Assessing Uncertainty in Urban Simulations Using Bayesian Melding

Hana Ševčíková[a], Adrian E. Raftery[b], Paul A. Waddell[c*]

Abstract: We develop a method for assessing uncertainty about quantities of interest using urban simulation models. The method is called Bayesian melding, and extends a previous method developed for macrolevel deterministic simulation models to agent-based stochastic models. It encodes all the available information about model inputs and outputs in terms of prior probability distributions and likelihoods, and uses Bayes's theorem to obtain the resulting posterior distribution of any quantity of interest that is a function of model inputs and/or outputs. It is Monte Carlo based, and quite easy to implement. We applied it to the projection of future household numbers by traffic activity zone in Eugene-Springfield, Oregon, using the UrbanSim model developed at the University of Washington. We compared it with a simpler method that uses repeated runs of the model with fixed estimated inputs. We found that the simple repeated runs method gave distributions of quantities of interest that were too narrow, while Bayesian melding gave well calibrated uncertainty statements.

Keywords: Bayesian melding, urban simulation, uncertainty analysis, stochastic models

1 Introduction

Patterns of land use and available transportation systems play a critical role in determining the economic vitality, liability, and sustainability of urban areas. Decisions

[a]Center for Urban Simulation and Policy Analysis, University of Washington, Box 353055, Seattle, WA 98115, USA; hanas@u.washington.edu

[b]Department of Statistics, University of Washington, Box 354322, Seattle, WA 98115, USA; raftery@u.washington.edu

[c]Daniel J. Evans School of Public Affairs, University of Washington, Box 353055, Seattle, WA 98115, USA; pwaddell@u.washington.edu

*Corresponding Author. Tel.: +1 206 221 4161; Fax: +1 206 616 6625

about such infrastructure investments as a new light rail system or freeway expansion, or major policy changes such as establishing an urban growth boundary, have consequences for decades to come. Furthermore, their development is typically controversial, not only because of the massive public expenditures involved, but also because the benefits and costs of the projects are contested. Not only do different stakeholders differ on what values are important, but there is also generally concern that costs or benefits are overstated, or risks of bad outcomes underestimated. In a recent study of transportation projects with a combined cost of $59 billion, in 9 out of 10 transit projects, transit ridership was overestimated by an average of 106%, and half of all roadway projects studied had errors in predicted usage of more than 20% (Flyvbjerg et al. 2005). These errors can be attributed to many causes, from cynical interpretations of models to achieve political aims of leveraging federal investment in desired projects, to uncertainty in model assumptions and errors in model specifications. For example, predictive simulation models usually provide point results and do not take into account stochasticity or uncertainty about input parameters. It is then almost impossible for stakeholders to make the right judgment about the confidence of such results.

The need for expressing uncertainty in simulation models is widely recognized (see for example Refsgaard and Henriksen 2004, van Asselt and Rotmans 2002). In fact, there has been a great deal of work done on this subject in high risk areas, for example in hydrology (Beven and Binley 1992a, Beven 2000, Christensen 2003, Neuman 2003, Korving et al. 2003) and whale management (Raftery et al. 1995, Poole and Raftery 2000).

Over the past few years, we have been involved in the development of UrbanSim (Waddell 2002, Waddell et al. 2003), a sophisticated simulation system to model urban development. Its goal is to help inform public deliberation and decision-making on major land use and transportation issues. However, in the present form, UrbanSim — like many other simulation systems — essentially provides only point predictions.

In this paper, we develop formal methods for assessing uncertainty for land use and transportation policy. This is an underdeveloped area, where little research has been done. We will apply and adapt previous work from other policy-related disciplines where possible. However, for several reasons that will be discussed later in the paper, the

previous work is not directly applicable to the present problem. One relevant prior study has examined the propagation of uncertainty in the context of UrbanSim (Krishnamurthy and Kockelman 2002). However, this study used only sensitivity analysis to a small sample of selected input values and explored the effect of these changes on outputs, in addition to simple stochastic simulation error from variation in random seeds. This work did not generate any statistical inferences, nor any new methodology to calibrate a stochastic model system with respect to uncertainty, which are the principal aims of this paper.

The paper is organized as follows. In Section 2 we briefly describe UrbanSim with its model components and review sources of uncertainty that can be found in simulation models. In Section 3 we develop a formal method for assessing uncertainty. In Section 4 the method is applied to the UrbanSim system. Section 5 presents our results. Finally, in Section 6 we discuss limitations of and possible improvements to our approach.

