NISS

Statistical Inference for Gravity Models in Transportation Flow Forecasting

Mike West

Technical Report Number 60
May, 1997

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

STATISTICAL INFERENCE FOR GRAVITY MODELS IN TRANSPORTATION FLOW FORECASTING

By

MIKE WEST
Institute of Statistics and Decision Sciences
Duke University, Durham NC 27708-0251

Abstract – Gravity models are a class of log-linear regressions that have been used in studies of traffic flows between geographical zones. Stochastic parameter variations on these models, and their Bayesian analyses via stochastic simulation, are explored here in connection with the development of approaches to studying variability questions in established traffic flow network equilibrium models. In addition to developing methods of statistical inference and prediction specific to gravity models, this paper provides discussion of general concepts of Bayesian modelling and stochastic simulation analysis that will be of wider interest to the transportation community.

This report represents research performed under the cross-disciplinary transportation project *Measurement, Modeling and Prediction for Infra-structural Systems,* run by the National Institute of Statistical Sciences (NISS), and sponsored by NSF-DMS-9313013.

STATISTICAL INFERENCE FOR GRAVITY MODELS IN TRANSPORTATION FLOW FORECASTING

By

MIKE WEST
Institute of Statistics and Decision Sciences
Duke University, Durham NC 27708-0251

1. CONTEXT AND BACKGROUND

In connection with current and anticipated developments in intelligent transportation systems, studies of problems of short-term traffic flow forecasting and management on urban road networks raise questions about patterns of variability of link flows and travel times, and dependencies amongst such quantities across collections of links. These kinds of issues are being studied as part of the collaborative project run by the National Institute of Statistical Sciences (NISS). As prelude to wider statistical modelling and exploration of variability and dependence issues, a part of the initial stage of this project focuses on exploration of the degrees of uncertainties about, and relationships between, equilibrium link travel times arising from static network equilibrium models. The network structure and flow models of the Advance project (Boyce et al, 1992; Berka and Boyce, 1994), based on a geographic zone structure in northeastern Illinois, provides relevant context, and associated zone-to-zone flow survey information from the Chicago Area Transportation Survey (CATS–Ghislandi, 1994) provides some relevant data. In connection with this exploratory study, variations on traditional gravity models for zone-to-zone flows are examined, with a view to a further stage of the project that may address questions of uncertainty about equilibrium link flows and travel times by repeatedly simulating average zone-to-zone flows and then running such replicates through existing network flow models, such as the Advance model. This way, replicated runs produce sampled equilibrium flows and times that incorporate and represent the uncertainties about average zone-to-zone flow rates captured in the statistical measures of uncertainties about the gravity model parameters. Patterns of dependency among equilibrium characteristics are similarly represented.

This program requires initial work on statistical inference for gravity models, and that is the subject of this article. In particular, the development here presents approaches to Bayesian inference in variations of traditional gravity models, and demonstrates how the Bayesian approach naturally delivers simulated values of the model parameters that can lead into the kinds of sensitivity studies in equilibrium flow models just described.

It should be noted that, in addition to the immediate goal of sensitivity analysis in static equilibrium flow models, the kinds of statistical developments here provide a formal basis for integrating Bayesian inference in gravity models (or other models for predicting traffic flow patterns) into micro-simulations of networks, whether viewed as static or, more generally and ultimately realistically, dynamic and stochastic. Simulated realisations of actual zone-to-zone flows are trivially generated within this framework, and represent parameter uncertainty as well as natural stochastic variation in flows about average values. It is expected that such developments with be forthcoming in future studies.

Some discussion of gravity model forms is given in the next section. Section 3 develops the framework for Bayesian inference and simulation of posterior (or post-data) distributions for gravity model parameters in the context of actual zone-to-zone flow data observed on a specified geographical zone structure. Covariate information about anticipated average travel times between zones is assumed available to provide impedances in the models. Section 4 exhibits some summaries from analysis of a small section of a survey data set from the Chicago area, on a small sub-network. The section concludes with further discussion, including questions of scaling up from survey sample to population levels.

In addition to the specific focuses in the gravity model context, it is expected that the conceptual and methodological aspects of the Bayesian approach will have wider utility in the transportation research community, so this paper serves, in part, to introduce and exemplify Bayesian inference in an accessible disciplinary context.

2. MODELS FOR ZONE FLOWS

The cornerstone of gravity modelling is the class of Poisson log-linear models for zone-to-zone traffic flows (Sen, 1986; Smith, 1987) Label the geographic zones of the area under study as $1, \ldots, n$, and consider a specified period of the day during which zone-to-zone trips arise at an assumedly constant rate. Write y_{ij} for the number of trips from origin zone i to destination zone j in the period; assume, initially, that these flows are to be observed precisely at some future time. The basic Poisson model assumes these quantities to be conditionally independently Poisson distributed with means t_{ij}, these means depending on characteristics of the origin and destination zones and the transportation network. The particular variant of gravity model discussed here adopts the form and notation $t_{ij} = a_i b_j f_{ij}$ where characteristics of the zones as origins and destinations are incorporated via zone-specific parameters a_i and b_j respectively, and the interaction term f_{ij} represents additional factors arising primarily from network characteristics.

For the technical reason of parameter identification, we assume and constrain the model so that $f_{ii} = 1$ for all zones i, and must also constrain one of the a_i or b_j parameters to a specified value; one simple way of doing this is to rewrite as

$$t_{ij} = m a_i b_j f_{ij},$$

where $a_n = b_n = 1$ so that the single parameter $m = t_{nn}$. (This is known as aliasing the parameters a_n and b_n, and can be effected in other ways; an equivalent alternative, but one that here leads to more technical complications, is to constrain, say, the geometric mean in this of the a_i to be unity, and the same for that of the b_j. The aliasing used here is similar to that in the GLIM system (Baker and Nelder, 1985).)

All quantities m, a_i, b_j and f_{ij} are positive, hence the log-linear representation $\log(t_{ij}) = \log(m) + \log(a_i) + \log(b_j) + \log(f_{ij})$. Further assume that the interactions depend on available covariates, a vector of covariate values x_{ij} being available for each zone pair. The simplest form of dependence is a regression $\log(f_{ij}) = g' x_{ij}$ for some uncertain regression parameter vector g. Coupled with the Poisson assumption, this specifies a log-linear model in the class of *generalised linear models* that is central to modern statistical regression analysis (McCullagh and Nelder, 1989; West, 1985). Often, a single covariate standing proxy for anticipated or estimated average travel

time between zone pairs is used, and this is the case in the study reported below. In such a case, assume the scalar x_{ij} represents expected travel time, so then g is a single parameter and the Poisson means become

$$t_{ij} = ma_i b_j \exp(g x_{ij}).$$

Assuming the common sense expectation that $g < 0$, the x_{ij} are sometimes termed traffic flow *impedances* as larger values reduce the expected flows. This, then, is the basic model studied here. The following development introduces some variants, and other variations that, for example, add further covariates, might similarly be studied and analysed using the methods of this paper.

The most interesting variant is introduced as a mechanism for relaxing the strict assumption of linear regression for the zone-by-zone interactions f_{ij}, and also provides a neat approach to inducing stochastic variations on the basic model that can be viewed as a means of relaxing the strict Poisson/log-linear structure for robustness reasons. That is, extend the above form to include positive *random interaction effects* h_{ij} so that

$$t_{ij} = ma_i b_j h_{ij} \exp(g x_{ij}) = ma_i b_j \exp(g x_{ij} + \log(h_{ij})).$$

This model is identified through the imposition of distributional assumptions for the h_{ij}. For example, suppose the h_{ij} are randomly generated from a log-normal or gamma distribution with mean unity, and this independently of the other model parameters m, a_i, b_j and g; then, conditionally on the remaining model parameters, $E(t_{ij}) = ma_i b_j \exp(g x_{ij})$, reducing to the original interaction form. This can be seen as an elaborated model that allows for *extra-Poisson* variation; indeed, with a gamma model for the h_{ij}, the Poisson is effectively replaced by the more diffuse negative binomial. Otherwise, and more pragmatically, introducing these random effects parameters allows for sensitivity and robustness studies relative to the basic Poisson model; estimation of the h_{ij} will indicate which zone-to-zone pairs are mostly consistent, and which are less consistent, with the gravity model simply by inferences on which of the h_{ij} are reasonably close to unity, and which are significantly larger than unity, respectively.

Previous works on inference in gravity models have developed maximum likelihood, and related, approaches to estimation of the a_i, b_j and g parameters (Sen, 1986; Smith, 1987). Here, in the context of the elaborated model, Bayesian inference is developed, to both extend and complement other approaches. In particular, Bayesian analysis as developed here delivers not only point estimates of all parameters, but accessible posterior distributions for these parameters that sufficiently summarise and describe the uncertainties about such parameters and the patterns of dependencies between parameters. Further, and especially of relevance in connection with uses of these models to impute or forecast zone-to-zone flows, the Bayesian approach centrally features the evaluation of predictive distributions for future forecasting purposes. The terminology here will be explained and elaborated below. For readers not fully conversant with the basic concepts and methods of Bayesian statistics, useful background reading can be found in Bernardo and Smith (1994).

