CHAPTER 2

A DESTINATION CHOICE MODEL FOR WORK TRIPS

Introduction

Workplace and residential location have often been treated as being outside the scope of transportation demand analysis. Land use models typically locate (primary) workplaces exogenously, and then distribute residences about workplaces in a manner that takes account of land use constraints, but that incorporates sensitivity to at most primitive measures of the attributes of travel. Short-run transportation analysis has usually treated work and residence locations as fixed, without sensitivity to transportation level-of-service variables. Further, changing demographic patterns have not been adequately accounted for in most intermediate-run transportation studies.

A model of the decisions determining work and residence location would be useful at two places in transportation demand analysis. First, a model that is sensitive to the demographic characteristics of the population can be used to forecast the total demand for work-trips. Combined with a behavioral mode-split model, this analysis could provide forecasts of aggregate numbers of trips and revenues. Second, a model that is sensitive to level-of-service attributes could be incorporated into a model system in which policy impacts including impacts on location are assessed. This chapter concentrates on a model developed to forecast demographic changes in the pattern of work and residence locations.

Location decisions have several aspects. First, employers locate establishments based on a variety of considerations, including land use restrictions, cost, ease of goods movement, and availability of labor. Workers then choose employers and residential locations, taking into account wages, other attributes of the work, residential neighborhood attributes, and transportation system attributes. Traditional land-use models assume the worker first chooses a work location, and then chooses a residential location treating the work location as given. This may be reasonable for some professional workers, but is probably not reasonable for secondary workers or for the large number of workers in occupations where numerous comparable employment opportunities exist. In the latter case, residential location is more likely to remain fixed, with a choice made among alternative work locations. In each case, wage rates or availability of positions will adjust to equate jobs demanded and jobs supplied at each location.

If the joint work and residence location choice can be described adequately by a multinomial logit model, then the conditional choice of work location given residential location will be multinomial logit. Furthermore, if the mean utility function is additively separable between attributes of the work location and attributes of the residential location and journey to work, then only the latter variables will enter the conditional multinomial logit probability for workplace choice.

Socioeconomic Determinants of Work Locations

In order to forecast total number of work trips taken between zone pairs at various forecast dates, we developed a model of workplace location splits as a function of household characteristics. While the model can in principle be given a choice model interpretation, our motivation was primarily to provide a data-analysis tool for auxiliary forecasting rather than to develop a behavioral model. The model structure, and its consequent interpretation, were severely constrained by data availability. Data restrictions also made it impractical to include transportation level-of-service variables, although with some effort the approach could be amended to allow such sensitivity.

The San Francisco Bay Area has 440 traffic analysis zones. It is impractical to attempt to estimate a destination choice model (with 440 alternatives) for each origin zone, even if an adequate calibration data base were available. Hence, the Bay Area was partitioned into larger units, termed districts. Table 28 lists the eight origin and twelve destination districts used in this study. For each origin district, the probability of working in each destination district was estimated as a function of household characteristics (interacted with destination-specific dummy variables). The geographical coverage of the Urban Travel Demand Forecasting Project surveys was too restrictive to permit use of this data source as a base for calibration. The 1965 BATSC survey provided a potential calibration base using individual household data, but was somewhat out of date and was known to have problems of representativeness that give some marginals inconsistent with census marginals. Census data for 1970 giving work destinations and major household characteristics is not available at the level of the individual, but is available at the level of the census tract.[1] The disaggregate model was developed with the objective of calibration using aggregate census tract data. Table 29 describes the household characteristics used in the analysis.

[1] The Census Public Use Sample provides household characteristics at the level of the individual, but fails to provide the work destination district.