2 UrbanSim

2.1 Description

UrbanSim is an urban simulation model operational in several urban areas in the United States (Waddell 2002, Waddell et al. 2003)). The system is implemented as a set of interacting models that represent the major actors and choices in the urban system, including households moving to a residential location, business choices of employment location and developer choices of locations and types of real estate development. It takes an extremely disaggregated approach by modeling individual households, jobs, and real estate development and location choices using grid cells of 150×150 meters in size. The modeling system microsimulates the annual evolution in locations of individual households and jobs as the results of choices by families and employers, and the evolution of the real estate within each individual grid cell as the results of actions by real estate developers.

In this section, we give a brief description of the main components of the system. Our emphasis is less on the details of the algorithms of the individual models, but rather on the information on which models base their procedures. For details about the algorithms see Waddell et al. (2003).

The core of UrbanSim (version 3.0) consists of nine models that run in a specified order once per simulated year. In Figure 1, the models are sorted from top to bottom according to their run order. Models that are placed on the same horizontal level are independent of each other.

[Figure 1 about here.]

A simulation is usually performed for a certain time period given in years, determined by the planning horizon in question, typically a thirty-year time frame for metropolitan transportation planning. It is assumed that at the beginning of the simulation, data reflect the true state in the starting year. We will call such data "base year" data. Each model uses a certain partition of the data as an input and modifies a certain partition of the data according to the results it produces. These relations are represented in Figure 1 by dashed arrows. Most of the models expect additional input parameters (see below) that are marked by solid arrows.

The simulation starts by running the **Accessibility model**. This model creates accessibility indices that summarize the accessibility from a given geographical unit to various activities. The indices are then used in Employment and Household location models where accessibilities are expected to influence household or business choice of location. The Accessibility model is loosely coupled with a Travel model which predicts congested travel times and costs (not included in the figure), treated here as an external model.

The **Economic transition model** simulates job creation and loss. It uses aggregate employment forecasts (control totals) that are obtained from external sources (state economic forecasts, commercial or in-house sources). The **Demographic transition model** simulates births and deaths in the population of households in a similar way. Here again, control totals from aggregated forecasts obtained from external sources are used as an additional input.

The **Employment mobility model** determines which jobs will move from their current locations during the simulated year. It uses annual mobility rates directly observed over a recent period. They are computed from longitudinally linked business

establishment files. Similarly, the **Household mobility model** simulates households deciding whether to move. The algorithm uses annual mobility rates estimated from the Census Current Population Survey.

The **Employment location choice model** is responsible for determining a location for each job that was either newly created or determined for moving. The **Household location choice model** chooses a location for each household that was either created or belongs to the mover set. Both models are based on a multinomial logit model calibrated to observed data. Thus, logit coefficients are required by the models as additional input parameters. These coefficients are usually estimated by external estimation procedures.

The **Land price model** simulates land prices of each grid cell as the characteristics of locations change over time. It is based on a hedonic regression (Waddell et al. 1993) using coefficients estimated externally from historical data.

The one year simulation is concluded by the **Real estate development model** that simulates developer choices about what kind of construction to undertake and where. It uses the multinomial logit model and here again, its coefficients are estimated externally using observed data.

2.2 Sources of uncertainty

In order to carry out a probabilistic analysis of a system, the first step is to identify the possible significant sources of uncertainty in the system (see e.g. Morgan and Henrion 1990, Dubus et al. 2003, Regan et al. 2003). The goal of the analysis is then to quantify as much of the identified uncertainty as possible. We now review the main potential sources of uncertainty for UrbanSim.

Measurement errors: As mentioned above, data that enter UrbanSim at the beginning of any simulation run (base year data) should reflect the state of the system, usually a metropolitan area, being modeled. The data are collected from different sources, such as census, tax assessor records or commercial sources. Such data are subject to errors and often contain missing values. For example, in parcel data, tax exempt

properties like government-owned properties tend to have very incomplete data in the assessor files, since there is no compelling reason for assessors to collect these items. These data are massive and rely on individual property assessors to input data, and the resulting databases are often riddled with missing data and data errors.