3. BAYESIAN ANALYSIS AND COMPUTATION

3.1. Bayesian framework and likelihood

Consider now the prospect of observing actual OD flows $y = \{y_{ij};\ i,j = 1,\ldots,n\}$. Interest lies in estimating all gravity model parameters $z = \{m, a_i, b_j, h_{ij}, g;\ i,j = 1,\ldots,n\}$, and in describing and representing resulting uncertainties about these parameters. From a Bayesian viewpoint, this is achieved by computing and summarising the posterior (or post-data) distribution based on a specific prior (or pre-data) distribution, and typically exploring the sensitivity of characteristics of the posterior to various changes in the prior and in the model (Bernardo and Smith, 1994).

Formally, this is performed via Bayes' theorem in terms of probability density functions, namely

$$p(z|y) = cp(z)p(y|z)$$

where c is simply a normalising constant, $p(z)$ represents the prior density for parameters and $p(y|z)$ represents the sampling model for the data conditional on the parameters; since y is here viewed as fixed once observed, $p(y|z)$ is the likelihood function for z. Under the gravity model specified, this is, as a function of z,

$$\begin{aligned} p(y|z) &\propto \prod_{i=1}^{n}\prod_{j=1}^{n} t_{ij}^{y_{ij}} \exp(-t_{ij}y_{ij}) \\ &= m^{y_{**}} \mathrm{e}^{gs} \left(\prod_{i=1}^{n} a_i^{y_{i*}}\right)\left(\prod_{j=1}^{n} b_j^{y_{*j}}\right)\left(\prod_{i=1}^{n}\prod_{j=1}^{n} h_{ij}^{y_{ij}}\right) \exp\left(-m\sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j h_{ij} \mathrm{e}^{gx_{ij}}\right) \end{aligned}$$

where

- $y_{i*} = \sum_{j=1}^{n} y_{ij}$ for $i = 1,\ldots,n$,
- $y_{*j} = \sum_{i=1}^{n} y_{ij}$ for $j = 1,\ldots,n$,
- $y_{**} = \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}$, and
- $s = \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}y_{ij}$

are some summary statistics. At this point, maximum likelihood based analysis would proceed by finding the parameter estimates that maximise this function, typically using some form of iterative mode search (eg. McCullagh and Nelder, 1989; Sen, 1986; Smith, 1987). Some more commentary on this appears below. The Bayesian approach requires a prior distribution $p(z)$, and the form currently used, and used in the data analyses reported below, is structured as follows.

3.2 Prior distributional structure

The prior is structured as

$$p(z) = p(m)p(g)\left(\prod_{i=1}^{n-1} p(a_i)p(b_i)\right)\prod_{i=1}^{n}\prod_{i=1}^{n} p(h_{ij}),$$

with implicit independence structure. Notice that, as $a_n = b_n = 1$, fixed, only the first $n-1$ of the origin and destination parameters a_i and b_j appear here. The components of this full joint prior density are described below. Throughout, gamma distributions are featured; here $w \sim Ga(w|r,q)$

denotes the gamma distribution with shape parameter r and scale parameter q; the density function is $p(w) = cw^{r-1}\exp(-qw)$ for $w > 0$, where c is simply the normalising constant $c = q^r/\Gamma(r)$. This distribution therefore has mean $E(w) = r/q$, mode $\max\{0, (r-1)/q\}$ and variance r/q^2.

- For the critical impedance parameter, a uniform prior is assumed; that is, $p(g) = 1/G$ for $-G < g < 0$, being zero otherwise, for some large and positive G.
- For $i, j = 1, \ldots, n-1$, each of the zone origin and destination parameters a_i and b_j has a gamma prior, namely $Ga(a_i|c_a a_{i0}, c_a)$ and $Ga(b_j|c_b b_{j0}, c_b)$. Here $c_a > 0$ and $c_b > 0$ are specified scale parameters, and the specified prior means are $E(a_i) = a_{i0}$ and $E(b_j) = b_{j0}$. An important special case is the reference or uninformative prior that arises by letting c_a and c_b tend to zero, producing $p(a_i) \propto 1/a_i$ and $p(b_j) \propto 1/b_j$.
- The quantity m is similarly gamma distributed, $Ga(m|c_m m_0, c_m)$, with specified prior mean $E(m) = m_0$ and shape parameter $c_m > 0$.
- The random effects h_{ij} are assumed gamma distributed with common prior $Ga(\cdot|c_h, c_h)$ for some specified constant shape parameter $c_h > 0$. This implies $E(h_{ij}) = 1$ and $V(h_{ij}) = 1/c_h$ for all i, j. Note that setting $h_{ij} = 1$ delivers the traditional gravity model, so that the prior on these random effects is therefore 'centred' about that traditional model. If c_h is very large, the gamma prior concentrates about the unit mean; otherwise, the prior is more diffuse, allowing for random deviations away from the traditional, or baseline, model.

It should be reaffirmed that our analysis is not specific to these priors, and other forms may be assumed and studied. Two comments are noteworthy, however. First, the assumed gamma forms lead to nice conditionally conjugate structure that is simplifying technically, and are traditional forms in Bayesian analysis for this reason, among others. Second, we use relatively diffuse, uninformative versions of these priors by taking small or zero values for the shape parameters; the prior for g is already uninformative in the uniform sense.

3.3 Conditional Posterior distributions

To begin, the combination of prior densities with the likelihood function to produce posterior densities is elaborated technically, followed by discussion of connections with point estimation methods based on maximum likelihood.

Directly from the mathematics of Bayes's theorem with the above priors, it is easy to see the structure of *conditional posteriors* for individual subsets of the parameter vector z, where the remaining parameters are assumed known (conditioned upon). This is a typical aspect of multi-parameter Bayesian analysis, and one that critically underpins analysis based on iterative methods of posterior simulation (Smith and Roberts, 1993; Müller, 1991).

To exemplify and develop the first such conditional distribution, focus on the parameter m, just the expected flow from zone n into itself under the specified model. It is clear that the posterior density for m conditional on y and all other parameters in z but m (written simply as $z\backslash m$) is simply proportional, as a function of m alone, to the full product $p(z)p(y|z)$. Ignoring all positive multiplicative factors not involving m, this simply reduces to

$$p(m|y, z\backslash m) \propto p(m)m^{y_{**}}\exp(-mk(z\backslash m)) \propto m^{c_m m_0 + y_{**} - 1}\exp(-m(c_m + k(z\backslash m)))$$

where

$$k(z\backslash m) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j f_{ij},$$

with $f_{ij} = h_{ij}\exp(gx_{ij})$. Hence,

- were $z\backslash m$ known, the posterior for m is simply a gamma distribution (updated from the gamma prior), namely $Ga(m|c_m m_0 + y_{**}, c_m + k(z\backslash m))$.

Similar ideas lead to the following conditional posteriors for all other parameters.

- For each $i = 1, \ldots, n-1$, a_i has distribution $Ga(a_i|c_a a_{i0} + y_{i*}, c_a + m\sum_{j=1}^{n} b_j f_{ij})$.
- For each $j = 1, \ldots, n-1$, b_j has distribution $Ga(b_j|c_b b_{j0} + y_{*j}, c_b + m\sum_{i=1}^{n} a_i f_{ij})$.
- For every pair $i, j = 1, \ldots, n$, the individual random effects parameter h_{ij} has the conditional distribution $Ga(h_{ij}|c_h + y_{ij}, c_h + m a_i b_j \exp(gx_{ij}))$.
- For the impedance parameter g, $p(g|y, z\backslash g) \propto \exp(gy_{**} - mk(z\backslash m))$, $(-G < g < 0)$; note that this function involves the argument g in through $k(z\backslash m)$.

Before proceeding, in the next subsection, to describe iterative simulation from these conditional distributions, it is illuminating to connect with maximum likelihood estimation by elaborating an iterative algorithm for computation of posterior modes. The mode of the full joint posterior $p(z|y)$ may be iteratively solved by sequencing through the above conditional distributions and, at each step, computing the mode of the individual conditionals. To tie explicitly with the MLE calculations, specialise to the reference prior case in which the prior scale parameters $c_m = c_a = c_b = 0$, and set $h_{ij} = 1$. Then each of the conditional distributions above (now excepting those for the h_{ij}) has mode, as a function of the other parameters, given below; the conditional means are noted too, for further use below:

- m has conditional mode $\max(0, (y_{**} - 1)/k(z\backslash m))$, and conditional mean $y_{**}/k(z\backslash m)$;
- each a_i has conditional mode $\max(0, (y_{i*} - 1)/\sum_{j=1}^{n} b_j f_{ij})$, and mean $y_{i*}/\sum_{j=1}^{n} b_j f_{ij}$;
- each b_i has conditional mode $\max(0, (y_{*j} - 1)/\sum_{i=1}^{n} a_i f_{ij})$, and mean $y_{*j}/\sum_{i=1}^{n} a_i f_{ij}$;
- the conditional mode of g is not available in closed form, but may be solved by numerically maximising the log posterior, i.e. solving the equation

$$y_{**} = m\sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j h_{ij} x_{ij} \mathrm{e}^{gx_{ij}}$$

over $-G < g < 0$. This is easily solved via the Newton-Raphson gradient algorithm.