TABLE 28

Origin Districts

1. Marin and Sonoma
2. Napa and Solano
3. Contra Costa
4. Northern Alameda County*
5. Southern Alameda County**
6. Santa Clara
7. San Mateo
8. San Francisco

Destination Districts

1. San Francisco CBD
2. San Francisco (remainder)
3. Oakland CBD
4. Oakland (remainder)
5. Alameda County (remainder)
6. San Mateo
7. Contra Costa
8. Marin and Sonoma
9. San Jose
10. Santa Clara (remainder)
11. Solano and Napa
12. Outside Bay Area

*Traffic Analysis Zones 123-186

**Traffic Analysis Zones 187-216

TABLE 29 Socioeconomic Variables Used in the SYNSAM Classification of Households

Variable	Symbol	Boolean Variable*	Categories	1970 PUS Tabulations
1. Workers	W	V^2 V^3	1. Zero workers 2. One worker 3. Two or more workers	P31
2. Family type	F	V^4 V^5 V^6	1. Husband and wife, head under 45 2. Husband and wife, head over 45 3. Other family 4. Primary individual	H70 and H72-73
3. Autos	A	V^7 V^8	1. Zero autos 2. One auto 3. Two or more autos	H60
4. Income	I	V^9	1. Above Bay Area median 2. Otherwise	H85-87 ($10,500 in 1970)
5. Persons	P	V^{10}	1. One or two person household 2. Otherwise	H12-13
6. Units	U	V^{11}	1. One unit attached or detached 2. Otherwise	H35
7. Race	B	V^{12}	1. Black 2. Otherwise	H71
8. Tenure	R	V^{13}	1. Renter occupied 2. Otherwise	H31-33
9. Mobility	M	V^{14}	1. Head moved in past five years 2. Otherwise	H90

*A Boolean variable is one if the household characteristic takes on the associated value, and is zero otherwise. For example, $V^2 = 1$ if the household has no workers, and $V^2 = 0$ otherwise. $V^1 = 1$ always.

In general, the relation between individual choice probabilities and the aggregate probabilities for the heterogeneous population in a census tract would be analytically complex, making calibration from aggregate data difficult. An exception is the linear probability model, which assumes that the probability of destination district δ, conditioned on origin district d and household characteristics σ, is a linear combination of the zero-one variables V_σ^k corresponding to σ (e.g., if σ denotes a household with one worker and zero autos, then $V_\sigma^3 = 1$ and $V_\sigma^8 = 1$). This model has the form

$$(1) \qquad P(\delta \mid d,\sigma) = b_{d\delta}^1 V_\delta^1 + \ldots + b_{d\delta}^k V_\sigma^k + \varepsilon_{d\delta\sigma} \quad ,$$

where $E\varepsilon_{d\delta\sigma} = 0$, $\sum_\delta b_{d\delta}^1 = 1$, and $\sum_\delta b_{d\delta}^k = 0$ for $k > 1$. The dependence of the coefficients $b_{d\delta}^k$ on destination district d could be reinterpreted as the result of interactions of the household characteristic boolean variables V_σ^k and destination-specific dummy variables. Note that the coefficients $b_{d\delta}^k$ are assumed to not depend on the census tract in which the household is located.

Now suppose the individual probabilities (1) are averaged over all the households in a census tract on traffic analysis zone τ. Let $P(\delta \mid d,\tau)$ denote the average probability for the tract, and $\overline{V}_\tau^k$ denote the average value of V_σ^k in the tract, or the proportion of the tract households with $V_\sigma^k = 1$. Then

$$(2) \qquad P(\delta \mid d,\tau) = b_{d\delta}^1 \overline{V}_\tau^1 + \ldots + b_{d\delta}^k \overline{V}_\tau^k + \varepsilon_{d\delta\tau} \quad .$$

Each of the variables in (2) is observed from census data at the tract level. Hence, the coefficients $b_{d\delta}^k$ can be fitted by applying ordinary least squares to data on the tracts in each origin district.[1] With eight origin districts and twelve destination districts, this requires 96 regressions, of which eight are redundant and can be

[1] The $\varepsilon_{d\delta\tau}$ are expected to be heteroscedastic for two reasons. First, in more populous tracts, the number of households in the census will increase, lowering the sampling variance in $P(\delta \mid d,\tau)$. Second, the requirement that $P(\delta \mid d,\tau)$ be contained in the unit interval will tend to lower its variance near values of zero or one. The first heteroscedasticity may be corrected by multiplying each observation by the number of households in the tract. This adjustment has the further advantage of preserving the aggregation property that the sum of calculated probabilities for all households in the district will equal aggregate shares. The second source of heteroscedasticity can be corrected only at the expense of making the model extremely sensitive to outliers (Domencich and McFadden (1975), Chapter 4). Hence, use of this adjustment is not recommended.

used as checks. These are from 93 to 287 census tracts in each origin district.