Systematic errors: Systematic errors are errors that bias the simulation results in a particular direction. This can arise, for example, from miscalibrated measurement tools or from sampling procedures that are not completely random. An example of the latter would be a situation where households belonging to a certain category, such as a certain race, are excluded or undersampled in the sampling process when creating the household database from census sources. Both the measurement errors and the systematic errors affect the quality of the base year data.

Uncertainty about model structure: Three sources of uncertainty about model structure can be distinguished. The first is the selection of variables in the statistical models embedded in UrbanSim, such as the multinomial logit model or the hedonic regression. The second is the choice of the statistical model itself. For example, processes modeled by the multinomial logit could be modeled instead by another discrete choice model, such as multinomial probit or mixed logit (Train 2003). The third source of uncertainty is the selection of the processes that are modeled by UrbanSim. Like any model, UrbanSim does not represent the complete set of processes that influence the evolution of households and job locations. The model is designed to represent the most important processes and ignore those that have little impact on outcomes, but there can be uncertainty about this choice.

Uncertainty about model input parameters: As described in Section 2.1, there are several sets of input parameters that enter UrbanSim and all of them are estimated by external models or procedures. Since these estimates are not exact, they contribute to overall uncertainty.

Stochasticity: An important source of uncertainty arises by using random numbers within models. In UrbanSim, simulations with different seeds of the random number generator produce different results, and this has to be taken into account in the uncertainty analysis. In UrbanSim, several models that are based on the multinomial logit model use a sampling procedure for sampling choice alternatives. Also, both mobility models use sampling for determining movers according to given rates.

For each model, Table 1 summarizes the information about the main component of the procedure, the estimation of the input parameters, the total number of input parameters, and whether or not random numbers are used by the model.

[Table 1 about here.]

3 Bayesian Melding Method

3.1 Notation

We denote the collection of model inputs about which there is uncertainty by Θ. Model inputs are quantities that we must specify for the model to run and about which we have some information. They can include model parameters and starting values of the system at the start time. The collection Θ is a subset of the set of model inputs, and does not include inputs that we take to be known or fixed.

We denote the collection of model outputs about which we have observed information by Φ. This will typically be a subset of all the outputs produced by the model, and can include values or summaries of the state of the simulated system at various times during the simulated time period. For now, we consider the situation where the model is deterministic, so that any two runs of the model with the same inputs yield the same outputs. In this case, we denote the mapping function that produces Φ by M_Φ, so that $\Phi = M_\Phi(\Theta)$. We will extend our analysis later to nondeterministic models.

We denote the quantities of policy interest by Ψ. These can be functions of model inputs, of model outputs, or of both, so that

$$\Psi = M_\Psi(\Theta, \Phi) = M_\Psi(\Theta, M_\Phi(\Theta)). \tag{1}$$

It follows from (1) that Ψ can be viewed as a function of the inputs alone. Often the quantities of policy interest will be values of the system at a future time.

Observed data that provide information about the outputs are often available, such as measurements of outputs at time points within the simulated time period. We denote the collection of such data by y.

3.2 Bayesian melding

Bayesian melding was proposed by Raftery et al. (1995) and Poole and Raftery (2000) as a way of putting the analysis of simulation models on a solid statistical basis. The basic idea is to combine all the available evidence about model inputs and model outputs in a coherent Bayesian way, to yield a Bayesian posterior distribution of the quantities of interest, Ψ. The method was developed initially for deterministic simulation models.

The first step is to encode the available information about model inputs and outputs in terms of probability distributions. We represent our information about the inputs, Θ, by a prior probability distribution, $q(\Theta)$. We specify a conditional probability distribution of the data y given the outputs Φ, and this yields a likelihood for the outputs

$$L(\Phi) = \text{Prob}(y|\Phi). \tag{2}$$

Because $\Phi = M_\Phi(\Theta)$, (2) yields a likelihood for the inputs also, since

$$L(\Theta) = \text{Prob}(y|M_\Phi(\Theta)). \tag{3}$$

We thus have a prior, $q(\Theta)$, and a likelihood, $L(\Theta)$, both defined in terms of the inputs. It follows from Bayes's theorem that we have a posterior distribution of the inputs given all the available information, namely

$$\pi(\Theta) \propto q(\Theta)L(\Theta). \tag{4}$$

In words, the posterior density is proportional to the prior density times the likelihood. The constant of proportionality is defined so that $\pi(\Theta)$ is a probability density, i.e. so that it integrates to 1. The collection, Ψ, of quantities of policy interest can be expressed as a function of the inputs by (1), and so, in principle, (4) yields a full posterior distribution,

$\pi(\Psi)$, of the quantities of interest. This combines all the available relevant information, and so provides a comprehensive basis for risk assessment and decision-making.