Sequencing through these equations, replacing each parameter by its current conditional mode at each step, leads iteratively to the approximate posterior mode. In cases where a y_{i*} or y_{*j} is zero, the corresponding a_i or b_j has a conditional posterior unimodal at zero and, in terms of solving for reasonable estimates of the parameters, the conditional means might be used instead; this happens in cases of zero zone-to-zone flows between some zones due to small samples or simply low flow rates. Note that this corresponds exactly to iteratively solving for the maximum likelihood values using standard methods (McCullagh and Nelder, 1989; Sen, 1986; Smith, 1987). In establishing useful starting values for Bayesian posterior simulations below, iterating between conditional means rather than modes is used to avoid the zero problem. From this Bayesian perspective, this method of evaluating point estimates is of interest only for initialising simulations, as full posterior descriptions are obtained from the simulations.

In cases of non-reference priors, exactly the same ideas apply. Further, incorporating the random effects parameters h_{ij} simply adds the computation of their modes or means, and they are directly available from the relevant conditional gamma distributions above, requires no further comment.

Note than, as n is typically fairly large, the posterior will be typically quite concentrated around the region corresponding to the mode for the single parameter g. Hence the standard Bayesian asymptotic normal approximation can be expected to provide a useful approximation to the true conditional posterior for g. This is important in simulation analysis, as a modified version of this normal approximation may be used as a Metropolis proposal or importance sampling distribution for generating g values. More on this below.

These iterative computations have been performed in models for various subsets of a real OD flow problem mentioned and used for illustration below. Custom software has been compared with the standard GLIM package (in the case of small n, i.e. $n = 50$, and random effects $h_{ij} = 1$), to verify correctedness.

3.4 Posterior simulations

Fully Bayesian inference requires more global descriptions of posterior uncertainties and relationships not easily available through traditional analytic methods in these models. Hence, consistent with the current revolution in applied statistics generally, simulation methods are used to effectively draw sets of sampled parameters from the true posterior $p(z|y)$, and then base posterior inferences on summaries of the samples (Smith and Robert, 1993). Calculation of posterior samples is feasible here via Markov Chain Monte Carlo methods, combining simple Gibbs sampling for most of the parameters with a Metropolis/rejection sampling step for the impedance parameter g. The techniques are quite standard in applied Bayesian work (Smith and Roberts, 1993; Müller, 1991) and there is a large applied literature describing such simulation methods in in many application areas.

The analysis parallels the MLE/posterior mode computation, repeatedly iterating through the individual conditional posteriors. Now, however, instead of computing a simple mean or mode at each step, a random draw is made from each conditional distribution (of course, if the posterior is very concentrated, such samples will lie close to the mode). Also, instead of simply iterating to convergence to a single modal value for z, successive sampled parameters are saved, and ultimately represent a (possibly very large) sample of z vectors from $p(z|y)$. There are issues of convergence and sensitivity to starting values. In these models, as many others, convergence is easily theoretically assured (using standard results in the above references). Starting values are taken as approximate posterior modes or means, as mentioned above, and it is usual to run such iterative simulations for some initial 'burn-in' steps before subsequent samples are saved, assuming convergence to sampling the true posterior is by then approximately achieved so that the dependence on starting values has dissipated. Analyses reported below burn-in for at least 1000 iterations, and repeat simulations for sensitivity analysis support assumed convergence.

In detail, the iterations proceed as follows. At a general iteration indexed by r, current values of all parameters are represented by the current value of $z = z^{(r)}$; thus

$$z^{(r)} = \{m^{(r)}, a_i^{(r)}, b_j^{(r)}, h_{ij}^{(r)}, g^{(r)};\ i, j = 1, \ldots, n\}$$

is (after burn-in) approximately distributed according to $p(z|y)$. Move to another sampled vector $z^{(r+1)}$ via the following sequence.

(1) Draw a new value of m, namely $m^{(r+1)}$, by simulating from the conditional gamma posterior $p(m|y, z\backslash m)$ with $z = z^{(r)}$; modify the parameter z simply by replacing $m^{(r)}$ by this new value.

(2) For each $= 1, \ldots, n-1$, draw a new, conditionally independent values of the a_i by simulating separately from the conditional gamma posteriors $p(a_i|y, z\backslash a_i)$ with z at the current value; call the resulting draws $a_i^{(r+1)}$, and modify z by replacing each $a_i^{(r)}$ by $a_i^{(r+1)}$.

(3) As in (2), but now independently simulating $b_1, \ldots, b_{n-1}$; update z with the new values.

(4) As in (2) and (3), but now independently simulating each of the h_{ij}. update z with the new values.

(5) Sample a value of g from $p(g|y, z\backslash g)$ with z set at the value so updated in steps (1)–(3); update z to include this sampled value $g^{(r+1)}$, and call the result $z^{(r+1)}$.

(6) Save $z^{(r+1)}$ as the latest sampled vector from $p(z|y)$. Update the index r by one, and return to step (1) to continue.

Only step (5) requires further comment, as sampling in all other steps simply involves (trivial) simulation from known gamma distributions. Current software uses routine from Numerical Recipes (Press et al, 1992)

At step (5), the univariate density for g is not of a standard form. It may be sampled many ways, the method of preference here being a standard Metropolis/independence chain method (Müller, 1991), for two reasons. First, a simple and excellent normal approximation is available, as described above, to provide 'candidate' samples, detailed further below; second, the method is computationally trivial to implement and embed in this iterative chain.

The basis of this step is to draw a 'candidate' value of g from a specified density $q(g)$ chosen as an approximation to the exact density $p(g|y, z\backslash g)$. Call this candidate value g^*. Then, save and record this candidate value as the new value for g, namely $g^{(r+1)} = g^*$, or reject it and revert to the previous value $g^{(r+1)} = g^{(r)}$. The candidate value is saved with Metropolis acceptance probability computed simply as

$$\frac{p(g^*|y, z\backslash g)q(g^{(r)})}{p(g^{(r)}|y, z\backslash g)q(g^*)}$$

(for more details, see Smith and Roberts, 1993, or Müller, 1991). Thus the sampled values of g mayor may not change between successive iterations, though convergence of the resulting $z^{(r)}$ vectors to samples from the true posterior is assured so long as the candidate density q is bounded and has the same support as the true posterior. In the context here, the true posterior is easily evaluated, up to a constant of proportionality, as described earlier, simply through $\log p(g|y, z\backslash g) = c + gy_{**} - mk(z\backslash m)$ over $-G < g < 0$. One of the key features of the Metropolis step is that the above acceptance probability involves the ratio of the posterior density at two points, so the proportionality constant is not required as it cancels.

As earlier mentioned, the recommended candidate density is based on a modified asymptotic normal approximation to the true posterior. The standard asymptotic normal approximation to $p(g|y)$ has mean $\hat{g}$ and variance $v^2(\hat{z})$, where $\hat{z}$ is the posterior mode and

$$1/v^2(z) = m \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j h_{ij} x_{ij}^2 \mathrm{e}^{gx_{ij}}$$

for all z. The recommended candidate density q is based on this normal approximation, but with a slightly inflated variance, i.e. a normal with mean $\hat{g}$ and variance $k^2v^2(\hat{z})$ where $k > 1$ a scale factor chosen to spread out this basic normal approximation. This recognises that the asymptotic normal posterior approximation, as the usual MLE approximation, will typically underestimate the uncertainty about g; in analyses reported below, the value $k = 5$ is used. One further detail arises in defining q, due to the fact that the true conditional posterior is truncated to the finite interval $-G < g < 0$; thus the above normal distribution is replaced by the same distribution truncated to this region, though this does not unduly effect computations as sampling normals truncated over an interval is trivial. This completes the specification of structure of the iterative posterior simulation algorithm.

4. CHICAGO AREA STUDY

The Household Travel Survey of the 1990 Chicago Area Transportation Study (Ghislandi et al, 1994) covers a six county region of northeastern Illinois and provides socio-demographic and travel information that forms part of the database used in traffic forecasting and network studies in connection with the Advance project (Boyce et al, 1992). Travel data from this survey are available over a collection of $n = 783$ zones during various daytime periods, and the example chosen here here is the late afternoon, two hour peak travel period. The useable returns from households surveyed across these zones total 19,314 households (Ghislandi et al, 1994, exhibit 4) out of the total census population of 2,760,200. The corresponding total sample number of trips during this period is $y_{**} = 22,759$.

Some illustration is provided initially on the basis of a selection of 50 zones, those labelled 651-700 inclusive in the CATS zone identification scheme. Figure 1 gives a perspective plot of the observed flows; note, in particular, that most OD pairs have no sampled trips and the non-zero flows are mainly within zones. The impedance factors x_{ij} are here taken as estimated average zone-to-zone travel time based on the Advance network equilibrium model outputs using old CATS OD flow data from earlier work as inputs (Boyce et al, 1992). Figure 2 provides a perspective plot of the values $-x_{ij}$ over these selected 50 zones.

