Title
Development of an estimation procedure for an activity-based travel demand model

Authors
Recker, Will W
Duan, J.
Wang, H.

Publication Date
2008

Development of an estimation procedure for an activity-based travel demand model

W. Recker*, J. Duan and H. Wang

Department of Civil and Environmental Engineering and Institute of Transportation Studies
University of California, Irvine, CA 92697, U.S.A.

Abstract: In this paper, we implement an estimation procedure for a particular mathematical programming activity-based model in order to estimate the relative importance of factors associated with spatial and temporal interrelationships among the out-of-home activities that motivate a household's need or desire to travel. The method uses a genetic algorithm to estimate coefficient values of the utility function, based on a particular multidimensional sequence alignment method to deal with the nominal, discrete, attributes of the activity/travel pattern (e.g., which household member performs which activity, which vehicle is used, sequencing of activities), and a time sequence alignment method to handle temporal attributes of the activity pattern (e.g., starting and ending time of each activity and/or travel). The estimation procedure is tested on data drawn from a well-know activity/travel survey.

Keywords: Activity-based travel analysis, Travel demand modeling, Mathematical programming

1. Introduction

An essential element in the development of conventional trip-based disaggregate travel demand models is the inferring the relative weights associated with potential components of the utility function that are determinants to a population's revealed selection of the decision variables (in the model estimation phase) with subsequent forecasts made using these weights in the application of the model. This particular aspect of the modeling process application of the activity-based research approach has remained a challenge.

In this paper, we propose and implement an estimation procedure for the Household Activity Pattern Problem (HAPP) model (a mathematical programming activity-based model offered by Recker, 1995) in order to estimate the relative importance of factors associated with the spatial and temporal interrelationships among the out-of-home activities that motivate a household's need or desire to travel. The procedure provides both the necessary constraint considerations on the household's decision alternatives within a utility-maximizing structure as well as a convenient mechanism for generating the set of feasible alternatives that are likely to be considered.

---

* Corresponding Author. wwrecker@uci.edu. Institute of Transportation Studies, University of California, Irvine, CA 92697-3600. Tel: 949.824.5642. Fax: 949.824.8385

The HAPP model is in the form of a Mixed Integer Linear Programming model (MILP), i.e., one comprising continuous variables (such temporal attributes of an activity pattern as the starting times of the associated activities) as well as discrete variables (e.g., attributes associated with the sequencing of activities, travel modes used, and persons performing the activities). Because of this complexity, an estimation approach based on a heuristic algorithm is the best (and, arguably, the only) available option for solution. From the collection of such heuristic algorithms, a Genetic Algorithm (GA) approach (Holland, 1992 a,b) was used because of its considerable advantages for this particular application. To deal with the comparison of discrete attributes between the actual activity pattern and the predicted activity pattern, we use a variation of the Multidimensional Sequence Alignment Method (Arentze, et al, 2002). The comparison of continuous attributes is based on a Time Sequence Method that was developed as part of this research.

The paper is organized as follows. We first provide a brief overview of the basic HAPP model as background for the current effort. We next state the general abstract form of the HAPP model and specify the activity-based HAPP model estimation problem. Then, we propose an estimation procedure using a GA. We describe in detail how to implement the proposed estimation procedure using the GA and how to compare discrete and continuous attributes between the actual activity pattern and the predicted activity pattern. We use data drawn from a well-known activity-travel survey to test the implementation of the proposed estimation procedure, and then list further research needed to extend the estimation procedure.

2. General Form of the HAPP Model and Its Estimation Problem

The general form of the HAPP mathematical program formulation of the travel/activity decisions for a particular household, say, *i*, during some time period is represented by:

$$\text{Minimize } \mathbf{Z}(\mathbf{X}_i) = \mathbf{B}_i' \cdot \mathbf{X}_i \qquad (1)$$

Subject to:

$$\mathbf{A} \cdot \mathbf{X}_i \le \mathbf{0}$$

where:

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{X}^{\mathbf{v}} \\ -- \\ \mathbf{H} \\ -- \\ \mathbf{T} \end{bmatrix},\ \mathbf{X}^{\mathbf{v}} = \left[X_{uw}^{v} = \begin{cases} 0 \\ 1 \end{cases} \right],\ \mathbf{H} = \left[H_{uw}^{\alpha} = \begin{cases} 0 \\ 1 \end{cases} \right],\ \mathbf{T} = \left[T_u \ge 0 \right],$$

The output $\mathbf{X}_i$ of the optimization of each household *i* are specified by the following decision variables:

X_{uw}^{v}: binary decision variable equal to unity if vehicle *v* travels form activity *u* to activity *w*, and zero otherwise.

H_{uw}^{α}: binary decision variable equal to unity if household member α travels from activity *u* to activity *w*, and zero otherwise.

T_u: the time at which participation in activity *u* begins.

In the above, $\mathbf{B}_i$ is a vector of coefficients that defines the relative contributions of each of the decision variables to the overall disutility of the travel regime to the household. Descriptively, the constraint sets $\mathbf{A} \cdot \mathbf{X}_i \leq \mathbf{0}$ for this MILP are classified into six groups: (a) routing constraints that define the allowable spatial movement of vehicles and household members in completing the household's activity agenda; (b) scheduling constraints specify the relationship of arrival time, activity begin time, and waiting time, and continuity condition along the temporal dimension; (c) assignment constraints that are applied to match the relations between activity participation and vehicle usage as well as activity performers (household members); (d) time window constraints that are used to specify available schedules for activity participation; (e) coupling constraints that define the relations between vehicle-related variables and member-related variables: and (f) side constraints including budget, capacity and rules for ride-sharing behavior. With the exception of the side constraints (i.e., classification "f" above"), these constraints capture the physical conditions that ensure that each member of the household, as well as each vehicle used by the household, have a consistent, continuous, path through time-space that results in all of the activities on the household's agenda being successfully completed. The reader interested in a detailed derivation and explanation of these constraints is referred to the original work by Recker (1995).

The solution vector, $\mathbf{X}_i^*$ to equations (1) represents the household's utility maximizing behavior, relative to the prescribed objective $Z(\mathbf{X}_i)$, with regard to completing its activity agenda. The solution patterns reveal personal travel behavior and activity participation within a household context, while preserving the concept that the need for travel originates from participation in activities, that travel constitutes the linkage between activities, and in which all of the required components are contained in the activity scheduling problem.

The methodology described above has been applied successfully to a number of transportation applications to explore issues relating to such areas as vehicle emissions, accessibility, trip-chaining, ride-sharing and travel time reduction (see e.g., Recker et al, 2001; Recker and Parimi, 1999). However, in all

of these applications, the specification of the objective function is ***prescribed*** by the analyst; e.g., the minimization of emissions produced by travel, rather than ***estimated*** from the revealed choices made by members of the household. The inference of the relative weights associated with potential components of the utility function that are determinants to a population's revealed selection of the decision variables is necessary before conventional application of the model to forecasts can be made; this particular aspect of the application of the activity-based research approach has remained a challenge.

Although the form of Equations (1) is generally similar to one that can be used to describe the general discrete choice problem in transportation (see Recker, 2001 for an example of such a formulation), there are some notable exceptions that greatly complicate its application in empirical demand analysis: 1) the set of feasible solutions (alternatives) in the system defined by Equations (1) is infinite, while that in traditional discrete choice analysis is countable (and, usually small), 2) the solution vector of Equations (1) comprises continuous, as well as discrete, variables, 3) while the overall solution represents a mutually exclusive choice, the solution vector itself is composed of components that are not generally mutually exclusive, 4) the components of $\mathbf{B}_i$, are not directly interpretable as utility weights of attributes, but rather are related to these weights through a transformation matrix, and 5) the complexity of the constraint spice of Equations (1) generally precludes the type of closed-form probability result.

The analyst can not directly observe $\mathbf{B}_i$; rather, an estimate, $\hat{\mathbf{B}}_i$, is sought that can be inferred from the observed behavior, $\tilde{\mathbf{X}}_i^*$. The goal, then, is to find the $\hat{\mathbf{B}}_i$ that minimizes some prescribed error $\boldsymbol{\varepsilon}_i$ between the solution vector $\mathbf{X}_i^*$ and the observed behavior $\tilde{\mathbf{X}}_i^*$. In the case of discrete choice analysis, this is usually accomplished with maximum likelihood estimation of $\hat{\mathbf{B}}_i \equiv \hat{\mathbf{B}}, \ \forall i$, i.e., assuming that the utility weights are common across observations, and assuming that the error terms are independently identically distributed (IID). The standard application of maximum likelihood requires that the model choice probabilities be a differentiable function of the parameters contained in $\mathbf{B}$. Without such stipulation, the maximization process could be accomplished using some heuristic search. In the case of the HAPP model (i.e., Equations 1), the only option available is a heuristic. The particular form of the HAPP model lends the accompanying estimation problem to solution by GA (Rudolph, 1994). Owing to the specification of $\mathbf{X}_i$ in Equation (1), we define a three-tuple

$$\boldsymbol{\mu}(\hat{\mathbf{B}}_i^p) = \left\{ \mu_1(\hat{\mathbf{B}}_i^p), \mu_2(\hat{\mathbf{B}}_i^p), \mu_3(\hat{\mathbf{B}}_i^p) \right\} \tag{2}$$

where its elements,

$$\mu_1(\hat{\mathbf{B}}_i^p) = err\left|\mathbf{T}_i(\hat{\mathbf{B}}_i^p) : \tilde{\mathbf{T}}_i \mid \tilde{\mathbf{X}}_i^v, \tilde{\mathbf{H}}_i\right|$$
$$\mu_2(\hat{\mathbf{B}}_i^p) = err\left|\mathbf{X}_i^v(\hat{\mathbf{B}}_i^p) : \tilde{\mathbf{X}}_i^v \mid \tilde{\mathbf{H}}_i, \tilde{\mathbf{T}}_i\right|,$$
$$\mu_3(\hat{\mathbf{B}}_i^p) = err\left|\mathbf{H}_i(\hat{\mathbf{B}}_i^p) : \tilde{\mathbf{H}}_i \mid \tilde{\mathbf{X}}_i^v, \tilde{\mathbf{T}}_i\right|$$

represent the error between the observed and predicted values of the respective components of the particular household observation's activity pattern, for any particular candidate solution $\hat{\mathbf{B}}_i^p$, $p = 1, \ldots, n_p$. We interpret $\boldsymbol{\mu}(\hat{\mathbf{B}}_i^p)$ as a multi-objective measure that indicates the fitness of any potential solution vector, $\hat{\mathbf{B}}_i^p$, in minimizing the discrepancies between the observed and model values. Then, using standard procedures in genetic algorithms, we generate an initial population of candidate solutions $\hat{\mathbf{B}}_i^{p^1}$, $p^1 = 1, \ldots, n_p^1$, and perform mating, crossover, and mutations (Holland, 1992a, Michalewicz, 1994) according to pre-described probability rules based on respective fitness scores to produce succeeding populations of candidate solutions, $\hat{\mathbf{B}}_i^{p^j}$, $p^j = 1, \ldots, n_p^j$; $j = 1, \ldots, J$,until the population converges to within sufficient tolerance of the observed values; i.e., $\hat{\mathbf{B}}_i = \hat{\mathbf{B}}_i^{p^j}$ ∋ $\lim_{j\to\infty}\left|\boldsymbol{\mu}(\hat{\mathbf{B}}_i^{p^j})\right| \le \varepsilon^*$, where ε^* is the prescribed error tolerance.

3. Implementation of the Estimation Procedure

The estimation procedure using the genetic algorithm described in the previous section was applied with one minor modification. To save computation time, we compressed the three-tuple into a two-tuple in the objective function by combining $\mu_2(\hat{\mathbf{B}}_i^p)$ for $\mathbf{X}_{uw}^v$ and $\mu_3(\hat{\mathbf{B}}_i^p)$ for $\mathbf{H}_{uw}^\alpha$ into a single measure $\mu_{23}(\hat{\mathbf{B}}_i^p)$. This is non restrictive because both $\mathbf{X}_{uw}^v$ and $\mathbf{H}_{uw}^\alpha$ in the general form of the HAPP model are (nominal) qualitative variables and, in the demonstration application reported here, weights on each of the three dimensions were arbitrary and assumed to be equal. We then used the Multi-Dimensional Sequence Alignment Method to compare multiple qualitative properties, $\mathbf{X}_{uw}^v$ and $\mathbf{H}_{uw}^\alpha$, under this common measure.