3.3 Simulating the posterior distribution

How can we evaluate the posterior distribution (4), and the resulting posterior distribution of the quantities of interest, $\pi(\Psi)$? The posterior density does not have an analytic form because it involves the mapping function $M_\Phi(\Theta)$, which is typically available only through computer evaluation of the model. Here instead, we use a Monte Carlo method introduced by Raftery et al. (1995) and Poole and Raftery (2000), based on the Sampling Importance Resampling (SIR) algorithm of Rubin (1987). It works as follows:

1. Draw a sample $\{\Theta_1, \ldots, \Theta_I\}$ of values of the inputs from the prior distribution $q(\Theta)$.

2. Obtain $\{\Phi, \ldots, \Phi_I\}$ where $\Phi_i = M_\Phi(\Theta_i)$.

3. Compute weights $w_i = L(\Phi_i)$. As a result, we get an approximate posterior distribution of inputs with values $\{\Theta_1, \ldots, \Theta_I\}$ and probabilities proportional to $\{w_1, \ldots, w_I\}$.

4. The approximate posterior distribution of the quantities of interest has values $\{\Psi_1, \ldots, \Psi_I\}$ where $\Psi_i = M_\Psi(\Theta_i, \Phi_i)$ and probabilities proportional to $\{w_1, \ldots, w_I\}$.

The method is illustrated in Figure 2.

[Figure 2 about here.]

The models that we are dealing here with are not deterministic. We consider here models that contain a stochastic component in the sense that they use random numbers and that runs with different seeds return different results. To include this source of uncertainty into the framework, we modify the above procedure as follows:

1. As before, draw a sample $\{\Theta_1, \ldots, \Theta_I\}$ of values of the inputs from the prior distribution $q(\Theta)$.

2. For each Θ_i, run the model J times with different seeds to obtain $\Phi_{ij}, j = 1, \ldots, J$.

3. Compute weights $w_i = L(\bar{\Phi}_i)$ where $\bar{\Phi}_i = \frac{1}{J}\sum_{j=1}^{J} \Phi_{ij}$. Here again, we get an approximate posterior distribution of inputs with values $\{\Theta_1, \ldots, \Theta_I\}$ and probabilities proportional to $\{w_1, \ldots, w_I\}$.

4. The approximate posterior distribution of Ψ now has $I \times J$ values $\Psi_{ij} = M_\Psi(\Theta_i, \Phi_{ij})$, with weights $w_{ij} = w_i/J$.

4 Application to UrbanSim

4.1 Data

We now illustrate Bayesian melding by applying it to a test case, using data from Eugene-Springfield, Oregon. The model is run starting in 1980, for which detailed information on the city's starting state are available. The goal is to use UrbanSim to predict the number of households in the year 2000 in each of the city's $K = 295$ Traffic Activity Zones, or in aggregations of zones. We will refer hereafter to a Traffic Activity Zone simply as a "zone." We have data for the year 2000, but here we use them only for assessing the method and verifying the predictions, not for making them.

In addition we have data on the number of households in each zone in 1994, denoted by $y = (y_1, \ldots, y_K)$. These data are used to form the likelihood. So the exercise is one of making predictions in 1994 for 2000, using detailed data from 1980 and less complete data from 1994. Thus the times used in the simulation are: "start": 1980; "present": 1994; "future": 2000, using the terminology of Figure 2.

4.2 Likelihood and posterior distribution

In order to compute the weights in Step 3 of Section 3.3, we need to define a likelihood function, as given by equation (3). Specifically,

$$w_i \propto p(y|\Theta_i) = \prod_{k=1}^{K} p(y_k|\Theta_i)\,. \tag{5}$$

The likelihood $p(y|\Theta_i)$ is based on the following model:

$$\begin{aligned} \Phi_{ijk} &= \mu_{ik} + \delta_{ijk}, \text{ where } \delta_{ijk} \overset{iid}{\sim} N(0, \sigma_\delta^2), \text{ and} \quad (6) \\ (y_k|\Theta = \Theta_i) &= \mu_{ik} + a + \epsilon_{ik}, \text{ where } \epsilon_{ik} \overset{iid}{\sim} N(0, \sigma_i^2). \quad (7) \end{aligned}$$

Here, μ_{ik} denotes the expected output corresponding to the i-th simulated input Θ_i and the k-th zone, δ_{ijk} and ϵ_{ik} denote model errors, and a denotes the overall bias in the model.