To the extent that census data is available on interactions between the socioeconomic variables, such interactions could be added as variables in (1) and then coefficients calibrated in the preceeding manner.

The destination choice model described above has two empirical shortcomings. First, much of the variation in household characteristics is known to be intra-tract. Fleet and Robinson (1968) found eighty percent of the variation in a selected set of socioeconomic variables to be intra-tract, and McFadden and Reid (1979) have found as much as ninety percent of the variance of some household characteristics to be intra-tract. This implies that there will be low variation in the averaged variables $\overline{V}_{\tau}^{k}$ over tracts relative to the variation of V_{σ}^{k} across individuals. Large statistical standard errors of the coefficients may result, and the averaged variables may exhibit considerable multicollinearity.

Second, the linear probability model does not force probabilities to lie in the unit interval, a particular problem when extrapolating the independent variables outside the range of the calibration data. Because we have argued that the range of tract averages is likely to be limited, and the values of the explanatory variables for a household will lie at an extreme, the problem of negative calculated probabilities is serious. At least within the confines of the linear probability model, there seems to be no fully satisfactory theoretical method for handling the problem. Two *ad hoc* procedures suggest themselves. First, negative calculated probabilities could be set to zero, and the remaining probabilities renormalized to sum to one. We have adopted this method in empirical analysis. Second, the calculated probability vector for a household could be averaged with the corresponding positive probability vector for the district to make the least component non-negative. I.e., let $P(\delta \mid d,\sigma)$ be the calculated probability for household σ , and let $P(\delta \mid d)$ be the positive observed probability for the district. Define $A(d,\sigma) = \{\delta \mid P(\delta \mid d) > P(\delta \mid d,\sigma)\}$, and $\theta = \max\left\{0, \max_{\delta\in A(d,\sigma)} \frac{-P(\delta \mid d,\sigma)}{P(\delta \mid d) - P(\delta \mid d,\sigma)}\right\}$. Define $P'(\delta \mid d,\sigma) = (1-\theta)P(\delta \mid d,\sigma) + \theta P(\delta \mid d)$. Then $P'(\delta \mid d,\sigma)$ is non-negative, with $\sum_{\delta} P'(\delta \mid d,\sigma) = 1$. Further, $P'(\delta \mid d,\sigma)$ summed over household types σ , weighted by the proportion of district house holds of type σ , equals the district probability $P(\delta \mid d)$. This procedure could also be applied using the observed probabilities for the tract or zone in which the household resides, $P(\delta \mid d,\tau)$.

However, in this case the final calculated probabilities $P'(\delta \mid d,\sigma,\tau)$ would depend on the residence tract.

The second adjustment method described above resembles a Hunt-Stein estimator,[1] which suggests that it might have more desirable statistical properties than the unadjusted estimator. However, it should be emphasized that these adjustments are essentially *ad hoc*, and must in the end be judged on empirical grounds.

[1] Hunt-Stein estimators exploit the fact that with samples from related populations, best estimates of population means utilize information on the overall mean of the samples as well as the individual sample means.

Empirical Results

As an illustration of calibrated probabilities, we consider one origin district--Contra Costa County, and the prototypical households given in Table 30. The calculated probabilities from the linear probability model are strongly different for the different consumer types, with a number of values outside the permissible zero-one range. This is due to extrapolation of the model to values of the explanatory variables well outside the range of the observed census tract averages on which the model is calibrated. One would then expect the linear probability model to signal correctly variations in destination patterns for a household type from the district averages, but to overstate the magnitude of the variation.

The results in Table 31 confirm the seriousness of the problem of negative probabilities. The two adjustment processes give substantially different results. Data is currently not available to evaluate either method. However, we conjecture that the second method will prove superior in most applications.