Figures 3 through 9 provide some summaries of posterior simulations based on the above development. The analysis uses reference priors for the parameters m, a_i, b_j, that is $c_m = c_a = c_b = 0$. For the random interactions h_{ij}, the shape parameter is $c_h = 2$, indicating the expectation that there will be some fair degrees of variation in the OD flow data away from the baseline Poisson model with $h_{ij} = 1$, simply allowing for such variation though not anticipating its nature. After running the mode/mean iterations for (a very generous and conservative) 5000 runs to convergence, the parameter values are approximately $\hat{m} = 0.3150, \hat{g} = -0.6393$, and the asymptotic normal posterior for g has approximate standard deviation $v(\hat{z}) = 0.0109$. (For comparison, the corresponding maximum likelihood values are $m = 0.2980$, $g = -0.6232$, and $v = 0.0190$.) Figure 3 provides a perspective plot of the values $\exp(\hat{g}x_{ij})$ over the 50 zones.

Starting at the posterior mode/mean mean values for z, the iterative simulations are burnt-in for 1000 runs and then a further 25,000 runs performed. Every 250 runs, the current parameter draws are saved resulting in a sample of 100 z vectors, approximately distributed according to $p(z|y)$. Each z vector is used to compute corresponding values of the Poisson means t_{ij}, thus generating a sample of size 100 from the posterior for these means. Figure 4 displays a plot of one

such sample, a representative from the posterior distribution of expected number of trips for the selected 50 zones. Figure 5 is a similar plot, but now representing predictions of future trips; for each zone pair, independent Poisson variates are simulated with the means displayed in Figure 4, resulting in a sample of actual flows from the posterior predictive distribution of such flows. Most of the flows occur within zones; to more clearly see the model implications within zones, Figure 6 plots actual within-zone flows, and adds the estimated flows based on the approximate posterior mean/modal values of the t_{ij} simply computed directly from $\hat{z}$. The agreement seems excellent. Figure 8 provides a similar plot, but now of all 100 posterior samples of within-zone expected flows. This gives some indication of posterior uncertainties about the t_{ii} for each zone i. Figure 9 displays a similar picture, but now of predicted flows, with the additional, purely random Poisson variation about the simulated t_{ii} added. Finally, Figure 7 provides an histogram display of the 25,000 draws from the posterior for the impedance factor g, representing the true posterior density. For contrast, the asymptotic normal approximation mentioned and used in the analysis is superimposed as a density curve; note that (as is typical) this asymptotic approximation is offset in location and, more importantly, underdispersed relative to the true posterior. A critical factor in the analysis is that the Metropolis candidate distribution $q(g)$ is rather more spread out and so the simulations capture regions of g values supported under the true posterior.

Repeat analysis constraining the interaction parameters h_{ij} to unity, or with priors very concentrated near unity, produce results that are broadly similar, although the more robust analysis with less constrained h_{ij} factors does protect against outlying counts and impedances, and other irregular or peculiar features in the data; this kind of modification of a baseline model is similar in concept and general terms to those in West (1995), though the technical structure differs. Further similarities exist with the random effects structures introduced in biased sampling models in West (1995) in an entirely different application context and with different models.

Turn now to analysis of the full 783 zone network of the CATS area study. Repeat analyses have been performed with the full data set on this very much larger network. It is found that the reasonable fit of the baseline model to the small 50 zone region does not scale up to the entire network. The critical limiting factor is the very limited and highly sparse data set of flows; there are only 8,941 $i-j$ pairs between which there are non-zero flows in the sample, of the total 613,089 zone pairs. Also, the extent to which the historical estimates of zone-to-zone expected travel times vary across zone pairs is rather small; in fact, many origin zones i have estimated times that are constant across ranges of destination zones j. Most of the larger flow rates are within zones, and most origin zones have very low average flows rates even though they may have one or two quite large observed flows. As a result, the baseline gravity model fit results in low values of the a_i and b_j parameters, and quite radically under-estimates the few much higher flows, notably those within-zones. This is to be expected from such extremely sparse data, and could only be improved if significantly larger surveys were mounted, producing many more zone-zone pairs with non-zero observed flows.

From a purely predictive viewpoint, however, the incorporation of the individual random effects h_{ij} provides a mechanism for improving the model fit by allowing for the larger flow rates through larger values of the h_{ij} for the zones in question. To do this automatically and objectively, the prior gamma distribution for these factors must be diffuse enough so that the posteriors adapt to much higher values. Exploratory analyses reveal that, in the context of the CATS data analysis,

the scale factor c_h of 2 above overly constrains the analysis, and a much lower value is needed. The remaining figures summarise results of analysis using $c_h = 0.01$, a very low value and one allowing for adequate adaptation.

In this analysis, $\hat{g} = 0.0098$ (with approximate standard deviation of 0.0007), a very much reduced value indicating the far lower explanatory power of the impedance factors in the full network. Figure 10 provides a perspective plot of the values $\exp(\hat{g}x_{ij})$ over just the 50 zones earlier selected; the reduced regression effect is clear here. Figure 11 displays a representative perspective plot of predictions from the full model, again restricted for display to the selected 50 zones. This is obviously quite good, though most of the agreement between predictions and actuals now comes through the individual random effects, especially along the 'diagonal' in this picture, i.e. for the within-zone flows. This can be seen in the final Figure 12 where, for these 50 selected zones, the actual h_{ij} values sampled from the posterior to produce the predictions exhibited are themselves displayed.

The above example analyses are based on recorded survey data. For eventual forecasting of expected or actual OD flows at the population level, analysis needs extension to scale-up from the survey to population levels. These kinds of scale-up problems can be approached in various ways, formally and informally, and using various additional sources of census and demographic information. Full discussion is not in order here, though the kinds of external modelling and imputation of sample-to-population level flows developed by Kim et al (1992) are suggestive of the kinds of more formal models that could be developed. In terms of the rather specific goals of the NISS project in connection with studies of uncertainties about, and dependencies amongst, flows, much simpler and cruder approaches should suffice. At a basic level, origin zone-specific scale up factors may be used, and the kinds of information sets needed for this are indeed available in the CATS area study. Suppose that the network region has census population counts of households, say N_{ci} total households for zone i. Suppose also that the survey returns provide N_{si} useable returns for zone i. Two further assumptions lead to simple scale-up: first, that non-responses to survey questionnaires are effectively random/non-informative, and second, that households within zones are sampled proportionately. Under such assumptions, the model scales-up simply, extrapolating the conditional Poisson model to the population level by increasing the expected flows t_{ij} to t_{ij}/p_i where $p_i = N_{si}/N_{ci}$ for each zone i (and noting that zones with no households are excluded, having zero outward flows). These methods may be quite reasonable in the context of preliminary studies of OD flow uncertainties and dependencies. They will not, however, necessarily adequately capture zone-to-zone variations in flow intensities that depend critically on socio-demographic characteristics of the zones; then more direct study of the relationships between such zone characteristics and trip flow rates, perhaps with Kim et al (1992) as a starting point, is in order. It should be remarked that, with such additional models in hand, their incorporation into the above analysis framework will be direct; at a general level, this will just extend the framework to include additional, uncertain scale-up factors as parameters, and add external information about those factors through appropriate prior distributions. Such extensions are anticipated in future.

Acknowledgements – Research partially supported by the NSF under grants DMS-9305699 and DMS-9313113. This work represents part of the NISS Transportation project funded under the latter grant. Useful discussions were held with Alan Karr and Ashish Sen, and the author received assistance with data management from Suresh Acharya, Giovanni Petris and Vonu Thakuriah.

REFERENCES

Baker, R.J., and Nelder, J.A. (1985) *GLIM Release 3.77,* Oxford University, Numerical Algorithms Group.

Berka, S., and Boyce, D.E. (1994) Implementation and solution of a large asymmetric network equilibrium model, *Working Paper*, Urban Transportation Center, University of Illinois at Chicago

Bernardo, J.M., and Smith, A.F.M. (1994) *Bayesian Theory,* Wiley, London.

Boyce, D. E., Kirson, A. M., and Schofer, J. L., (1992) ADVANCE: The Illinois Dynamic Navigation and Route Guidance Demonstration Program, in *Advanced Technology in Transport: IVHS and ATT*, ed: I. Catling, Artech House, London.

Ghislandi, A.C., Fijal, A.R., and Christopher, E.J. (1994) CATS 1990 Household Travel Survey: A Methodological Overview, *Working Paper 94-05*, Chicago Area Transportation Study, Chicago Information Services Division

Kim, H., Li, J., Roodman, S., Sen, A., and Sööm, S., (1992) Factoring the household travel surveys, *CATS Interim Report, Task 5*, Urban Transportation Center, University of Illinois at Chicago

McCullagh, P., and Nelder, J.A. (1989) *Generalised Linear Models (2nd Edition),* Chapman Hall, London and New York.

Müller, P. (1991) Metropolis based posterior integration schemes, *Journal of the American Statistical Association,* (to appear)

Press, W.H., Teukolsky, S.A. Vettering, W.H. and Flannery, B.P. (1992) *Numerical Recipes in Fortran (2nd Edition),* Cambridge, Cambridge University Press.

Sen, A. (1986) Maximum likelihood estimation of gravity model parameters, *Journal of Regional Science,* 26, 461-474.

Smith, T.E. (1987) Poisson gravity models of spatial flows, *Journal of Regional Science,* 27, 315-340.

Smith, A.F.M., and Roberts, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society, (Ser. B),* 55, 3-23.

West, M. (1985) Generalised linear models: outlier accommodation, scale parameters and prior distributions, in *Bayesian Statistics 2,* eds: J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, North-Holland, Amsterdam, and Valencia University Press.

