4

Analysis of Population Mobility Data

4.1 Methods

4.1.1 Data

The Exposure Factors Handbook (EFH) has three key studies for population mobility. Each study uses a unique approach to define and estimate residence time. Israeli and Nelson (1992) work with current residence time (time since moving into the current residence) and total residence time (time between moving into and out of a residence). Current residence time does not seem to be directly relevant to risk assessment because it is censored; that is, the unobserved residence time is ignored. Total residence time is more relevant, but Israeli and Nelson (1992) apparently estimate it in a way that allows frequent movers to contribute more times than infrequent movers. The result is a residence time distribution that tends to be much shorter than those from the other two key studies; that is, median is 1.4 years versus a median of 9 years for each of the other two key studies.

The second key study is based on a national survey by the U. S. Bureau of the Census (1993) of 55,000 housing units that yielded 93,147 residence times. Residents were asked about time lived at current and past residences.

Johnson and Capel (1992) used a simulation model to estimate the distribution of residential occupancy periods based on a methodology described in Price et al. (1991a,b) and the EFH. Occupancy period is the time between a person moving into a residence and moving out or dying. Census data were used for the dynamics of mobility. Data from the National Center for Health Statistics were used for mortality.

Table 4-1 contains estimates of selected percentiles from the three key studies. For Israeli and Nelson (1992), total residence time is used.

The residence time distributions for the second and third key studies are fairly similar at the 25th, 50th, and 75th percentiles, but the distribution of Johnson and Capel (1992) has a shorter right tail than the Census Bureau (1993) distribution. Times from Israeli and Nelson (1992) tend to be much shorter.

The first of two relevant studies (National Association of Realtors, 1993) estimated an average occupancy period of 7.1 years for homeowners. However, the response rate was only 12%. The second relevant study (Lehman, 1994) estimated average residence times as 14.3, 13.4, and 12 years for 1991, 1992, and 1993, respectively. Apparently, residence times are decreasing. The 12-year average is similar to the estimate of Johnson and Capel (1992).

Based on discussions and comparisons of the studies, Johnson and Capel (1992) seem to provide the most representative summary for EPA risk assessment purposes. They are the only source of age-specific distributions, and age is clearly a relevant factor. The analysis of population mobility data will therefore focus on the age-specific distributions of EFH Table 14-159, taken from the simulation study of Johnson and Capel (1992).

4.1.2 Statistical Methods

Models were fit to the 30 different age groups of EFH Table 14-159, which includes simulated averages and six percentiles for each group. The data of Johnson and Capel (1992) from EFH Table 14-159 are shown in Table 4-2. The simulation sample size was 0.5 million, or about 17,000 per age group. However, because their data came entirely from Monte Carlo simulations, it did not seem appropriate to treat them as if they had come from a sample survey of a “real” population. Accordingly, the weighted least squares (WLS) regression methods were used to estimate models whose cumulative distribution functions (CDFs) came as close as possible to the nominal probabilities at the tabulated percentiles. The models used were the generalized gamma and its three two-parameter special cases (gamma, lognormal, and Weibull). The adequacy of fit of the two-parameter models was evaluated by comparison with the fit of the generalized gamma distribution, using an F test with one degree of freedom for the numerator and three degrees of freedom for the denominator. This is a GOF test relative to the three-parameter model.

4.2 Results

Table 4-3 summarizes results. For each age group, the best fitting two-parameter model is indicated in column 2. Columns 3 through 8 contain the values of the estimated CDFs for these models at the tabulated quantiles from EFH Table 14-159. As for tap water consumption, the goal is to estimate the CDF in order to come as close as possible to the nominal cumulative probabilities of 0.25, 0.50, 0.75, 0.90, 0.95, and 0.99. Columns 9 and 10 contain the estimated mean for the fitted model and the simulated mean from EFH Table 14-159. The next to last column contains the F test *p*-value, PGOF, for goodness-of-fit (GOF) of the selected model relative to the three-parameter generalized gamma model. The generalized gamma distribution improves significantly on the best fitting two-parameter model at the 5% significance level whenever PGOF <0.05. This occurs in 6 of 30 cases. In 20 of 30 cases, the best fitting two-parameter model was the Weibull model.

4.3 Uncertainty Analysis

Information on parameter uncertainty distributions can be summarized as for tap water consumption in Section 3, using parameter estimates and the asymptotic covariance matrix produced by the SAS nonlinear regression (NLIN) procedure. For the gamma and Weibull models, logarithms of the usual positive parameters should be used. For the lognormal model, the parameters should be the mean and logarithm of variance of the logarithm of residence time.

Work is in progress to develop parameter uncertainty distributions for population mobility.

4.4 Conclusions

Given that all three types of the basic two-parameter models are needed to adequately fit the population mobility data, it might appear simpler just to tabulate the best fitting generalized gamma distributions. However, this would somewhat complicate the uncertainty analysis, which would require the use of a trivariate normal distribution with some parameters very highly correlated, or the use of one of the other uncertainty methods.

Another promising approach to population mobility as well as tap water consumption involves the use of generalized gamma regression models (Section 2.2.6).

The analysis of population mobility data focused on the age-specific distributions of EFH Table 14-159 taken from the simulation study of Johnson and Capel (1992). However, Israeli and Nelson (1992) provide results for geographic regions, farms, urban versus rural, and renters versus owners. These factors are also relevant. Efforts are under way within EPA to develop region-specific distributions for residence time.

Extensive information on population mobility is available on the worldwide web (http://www.census.gov/prod/1/pop/p20-485.pdf). We recommend that this information be reviewed to determine its applicability to estimation of population mobility distributions.