The steps in the estimation are as follows:

1. For each household *i*, define the objective function as Minimize $\boldsymbol{\mu}(\hat{\mathbf{B}}_i^p) = \left\{\mu_1(\hat{\mathbf{B}}_i^p), \mu_{23}(\hat{\mathbf{B}}_i^p)\right\}$.

   $\boldsymbol{\mu}(\hat{\mathbf{B}}_i^p)$ indicates the fitness of any potential solution vector, $\hat{\mathbf{B}}_i^p$, in achieving the objective of minimizing the error between the observed activity pattern (OAP) for household *i* and the predicted

activity pattern (PAP) resulting from the HAPP model. We define $\boldsymbol{\mu}(\hat{\mathbf{B}}_i^p)$ as the "edit distance" between PAP and OAP. The component $\mu_{23}(\hat{\mathbf{B}}_i^p)$ is defined as the "Sequence Distance" between PAP and OAP, and is measured by the number of steps needed to equalize the nominal (discrete) aspects of the two sequences, related only to such qualitative properties of activity pattern as: which activity to do at any particular point in the sequence of activity participation, which vehicle to use, which person to be assigned to the activity. We calculate Sequence Distance using the Multi-Dimensional Sequence Alignment Method (MDSAM) (Arentze, *et al*, 2002). Alternatively, $\mu_1(\hat{\mathbf{B}}_i^p)$ is defined as the "Temporal Similarity" between PAP and OAP, and indicates how similar the two activity sequences are in their temporal dimension, i.e., how much specific time usage in the sequence determined by the HAPP model overlaps that of the actual, observed, pattern. Temporal Similarity is related to such continuously varying properties of the activity pattern as starting time and ending time of each particular activity and/or travel that comprises the activity pattern. We compute Temporal Similarity using the Time Sequence Comparison Method (TSCM) described below. The "edit distance" goal is to equalize the observed and predicted activity patterns using the least editing effort; so, we define, for any given household *i*,

$$\underset{\hat{\mathbf{B}}_i^p}{Min}\, \boldsymbol{\mu}(\hat{\mathbf{B}}_i^p) = a\,\mu_{23}(\hat{\mathbf{B}}_i^p) - b\,\mu_1(\hat{\mathbf{B}}_i^p) + c \;;\; a,b,c \text{ positive constants} \qquad (3)$$

as the objective function; i.e., select the utility weights comprising $\hat{\mathbf{B}}_i^p$ that minimize the weighted sum of the number of steps required to align the respective activity sequences (PAP and OAP), i.e., actors performing the activities, and vehicles used to access the activities, while maximizing the temporal overlap between the two patterns. This implies that we minimize $\mu_{23}(\hat{\mathbf{B}}_i^p)$ and maximize $\mu_1(\hat{\mathbf{B}}_i^p)$. The constant *c* is used to scale the objective function to some range such as [0,1].

For a sample of *N* households, each household has its own PAP and OAP; correspondingly, each pair of PAP and OAP will have its own "edit distance" for any particular $\hat{\mathbf{B}}_i^p$. In applying Equation (3) as the kernel for estimating the optimal common utility weights (in the sense of minimizing the total error between the respective OAPs and PAPs for the sample), i.e., $\hat{\mathbf{B}}_i \equiv \hat{\mathbf{B}}$, $\forall i$, for the sample of N households, we form the objective function for determining $\hat{\mathbf{B}}$ as:

$$\underset{\hat{\mathbf{B}}}{Min}\,\boldsymbol{\mu}(\hat{\mathbf{B}}) = \sum_{i=1}^{N} w_i \cdot \boldsymbol{\mu}(\hat{\mathbf{B}}_i)$$
$$= \lim_{j\to\infty} \sum_{i=1}^{N} w_i \cdot (a\,\mu_{23}^{i}(\hat{\mathbf{B}}_i^{p^j}) - b\,\mu_1^{i}(\hat{\mathbf{B}}_i^{p^j}) + c) \qquad (4)$$

where w_i is a weighting factor that can be assigned to account for differential importance placed on the error contribution of household i to the total error, $a, b, \text{and } c$ are positive constants, and where $\mu_{23}^{i}(\hat{\mathbf{B}}_i^{p^j})$ and $\mu_1^{i}(\hat{\mathbf{B}}_i^{p^j})$ are scaled to $[0,1]$.

2. Generate a population, say the jth, of n_p^j candidate solutions $\hat{\mathbf{B}}_i^{p^j} = \hat{\mathbf{B}}^{p^j}, \forall i,\ p^j = 1, \ldots, n_p^j$.

   Since the signs of the coefficients of $\hat{\mathbf{B}}_i^{p^j}$ typically are known, the utility components can always be specified in such a manner that the elements of $\hat{\mathbf{B}}_i^{p^j}$ are strictly non-negative; and, since the solution is unaffected by the scaling of the utility coefficients, they can be scaled to be in a given range. We first encode each coefficient in $\hat{\mathbf{B}}_i^{p^j}$ as a binary (0, 1) string. Then, all of these strings are concatenated together in sequence to form a chromosome; the elements of each chromosome are referred to as genes.

   We encode the integer part and decimal fraction of each coefficient separately. For the integer part, we define a range of [0,2^m], where m is the length of the binary string for the integer part. For the decimal fraction, we define a range of [0, 2^n], where n is the length of the binary string for decimal fraction. So, the first m bits of an m+n binary string represent the integer part of a coefficient and its last n bits represent the decimal fraction of the coefficient. Decoding is accomplished by separating the chromosome into two parts. The first part (length = m) is for integer part, which can be transformed directly from binary string to decimal; the second part (length = n) is for decimal fraction, which is transformed by a two-step process: 1) transform the binary string to decimal integer, 2) divide the decimal integer obtained from the fist step by 2^n.

3. Each chromosome, $\hat{\mathbf{B}}^{p^j}$, $p^j = 1, \ldots, n_p^j$ in the population is assigned a "fitness score"

$$\mu(\hat{\mathbf{B}}^{p^j}) = \sum_{i=1}^{N} w_i \cdot \boldsymbol{\mu}(\hat{\mathbf{B}}_i^{p^j}) = \sum_{i=1}^{N} w_i \cdot (a\,\mu_{23}^{i}(\hat{\mathbf{B}}_i^{p^j}) - b\,\mu_1^{i}(\hat{\mathbf{B}}_i^{p^j}) + c);\ p^j = 1, \ldots, n_p^j \qquad (4)$$

Here, for demonstration purposes, both $\mu_{23}^{i}(\hat{\mathbf{B}}_{i}^{p^{j}})$ and $\mu_{1}^{i}(\hat{\mathbf{B}}_{i}^{p^{j}})$ are scaled to [0, 1]. In order to scale $\mu_{23}^{i}(\hat{\mathbf{B}}_{i}^{p^{j}})$ to [0, 1], the maximum possible value of the Sequence Distance, as computed by the MDSAM method, is required. The results reported here are based on the assumption that insertion and deletion costs are 1, while the substitution cost is 2 (effectively a deletion followed by an insertion). We also assume that each discrete attribute may have its own relative cost weight; we label the largest cost weight among the attributes as $\omega_{\max}$. If, in addition, the OAP and PAP sequences have different lengths m and n (m>n), the maximal MDSAM operation cost will be $\omega_{\max} * [2n + (m-n)] = \omega_{\max} * (m+n)$. Therefore, the MDSAM result scaled to [0, 1] could be obtained through normalization by $\omega_{\max} * (m+n)$. Correspondingly, the normalization value for $\mu_{1}^{i}(\hat{\mathbf{B}}_{i}^{p^{j}})$ will be the maximum time coverage, or the sum of all the durations in the observed sequence.

Scaled as above, the Sequence Distance and Time Coverage results can be combined to represent the integrated "edit distance" between the predicted and the observed sequences for any given household *i* as:

$$\mu(\hat{\mathbf{B}}_{i}^{p^{j}}) = a\,\mu_{23}^{i}(\hat{\mathbf{B}}_{i}^{p^{j}}) - b\,\mu_{1}^{i}(\hat{\mathbf{B}}_{i}^{p^{j}}) + c = \frac{(\text{Sequence Distance} - \text{Time Coverage}) + 1}{2} \qquad (5)$$

The range of this result will be [0,1].

4. Each chromosome in population *j* is assigned a series of three "probability of reproduction" values, each of which is inversely proportional to its fitness $\mu(\hat{\mathbf{B}}^{p^{j}})$ relative to the other chromosomes in the population i.e., the lower the value of the fitness measure (distance to the observed), the higher the probability of reproduction.

5. According to the assigned probabilities of reproduction, a new population, *j*+1, is generated by "mating" a selection drawn according to the probability.

6. The selected pairs of chromosomes generate offspring via the use of such specific genetic operators as crossover and gene mutation. Crossover is applied to two chromosomes (parents) and creates two new chromosomes (offspring) by selecting a random position along the coding and then splicing the section that appears before the selected position in the first string with the section that appears after selected position in the second string, and vice versa.

7. The process is halted if a suitable solution is found. Otherwise, the process returns to step 3, where the new chromosomes are scored and the procedure iterates.

To initiate the GA procedure, we generate an initial population randomly. Subsequently, new populations are generated in the mating pool using the GA operations described above, based on the given probabilities.

For the results presented here, the HAPP Mixed-Integer Linear Programming model is solved using the CPLEX algorithm in the GAMS software package developed by the World Bank. Under the current implementation, the algorithm takes as input the composition (e.g., number of household members, number of vehicles by type), activity plan (e.g., number, location and type of activities to be completed) and constraints (e.g., activity time windows, pre-assigned roles) faced by a particular household, and a set of model coefficients (i.e., candidate utility weights), which are decoded from the chromosome produced by the GA. (Each chromosome includes all coefficients in the HAPP model, represented as a binary string.) Based on this information, the HAPP model outputs the predicted activity pattern that maximizes the objective utility function for the particular utility weights.

The first stage of the process optimally (i.e., minimum effort) "re-constructs" the nominal parameters (i.e., integer variables of the HAPP model) of the predicted household activity sequence to perfectly align with the observed sequence of the particular household (i.e., to match the observed assignments of household member and vehicle to each activity in the household's activity plan, as well as the observed sequence in which the respective activities were performed). This process uses the Multi-Dimensional Sequence Alignment Method to compare the two activity sequences, and produces the minimal effort required, $\mu_{23}^{i}(\hat{\mathbf{B}}^{p^j})$, to perfectly align the two activity sequences, as follows:

1. Calculate the Levenshtein distance between each pair of uni-dimensional sequences in the source and target multi-dimensional sequences, using the Uni-Dimensional Sequence Alignment Method (UDSAM) of MDSAM.

2. Construct a ***tri-branch optimum trajectory tree*** for each pair of uni-dimensional sequences, according to Levenshtein distances above.

3. Search the trees to find all optimum trajectories (paths) for each pair of uni-dimensional sequences and record the operation set of each trajectory. Delete, substitute, and insert operations are recorded, but identity operations are ignored because they are cost-free. All trajectories are stored in a dynamic vector**.**

4. Produce all possible element-operation-based combinations from all optimum trajectories of every pair of sequences, and then rearrange each element-operation-based combination to segment-operation-based combination. To calculate more efficiently, hash tables are used to store all possible segments (no duplicate) in each trajectory.

5. Calculate the costs of each segment-operation-based combination above.

6. Compare all costs, and get the optimum combinations with the lowest cost.

Following this stage in the process, the predicted activity pattern (PAP) of any particular household in the data set will be sequentially aligned with the observed (OAP), but the predicted patterns corresponding activity start times will generally be different from those observed.

The second stage of the process uses a Time Sequence Comparison Method to compare the two activity sequences, and determine the similarity **S** (overlap time) between identical activities in the two activity sequences. This second stage has two distinct components: 1) global temporal shift of the entire pattern, and 2) local temporal comparison of each activity.

The steps in this procedure are as follows:

1. Re-adjust the predicted sequence according to the observed sequence performed by the same person, i.e., re-shuffle the predicted activity sequence to the same order as the observed sequence. During the reshuffle, if an activity in the observed sequence is missing in the predicted sequence, then insert this activity into predicted sequence, but with start time = -1. If there is an activity in the predicted, but missing in observed sequence, simply delete it from predicted sequence

2. Using the predicted activity sequence in step 1 as a base, produce the corresponding starting time and ending time for each activity in the predicted and observed patterns. Activities in the two lists have a one-to-one relationship relative to the pair (Activity, Person).

