TO:	Dr. Gregory Macfarlane
FROM:	Hayden Atchley, Steven Goodsell, Austin Nichalson, Tristan Parker
DATE:	September 17, 2021
SUBJECT:	Lab 1: Building Blocks

As civil engineers, it isn't sufficient for us to only collect data for future demographic fields such as total number of jobs, households, and residents. It is our job to both comprehend and transform this data into reports in order to better understand and serve the future needs of civilians. In this lab, we were able to analyze household classification models that turn raw counts of households and persons into distributions of households by size, number of workers and vehicle ownership. We began by smoothing out the curves of collected data to obtain the marginal distribution of household size, vehicles and workers based on the average in a given zone. We were then able to confirm the accuracy of our joint distribution by comparing it with the CTTP trends provided by AASHTO and Census Bureau. The following is a summary of the work we performed that includes analysis and conclusions that were made during this process.

Fixing the Distribution Curves

We needed to adjust the given marginal distribution tables (graphed in Figure 1) so that the graphed data had smooth curves that reflected more realistic scenarios. Each cell contained the proportion of total households that fit in the distribution, so each row needed to sum to one. The proportions at the top and bottom of the table did not make sense, but the data in the middle seemed realistic. It was more difficult than we anticipated to smooth the curves so that they made sense, as well as make sure all the curves summed to one for a given input.

The easiest way to do it was to start with the left column and right column by determining the starting distributions and increasing or decreasing values until they aligned with the middle rows data. Then we could adjust the middle two columns based on the middle rows data and try to sum each row to one. This was an iterative process that took a lot of thought to make work. Eventually our graphs were smooth enough we felt we could use them to determine the joint distribution. The graphs are shown in Figures 2, 3, and 4.

