A Reinforcement Learning Model with Endogenously Determined Learning-Efficiency Parameters: Applications to Route Choice Behavior in Congested Networks

Toshihiko Miyagi
Department of Regional Studies, Gifu University
1-1 Yanagido, Gifu 501-1193 JAPAN
miyagi@cc.gifu-u.ac.jp

***Abstract*-This paper proposes a new algorithm for finding disaggregate user equilibria on a congested network where a driver is assumed to be an agent who performs reinforcement learning to get maximal payoff (minimum loss) under limited route information. A reinforcement learning with endogenously determined leaning- efficiency parameters is presented and its relation to the user equilibrium is also explored.**

Key words: *reinforcement learning, user equilibrium, route choice behavior, ITS, bounded rationality*

I. INTRODUCTION

To design and provide an efficient Intelligent Transportation Systems (ITS) or optimal route guidance strategies in transport networks requires reliable traffic forecast. For this problem, one of the fundamental problems not yet solved is driver's response to travel information. Without ITS technology, a driver would try to select his best route under limited travel information. Such human decision-making on route choice can be seen as a sort of leaning behavior under iterated tasks.

Until then, route choice behavior in congested networks has been modeled as mathematical optimization programs [1,2] or variational inequality problems [3,4]. Those models have been built upon the assumptions of rational behavior of drivers, of aggregated decision-makers and of mathematically tractable performance functions of roads. In those approaches, it is assumed that individual driver has complete information on all routes and can make an optimal decision on his or her route choice.

Recent approaches [5-7] are directing toward to modeling drivers' learning behavior based on bounded rationality under limited travel information. To examine the route choice behavior, Selton, et al. [5] have carried out a psychologically designed decision-experiment, where all players had to repeatedly decide between two alternative roads with different road capacities and should try to maximize their resultant payoffs. Two treatments had been repeatedly tested for N test persons: while in treatment 1 all player are known only their own payoffs (previously experienced), treatment 2 is designed for players to be informed about payoffs of two alternatives after each trial. They found that the experiments showed test persons' behavior well suited to the user equilibrium and that a simulation model based on reinforcement learning could trace the behavior of tested persons. The reinforcement leaning model they adopted is a similar one with those that have been used in cognition science [8,9], machine learning [10] and an economic agent model [11].

Helbing et al. [6] has added further experiment to explore the volatile decision dynamics observed in the experiment by Selton, et al.[5]. They found out through the experiment that decision guidance by means of user-specific recommendations can increase the adaptation of players and reduce the deviation from the time-dependent user equilibrium, thereby enhancing the average and individual payoffs.

Independently of those works, Miyagi [7] developed the reinforcement model for studying interactions between traveler information and individual driver's route choice behavior. In his reinforcement model, it is assumed that the reference cost, which is a dynamic route information indicating the level of network congestion, is provided by a network administrator and that drivers' decision-makings are mutually influenced by congestion effect of overlapping routes. He reported that in calculation of the disaggregate user equilibrium the reinforcement model has the full adaptability for ill-defined performance functions such as asymmetric link cost and discontinuous link cost functions.

The object of this paper is two-fold. The first, we propose a new reinforcement learning model in which individual-specific stepsize parameter that combines ex-ante propensities to choice with obtained payoffs is endogenously determined, thereby describing autonomous route choice behavior of individual driver. The second is to examine the behavioral assumptions under which the volatile user-equilibrium or more stable user-equilibrium occurs.

II. MODEL FORMULATION

A. *Notation on Network and Value function*

Consider a single origin-destination (O-D) pair connected by paths (or routes) denoted by positive integers, $p \in \mathbf{P}$, in which $\mathbf{P}$ denotes a set of paths, $\{1, \cdots, p \cdots, M\}$. Path flows are denoted by $h = (h_1 \cdots, h_p \cdots h_M)$. Each driver is identified by

$i \in \mathbf{I} = \{1, \cdots, i \cdots, N\}$. Thus, N represents the number of O-D trips as well. Let $\mathbf{C}_i$ and $\overline{\mathbf{C}}_i$ denote a set of paths used and a set of unused paths by driver i. Cleary, it holds that $\mathbf{P} = \mathbf{C}_i \cup \overline{\mathbf{C}}_i \ \ for\ \ all\ \ i \in \mathbf{I}$. Let $x_{ip}\ (i \in \mathbf{I}, p \in \mathbf{P})$ denotes a choice probability of path p of driver i. Then, we have the following relations:

$$\sum_{p\in\mathbf{P}} x_{ip} = 1,\ \ for\ \ all\ \ i \in \mathbf{I} \tag{1a}$$

and

$$\sum_{i\in\mathbf{I}} x_{ip} = h_p,\ \ for\ all\ \ p \in \mathbf{P}. \tag{1b}$$