3. Global alignment to decide the starting point of local comparison in the two lists.

```
Do Loop
{
Move pointers forward in two lists one by one;
If (two activities in current pointer position are same)
{
        If (both of start times > 0)
                Do Local Time Sequence Comparison;
}
}while (all activities are scanned)

Return optimal result based on a given dimension (see below)
```

4. Local Time Comparison. For each dimension (e.g., person performing, vehicle mode in use, etc.), calculate the time overlap of the predicted activity pattern with its corresponding observed activity pattern—the common time in this comparison is a measure of the time similarity for this particular pair of activities. Return to the global alignment (step 3) to continuously compare other pair of activities in the activity sequence.

4. Initial Tests

To demonstrate the estimation procedure, we randomly selected household activity patterns 65 two-adult-member households extracted from the Southwest Washington and Oregon Area 1994 Activity and Travel Behavior Survey, which contains sufficiently detailed information, including comprehensive travel/activity diaries (with mode availability) and regional transportation network model, to support an application of the HAPP model used in this study. The activities considered in the analysis are all out-of-home activities and in-home meals for the first day of the two-day activity diary; in-home activities other than meals are assumed to be discretionary and flexible. The mean number of activities per household is 6.56 with a standard deviation of 2.23; the mean number of activities performed by all licensed drivers within a household is 5.84 with a standard deviation of 1.87. About 62 percent of the activities were out-of-home activities that required travel (17.3% work, 14.6% general shopping, and 23.9% from the remaining categories), with an average travel time per activity requiring travel of approximately 18 minutes. The mean total travel time of those individuals in the household who traveled is 0.95 hours. The mean duration of a meals activity is 1 hour and the corresponding mean duration of work and shopping activities is 6 hours and 0.8 hours, respectively. Approximately 75 percent of the households have two vehicles and about 93 percent of the households have more than one vehicle; 90 percent of the households had two licensed drivers, while 8 percent had only one licensed driver.

Since the HAPP model represents activity/travel behavior of the collective members of a household, detailed information is required on travel and activity participation for each member, as well as transportation supply information (including household vehicle holdings and network travel times). To help define the time window constraint space, the average activity starting and ending times for each activity type were computed for the whole sample to provide benchmark information on the temporal flexibility of the activities. Shortest path travel times between all activity locations of each household have been generated for all the households in the sample using TRANSCAD. This procedure allows for the exploration of all possible activity/travel linkages for each household, which is fundamental to the optimization procedure.

Following Equation (1), for each of the households drawn from the sample for this initial test, a common disutility function was assumed of the form:

$$\text{Minimize } \mathbf{Z}(\mathbf{X}_i) = \mathbf{B}_i' \cdot \mathbf{X}_i \tag{6}$$

where $\mathbf{B}_i$ is a vector of unknown parameters (to be estimated), and

$$\mathbf{X}_i = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} = \begin{bmatrix} \sum_{v\in\mathbf{V}}\sum_{u\in\mathbf{N}}\sum_{w\in\mathbf{N}} c_{uw}^{v} \cdot X_{uw}^{v} \\ \sum_{v\in\mathbf{V}}\sum_{u\in\mathbf{N}}\sum_{w\in\mathbf{N}} t_{uw}^{v} \cdot X_{uw}^{v} \\ \sum_{u\in\mathbf{P}^+} (T_{u+n} - T_u - s_u - t_{u,u+n}^{v}) \\ \sum_{u\in\mathbf{P}^+} (T_u + s_u - b_u) \\ \sum_{u\in\mathbf{P}^-} (T_u - b_u) \\ \sum_{v\in\mathbf{V}} (T_{2n+1}^{v} - T_0^{u}) \\ \sum_{v\in\mathbf{V}}\sum_{u\in\mathbf{P}^+} c_v \cdot X_{0u}^{v} \\ \sum_{u\in\mathbf{P}^+} \left(T_u - \sum_{v\in\mathbf{V}}\sum_{w\in\mathbf{N}} (T_w + t_{wu}^{v}) \cdot X_{wu}^{v} \right) \end{bmatrix} \tag{7}$$

The notation used in Equation (7) is the same as that in Recker (1995). Succinctly, c_{uw}^{v} and t_{uw}^{v} are respectively the travel cost and time for a trip from the location of activity *u* to the location of activity *w* by vehicle *v*; s_u is the duration of activity *u*; b_u is the close of the availability of activity *u*; $\mathbb{V}$ is the set of vehicles available to the household; $\mathbf{P}^+$ is the set of *n* out-of-home activities; $\mathbf{P}^-$ is the set of *n* return-home activities; nodes 0 and $2n+1$ are respectively the beginning- and end-of-day locations; and $\mathbf{N} = \{0, \mathbf{P}^+, \mathbf{P}^-, 2n+1\}$. A qualitative interpretation of the variables comprising the assumed disutility function is provided in Table 1.

Based on this objective function, solutions to the corresponding HAPP model for each household were obtained for an initial population of 20 chromosomes, i.e., $\hat{\mathbf{B}}_\mathbf{p}, p = 1,\ldots,20$, each having 8 genes corresponding to the 8 unknown coefficients of the x_i. The resulting fitness values were determined according to the two-stage process described above (i.e., minimizing the Levenshtein distance between the observed and model-predicted discrete sequence components, followed by the time sequence adjustment phase), and input to the genetic algorithm. The parameters used by the genetic algorithm are displayed in Table 2.

4.1 Demonstration Application Inferring $\mathbf{B}_i$ for a Single Household

In this section, we provide some detailed results of application of the estimation procedure to a series of single, independent, households; i.e., those for which we did not invoke the standard constraint of having common coefficients, $\hat{\mathbf{B}}_i^{p^j} = \hat{\mathbf{B}}^{p^j}, \forall i,\ p^j = 1, \ldots, n_p^j$. Shown below, in Figures 1, 7 and 9, are schematic representations of the actual observed activity patterns of three households, each of which is typical of those in the data set that are comprised of two persons of driving age and having two vehicles available to the household.

Figure 2 shows the convergence of the genetic algorithm for household observation #1; displayed are statistics on the overall fitness values for five generations of the population. (Since the scale on the fitness axis is arbitrary, the numerical values are not displayed.) Because the constraints on the problem limit the number of feasible patterns, and feasibility is a necessary condition, even the initial population (as well as subsequent generations) contained one or more "solutions" that matched the observed pattern of Figure 1.

The algorithm converged to the actual, observed, pattern in only four generations. The predicted patterns associated with the worst (i.e., greatest positive) fitness values for generations #1 and #2 are displayed in Figures 3 and 4, respectively. In these figures, errors in temporal values (e.g., starting times) have been identified by placing the correct, observed, value in parentheses next to the predicted values; errors in discrete variables, such as paths and vehicles used, have been designated with the symbol ⊘. We note that the predicted patterns (i.e., as output by the HAPP model) have an expanded version of the "home" node (as required by the HAPP formulation) that includes the "virtual" activities corresponding to the initial departure from home (labeled "ACT 0"), the return-to-home activity for each real activity $i = 1, \ldots, n$ on the household's activity agenda (labeled "ACT $i + n$"), and the end-of-day activity (labeled "DNP1"). The observed pattern for observation #1 (Figure 1) indicates that person #1 departed from home in vehicle #1 at 7.00 hours (time is in military decimal), arriving at the destination to perform activity #1 at 7.17, where participation begins immediately upon arrival. Immediately following completion of the activity, person # 1 returns home. Person #2 leaves home at 9.91 in vehicle #2, and trip-chains activities #2 and #3 before returning home. The predicted pattern in Figure 3, which corresponds to the HAPP pattern with the worst fitness score relative to the observed pattern, exhibits both activity linkage errors and temporal deviations from the observed (Figure 3). It fails to capture the trip chaining behavior of person #2, as well as not correctly matching the initial departure from home for person #2 and the start and end times for activity #2; this pattern is eliminated after the first generation by the genetic algorithm.

The only discrepancy in the pattern with the worst fitness score in generations 2 and 3 is in the predicted vehicle used by persons #1 and #2 in completing their activities (Figure 4). By generations 4 and 5, all ten of the predicted patterns in each of these generations matched the observed patterns exactly (Figure 5).

The corresponding convergence in the utility coefficients to produce an exact match between the predicted and observed patterns is displayed in Figure 6, where the mean value of the estimated coefficients, together with their standard deviation error band are plotted for each generation. (We note that, as with the case in traditional demand modeling, the set of model coefficients that replicates the observed outcome generally is not unique.) After five generations, most coefficients have converged to a single value for the entire population; those that have not differ from each other by relatively small amounts.

Shown in Figure 7 is the observed pattern for a second household, Household #2. The activity pattern of this household included three specific "in-home" activities (meals): Activity #1 (Person #1's breakfast, and Activities labeled #3 and #5 (joint dinner). Shown in Figure 7 is the corresponding predicted pattern with the best fitness score after five generations. By the fifth generation the estimated utility coefficients produced by the genetic algorithm resulted in a prediction that varied from the observed only in the timing of Activity #4; Activity #4 is predicted to commence 0.17 hours later than observed, and following its completion the prediction has person #2 waiting at the location of Activity #4 for 1.5 hours before returning home to participate in Activity #4. (There is no "penalty" in the specified utility function for wait time incurred at the location of the previous activity; only for wait time incurred at the location of the current activity, prior to its commencement.)

The final household considered in this initial test (Household #3) is distinct from the previous two in that the observed pattern, as reported in the survey, contained reporting errors that rendered the reported pattern infeasible (Figure 9). Specifically, the pattern for this household includes carpooling to a joint out-of-home activity (labeled Activity #1 and #2 for Person #1 and #2, respectively) that has been reported incorrectly in the survey. Although using the same vehicle (Vehicle #1) and arriving at the same destination, Person #2 reports an activity start time and return home time that differ from that reported by Person #1 by 0.12 hours. After five generations, the genetic algorithm failed to identify a set of coefficients that would result in the HAPP model capturing the observed (albeit misreported) carpooling behavior (Figure 10). Nonetheless, the model was able to replicate the reported times for Activities #1 and #2 exactly; the misclassification of the person and vehicle assigned to these two activities is a semantic difference since the two activities are, in reality, the same. The only other difference between

the "best" prediction and the observed is a 0.29 hour difference in the schedule times associated with Activity #3, and the vehicle used to complete that activity.

4.2 Results of Estimation of Utility Coefficients on a Sample of 65 Households

In this section, we give summary results of application of the estimation procedure to the sample of 65 randomly selected two-member households under the usual assumption of homogeneity, i.e., $\hat{\mathbf{B}}_i = \hat{\mathbf{B}}, \forall i$. The results are based on 25 generations of populations of size 20, using the genetic algorithm described in the previous sections, and based on GA parameters shown in Table 2. Because there are no well-defined theoretical distributional properties related to the error of the estimation, we rely on several "goodness-of-fit" indicators to capture the efficacy of the estimation: 1) the mean edit distance $\boldsymbol{\mu}(\hat{\mathbf{B}})$ $[0,1]$ for the sample, 2) the percentage of activity sequence orderings predicted correctly, and 3) the total difference between the starting times for all activities in a household member's observed and predicted activity patterns. Shown in Figure 11 are the distributions of these measures across the entire population covered by the random initial population and subsequent 25 generations (i.e., based on 26x20=520 sets of parameters, $\hat{\mathbf{B}}$).

These results tend to indicate that, for the sample considered, constraints play a dominant role in the execution of the observed activity patterns—relatively high agreement between the observed and predicted patterns is achieved over a broad range of ascribed utility weights represented by the components of $\hat{\mathbf{B}}$. This observation is further evident in the trace of the convergence of utility coefficients shown in Figure 12, which shows the maximum and minimum values for each coefficient estimate at each generation, together with its mean across the population of twenty. (Note: In these results, without loss in generality, the coefficient values have been scaled [0,1].) We note in particular that the relatively sharp convergence shown in the previous examples of single households in which the utility weights are free to vary across individual households, i.e., expressed by $\hat{\mathbf{B}}_i$, in most cases is not apparent under the homogeneity restriction. The implication here is that an assumption more akin to the heterogeneity assumption in the coefficients that underpins mixed logit may be a more attractive option.