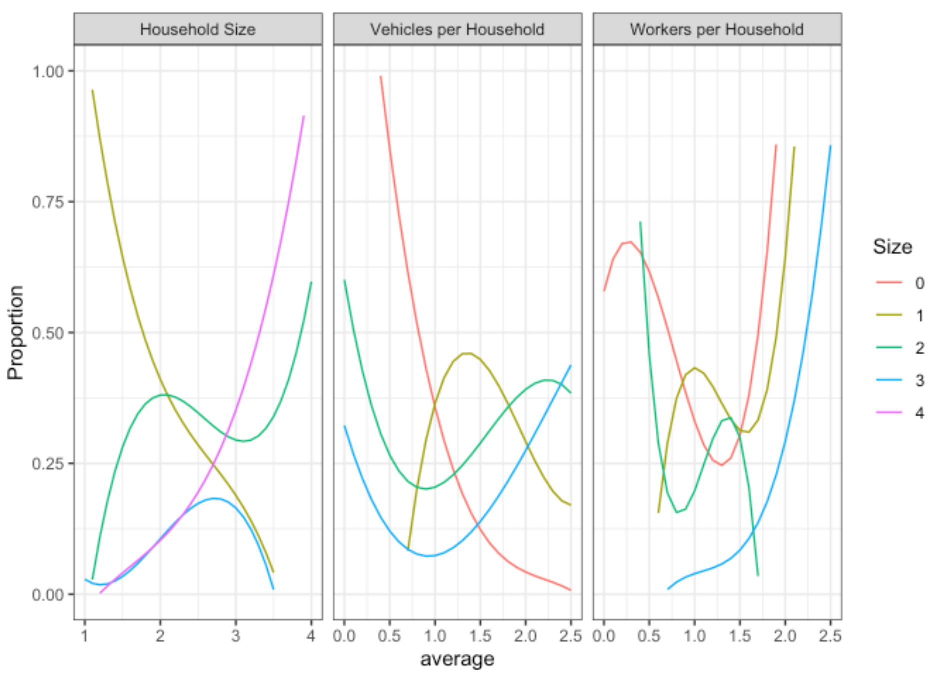

Figure 1: Graphs of original distributions

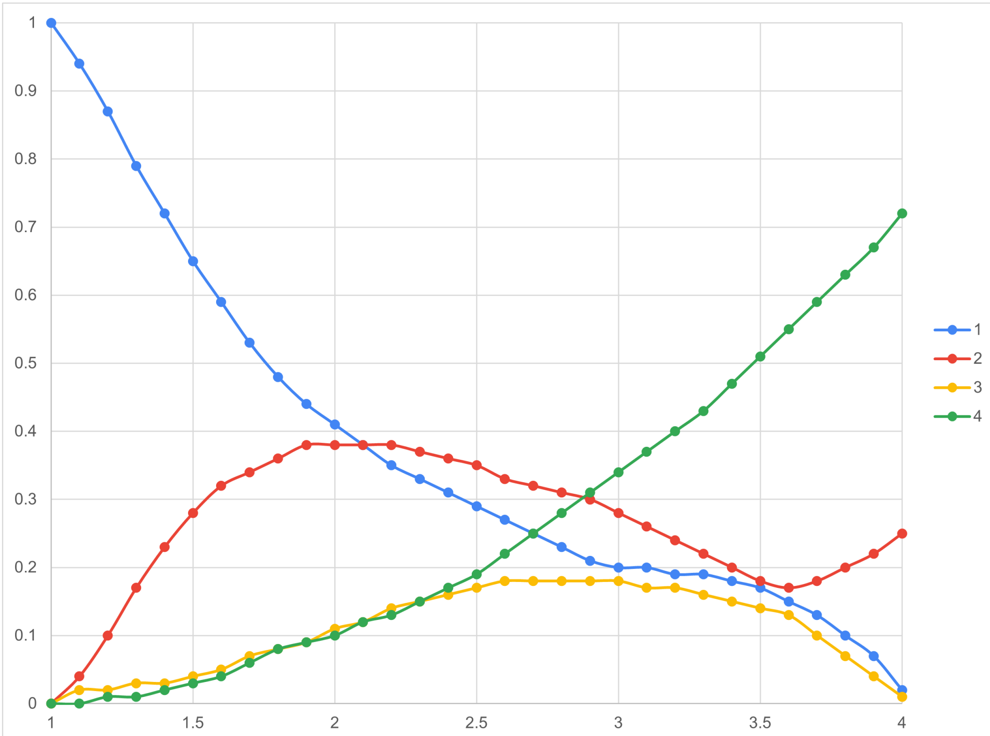

Figure 2: Graph of distribution of households

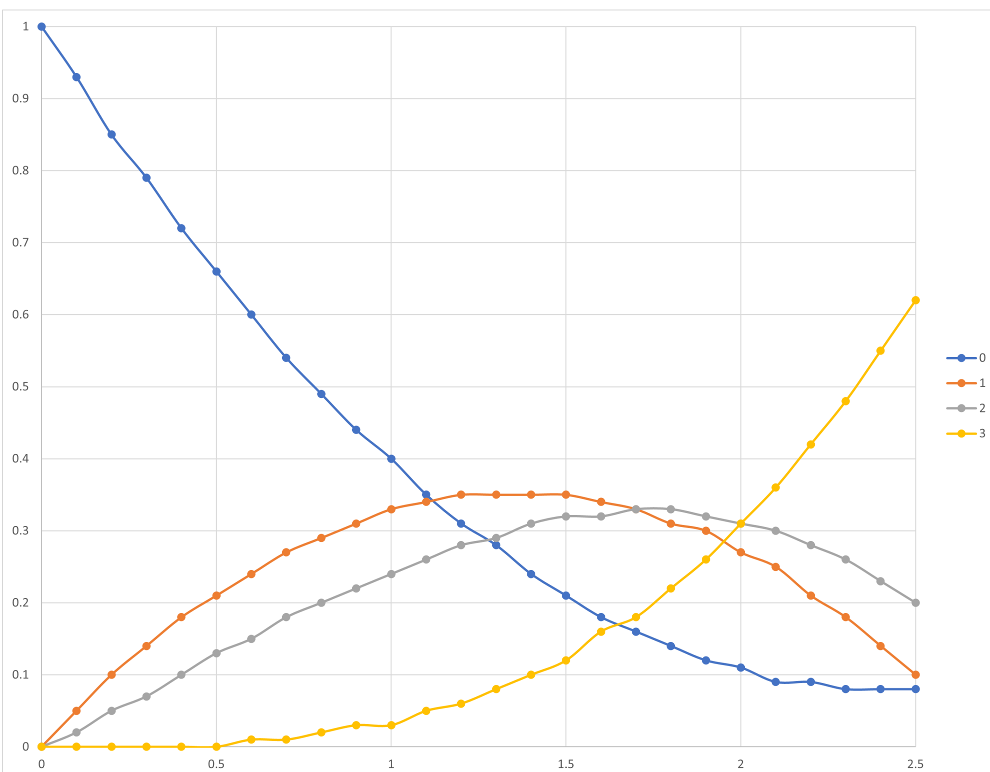

Figure 3: Graph of distribution of vehicles

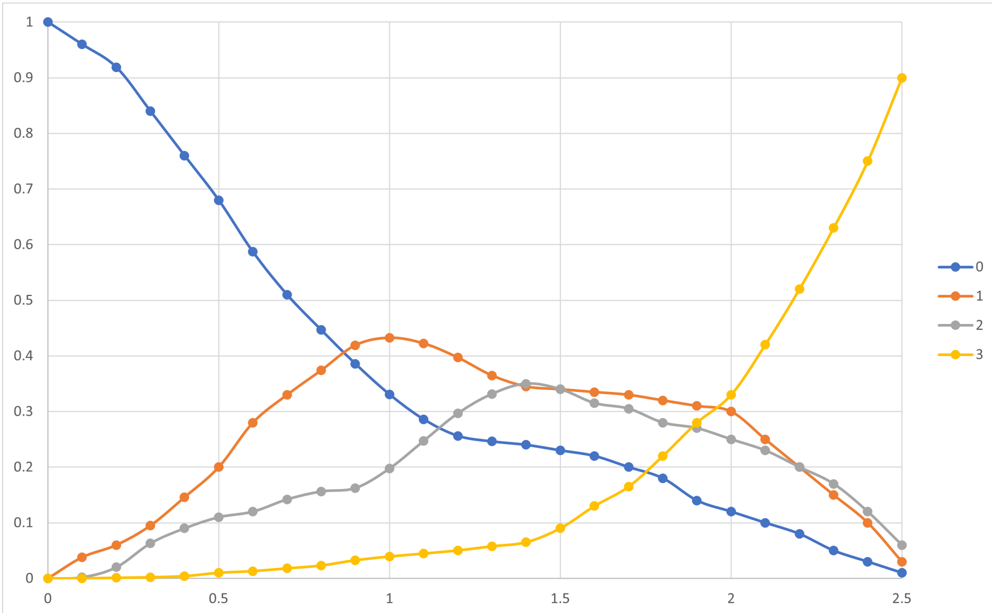

Figure 4: Graph of distribution of workers

Creating the IPF Seed

In order to run the IPF process in Cube, we need to create a seed. This seed is a 3-dimensional joint distribution of workers, household size, and vehicles. This data was obtained from the Census Transportation Planning Package (CTPP) and provided to us. We were also given an R script to run which put the data into a .dbf file Cube will recognize. These values are provided in Table 1 below.

Table 1: Seed for IPF in Cube

Persons	Workers	0 Vehicles	1 Vehicles	2 Vehicles	3+ Vehicles
1	0	0.0424577	0.0937149	0.019359	0.0040148
1	1	0.0129243	0.1195635	0.0273885	0.0073146
1	2	0	0	0	0
1	3	0	0	0	0
2	0	0.0079746	0.0269485	0.0441076	0.0211739
2	1	0.0088545	0.036958	0.0531271	0.0283235
2	2	0.0019249	0.0120993	0.074246	0.037893
2	3	0	0	0	0
3	0	0.0031348	0.0067096	0.0040148	0.0029698
3	1	0.0041248	0.019249	0.017874	0.0114394
3	2	0.000825	0.0051697	0.0323383	0.0255736
3	3	0.000275	0.00055	0.0038498	0.0146842
4	0	0.0015949	0.0033548	0.0010999	0.0024749
4	1	0.0025299	0.0140792	0.0231537	0.0153992
4	2	0.0015399	0.0087445	0.0382779	0.0257936
4	3	0.000154	0.0021999	0.0027499	0.0237037