From (1a) we observe that each driver's choice probabilities are defined on a simplex S^{M-1}:

$$\mathrm{X} \in \mathbb{S}^{M-1} = \{\mathrm{X} \in \Re^{M}_{+} \mid x_{ip} \geq 0, \sum_{p\in\mathbf{P}} x_{ip} = 1\}$$

The number of O-D trips and link flows are related with path flows by

$$\sum_{p\in\mathbf{P}} h_p = N \tag{2}$$

and

$$\sum_{p\in\mathbf{P}} \delta_{ap} h_p = f_a,\ a \in \mathbf{A}, \tag{3}$$

where $\mathbf{A}$ denotes a set of links, f_a and δ_{ap} are respectively flow on link $a \in \mathbf{A}$ and an element of link-path incidence matrix. Denoting link travel time on $a \in \mathbf{A}$, be $T_a(\mathbf{f})$, we have path travel time of $p \in \mathbf{P}$ as:

$$u_p(\mathbf{h}) = \sum_{a\in A} \delta_{ap} T_a(\mathbf{f}(\mathbf{h})) \tag{4}$$

We assume that link travel time functions is nondecreasing of $\mathbf{f}$. Each driver appreciates path travel time in a different way because of the variation of value of time. To reflect this, we put the perceived cost of path p for driver i to

$$u_{ip}(h) = w_i u_p(h),$$

where w_i represents the value of time of driver i.

We introduce the reference cost which is a sort of travel information provided by a network administrator and, is referred by all drivers like a market price. The reference cost is not information for the sake of route guidance of drivers, but a sign indicating the congestion level of the current network state. Drivers know their present states by getting this information. We assume a maximum path cost as the reference cost:

$$\lambda_i(h) = \max_{p\in\mathbf{P}}\ u_{ip}(h) \tag{5}$$

The reference cost is also considered as the perceived cost for each driver because of the difference of value of time. A driver knows his present state by comparing with the reference cost:

$$r_{ip}(\mathbf{X}) = \lambda_i(\mathbf{X}) - u_{ip}(\mathbf{X}). \tag{6}$$

We call (6) a payoff function of driver i. If a driver chooses a minimum cost route, then he receives a payoff (or a reward) as long as the route he selected is not the maximum-cost route. If no driver can get any payoff whenever changing his or her route, it implies that the system is in equilibrium in the sense that no one can get any benefit by unilaterally changing his or her route.

While a driver can know his travel cost at the current day through his experience, he cannot recognize the travel costs of routes that he didn't select. Under the assumption of the conplete information, every drivers are informed of payoffs on all routes, however, some drivers may not have communication tools to get the travel information on routes or they don't take care about the travel information of rarely used routes. We can consider of a variety of contexts where the assumption of the incomplete information has validity.

Let me introduce a value function of driver i directly defined by the expected payoff function:

$$v_i(\mathbf{X}) = \sum_{p\in\mathbf{C}_i} r_{ip}(\mathbf{X}) x_{ip} \tag{7}$$

Note that the expectation is measured over only used paths. Then, to find the user equilibrium it may be appropriate to consider the following minimization problem of the value function for each driver:

$$\min_{\mathbf{X}\in S^{M-1}}\ v_i(\mathbf{X})\ \ for\ \ all\ \ i \in \mathbf{I} \tag{8}$$

Unfortunately, however, even if we assume that each of link cost functions is convex, we encounter the difficulty that the value function is neither convex nor differentiable. Therefore, the ordinary optimization procedures can not be applied to this problem. The optimality condition for the problem with non-smooth functions like (8), which is described in terms of the generalized gradients, has been studied by Clarke[12] in the locally Lipscheitz case.