The overall best fitness score among the candidate $\hat{\mathbf{B}}$'s at any particular generation, followed a pattern similar to that displayed for the individual households of the previous section—most of whatever improvement achieved was captured within the first five generations. This pattern is exemplified by the results for "edit distance" shown in Figure 13; here, improvement in the minimum value was obtained during the first five generations, followed by no further improvement until the 24th generation.

After these 25 generations, the "best" estimates for the components of $\hat{\mathbf{B}}$ were determined as:

$$\hat{\mathbf{B}}' = \begin{bmatrix} 0.210 & 0.221 & 0.770 & 0.619 & 0.875 & 0.393 & 0.839 & 0.306 \end{bmatrix} \quad (8)$$

Applying these common utility weights to the utility functions of the 65 households (130 individuals) in the sample produced the following distribution of the three "goodness-of-fit" measures across the sample:

We judge these errors to be surprisingly good, considering the complexity of the problem being addressed; i.e., the prediction of the combined 1) assignment of activities to members within the household, 2) the start times of those activities, 3) the order of performance of the activities, and the travel linking those activities (e.g., number of tours, particular sojourns comprising tours, etc.). The mean edit distance for the sample was about 0.3; the vast majority of activity sequences were predicted correctly; well over 50% of the sample had little or no error in the predicted start times of their various activities.

5. Concluding Remarks

Most probably because of the inherent overwhelming complexities of treating the "whole" of travel, the so-called "activity-based" approach has not been embraced by the mainstream of transportation researchers as offering a viable practical paradigm for travel demand modeling. Despite efforts over the past two decades, activity-based modeling advancement has been relegated largely to either descriptive or proscriptive study, falling well short of being able to be used in the context of actually forecasting changes in travel behavior.

Although we believe that the procedures reported here represent an important step in moving activity-based travel approaches beyond descriptive analysis toward a foundation for practical demand modeling, there nonetheless remain significant challenges. In the current implementation we have addressed the inference of utility coefficients as a "common" set of coefficients that can be applied to an entire sample of activity patterns with the result that the predicted activity patterns are closest to the actual activity patterns, as determined by a set of three distance measures. This was accomplished by defining a fitness score as some weighted average of "edit distance" and "temporal alignment" across all sample points. Because the total

number of possible multidimensional operation sets is equal to the product of the numbers of all optimal paths in the Uni-dimensional Sequence Alignment Method for all attributes, the Multidimensional Sequence Alignment Method may be very time-consuming when the number of activities considered is larger than those considered in the test cases reported here; assuming that the operation cost for one insertion and one deletion is the same as one substitution in the Uni-dimensional Sequence Alignment Method, the number of all optimal paths for a pair of sequences with length m will have the complexity of $o(m^m)$. Additionally, the Time Sequence Comparison technique that has been employed, which simply offsets activities along the time dimension to maximize the overlapped time or number of overlapped activities, is quite simplistic. For purposes of illustration, we have used a fitness score that is simply the difference between the "sequence distance" (i.e., the number of steps to equate two sequences) and the degree of overlap along the temporal axis. Though the fitness score is a scaled value, it has no practical meaning. Moreover, there are no well-defined statistical properties for any of the measures that would allow statistical confidence estimates for the results.

Finally, despite its complete description (and analytical management) of the "hard" physical constraints imposed on a household's collection of activity/travel decisions, the HAPP model has no explicit representation of such "soft" constraints as each member's role in the household, or their particular predilections and idiosyncrasies, that may be imposed on top of these physical constraints. One promising avenue for incorporating these considerations may be in adapting some of the procedures identified with the relatively new field of uncertain programming (Liu, 1999).

Acknowledgements

This research was supported, in part, by a grant from the USDOT University Transportation Center of the University of California, with matching funds provided by the Division of Research and Innovation of the California Department of Transportation. Their financial support is gratefully acknowledged.

References

Arentze, T.A., F. Hofman, C.-H. Joh, and H. Timmermans (2002). Activity pattern similarity: a multidimensional sequence alignment method. *Transportation Research Part B: Methodological*, 36, pp. 385-403.

Holland, J.H. (1992a). *Adaptation in Natural and Artificial Systems*, 2nd Ed., MIT Press, Cambridge, MA.

Holland, J.H. (1992b). Genetic algorithms. *Scientific American*, July, pp. 66-72.

Liu, B. (1999). *Uncertain Programming*. John Wiley & Sons, Inc. New York.

Michalewicz, Z. (1992). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, New York.

Recker, W.W. (1995). The household activity pattern problem: general formulation and solution. *Transportation Research*, 29B, 1, pp. 61-77.

Recker, W.W. and A. Parimi (1999). Development of a microscopic activity-based framework for analyzing the potential impact of TCMs on vehicle emissions. *Transportation Research, Part D: Transport and Environment*, 4, 357-378.

Recker, W.W., C. Chen and M.G. McNally (2001). Measuring the impact of efficient household travel decisions on potential travel time savings and accessibility gains. *Transportation Research, Part A: Policy and Practice*, 35, 339-369.

Recker, W.W. (2001). A bridge between travel demand modeling and activity-based travel analysis. *Transportation Research , Part B: Methodological*, 35, 481-506.

Rudolph, G. (1994). Convergence properties of canonical genetic algorithms. *IEEE Transactions on Neural Networks*, 5:1, 96-101.

List of Tables

Table 1. Qualitative Interpretation of Components of Disutility Function

Variable	Interpretation
x_1	Total cost of travel to the household
x_2	Total travel time expended by members of the household
x_3	The difference between the time that a household member arrives home from a particular activity and the time he/she started the activity; a measure of delay time in returning home from an activity because he/she went to another activity prior to returning home (i.e., trip chained).
x_4	A measure of how close a household member comes to being unable to complete an activity before its window closes, i.e., the risk of not being able to perform an activity if some stochastic delay (e.g., congestion) had arisen.
x_5	Same as x_4 above, but with respect to returning home from an activity no later than that needed/desired.
x_6	Total extent of the travel day; i.e., how spread out the travel is.
x_7	The "cost" associated with using more vehicles than absolutely necessary; e.g., using two different vehicles, rather than one.
x_8	Total waiting time (at destination) before performing activities.

Table 2. Genetic Algorithm Parameters

Genetic Parameter	Value
Chromosome Length	80
Alphabet Size	2
Population Size	20
Stop Criterion	25 Generations
Crossover Probability	1.0
Mutation Probability	0.05
Selection Method	Roulette Wheel
Crossover Option	Two Point Crossover
Constants *a*, *b, c*	0.5, 0.5, 0.5