Estimation of μ_{ik}, σ_δ^2, σ_i^2, and a can be done by approximate maximum likelihood (see Appendix A). We denote these estimates by $\hat{\mu}_{ik}$, $\hat{\sigma}_\delta$, $\hat{\sigma}_\epsilon^2$ and $\hat{a}$.

This yields a predictive distribution of our quantity of interest:

$$y_k|\Theta_i \sim N(\hat{a} + \hat{\mu}_{ik}, v_i) \quad \text{with } v_i = \hat{\sigma}_i^2 + \frac{\hat{\sigma}_\delta^2}{J}\,. \quad (8)$$

(see Appendix B for details).

We then have

$$\begin{aligned} w_i &\propto p(y|\Theta_i) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi v_i}} \exp\left[-\frac{1/2(y_k - \hat{a} - \hat{\mu}_{ik})^2}{v_i}\right], \quad \text{and} \quad (9) \\ \log w_i &\propto -\frac{K}{2}\log(2\pi v_i) - \frac{1}{2v_i}\sum_{k=1}^{K}(y_k - \hat{a} - \hat{\mu}_{ik})^2\,. \quad (10) \end{aligned}$$

Given that $\hat{\sigma}_\delta$, $\hat{\sigma}_\epsilon^2$ and $\hat{a}$ were estimated at the "present" time $t_1 = 1994$, the marginal distribution of the quantity of interest, Ψ_k, the number of households in the k-th zone in the year $t_2 = 2000$, is given by a mixture of normal distributions, as follows:

$$\begin{aligned} p(\Psi_k) &= \sum_{i=1}^{I} w_i N(\hat{a}b_a + m_{ik}, (\hat{\sigma}_i^2 + \frac{\hat{\sigma}_\delta^2}{J})b_v)\,, \quad k = 1, \ldots, K, \quad \text{where} \quad (11) \\ m_{ik} &= \frac{1}{J}\sum_{j=1}^{J} \Psi_{ijk}\,. \end{aligned}$$

Here, b_a and b_v denote propagation factors of the bias and the variance over the time period $[t_1, t_2]$.

4.3 Prior on inputs

Input parameters for several of the UrbanSim model components were estimated by multinomial logistic regression or by hedonic regression; see Table 1. These have known standard errors (SE) and covariance matrices (C). For these parameter sets we used the multivariate normal distribution $\mathrm{MVN}(\hat{\Theta}, \mathrm{C}((\hat{\Theta}))$, where $\hat{\Theta}$ is the estimator of Θ. The off-diagonal elements of the covariance matrices were very small, so we used only diagonal matrices, $\mathrm{diag}\Big(\mathrm{SE}(\hat{\Theta})^2\Big)$.

For each mobility rate r in the employment and household mobility models we used the normal distribution $\mathrm{N}(\hat{r}, (\frac{\hat{r}(1-\hat{r})}{n})^2)$, truncated at zero, where $\hat{r}$ is an estimate of the rate r and n is the number of observations from which $\hat{r}$ was obtained.

Prior distributions of the control totals for the economic and demographic transition model are based on forecasts for the end year of the simulation made long before. In order to estimate the standard error of these totals, we obtained several pairs of prediction and observation values from the historical data (see Table 2). From the four triples (prediction, observation, time difference) denoted by $(\hat{c}_i, c_i, d_i)$ where $i = 1, \ldots, 4$, we calculated variance per year as

$$V(\hat{c}) = \frac{1}{4} \sum_{i=1}^{4} \frac{\left[\log\left(\frac{\hat{c}_i}{c_i}\right)\right]^2}{d_i}.$$

[Table 2 about here.]

Using the data in Table 2, we obtained $V(\hat{c}) = 0.000647$. Our prior distribution for the control total is then the normal distribution $\mathrm{N}(\hat{c}, 20V(\hat{c}))$, where the variance is multiplied by the 20 prediction years, 1980–2000. In this example, we used the actual value of c for 2000 as the mean $\hat{c}$ of the prior distribution, purely for illustrative purposes. In practice, of course, this would not be available for forecasting purposes, and one would instead use a current forecast. In our case, the prior distribution was quite spread out, and so the results would not be very sensitive to precisely what prior mean is used. This approach to assessing uncertainty about forecasts is similar to an approach used in demographic population projections called ex post analysis (Keyfitz 1981, Stoto 1983, Smith and Sincich 1988).