West,M. (1995) Inference in successive sampling discovery models, *Journal of Econometrics,* (to appear)

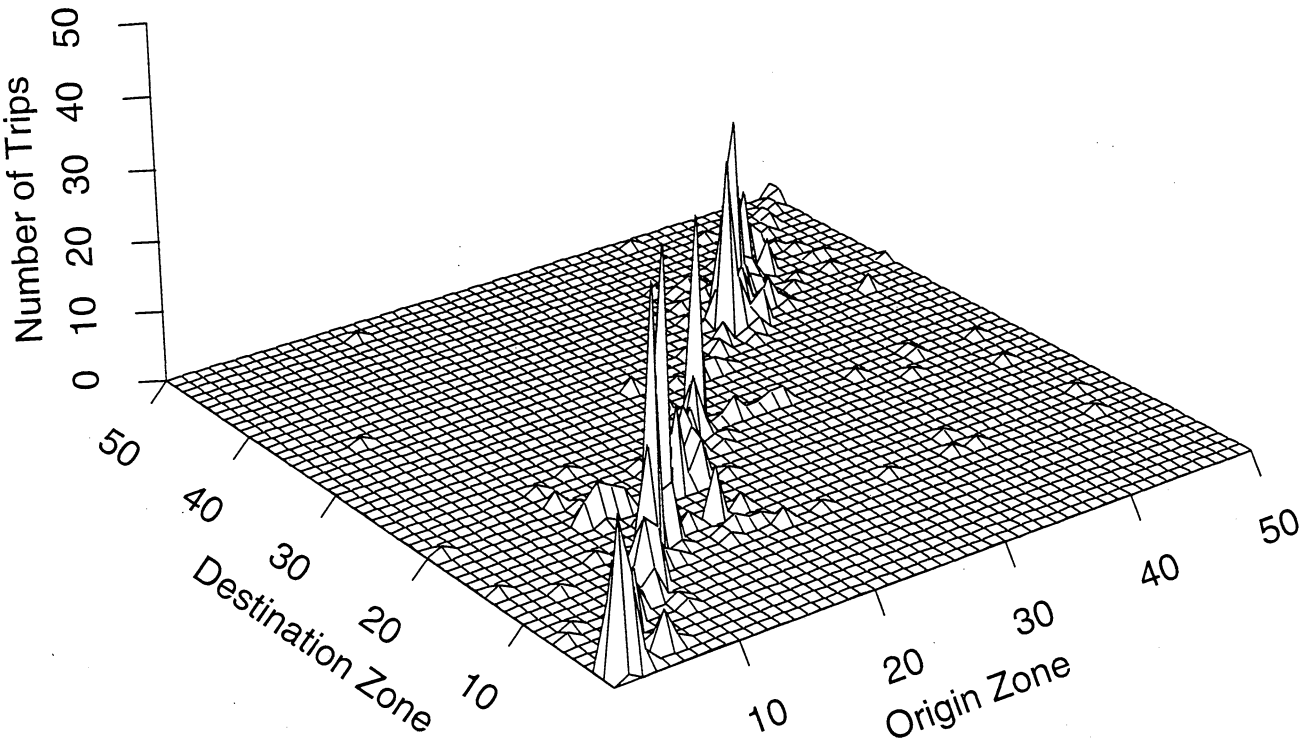

Figure 1. Observed trips for the selected 50 zones from CATS 1990 survey.

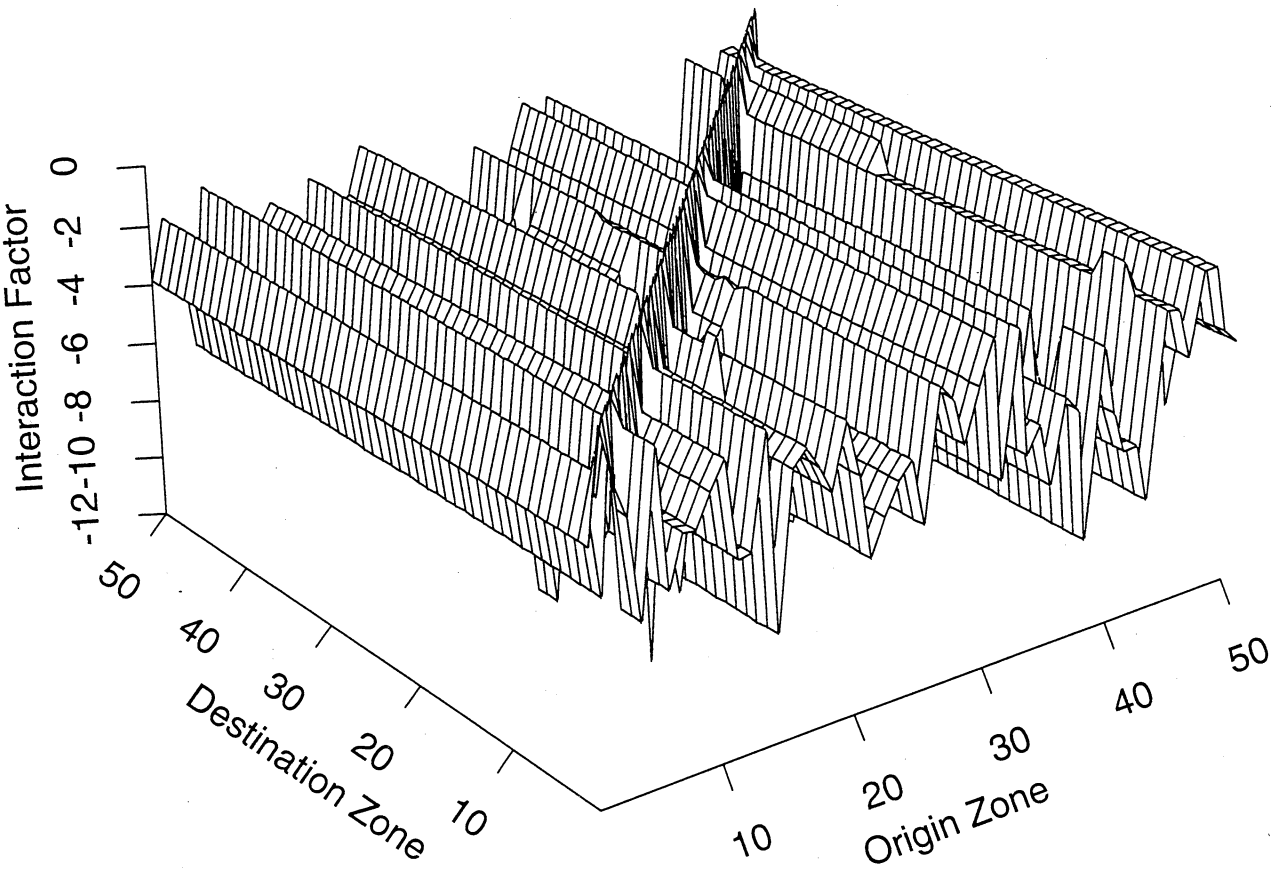

Figure 2. Negative impedances $-x_{ij}$ over all zone pairs for the 50 selected zones.