List of Figures

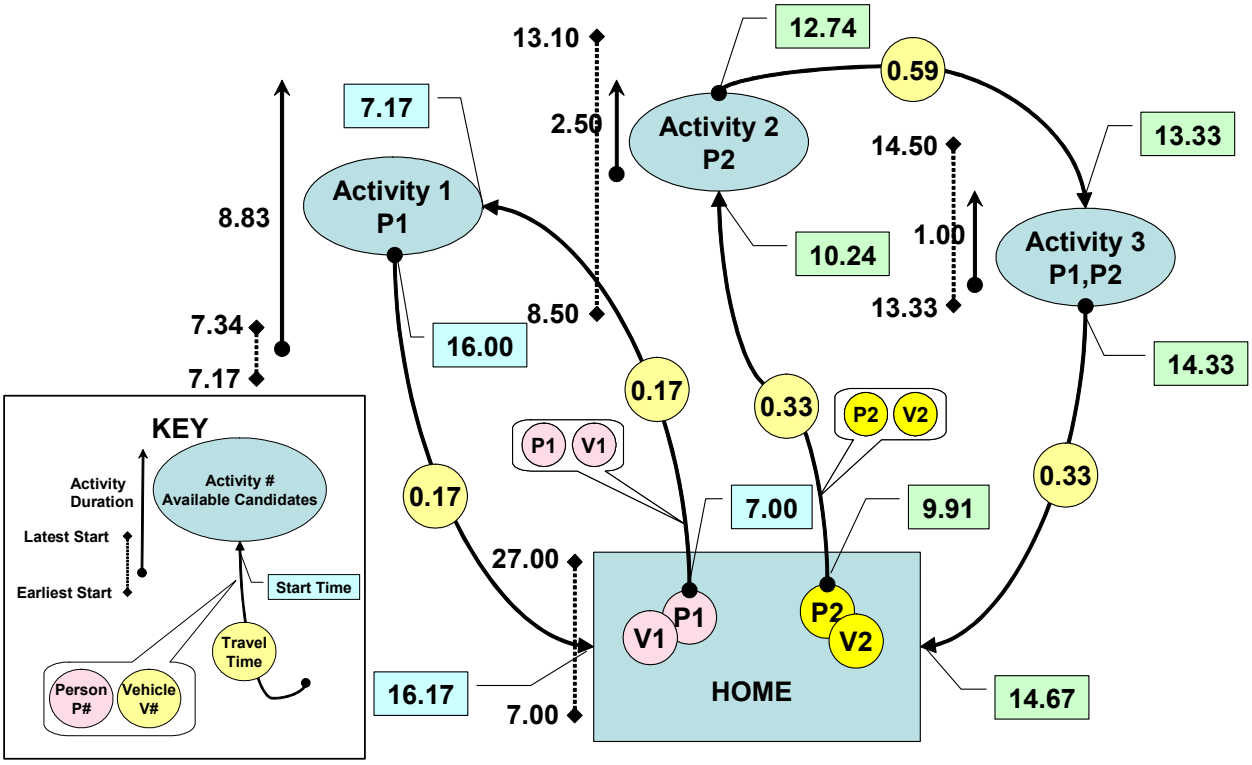

Figure 1. Observed Activity Pattern for Household #1

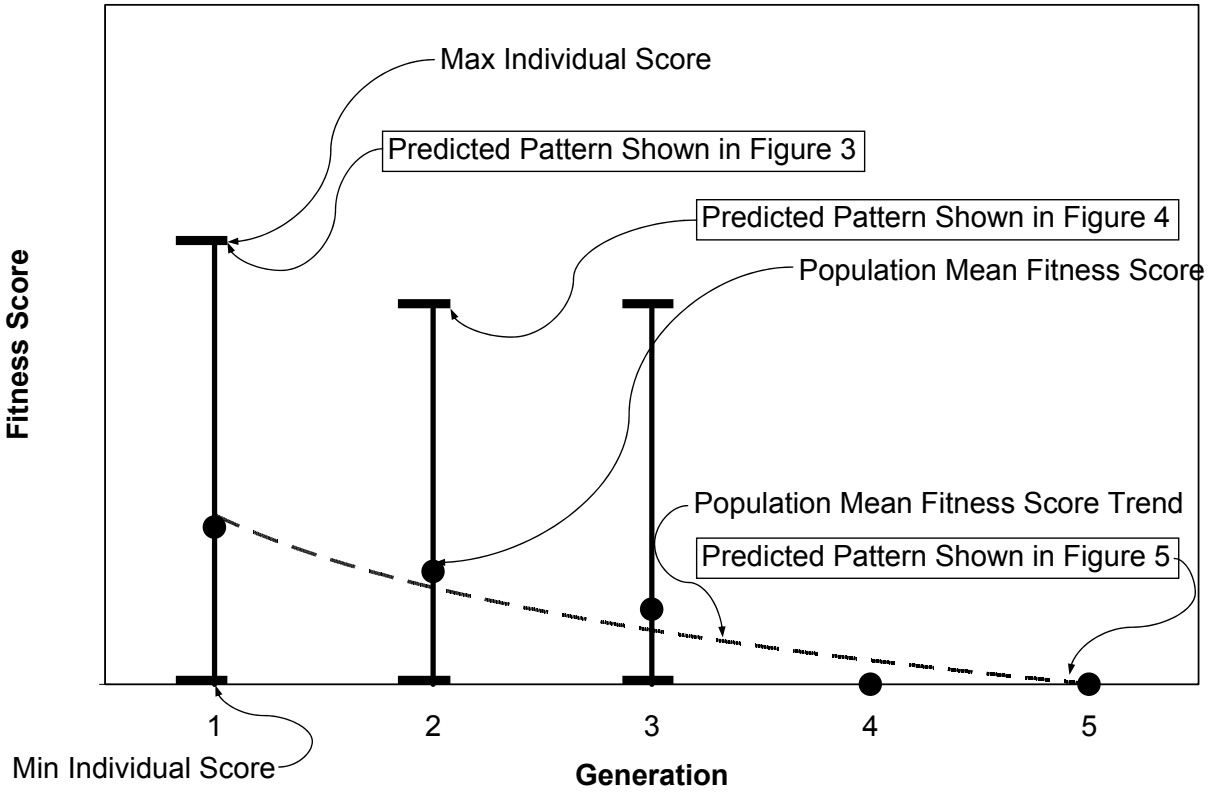

Figure 2. Convergence of Genetic Algorithm for Household Observation #1

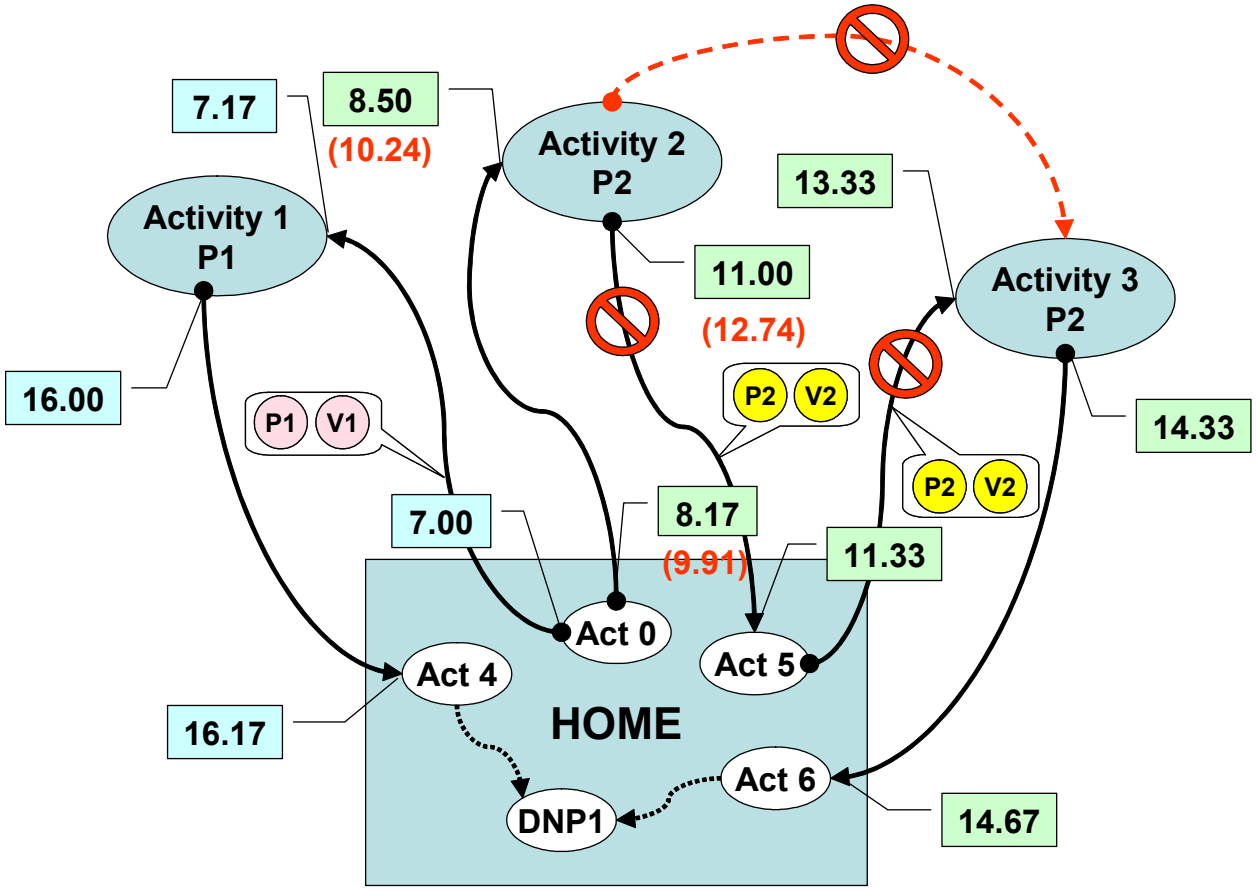

Figure 3. HAPP Predicted Pattern for Household #1, ”Worst Fitness” Value: Generation 1

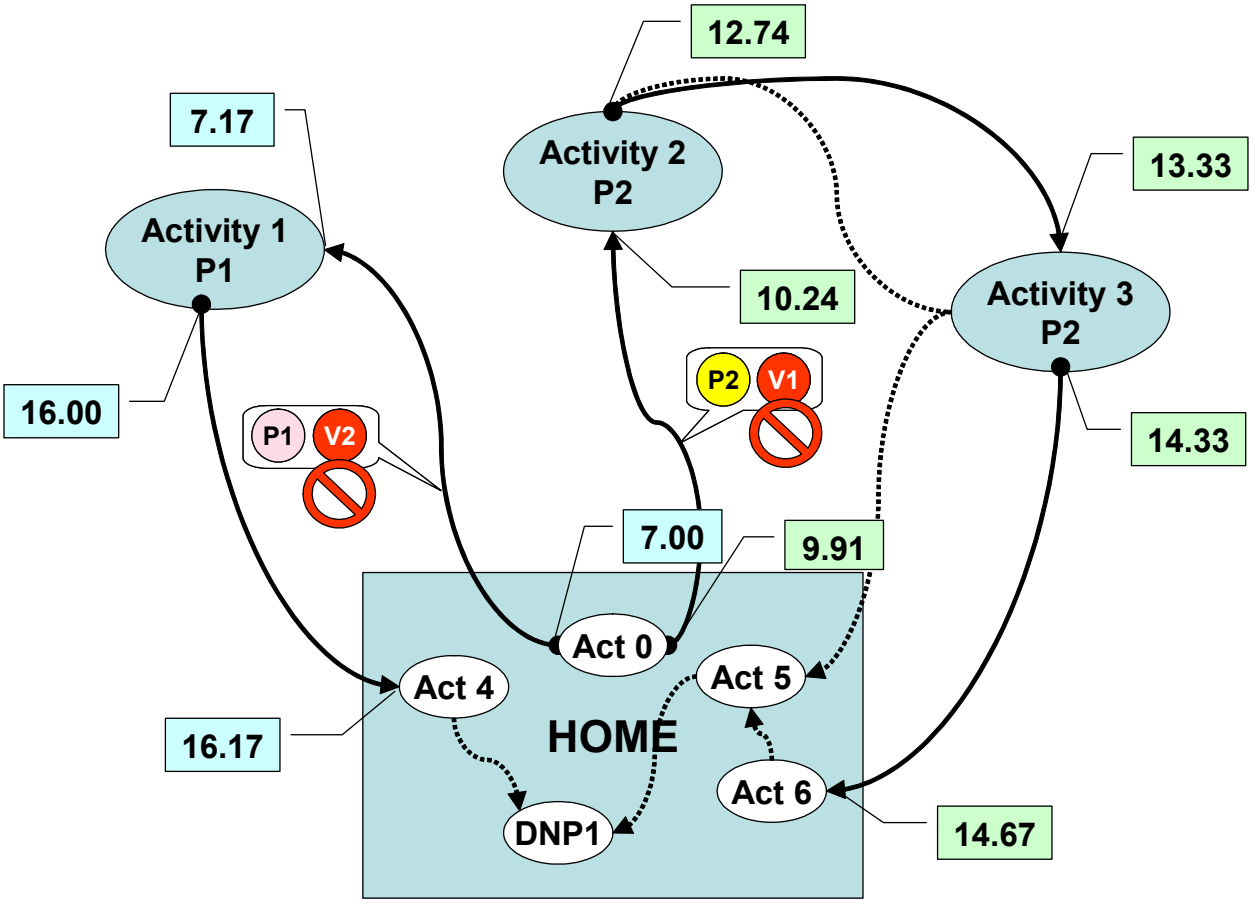

Figure 4. HAPP Predicted Pattern for Household #1, ”Worst Fitness” Value: Generation 2

