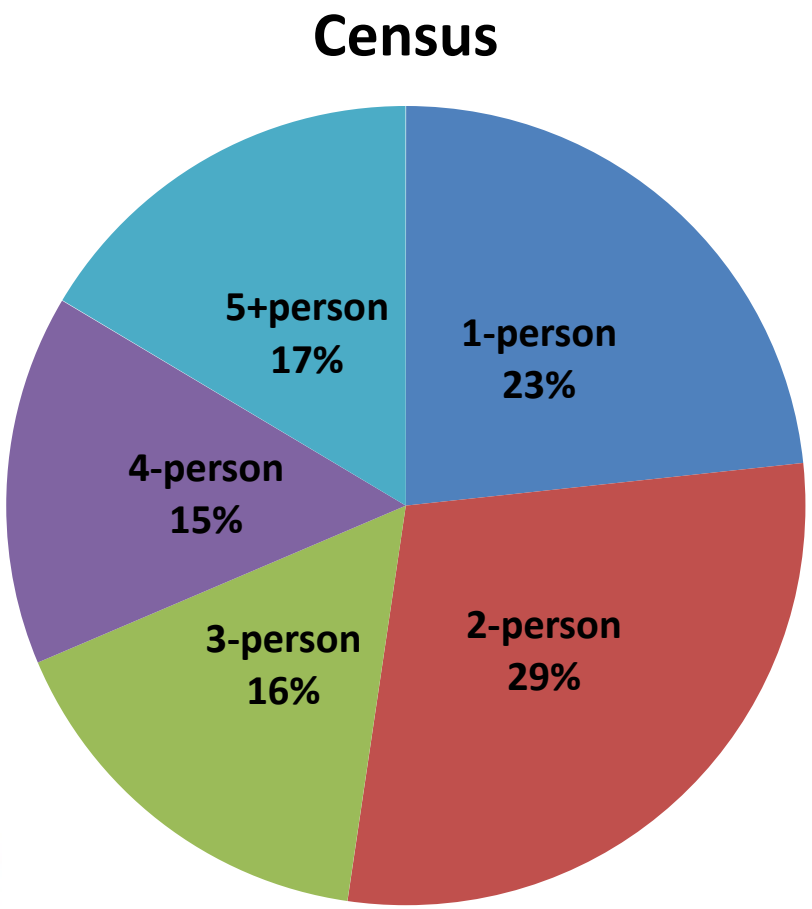


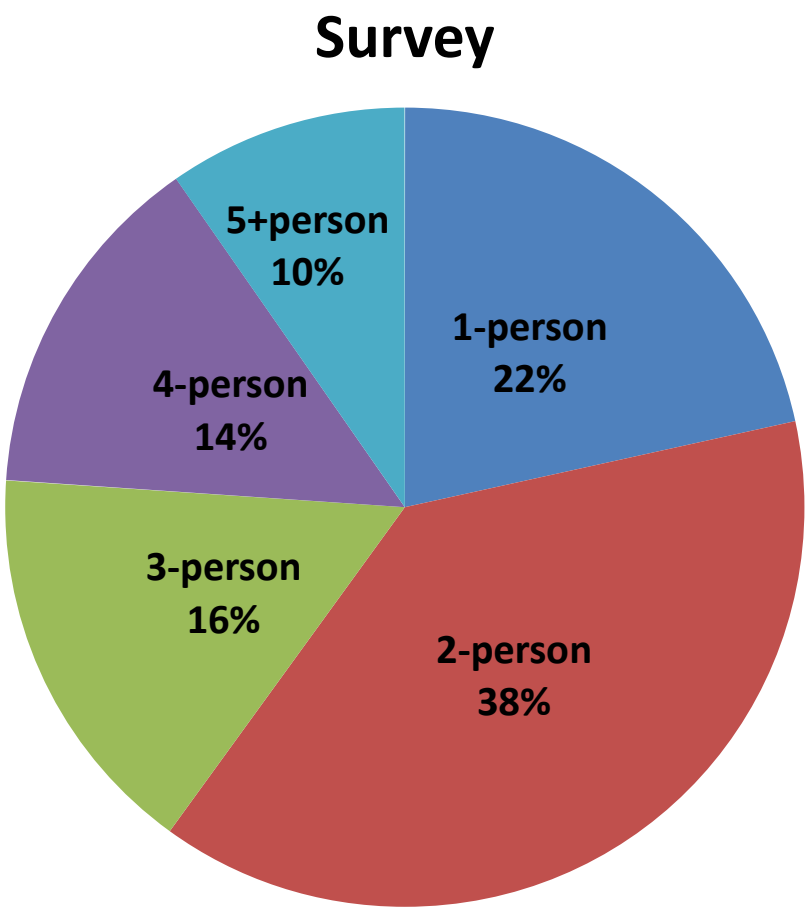
Workers in Household





Household Size





Imputation for Missing Data

- Tenure
- Age
- Race/Ethnicity

- “Hot Deck” Imputation

Next Steps

- Trip Linking, Trip Chaining, Travel Tours Procedures
- Trip Correction Factors based on GPS Datasets
- Reporting Aggregate Travel Characteristics
- Disaggregate Model Estimation

For More Information

- Shimon Israel, MTC (sisrael@mtc.ca.gov)
- David Ory, MTC (dory@mtc.ca.gov)
- Chuck Purvis, MTC-RA (cpurvi@mtc.ca.gov)
- MTC Analytical Modeling Wiki: http://analytics.mtc.ca.gov/foswiki/Main/HouseholdSurvey2012Weights

Discussant Take-Aways

Statistical Use of Data

Project Management

- Write a detailed contract with interim deliverables, robust QA/QC process,
- Provide staff resources to adequately manage project, review deliverables
- Choose payment method thoughtfully

Data Collection

Post Survey Processing

Discussant Take-Aways

Project Management

- Include data users in identifying data desired from survey, followed by estimated collection costs to prioritize what to collect
- Prioritize data uses, will need this for weighting
- Monitor progress closely
- Localize survey to extent possible

Data Collection

Discussant Take-Aways

Data Collection

- Weighting is very important, but never perfect; must be sufficient to meet priority needs
- Provide resources for post survey processing – trip linking, tours, database augmentation

Post Survey Processing

- Prepare to handle confidential data for requests

Discussant Take-Aways

Post Survey Processing

- Provide resources for model estimation
- Provide resources for model validation

Statistical Use of Data

- Provide resources for documentation and reporting overall results with general public and decision makers

Growing body of knowledge

Project Management

Contact peers to gain insight

Statistical Use of Data

Data Collection

Post Survey Processing

Avoid making our mistakes…

For More Information Contact:

Christi McDaniel-Wilson, P.E.

christina.a.mcdaniel-wilson@odot.state.or.us

Becky Knudson

rebecca.a.knudson@odot.state.or.us

Transportation Development Division

Transportation Planning Section

Transportation Planning Analysis Unit

TMIP Updates

For future webinar announcement, please sign up for GovDelivery at http://www.fhwa.dot.gov/planning/tmip/ if you have not done so.

TMIP Contacts

If you have any questions or comments about today's presentation or TMIP, or if you are interested in sharing your experience, please contact me at:

sarah.sun@dot.gov or feedback@tmip.org.