B. Reinforcement Learning Model with Endogenously Determined Learning-Efficiency Parameters

We are interesting in day-to-day variation of drivers' path choice behavior. Let t be time describing a certain day. Let Q^{t}_{ip} be an ex-ante propencity to choose path p of driver i at time t. According to the propensities, a choice is then selected from the choice set, $\mathbf{P}^{t}$, which contains all the available paths at time t. The probability of choosing a path $p \in \mathbf{P}^{t}$ by individual $i \in \mathbf{I}$ is given as:

$$x_{ip}^{t+1} = \frac{Q_{ip}^{t+1}}{\sum_{p \in \mathbf{P}} Q_{ip}^{t+1}} \tag{9}$$

Let Q_{ip}^{0} be a prior propencity to choice for path p of driver i. After experiencing a trip, he can observe the partially available state of its environment and may change his choice on the next day. It depends on whether the driver has obtained payoff or not. If there is a payoff, the driver reinforces the ex-ante propensity for path p using the following updating formula:

$$Q_{ip}^{t+1} = (1-\alpha_i^t)Q_{ip}^t + \alpha_i^t R_{ip}^t, \ where\ R_{ip}^t = r_{ip}^t x_{ip}^t \tag{10}$$

where $\alpha_i^t \in (0,1]$ is a stepsize parameter or a learning-efficiency parameter. Such feedback-typed tasks are continued until no payoff is generated. Equation (10) can be rewritten as:

$$Q^t = \prod_{k=1}^{t}(1-\alpha^{k-1})Q^0 + \sum_{k=1}^{t}\alpha^{k-1}R^{k-1}\prod_{j=i}^{t-1}(1-\alpha^j) \tag{11a}$$

In case of a constant α regardless of time sequence, it reduces to

$$Q^t = (1-\alpha)^t Q^0 + \sum_{k=1}^{t}\alpha(1-\alpha)^{t-k}R^{k-1} \tag{11b}$$

For a large t, the sum of leaning-efficiency parameters equals one. This implies a path-propensity Q_{ip}^{t} is a weighted average of the expected payoffs $\{R_{ip}^t\}$ obtained at each trip. For this reason, the reinforcement rule expressed by (11) is called a recency weighted method. We may call the reinforcement rule, (10), the knowledge-base rule in the sense that the propensities to choice are determined depending on the learning process and the obtained travel information.

We define individual-specific learning-efficiency parameter by:

$$\alpha_i^t = \begin{cases} \sum_{p\in\mathbf{P}} r_{ip}^t x_{ip}^t / V_i^t, & if\ \ 0 < \sum_{p\in\mathbf{P}} r_{ip}^t x_{ip}^t / V_i^t \le 1 \\ \gamma \ \in U(0,1), & otherwise \end{cases} \tag{12}$$

where γ is a random number drown from an uniform distribution, $U(0,1)$, and V_i^t is the value function at time t defined by

$$V_i^t = \sum_{p\in\mathbf{P}} Q_{ip}^t \tag{13}$$

We call the reinforcement learning model with endogenously determined learning-efficiency parameters, a set of equations consists of (9),(10),(12) and (13), RL-EDLE for a short.

III. CHARCTERIZATION OF EQUILIBRIA

A. Algorithm

RL-EDLE is implemented as follows:

Step 1. Set t=0. Generate the initial payoff matrix $\mathbf{r}^0$ and the prior propencity distribution, $\mathbf{Q}^0 = \mathbf{r}^0$, according to unform distribution ρU , where ρ is a arbitarary positive constant. Then, the initial choice probability $\mathbf{X}^0$.

Step 2. Set t=t+1. A learning-efficiency parameter for each driver $\{\alpha_i^t\}$ is determined by (12), then the ex-post propensity distribution $\mathbf{Q}^t$ at time t is calculated. The choice probability matrix $\mathbf{X}^t$ is given, then, updated payoff matrix $\mathbf{r}^t$ is obtained.

Step 3. If the stop rule, $\max_{p\in\mathbf{C}_i, i\in\mathbf{I}} \left| x_{ip}^{t+1} - x_{ip}^t \right| \le \varepsilon$ (ε is a positive small constant), is satisfied, then stop. Otherwise, return to Step 2.

The above algorithm consisits of a simple iterartion procedure. In spite of its simple strucure, it can apply to a complex route choice problem in congested networks.

B. Characterization of Equilibria

Analytically, the algorithm terminates if the following conditions are satisufied:

$$Q_{ip}^{t+1} - Q_{ip}^t = 0,\ for\ all\ \ p \in \mathbf{P}, i \in \mathbf{I} \tag{14}$$

because that if the above conditions are achieved, then the stop rule in RL-EDLE is also satisfied.