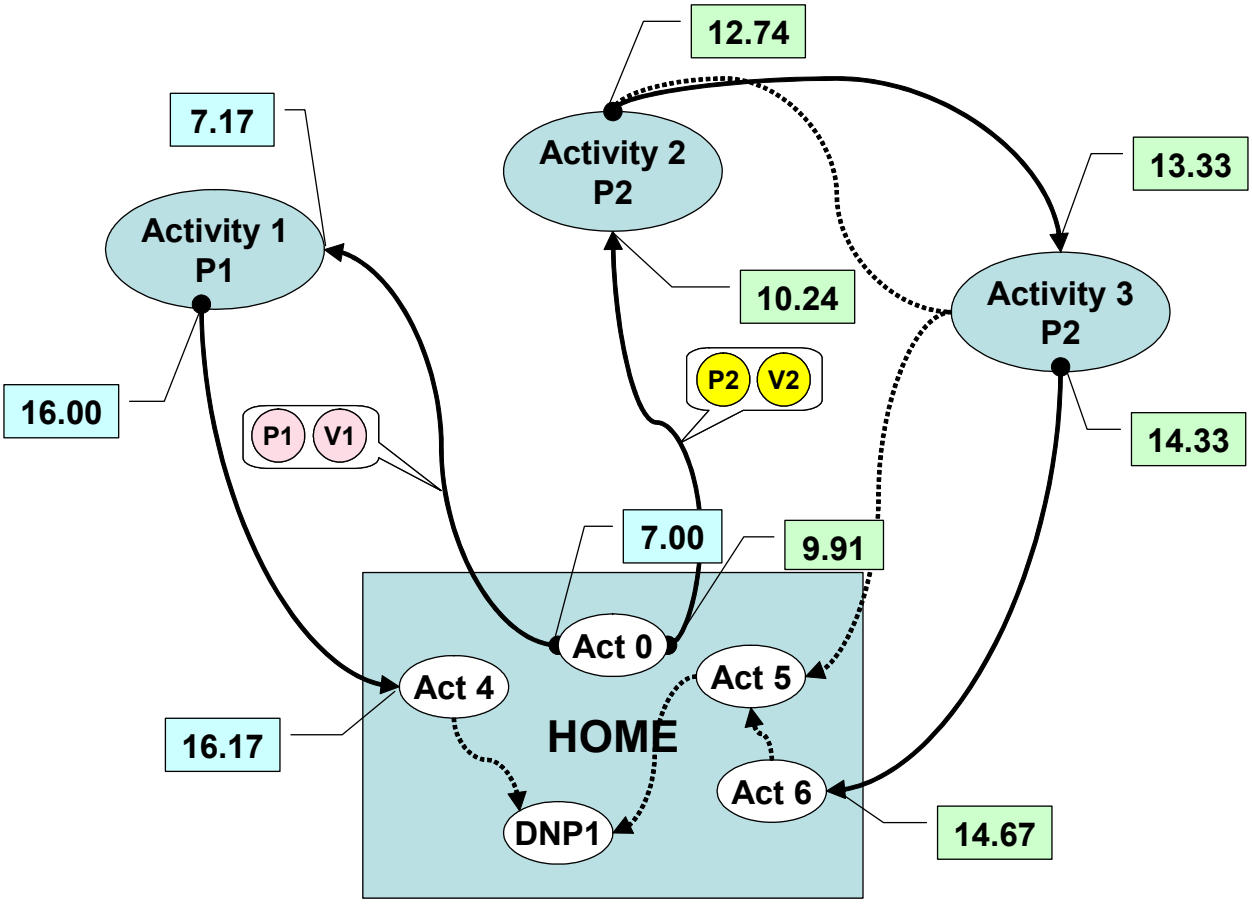

Figure 5. HAPP Predicted Pattern for Household #1, Generations 4 and 5

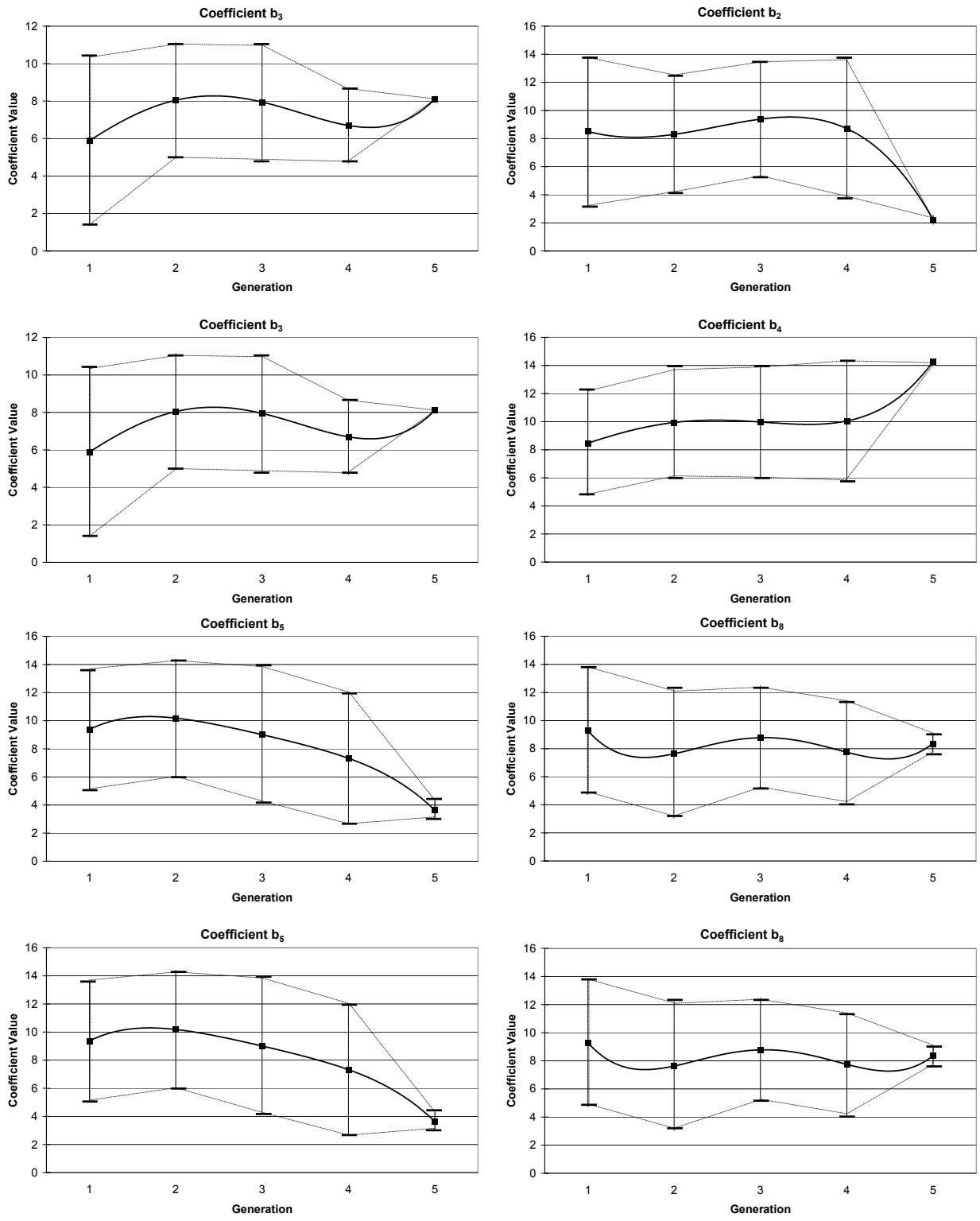

Figure 6. Convergence of Utility Coefficients for Household #1

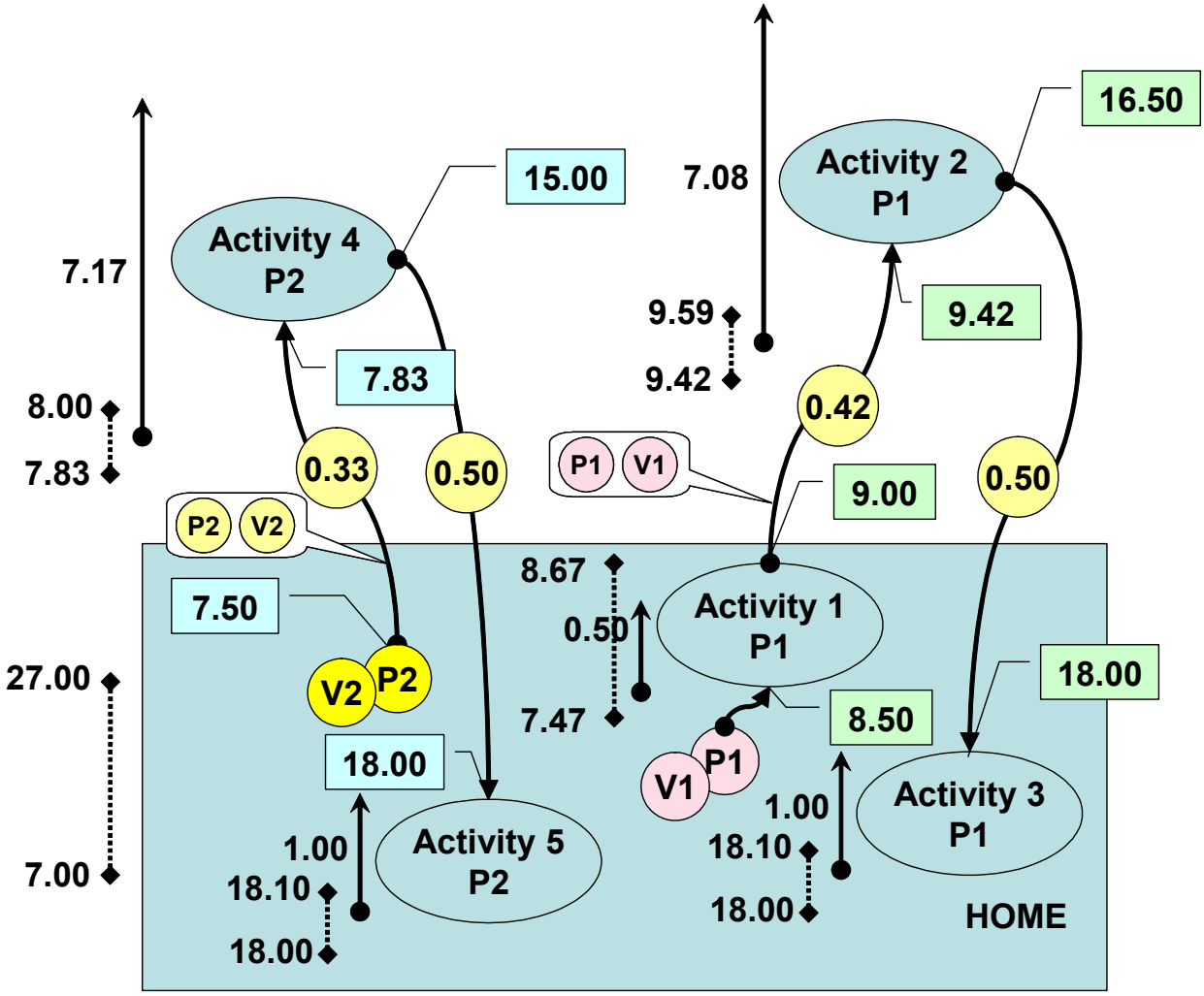

Figure 7. Observed Activity Pattern for Household #2

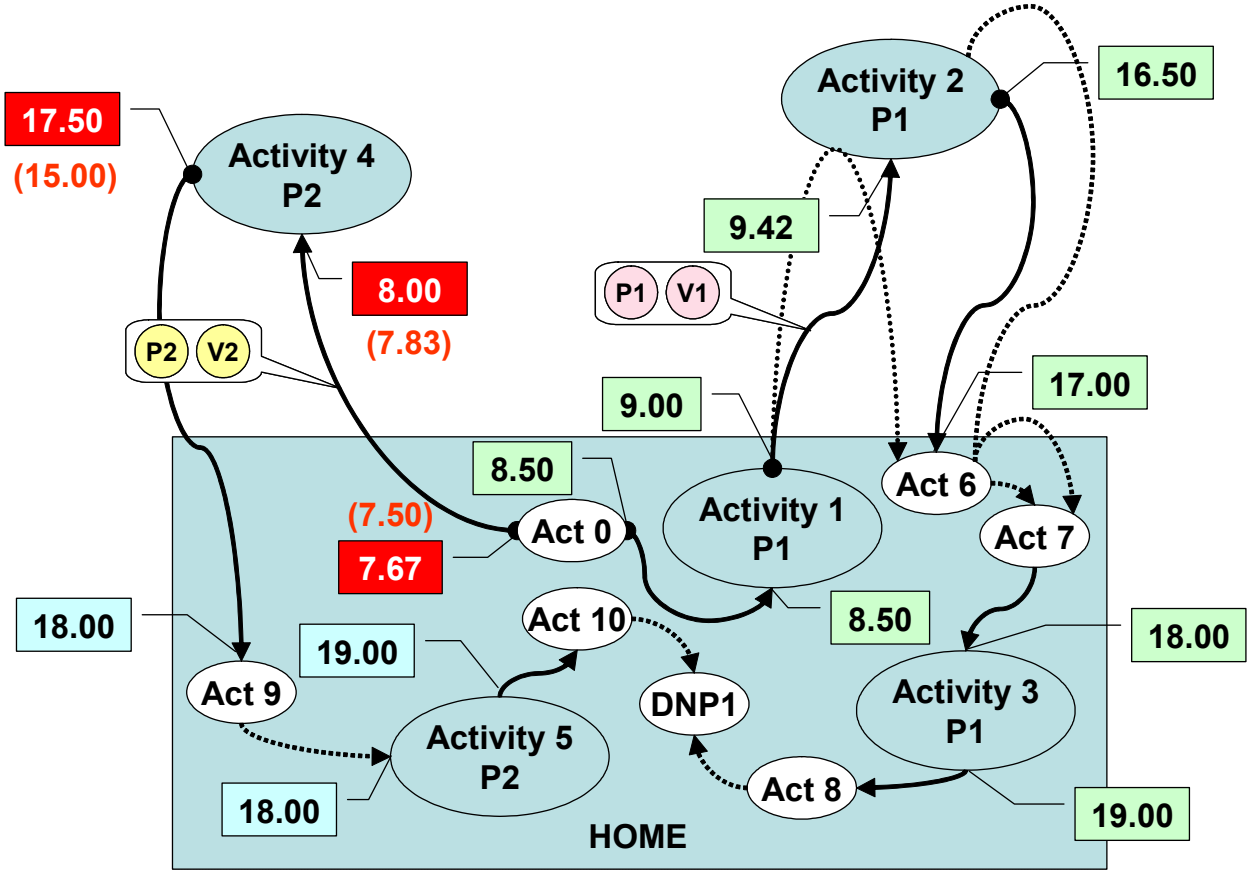

Figure 8. Best HAPP Predicted Pattern for Household #2, after Generation 5