Examining the CTTP data

Once we had completed adjusting the marginal distribution tables, we compared our modified data with the statistics provided by AASHTO and Census Bureau. Specifically, we analyzed the CTPP graphs and tables under the A112305 section. This particular data set contained 5 years worth of data ranging from the years 2012 to 2016. It compiled information about: household size (5), number of workers within a household (6), the vehicles available (5), and total number of households for each state and county within the respective state. We analyzed the figures of each team member's hometown county (Utah County and Ada County) and compared them with the overall trends and averages provided in the marginal distribution tables. Upon initial inspection of the data, it was shocking to see the number of households with zero workers that owned vehicles. Additionally, there are more households, in both Ada and Utah Counties, with zero workers that own at least 1 or more vehicles than those who don't own any. Given that the primary mode of transportation in both counties is a personal vehicle, it's logical that even households without any workers would own at least one vehicle. This could lead us to assume that the majority of the households that have zero workers are elderly, and therefore retired. However, because we do not know more about the socioeconomic status of these households with zero workers, we are unable to make any definitive conclusions about the demographics of these counties.

As far as correlating trends that we observed between our marginal distribution tables and the AASHTO and Census Bureau's data, as the number of workers in a household increased, so did the number of vehicles for that household. This follows the pattern we observed in the marginal distribution tables, that the number of workers per household is generally the average of vehicles that the household owns; the only outlier again being the zero worker household. The numbers expectedly varied between Utah and Ada County due to the overall population difference; however, the trends and observations aforementioned held consistent between the two counties. This could be true for a number of various reasons and it would be a safe assumption that it mainly has to do with the heavy reliance on personal vehicles in both areas. Although one county is much more rural and less developed than the other, the lack of well established public transportation in both regions makes owning at least one vehicle a necessity.

Running IPF in Cube When we first ran the CUBE model, it generated nonsensical data. The total number of three workers with zero vehicles, zero workers with three vehicles, and two persons with two vehicles households were identical in every district. Not only that, but other metrics like the proportions of different household sizes gave impossible answers that accounted for more or less than 100% of the population in different situations.

Using the data from our adjusted marginal distribution curves, we were able to get sensible results from CUBE. The table presenting the total number of households per district with the worker, vehicle, and person distributions mentioned above is presented in Table 2 below. These numbers better reflect the likely probability that a minority of homes have three or more workers and no vehicles, more homes have no workers and three or more vehicles, and many more homes have two people and two vehicles. We didn't suffer much in processing time within the model as we tried multiple different IPF iterations in the classification model (possibly because this is a relatively small project). Considering the circumstances of this project, it seems best to run 150 iterations to ensure more accurate results. When working on a larger project, however, it may be ideal to run 50 or 100 iterations to save processing time.

Table 2: Distribution of various types of household

	Sum(W3V0) All	Sum(W0V3) All	Sum(P2V2) All
DISTRICT	232.74	2995.79	16285.65
<0.5	–	–	–
1	24.57	553.75	2487.15
2	4.51	151.91	465.78
3	18.08	287.24	1507.93
4	13.21	429.87	1490.54
5	4.85	309.52	895.5
6	10.96	243.04	838.56
7	14.32	165.67	862.53
8	28.29	211.04	1477.22
9	31.12	38.47	945.52
10	27.88	250.95	1724.98
11	25.36	190.05	1899.17
12	29.59	164.28	1690.77
>12.5	–	–	–

Conclusion

Data gathered to help planners determine trip generation for households needs to be processed before it can be relied on. A model running the original marginal distribution curves will not reflect a realistic model well. Only after the data has been smoothed to better reflect the real world can a joint distribution be put together to create a better Cube model. The model then iterates with the probability to create a more accurate simulation of trip generation for the region.