The next three propositions characterize the equilibrium achieved by the algorithm.

Proposition 1. For $p \in \mathbf{P}\ and\ i \in \mathbf{I}$, the equilibirum conditions (14) are satisfied if and only if the following conditions hold:

$$i)\ r_{ip}^t x_{ip}^t = 0,\ for\ \ p \in \mathbf{P}, i \in \mathbf{I}\ \ or$$
$$ii)\ r_{ip}^t = \beta_i > 0\ \ and\ \ r_{ip}^t x_{ip}^t = Q_{ip}^t,\ \ for\ \ p \in \mathbf{C}_i, i \in \mathbf{I}$$

Proof) Since from (10) we have $Q_{ip}^t - Q_{ip}^t = \alpha_i^t(R_{ip}^t - Q_{ip}^t)$, if we assume the following conditions

$r_{ip}^t x_{ip}^t = 0 (\Rightarrow \alpha_i^t = 0)$ for all $p \in \mathbf{P}\ and\ i \in \mathbf{I}$ or

$r_{ip}^t \ne 0\ and\ r_{ip}^t x_{ip}^t = Q_{ip}^t (\Rightarrow \alpha_i^t = 1)$ for $p \in \mathbf{C}_i, i \in \mathbf{I}$,

then (14) is satisfied.

Inversely, in order to satisfy the conditions (14), it should be either

$\alpha_i^t = 0$ or $r_{ip}^t x_{ip}^t = Q_{ip}^t$ for all $p \in \mathbf{P}$ *and* $i \in \mathbf{I}$. So the first conditons i) in the proposition is derived in a straightfoward way. Furthermore, from the expression, $Q_{ip}^t - Q_{ip}^t = \alpha_i^t x_{ip}^t (r_{ip}^t - V_i^t) = 0$, it follows that if $x_{ip}^t > 0$, then it should be $r_{ip}^t - V_i^t = 0$. This means that regardless of path selected, payoffs take the same value: $r_{ip}^t = \beta_i, p \in \mathbf{C}_i, i \in \mathbf{I}$. On the other hand, since there exisits the most expensive path among avairable paths, from the definition of payoff function, (6), there exists at least one of the paths whose payoff should vanish. Without loss of generality, we degignate the most expensive path q for driver i. Then it follows that

$$
\begin{aligned}
r_{iq} &= \lambda_i - u_{iq} = 0, \;\; q \in \overline{\mathbf{C}}_i \\
r_{ip} &= \lambda_i - u_{ip} = \beta_i \geq 0, \;\; p \in \mathbf{C}_i
\end{aligned}
\tag{15}
$$

Putting $\beta_i = 0$ leads to $x_{ip}^t = 0$. This contradicts the assumption of positive choice probabilities, so it should be $\beta_i > 0$.

Proposition 2. For $p \in \mathbf{C}_i, i \in \mathbf{I}$, the equilibirum conditions (14) are satisfied if and only if it does hold that

$$r_{ip}^t(\mathbf{X}^t) = 0, \;\; p \in \mathbf{C}_i, i \in \mathbf{I} .$$

This propostion says that if a set of paths $\mathbf{P}$ does not contain unused paths, then only the first conditions, i) , in proposition 1 hold. Proof is parallel to the one as shown in proposition 1. So we omit it. Proposition 2 ensures that the value function defined by (13) is minimized ultimately and that RL-EDLE will solve the non-smooth problem (8) if it converges.

Proposition 3. The equilibrium state described by proposition 1 or 2 is the user equlibrium.

The conditions i) in proposition 1 represent the well-known complementarity conditions for the user equilibrium. Similarly, the conditions ii) in proposition 1 lead to the expressions (15), implying that all used paths must have equal travel costs and travel costs of unused paths are at most equal to those of used paths. In addition, the conditions ii) lead to the expressions $r_{ip}^t - \sum_{p \in C_i} r_{ip}^t x_{ip}^t = 0$, which derive another definition of the equilibrium that at the equilibrium payoff obtained from each route is the same and equal the expected payoff over used paths.