5 Results

5.1 Simulation

We ran UrbanSim for 100 different inputs, each of them twice with a different random seed. This gave 200 runs in total, with $I = 100$ and $J = 2$. We confirmed our results by a larger simulation with $I = 1000$ and $J = 3$ (see Section 5.6).

The simulation was started in 1980 and run until 2000. The results Φ_{ijk} and Ψ_{ijk} were saved for all $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, where Φ_{ijk} and Ψ_{ijk}, are the numbers of households in zone k for the j-th run with the i-th set of input values in 1994 and 2000, respectively.

5.2 Transformation

In equations (6) and (7) we assume that our model errors have constant variances σ_δ^2 over all i and k, and σ_1^2 over all k. In Figure 3, we examine the relationship between the estimate of Φ_{ijk}, denoted by $\hat{\mu}_{ik}$, and its standard deviation, $\mathrm{sd}(\Phi_{ijk})$. The lowess curve is also shown; this is a nonparametric smooth estimate of the relationship between the estimate and the standard deviation (Cleveland 1979).

[Figure 3 about here.]

It is clear that $\mathrm{sd}(\Phi_{ijk})$ increases with increasing $\hat{\mu}_{ik}$. The relationship is approximately linear on the log-log scale, as shown in the right panel of Figure 3. A detailed analysis of the slope suggested transforming the data by taking the square roots of the number of households, so as to obtain a more nearly constant variance.

Figure 4 shows the scatter plots from Figure 3 after the square root transformation. The relationship was much weaker and the squared correlation coefficient was now only $R^2 = 0.03$. This suggests that the variance is close to being constant when the data are transformed to the square root scale. We excluded zones from the analysis for which either $\hat{\mu}_{ik}$ or $\mathrm{sd}(\Phi_{ijk})$ was equal to zero; these were typically zones with no residential units.

[Figure 4 about here.]

The relationship between the absolute error $|y_k - \hat{\mu}_{ik}|$ and $\text{sd}(\Phi_{ijk})$ after the transformation is shown in Figure 5. Here the squared correlation coefficient is $R^2 = 0.01$. This provides further evidence that the square root transformation has stabilized the variance. We therefore applied the square root transformation to Φ_{ijk} and y_k for all i, j, k, prior to any computation.

[Figure 5 about here.]

5.3 Computing the weights

From the results Φ_{ijk} we estimated the quantities needed to compute the weights given by equations (9) and (10). They are summarized in Table 3. The table shows the mean and standard deviation of $\hat{\sigma}_i^2$ and v_i computed over all i. It is clear that $\hat{\sigma}_i^2$ is the dominant component in the variance v_i. We excluded zones where $\hat{\mu}_{1\cdot} = 0$ from the analysis. These were mostly zones in which up to 1994 no residential units were placed. This reduced the number of zones, and hence the dimension of the outputs, to $K = 265$.

[Table 3 about here.]

Table 4 shows the resulting seven highest weights and the minimum weight for the 100 simulation runs. Each weight reflects how well the corresponding outputs predicted the values observed in 1994. For example, the simulation for simulated input $i = 64$ had the largest weight among all the simulated inputs, and yielded outputs whose correlation coefficient with the observed data is $R = 0.93$. On the other hand, simulated input $i = 33$ had the minimum weight, and produced outputs whose correlation with the observations was considerably smaller, at $R = 0.84$.

[Table 4 about here.]

5.4 Posterior distribution

In order to compute the posterior distribution of our quantity of interest given by Equation (11), we set the propagation factors b_a and b_v to 20/14, equal to $\frac{2000-1980}{1994-1980}$.

The results for one zone are shown in Figure 6. The left panel shows a histogram of the results from the 200 UrbanSim runs without applying Bayesian melding. The right panel shows the Bayesian melding distribution of the number of households in the zone. It can be seen that the Bayesian melding distribution is wider than the range of the simple multiple runs, and for this zone the observation (bold vertical line) fell outside the 90% interval defined by the multiple runs, but well inside the Bayesian melding 90% interval (marked by dotted vertical lines).