TABLE 30 Prototypical Households

	I	II	III	IV	V
	Older Upper Class Family	Older Lower Class Family	Young Family With Children	Young Family Without Children	Primary Individual
Workers--W	2	3	2	3	2
Family Type--F	1	1	2	2	4
Autos--A	3	2	2	2	2
Income--I	1	2	2	2	2
Persons--P	1	1	2	1	1
Units--U	1	2	1	2	2
Race--B	2	2	2	2	2
Tenure--R	2	1	1	1	1
Mobility--M	2	2	1	1	1

TABLE 31 Calculated Destination District Probabilities from Contra Costa County[1]

Destination District	District Average Probability	I Older Upper Class Family			II Older Lower Class Family			III Young Family With Children			IV Young Family Without Children			V Primary Individual		
		(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1. SF CBD	.053	-.210	0	.043	-.020	0	.050	.451	.121	.064	.381	.152	.072	.278	.151	.083
2. SF (other)	.052	.271	.067	.060	.342	.106	.065	-.051	0	.049	.253	.100	.064	.017	.009	.047
3. OAK CBD	.021	.127	.032	.025	.148	.046	.027	.063	.017	.022	.222	.088	.033	-.047	0	.012
4. OAK (other)	.083	1.049	.260	.119	1.153	.358	.132	-.338	0	.071	.940	.374	.132	-.407	0	.018
5. Alameda County (other)	.132	2.031	.503	.202	1.267	.394	.184	-1.731	0	.081	.587	.233	.159	-.269	0	.079
6. San Mateo	.014	.270	.067	.024	.021	.007	.014	-.275	0	.006	-.133	0	.006	-.004	0	.006
7. Contra Costa	.617	-2.736	0	.493	-2.133	0	.491	3.063	.821	.684	-1.249	0	.509	1.528	.828	.737
8. Marin & Sonoma	.006	.041	.010	.007	.083	.026	.010	-.050	0	.004	.061	.024	.009	-.039	0	0
9. San Jose	.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10. Santa Clara (other)	.008	.245	.061	.167	.089	.028	.012	-.286	0	0	-.131	0	0	-.040	0	.002
11. Solano & Napa	.011	-.011	0	.010	.114	.036	.015	.048	.013	.012	.055	.022	.013	.005	.003	.010
12. Outside Bay Area	.003	-.080	0	0	-.064	0	0	.104	.028	.006	.016	.006	.004	.018	.010	.005

[1] Column (1) is the calculated probability from the linear probability model.
Column (2) is the adjusted probability obtained by setting the negative calculated probabilities to zero and normalizing the remainder to sum to one.
Column (3) is the adjusted probability obtained by an average of column (1) and the district average probabilities, with the least weight on the latter consistent with non-negative probabilities.

Extension of the Model to Include Level-of-Service Variables

The model described in the preceding section omitted transportation level-of-service variables. While the effects of level-of-service are captured in the baseline model by the alternative-specific parameters, the model is not sensitive to changes in level-of-service, and hence cannot be used to answer policy questions as to the effect of level-of-service changes on work destinations. (Of course, a complete picture of the impact of level-of-service changes also requires a description of the supply of jobs, and the equilibration process that equates demand and supply at various locations.) Further, to the extent that work destination choice is sensitive to level-of-service variables, and that these variables change over time without being accounted for in the model, erroneous forecasts will result.

District-to-district travel times and costs could be added to the explanatory variables in column (1), and parameter estimates obtained from the regression analysis. A further refinement could be made by using times and costs from the observed tract or zero to the destination district. The definition of these times and costs will in either case involve construction of indices. To the extent that the linear probability model is realistic, these indices can be defined as weighted averages of zone-to-zone variables, with weights proportional to the number of trips for each zone pair. If a multinomial logit model is more appropriate at this level, then the indices should have the form of "inclusive prices", as defined in Part I, Chapter 2.

An overall refinement to the destination choice model described in this section would be to replace destination-specific variables with generic measures of accessibility and attractiveness of destination districts from each origin district. Such a model would move in the direction of land use models currently calibrated by simulation methods with limited data.