C. *Behavioral Assumptions on Incomplete Information*

In this paper, we will conduct simulation under the assumption that while drivers have exact information on the route he selected, they have no information or have uncertainty in travel times on routes that they rarely use. In case of no information, we further assume that drivers evaluate unused paths with their prior belief, but, at the same time it is assumed that oblivion effect of past memory works. Those behavioral assumptions (Assumption I) are expressed as follows:

$$R_{ip}^0 = \rho U(0,1),\ where\ \rho\ is\ a\ given\ positive\ cons\tan t.$$

$$R_{ip}^t = \begin{cases} r_{ip}^t x_{ip}^t,\ if\ \ p\, is\, the\ path\ with\ \max imum\ choice\ probability \\ \gamma R_{ip}^{t-1},\ otherwise,\ where\ \gamma \in N(0,1) \end{cases}$$

If path p is rarely used path, then at each iteration the propensity to choose that path is gradually fading away at the rate of γ , a random number drown from a normal distribution.

The next behavioral assumption that we want to examine is mathematically expressed as follows (Assumption II):

$$R_{ip}^t = \begin{cases} r_{ip}^t x_{ip}^t,\ if\ \ p\, is\, the\ path\ with\ \max imum\ choice\ probability \\ \gamma r_{ip}^t x_{ip}^t,\ otherwise,\ where\ \gamma \in N(0,1) \end{cases}$$

The expressions reflect the assumption that a driver is informed of the travel times of paths available, but he does not take care about some routes because of rarely used paths for him.

Note that updating rules mentioned above is carefully designed to guarantee the convergence conditions stated in proposition 1 or 2 to realize the user equilibrium. It is not difficult task to put assumptions leading to non user-equilibrium states. For example, if we assume that drivers are persistent in their initial propensities, the system never converges to the user equilibrium.

IV. NUMERICAL EXAMPLES

A. *Network and Link Cost Function for Simulation*

A network for simulations is quoted from Braess [11] is depicted in Fig. 1. It is assumed that eight trips between origin-destination pair have three alternative paths. Assume that the link travel cost functions are:

$$
\begin{aligned}
&c_1(f_1) = f_1 + 50,\;\; c_2(f_2) = f_2 + 50 \\
&c_3(f_3) = 4f_1,\;\; c_4(f_4) = 4f_4,\;\; c_5(f_5) = f_5 + 10
\end{aligned}
$$

The user equilibrium principle generates the flow pattern $\mathbf{h}_1 = (0,0,8)'$ with the path costs: $\mathbf{u}_1 = (82,82,82)'$. The total travel cost is 656. A toll pricing policy imposing 21-unit fare on link 5, however, changes the previous flow pattern to $\mathbf{h}_2 = (3,3,2)'$ with the path costs $\mathbf{u}_2 = (73,73,73)'$, reducing the total cost to 584.

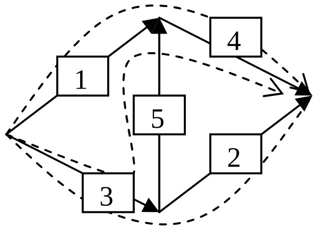

Figure 1. A single O-D connected by three paths: Numbers in parentheses indicate link numbers.

B. Oblivion effect

Figure 2 shows the simulation results under the assumption I. First three figures show frequency distributions of realized flows on three paths and the final figure in the bottom-corner represents the flow variation with iterations of calculation. Two doted lines indicating in the final figure show the analytical user equilibrium solution: $\mathbf{h} = (3,3,2)$. After 60 iterations the algorithm converges to the user equilibrium and the first three figures show that the equilibrium path flow occurs at extensively high frequency. However, the flow patterns generated in this simulation largely depend on relative magnitude between the initial propensities and the oblivion effects. As the ρ-parameter concerning with the initial propensities increases, different flow patterns from Figure 2 could occur.

C. Volatile User-Equilibrium

Figure 3 shows the simulation results under the assumption II. Like the previous case, the user equilibrium flow pattern is generated at high frequency; however, this time, we can observe the fluctuation around the user equilibrium until a certain time period. After this volatile user-equilibrium, the system converges to the user equilibrium in the end. This result partially accounts for the volatile user-equilibrium observed by Selten, et al. [5].

B. Concluding Remarks

One of the remarkable characteristics of the method presented in this paper is that it can deal with various drivers' route choice decisions based on the different behavioural assumptions. In addition, the method can apply to network equilibrium problems with ill-defined performance functions such as asymmetric and discontinuous cost functions. Those properties are suited with simulating of drivers' route choice behaviours.