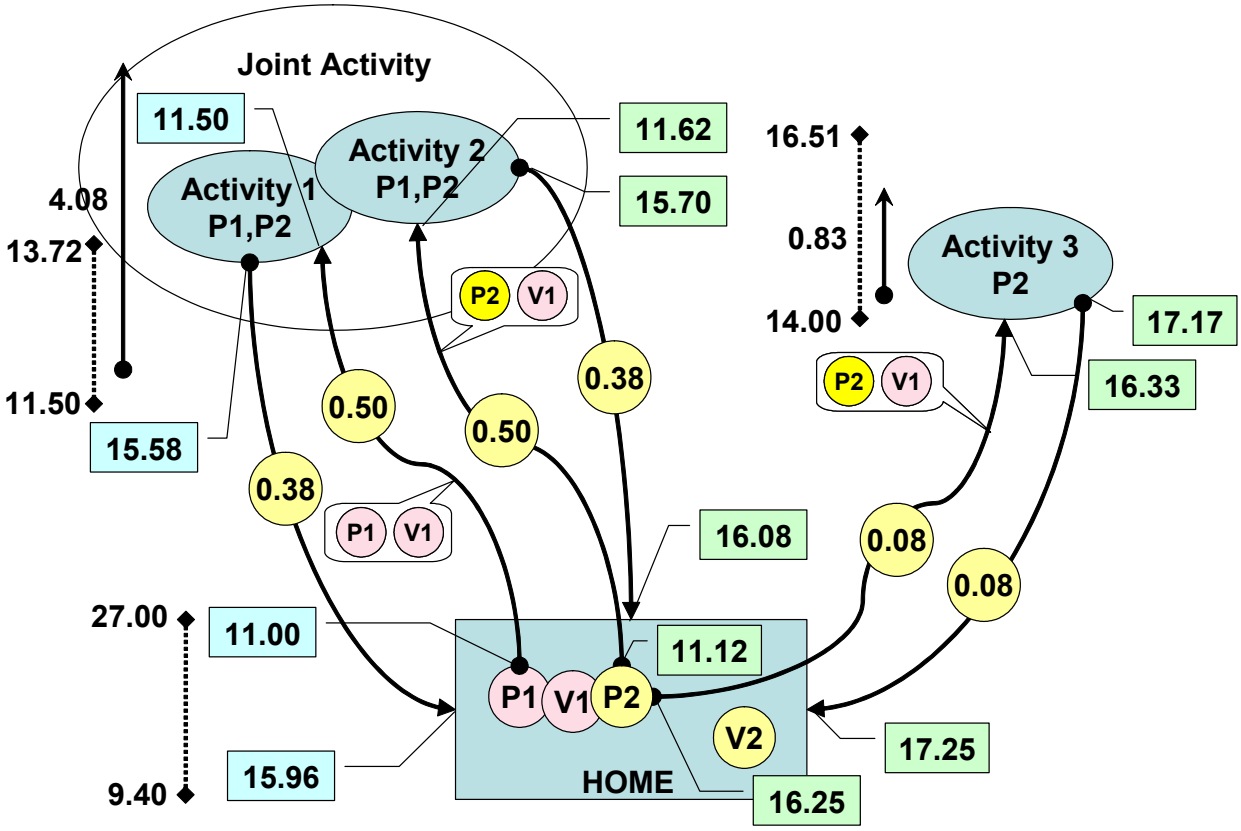

Figure 9. Observed Activity Pattern for Household #3

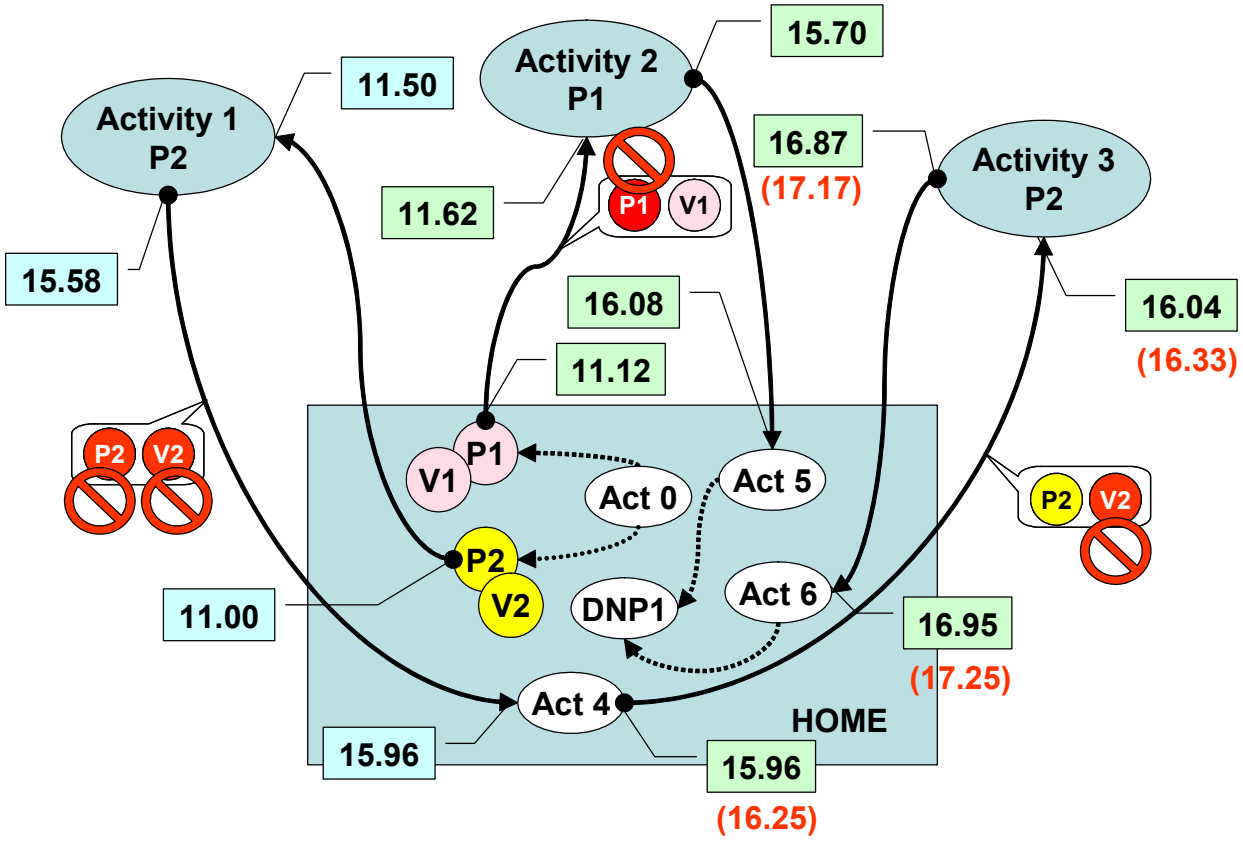

Figure 10. Best HAPP Predicted Pattern for Household #3, after Generation 5

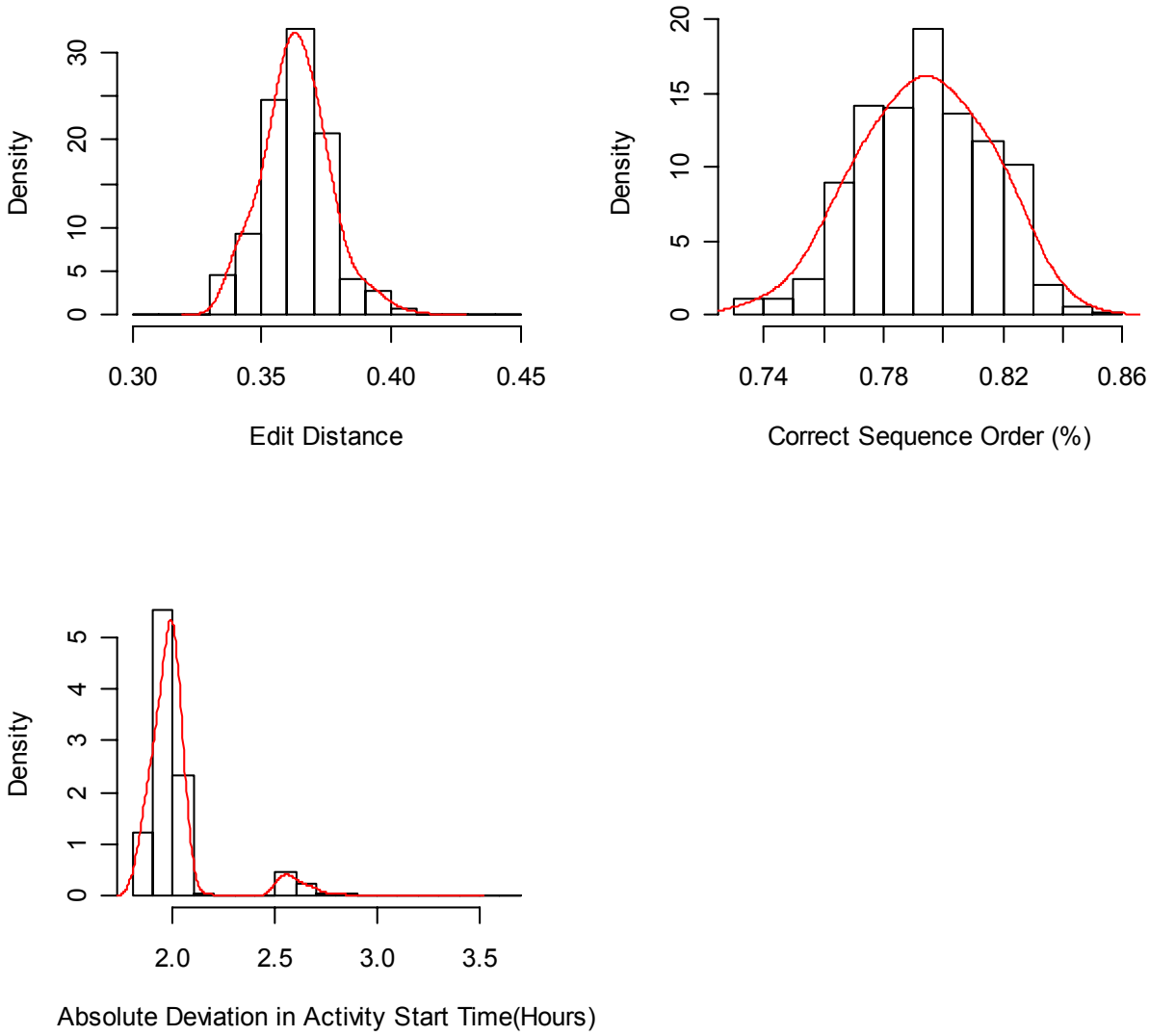

Figure 11. Histogram of Statistical Results Obtained from the GA

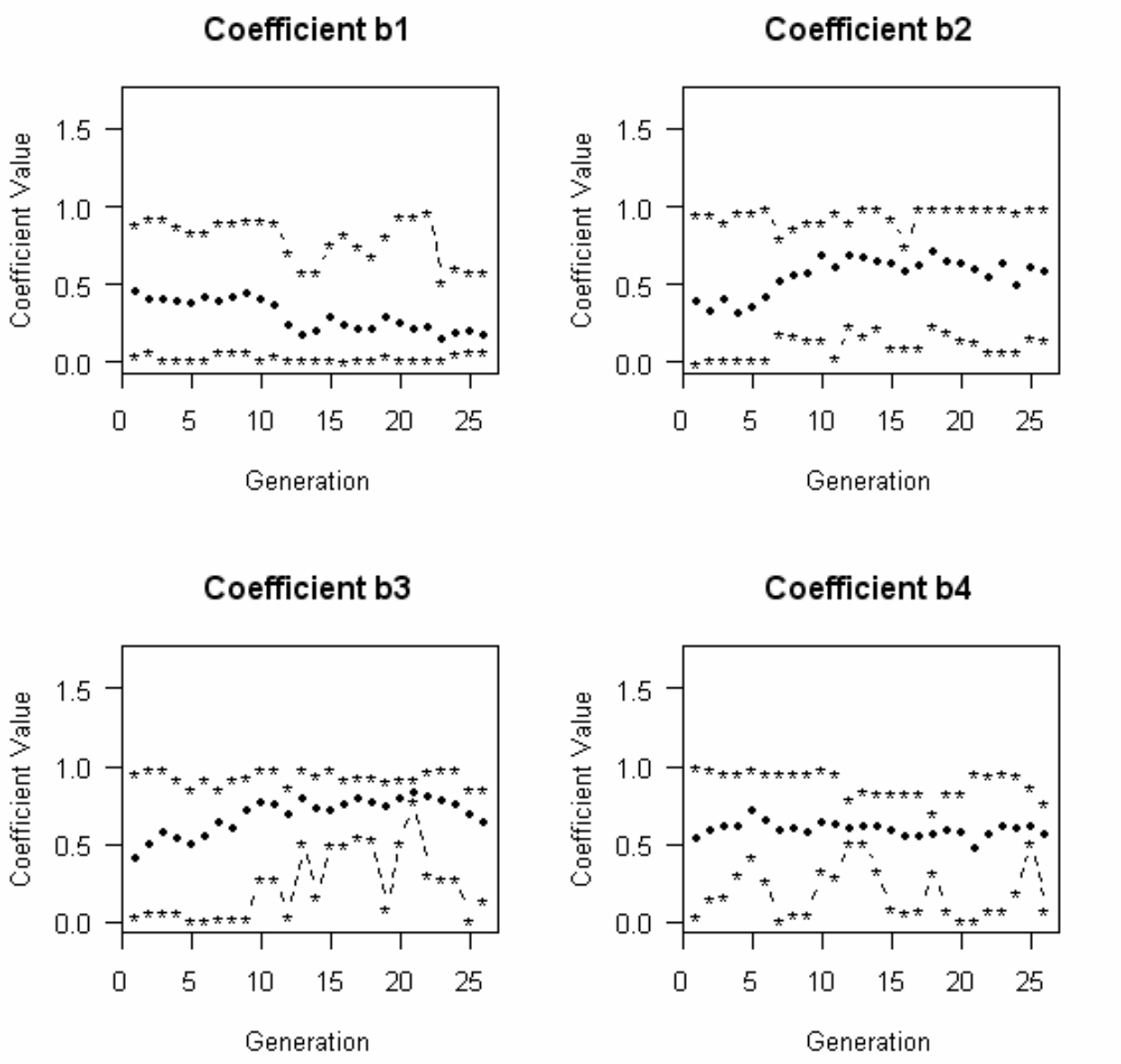