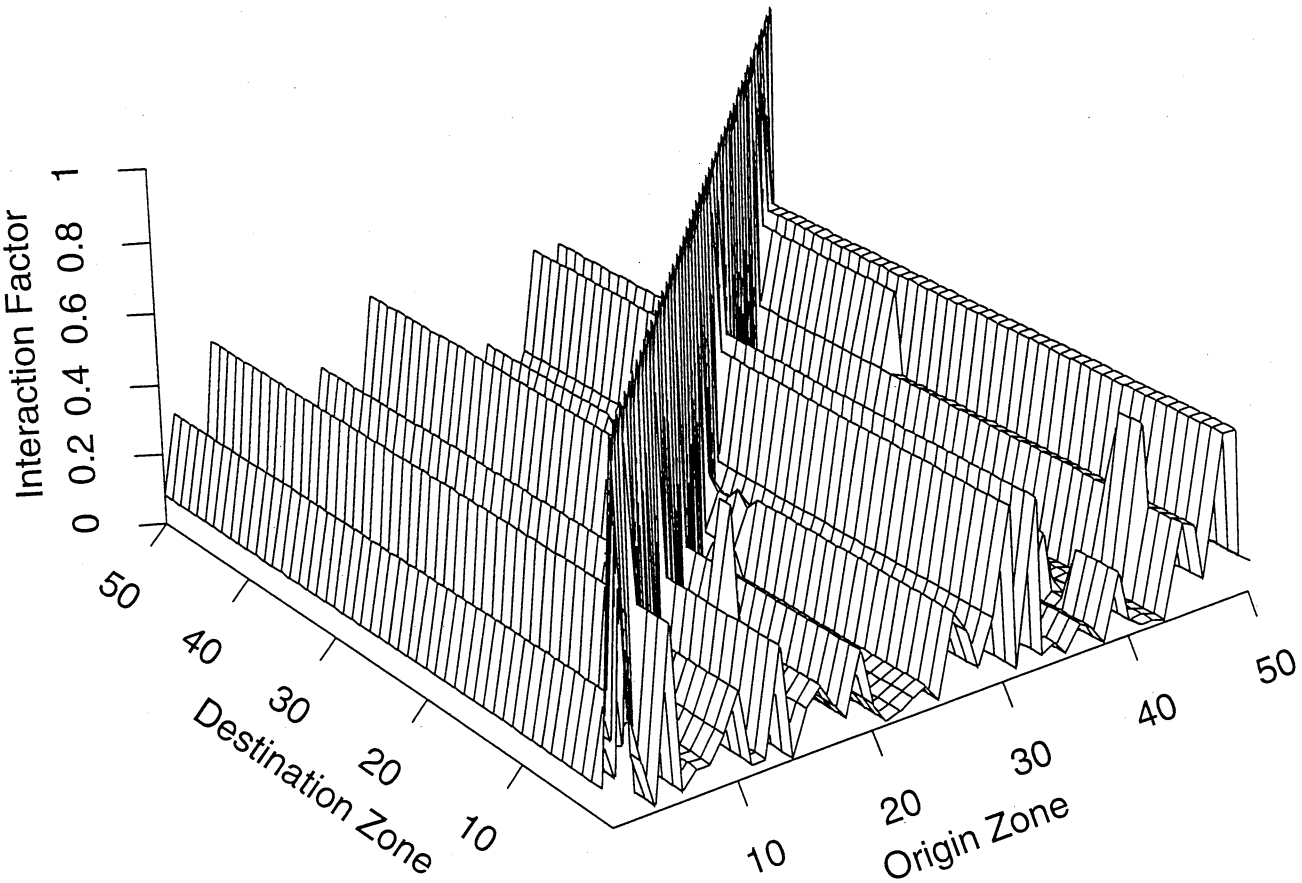

Figure 3. Estimated exponential impedances $\exp(\hat{g}x_{ij})$ over all zone pairs for the 50 selected zones.

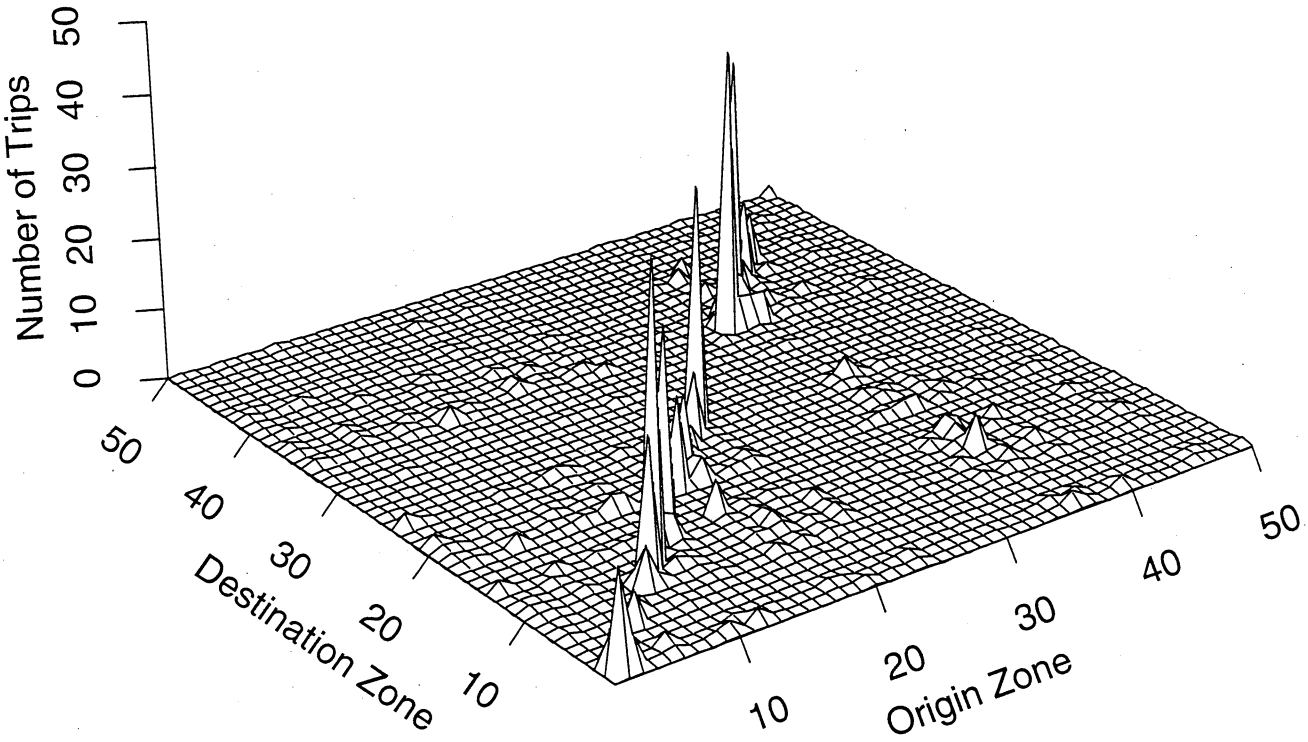

Figure 4. A sample from the posterior distribution of expected number of trips for the selected 50 zones from CATS 1990 survey.

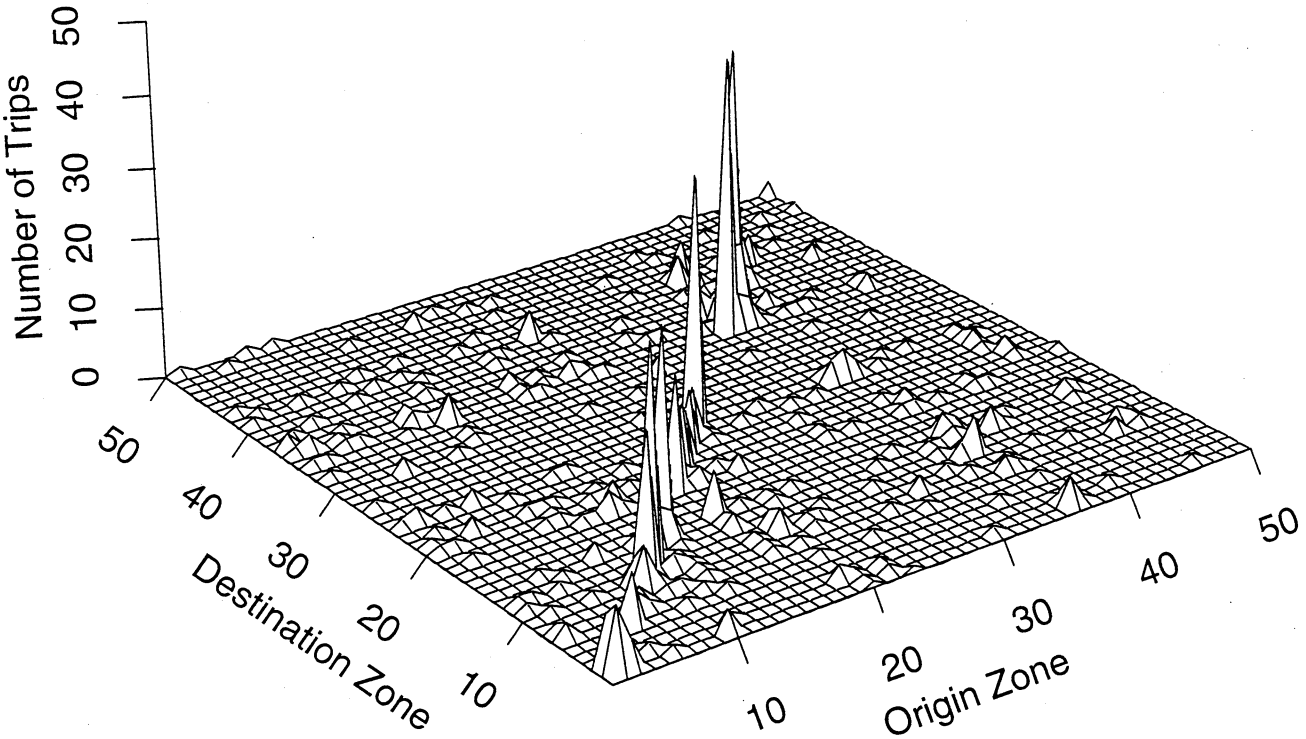

Figure 5. A sample from the posterior predictive distribution of actual trips for the selected 50 zones from CATS 1990 survey.