[Figure 6 about here.]

In fact this phenomenon was quite general: the multiple runs interval missed the true observation far more often than the Bayesian melding interval. This can be seen in Table 5 which shows the number of missed cases and the coverage for both procedures. The multiple runs interval included the actual observation less than 40% of the time, compared with its 90% nominal level. In comparison, the Bayesian melding 90% interval contained the observation 88% of the time, which is quite satisfactory. Similar results are obtained from the 1000×3 simulation (see Section 5.6). The table suggests that Bayesian melding is much better calibrated than a distribution provided by simply taking results from multiple Monte Carlo runs.

[Table 5 about here.]

We did a more complete check of the calibration of the methods by calculating their verification rank histograms. The verification rank histogram (Anderson 1996, Talagrand et al. 1997) was introduced by meteorologists to evaluate probabilistic weather forecasts and is a useful tool for assessing predictive distributions. For each zone, we simulated 99 values from the Bayesian melding predictive distribution, and then we computed the rank of the observation within the set of $99 + 1 = 100$ numbers consisting of the 99 simulated values and the observation. If the predictive distribution is well calibrated, the distribution of the resulting ranks should be close to uniform. The histogram of the ranks is called the verification rank histogram, and large deviations of the histogram from uniformity indicate that the predictive distributions are not well calibrated.

Figure 7 shows the verification rank histograms for multiple runs of UrbanSim and for Bayesian melding. It is clear that the verification rank histogram for the multiple runs is very far from being uniform, showing that this method is poorly calibrated. The verification rank histogram for Bayesian melding is much closer to being uniform.

[Figure 7 about here.]

The same information is shown in a different way in Figure 8, which shows the cumulative distribution functions (CDFs) of the ranks shown in Figure 7. If the verification rank histogram is uniform, then the CDF should lie on or close to the 45° line, or line of equality, $y = x$. It is clear that the CDF for the multiple runs is far from this line. The CDF for Bayesian melding is quite close to the line of equality, indicating once again that Bayesian melding is much better calibrated.

[Figure 8 about here.]

5.5 Aggregated results

Often we are interested in a higher level of aggregation than the basic zones for which results come out of the model, such as the zones of Eugene-Springfield. For example, we might be interested in the central business district (CBD). The posterior distribution of a spatially aggregated output, such as the number of households in the CBD, can be found by aggregating the results for the smaller geographical units. In our case study, the posterior distribution of the number of households in a collection of zones is that of a sum of random variables, each of which has a distribution that is a mixture of several truncated normal components. Although this is complicated to find analytically, it is easy to evaluate by simulation.

Suppose we are interested in the number of households that live within a certain radius of the CBD. Figure 9 shows histograms of results for four different radii, measured in time distance, each of them simulated by 1000 values. In each case the observation (bold line) fell close to the mean of the posterior distribution.

[Figure 9 about here.]

5.6 Number of simulations

The results presented were obtained with 200 runs in total (with $I = 100$ and $J = 2$). Since an UrbanSim simulation of large size is not yet feasible in practice due to the long run time, we explored whether results obtained by using more runs are significantly different. To check this, we ran UrbanSim with $I = 1000$ and $J = 3$, i.e. with 15 times as many runs.

Table 6 compares coverage from Table 5 for the two different simulation sizes. It can be seen that there is almost no difference in coverage between the two cases.

[Table 6 about here.]

Figure 10 compares simulated values from the resulting posterior distribution $p(\Psi)$ given by equation (11), for the two simulation sizes. In this experiment, we simulated from each of the two distributions 100 values for each zone and computed their mean value. In Figure 10, means resulting from the smaller simulation are plotted on the y-axis, whereas means coming from the larger simulation are plotted on the x-axis. The fact that almost all the values lie very close to the diagonal suggests that the initial number of replications (200) was adequate to obtain accurate results.

[Figure 10 about here.]

6 Discussion

We have introduced a new method for assessing uncertainty in the output from urban simulation models, called Bayesian melding. Bayesian melding was introduced in another field and applied to deterministic simulation models, but the urban simulation models we are concerned with are stochastic, and so we extended Bayesian melding to deal with stochastic simulation models. Bayesian melding combines all the available information about model inputs and outputs and combines them in a Bayesian way, to provide a posterior distribution of quantities of interest that provides a fully assessment of uncertainty and risk. The method gave well calibrated results in a case study.