Figure 12. Convergence Properties of Utility Coefficients via the GA

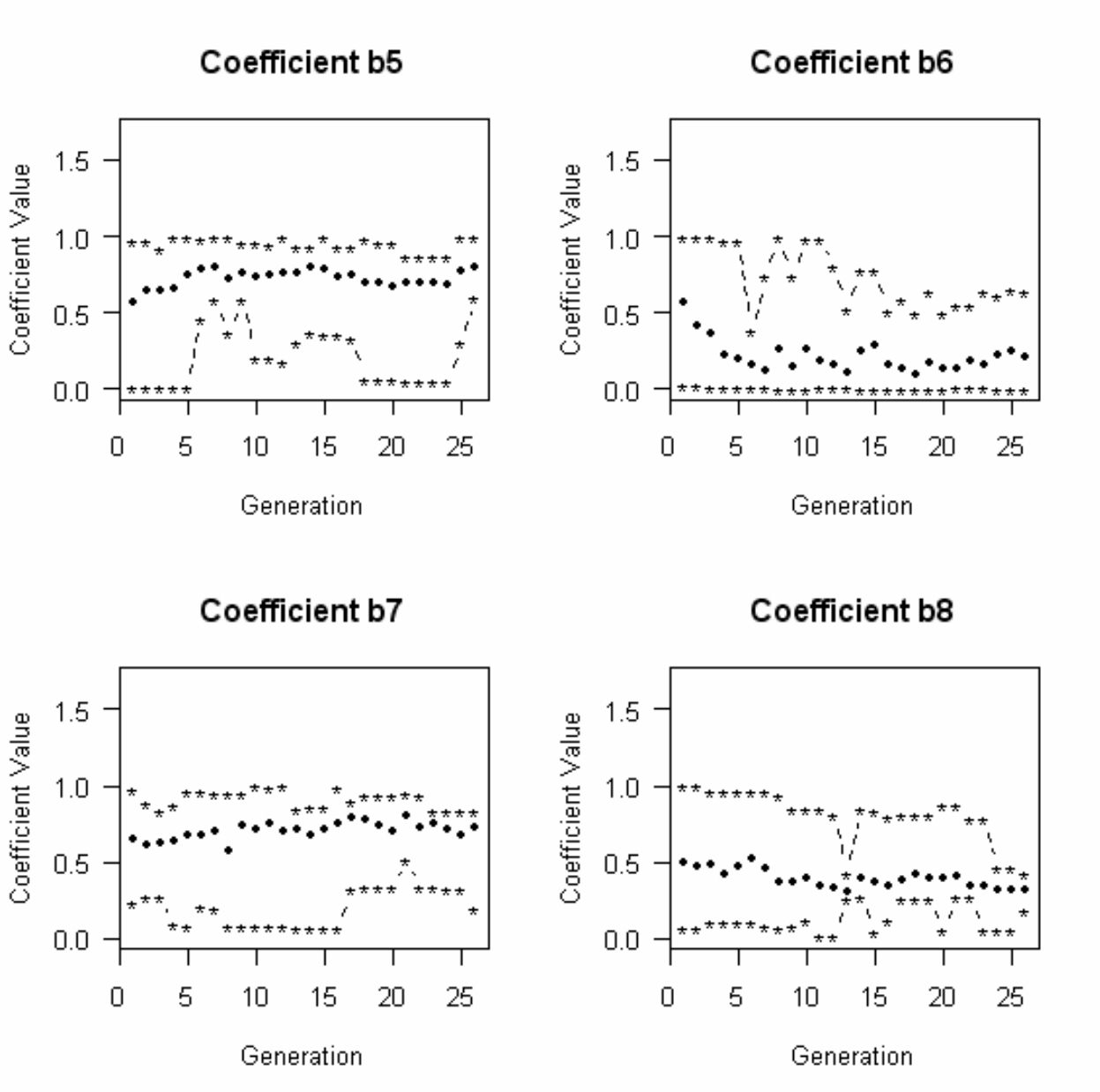

Figure 12 (continued). Convergence Properties of Utility Coefficients via the GA

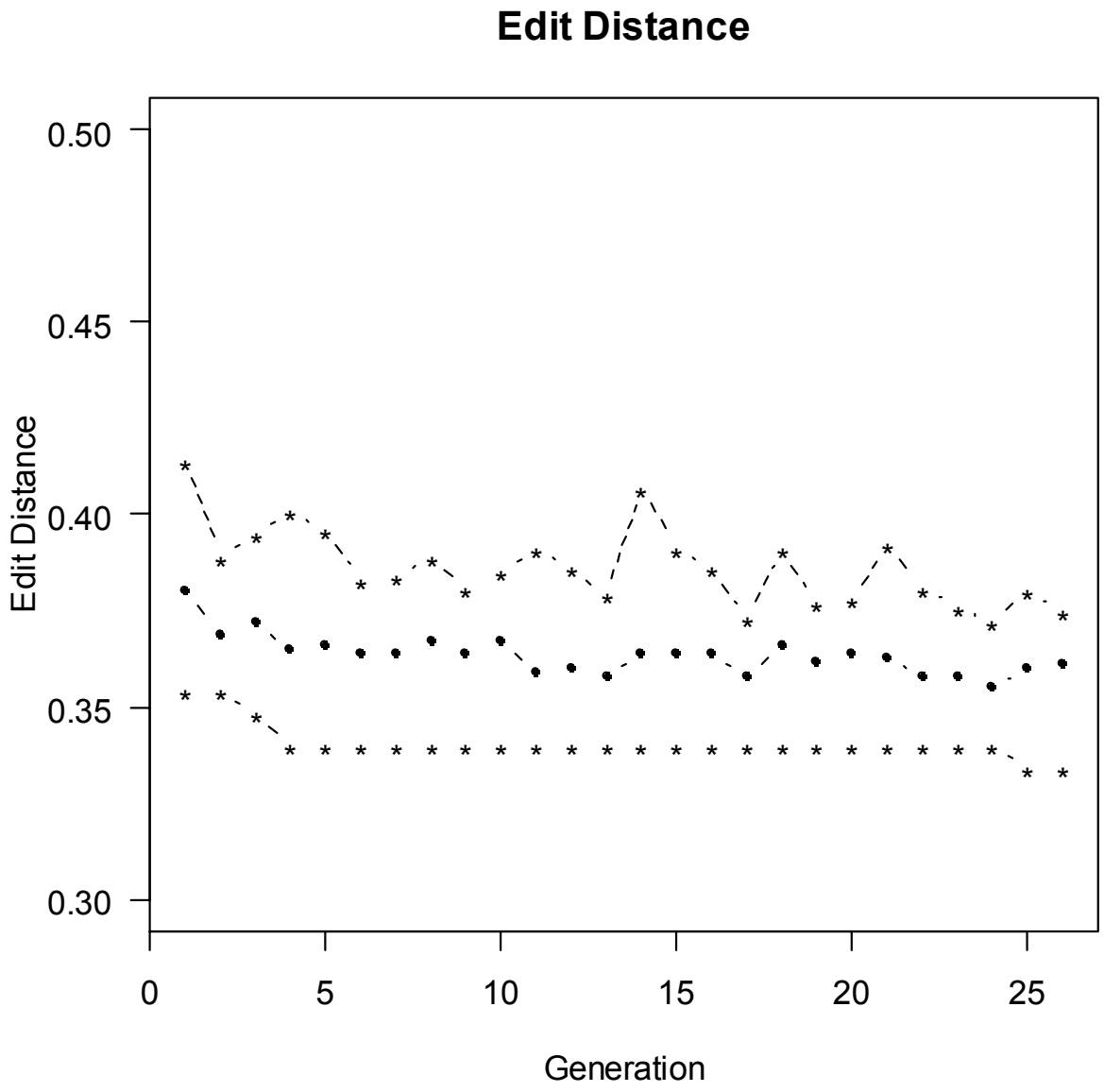

Figure 13. Convergence of Fitness Score for "Edit Distance" to its Minimum Value

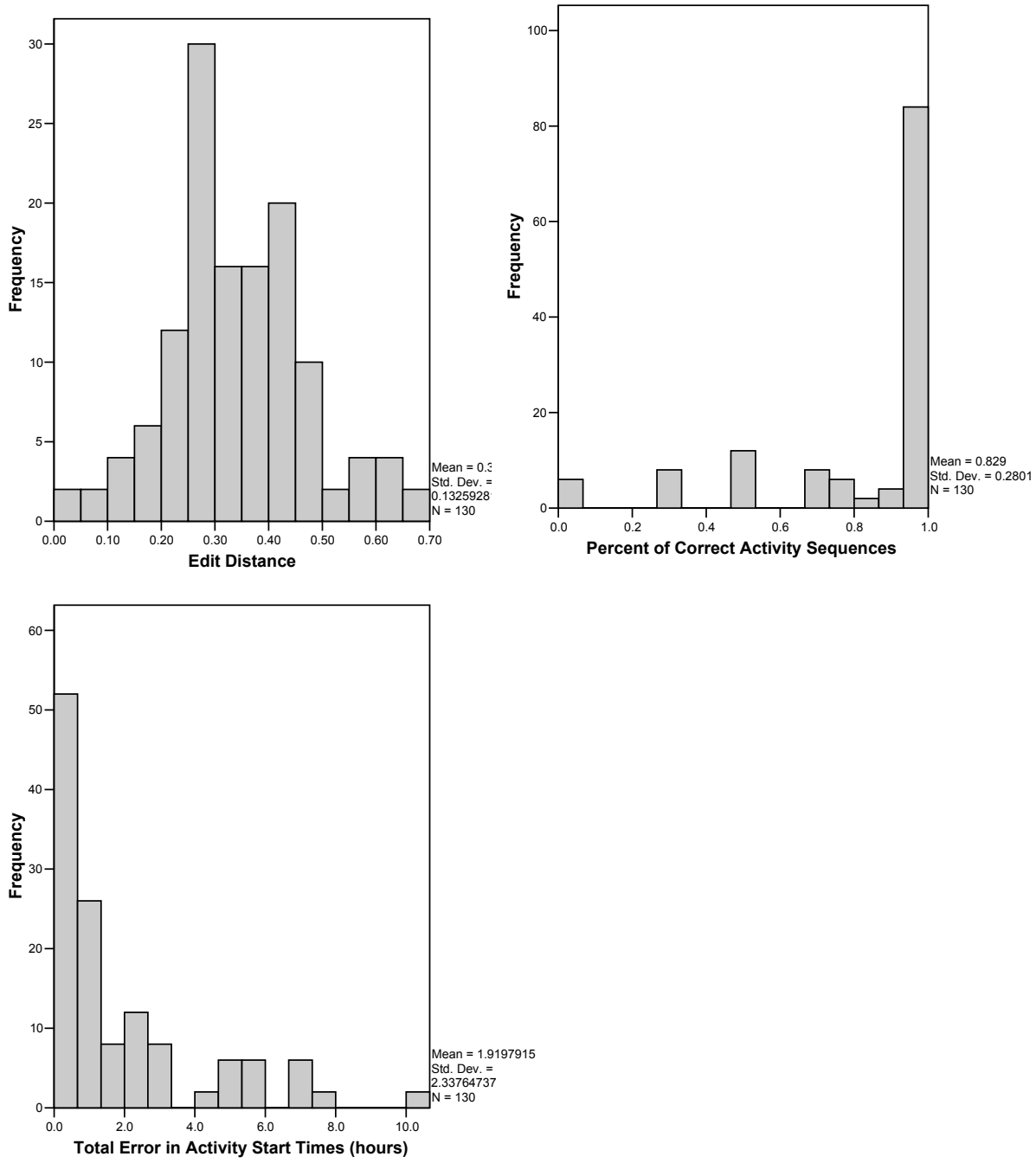

Figure 14. Distribution of Errors in Estimation across Sample