For forecasting flows on congested networks, it is very important to improve our understanding of the conditions under which human adaptation deviates from expected value maximization. For this purpose, we need to develop a model consistent with cognitive science. Reinforcement leaning model is based on the simple principle that the probability of successful responses tends to increase with time. Human behaviour is so complex and many hidden parameters may affect their decision-makings. Our model also suggests that various flow patterns can occur according to different parameter-settings. We also examined that oblivion effect was one the important factors for achieving the user equilibrium. The relationship between the convergence property and the oblivion effect should be confirmed analytically.

References

[1] Beckmann M., C.B. McGuire, and C.B. Winsten, Studies in the Economics of Transportation, Yale University Press, NH,1956.

[2] Sheffi Y., Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming, Prentice-Hall, NJ, 1985.

[3] Smith M.J., The existence, uniqueness and stability of traffic equilibria, Transportation Research, 13B, pp.295-304,1979.

[4] Patriksson M., The Traffic Assignment Problem: Models and Methods, VSP PV, Utrecht, The Netherlands,1994.

[5] Selten R., M. Schreckenberg, T. Pitz, T. Chmura and S. Kube, Experiments and Simulation on Day-to-Day Route Choice-Bahaviour, CESifo Working Paper No.900,2003.

[6] Helbing D., M. Schonhof and D. Kern, Volatile decision dynamics: experiments, stochastic description, intermittency control and traffic optimization, New Journal of Physics 4, 33.1-33.16, 2002.

[7]Miyagi T., A modeling of route choice behaviour in transportation networks: An approach from reinforcement leaning, Urban Transport X, WIT press, UK, pp.235-244, 2004.

[8] Roth, A.E., and Erev I., Learning in extensive form games: Experimental data and simple dynamic models in the intermediate term, Games and Economic Behavior 8, 164-2122, 1995.

[9] Roth, A.E. and Erev, I., Predicting how people play games: Reinforcement learning in games with unique strategy equilibrium, American Economic Review 88, 848-881,1998.

[10] Grefenstette, J., "Credit assignment in rule discovery systems Based on genetic algorithms", Machine Learning, **3**, pp. 225-245, 1988.

[11]Arthur, W. B., Designing economic agents that act like human agents: A behavioral approach to bounded rationality, Amer.Econ.Rev. Papers Proc. 81,233-359, 1991.

[12] Clarke F.H., Optimization and Nonsmooth Analysis, SIAM, US, 1990 (Originally published in Wiley,1983).