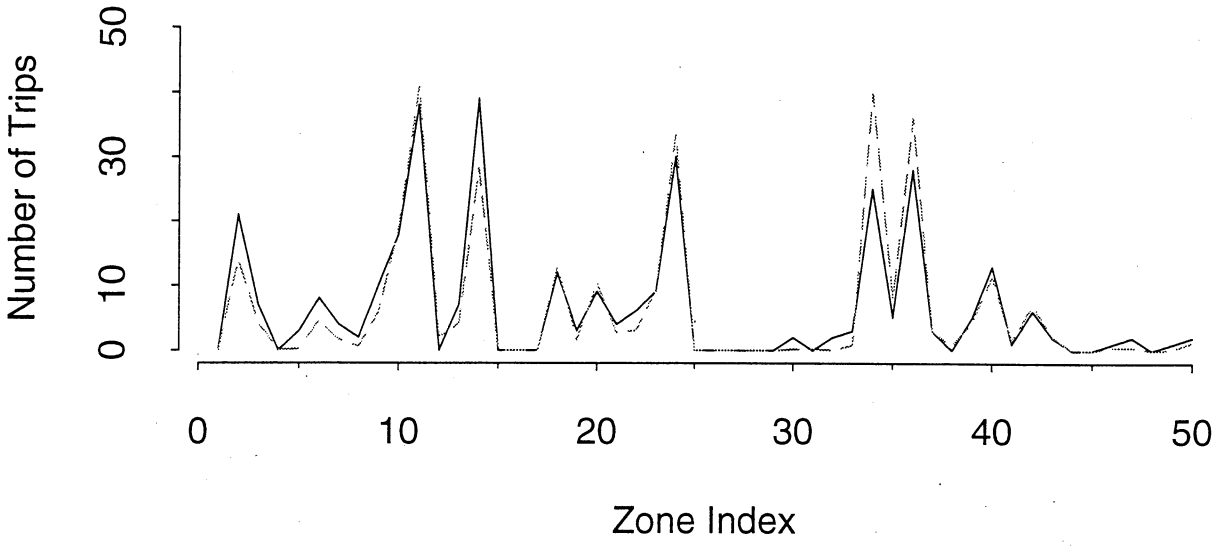

Figure 6. Actual (in full line) and expected (in broken line) within-zone flows over 50 zone region.

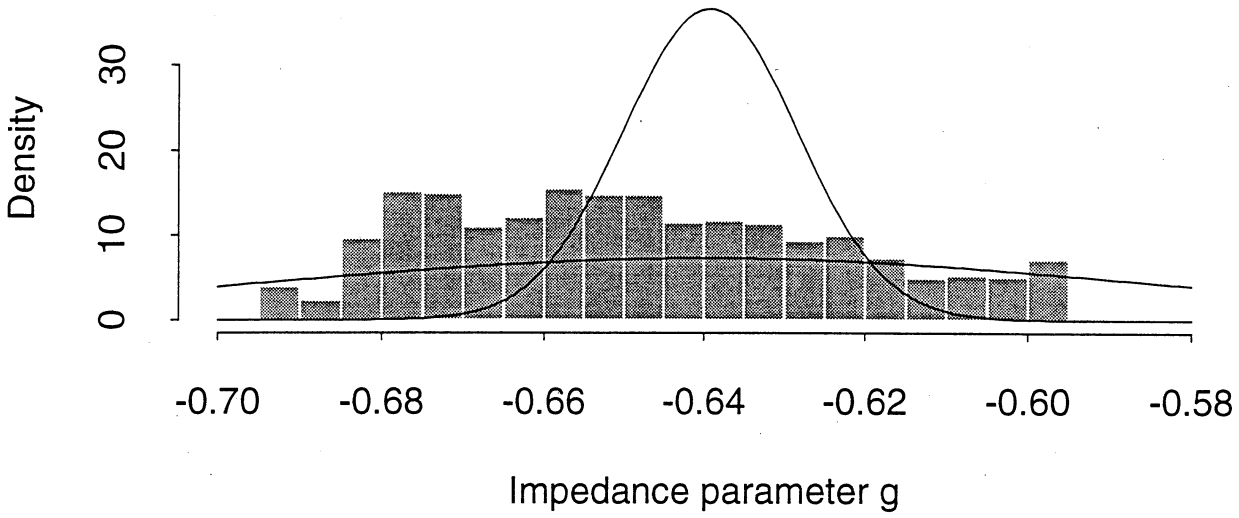

Figure 7. Histogram representing posterior density for the impedance parameter g in analysis of 50 zone data set. The more peaked density curve superimposed in the asymptotic normal approximation, and the more diffuse density curve that with a standard deviation inflated by a factor of 5 and used in the Metropolis sampling steps of analysis.

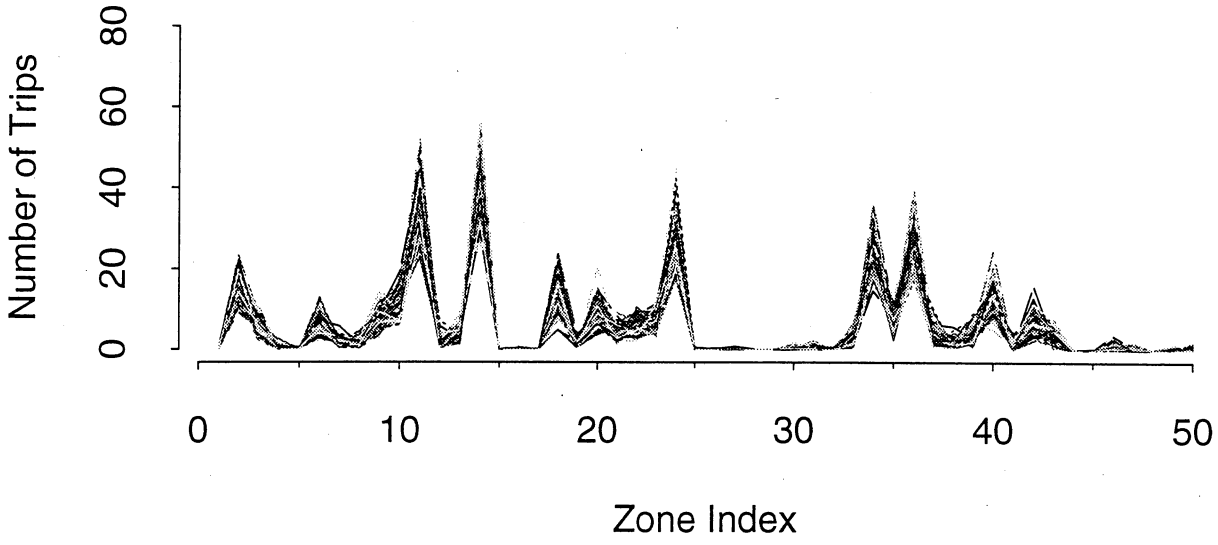

Figure 8. 100 posterior samples for the expected within-zone flows over 50 zone region.

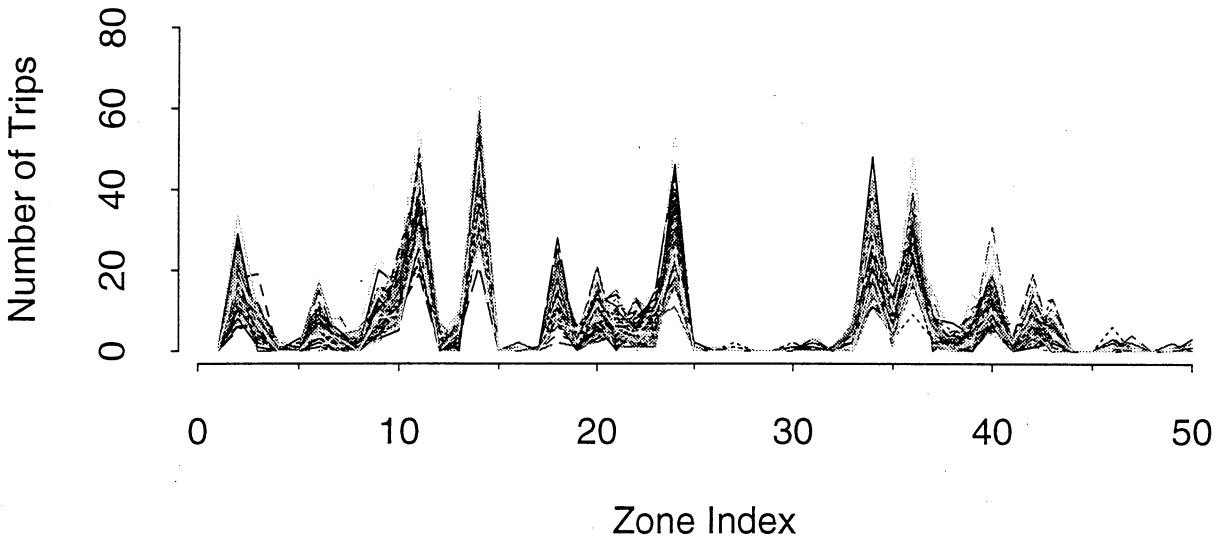

Figure 9. 100 posterior predictive samples for actual within-zone flows over 50 zone region.