Johnson and Capel (1992) developed a methodology to determine the distribution of residential occupancy periods in a simulated population of 500,000 individuals. The raw data used by Johnson and Capel (1992) and the assumptions made in their analysis are considered representative of the U.S. population. Therefore, the distributional summaries in table 4.3 are also considered representative of the general U.S. population. The distributions presented in Table 4-3 may be used when the general population is the population of concern and for the age groups presented.

Table 4-1. Selected Percentiles of Residence Times in Years from Three Key Studies

Statistic	Israeli & Nelson (1992)	Census Bureau (1993)	Johnson & Capel (1992)
25th percentile	0.5	4	4
50th percentile	1.4	9	9
75th percentile	3.7	18	16
90th percentile	12.9	32	26
95th percentile	23.1	40	33
Average	4.6	N/A	11.7

Table 4-2. Residence Time[a] Distributions in Years from Johnson and Capel (1992)

Age Group	Mean Years	Percentile					
		25th	50th	75th	90th	95th	99th
00-03	6.50	3	5	8	13	17	22
04-06	8.00	4	7	10	15	18	22
07-09	8.90	5	8	12	16	18	22
10-12	9.30	5	9	13	16	18	23
13-15	9.10	5	8	12	16	18	23
16-18	8.20	4	7	11	16	19	23
19-21	6.00	2	4	8	13	17	23
22-24	5.20	2	4	6	11	15	25
25-27	6.00	3	5	8	12	16	27
28-30	7.30	3	6	9	14	19	32
31-33	8.70	4	7	11	17	23	39
34-36	10.4	5	8	13	21	28	47
37-39	12.0	5	9	15	24	31	48
40-42	13.5	6	11	18	27	35	49
43-45	15.3	7	13	20	31	38	52
46-48	16.6	8	14	22	32	39	52
49-51	17.4	9	15	24	33	39	50
52-54	18.3	9	16	25	34	40	50
55-57	19.1	10	17	26	35	41	51
58-60	19.7	11	18	27	35	40	51
61-63	20.2	11	19	27	36	41	51
64-66	20.7	12	20	28	36	41	50
67-69	21.2	12	20	29	37	42	50
70-72	21.6	13	20	29	37	43	53
73-75	21.5	13	20	29	38	43	53
76-78	21.4	12	19	29	38	44	53
79-81	21.2	11	20	29	39	45	55
82-84	20.3	11	19	28	37	44	56
85-87	20.6	10	18	29	39	46	57
88-90	18.9	8	15	27	40	47	56
Adult	16.2	8	14	22	30	36	48
All	11.7	4	9	16	26	33	47

[a]Number of years between the date that a person moves into a new residence and the date that a person dies or moves out of the residence.

Table 4-3. Results of Statistical Modeling of Population Mobility Data[a]

Age Group	Best Model	P25	P50	P75	P90	P95	P99	Model Mean	Data Mean	PGOF
00-03	Wei2	.287	.485	.709	.904	.965	.991	6.3	6.5	.088
04-06	Wei2	.257	.513	.718	.909	.959	.988	7.7	8.0	.597
07-09	Wei2	.248	.491	.758	.910	.949	.986	8.8	8.9	.299
10-12	Wei2	.217	.527	.779	.894	.940	.989	9.3	9.3	.096
13-15	Wei2	.252	.492	.755	.906	.946	.989	8.8	9.1	.720
16-18	Wei2	.252	.495	.744	.911	.957	.985	8.1	8.2	.745
19-21	Gam2	.260	.480	.752	.905	.956	.987	5.8	6.0	.949
22-24	Log2	.237	.542	.721	.904	.953	.989	5.2	5.2	.782
25-27	Log2	.245	.509	.752	.894	.950	.991	6.4	6.0	1.00
28-30	Log2	.222	.552	.744	.890	.948	.989	7.3	7.3	.336
31-33	Log2	.241	.520	.745	.894	.951	.991	8.9	8.7	.751
34-36	Log2	.259	.495	.739	.901	.953	.991	10.6	10.4	1.00
37-39	Log2	.235	.515	.757	.904	.949	.986	11.8	12.0	.011
40-42	Gam2	.255	.501	.740	.896	.956	.991	13.5	13.5	.192
43-45	Gam2	.248	.513	.732	.905	.953	.989	15.3	15.3	.872
46-48	Wei2	.261	.497	.736	.900	.954	.991	16.4	16.6	.055
49-51	Wei2	.257	.486	.754	.902	.951	.989	17.4	17.4	.882
52-54	Wei2	.241	.501	.759	.903	.951	.987	18.0	18.3	.030
55-57	Wei2	.246	.498	.756	.902	.952	.987	18.9	19.1	.123
58-60	Wei2	.245	.498	.762	.900	.947	.989	19.7	19.7	.204
61-63	Wei2	.235	.517	.749	.904	.949	.988	20.1	20.2	.118
64-66	Wei2	.233	.516	.755	.900	.949	.988	20.9	20.7	.042
67-69	Wei2	.231	.507	.767	.904	.950	.985	21.2	21.2	.002
70-72	Wei2	.253	.493	.755	.896	.952	.990	21.2	21.6	.763
73-75	Wei2	.254	.491	.751	.905	.950	.989	21.7	21.5	.933
76-78	Wei2	.251	.486	.762	.904	.953	.986	21.1	21.4	.413
79-81	Wei2	.229	.520	.754	.904	.951	.986	21.3	21.2	.048
82-84	Wei2	.243	.510	.751	.895	.952	.990	20.6	20.3	.518
85-87	Wei2	.239	.498	.768	.903	.951	.986	20.5	20.6	.039
88-90	Wei2	.244	.483	.770	.919	.956	.981	18.6	18.9	.259

[a]Tabulated probabilities are obtained by evaluating the best fitting two-parameter CDF at the percentile value of Johnson and Capel (1992), from Table 4-2.