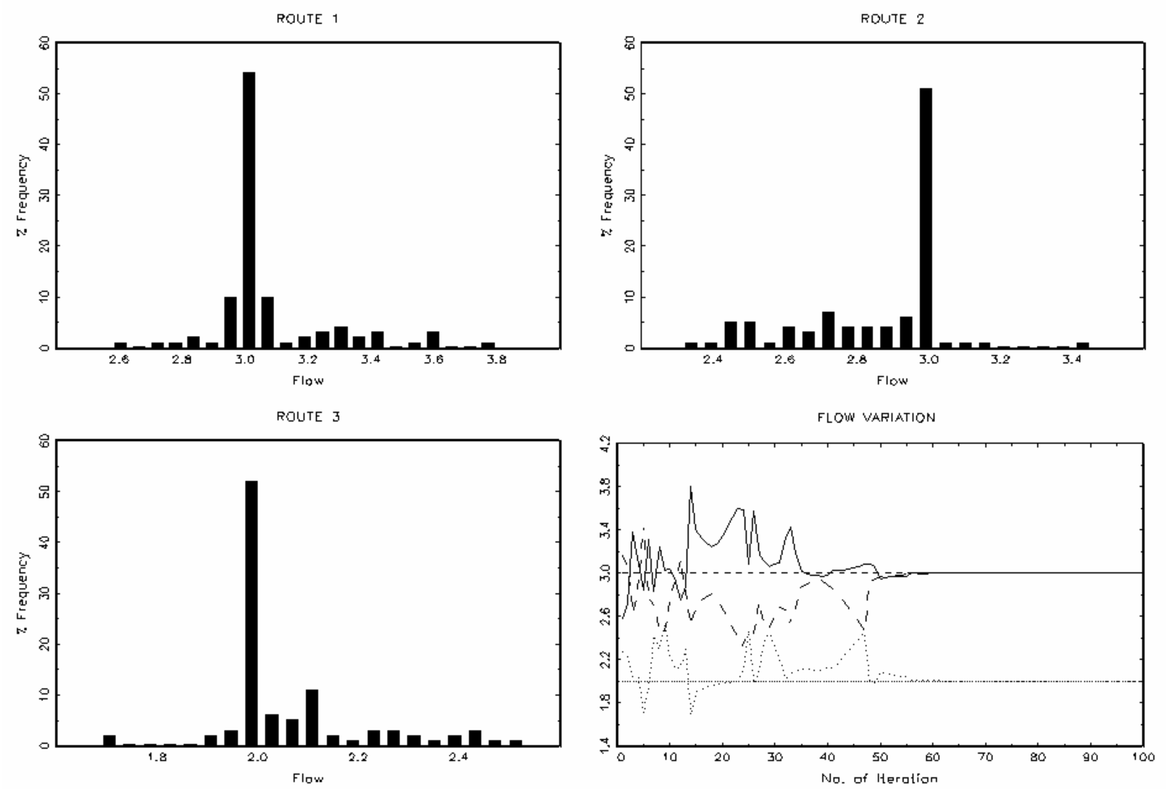

Figure 2. Frequency Distribution and Variation of Flow on Each Route: Oblivion Effect of Prior propensities

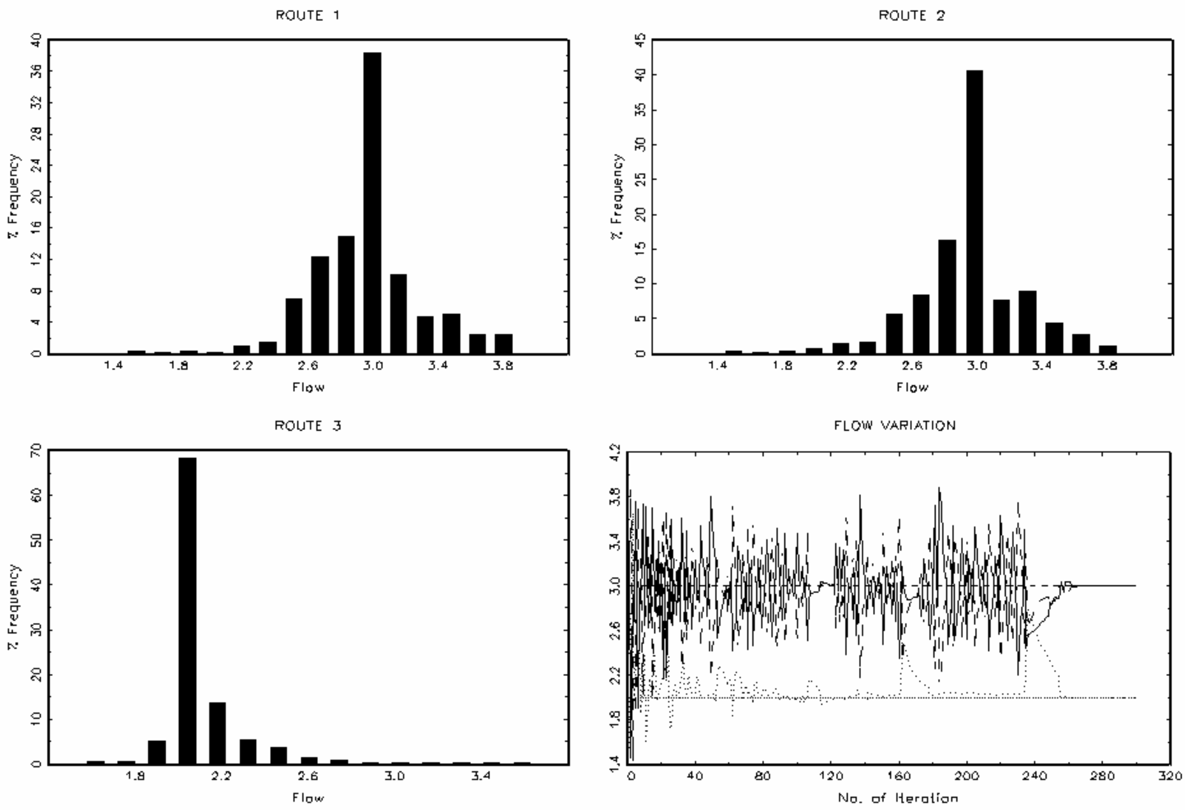

Figure 3. Frequency Distribution and Variation of Flow on Each Route: Volatile User Equilibrium