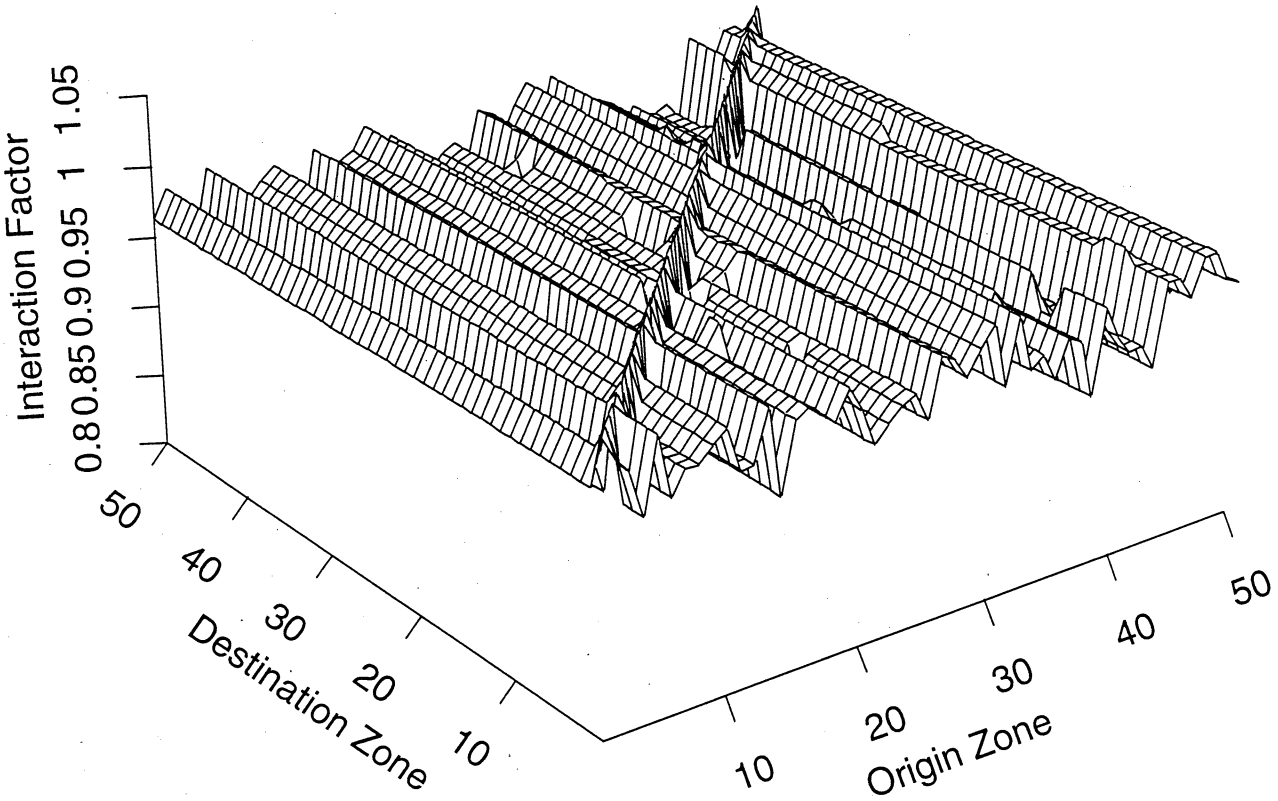

Figure 10. Estimated exponential impedances $\exp(\hat{g}x_{ij})$ over all zone pairs for the 50 selected zones but now based on the full network analysis using data across all $n = 783$ zones.

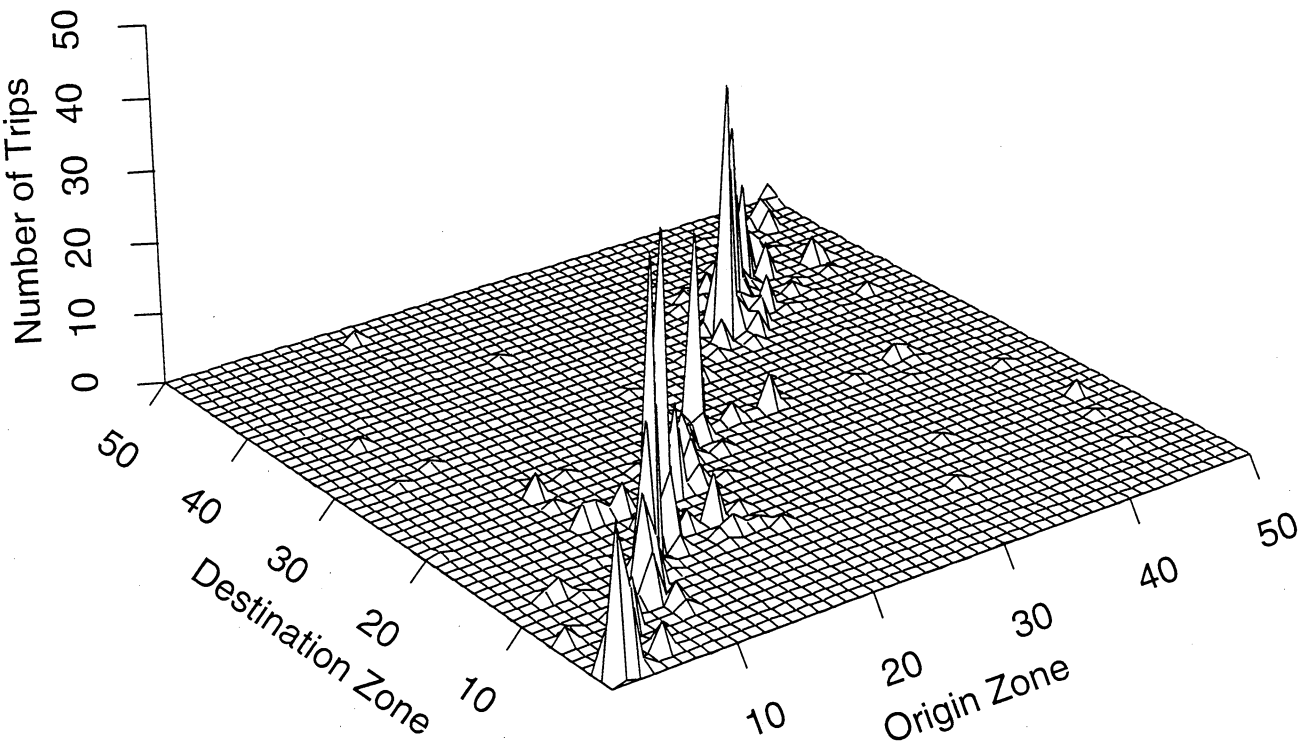

Figure 11. A sample from the posterior predictive distribution of actual trips for the selected 50 zones, but now based on the full network analysis using data across all $n = 783$ zones.

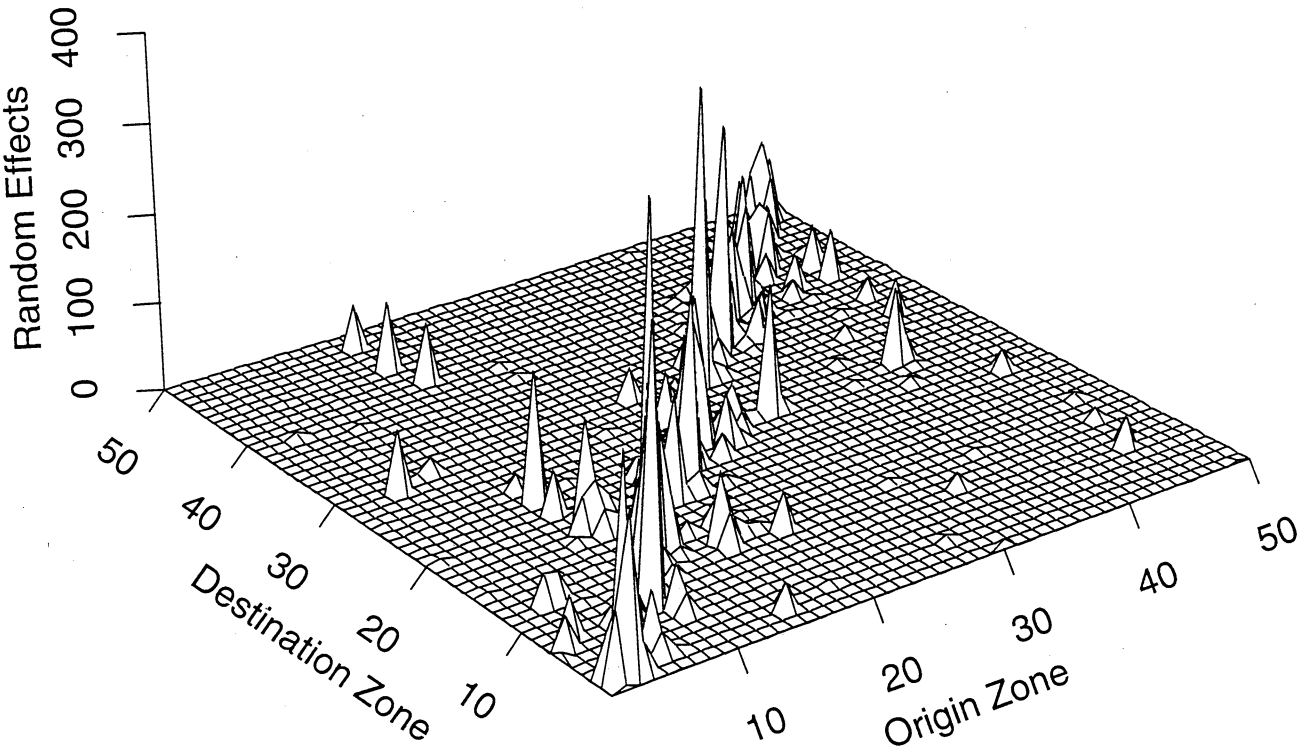

Figure 12. A sample from the posterior distribution of the random effects/interaction parameters h_{ij} for the selected 50 zones, based on the full network analysis using data across all $n = 783$ zones. In this analysis, the prior gamma distribution for these quantities is so diffuse as to permit a high degree of adaptation of the posteriors to the data, as is evident particularly along the diagonal of this figure where zone flows are higher and the baseline gravity model inadequate.