Our approach could be improved in various ways. In our case study we were modeling counts of households, and we modeled these using a transformed truncated normal distribution. This worked well, but it may be more appropriate to build a model on the Poisson distribution, which is designed for counts. Also, for simplicity we ignored spatial correlation in the prediction errors, even though it is quite possible that spatial correlation is present. It would be possible to model spatial correlation by replacing the independence assumption in equation (7) by a joint distribution with dependence specified by a geostatistical correlation function. Something like this was done for probabilistic weather forecasting by Gel et al. (2004).

One practical problem with our approach is that for larger cities UrbanSim can take a long time to run, and so our method could be extremely expensive computationally, as it requires several hundred runs of the model. There are various ways in which it may be possible to reduce the time required.

A first approach is simply to use a more efficient importance sampling function than the prior on inputs, $q_1(\theta)$, such as the pre-model distribution of the inputs, proportional to $q_1(\theta)L_1(\theta)$. A second method is to use adaptive sampling importance resampling (Givens and Raftery 1996), in which SIR is run a first time, and then the importance sampling function is replaced by a more efficient one based on the first results, and used in a second application of SIR.

A third possible approach would be to use Latin hypercube sampling(McKay et al. 1979) in place of simple random sampling in the Bayesian melding algorithm. This would be likely to provide a more efficient algorithm. However, this too is unlikely to be a full solution on its own, because the number of model runs is still likely to be substantial. A fourth method is the three-point Gauss-Hermite quadrature method of Raftery and Zeh (Raftery and Zeh 1993). This can be extremely efficient in terms of number of model runs, but requires some initial knowledge of where the best values of the inputs are likely to be. It has been used with success in hydrogeological applications (Levy et al. 1998, Levy and Ludy 2000).

A fifth approach, used for the extremely high-dimensional problems in atmospheric sciences, is based on model ensembles. A fairly small ensemble of runs (typically 5–50)

from different inputs is used; these are not necessarily drawn from a probability distribution. Records of predictions and observations are typically used to postprocess the output to correct biases and produce approximate posterior distributions. One method for postprocessing small ensembles that has worked well for probabilistic weather forecasting is the recent Bayesian model averaging approach of Raftery et al. (2005); this can yield a calibrated posterior distribution.

Finally, another promising idea is the combination of Bayesian melding with the notion of a model emulator (Sacks et al. 1989). A model emulator is a fast approximation to the model, which, when provided with the model inputs, yields approximate values of the model outputs, and takes much less computer time to run than the full model. The most prevalent model emulators work by performing a small number of model runs at strategically chosen input values, such as those resulting from Latin hypercube sampling. The emulator then interpolates between the model runs, with kriging often being used as the basis for the interpolation (Sacks et al. 1989, Currin et al. 1991).

Here is a composite strategy combining several of these ideas to yield a more efficient approach: Suppose, we have a limited "budget" of B runs. Then we can:

1. Use Latin Hypercube Sampling to draw a sample of size $B/3$ from the prior distribution of the inputs. We then run the model for each one of these inputs.

2. Cut down on the dimension of the input space, for each input parameter by testing whether the outputs vary significantly with that input. If this is not the case, we disregard uncertainty due to that input and keep it fixed. Suppose that the number of parameters remaining is N_{param}. N_{param} could be constrained to be no more than some upper limit, which could be a function of B, for example $\sqrt{B/3}$.

3. Find the approximate posterior mode by using $B/3$ iterations of a Nelder-Mead optimization algorithm, starting at a subset of the initial input values sampled by Latin Hypercube Sampling.

4. Use approximate three-point iterated Gauss-Hermite quadrature centered at the approximate posterior mode, to carry out the required integrations. This requires

$O(N^2_{param})$ model runs.

Several other approaches, some also using Monte Carlo methods and Bayesian ideas, have been proposed. Generalized likelihood uncertainty estimation (GLUE) (Beven and Binley 1992b) has been applied to a wide range of hydrology and water quality models; see Beven et al. (2000) for a review. This is similar to Bayesian melding, but the "likelihood" function defined is typically derived from a measure of goodness of fit of model predictions to observations, rather than from a probability model. "Nonbehavioral" realizations are rejected, and the remaining realizations are weighted by the likelihoods; the threshold for this decision is subjectively chosen by the user. As a result, GLUE has no clear Bayesian interpretation.

In Section 2.2, we reviewed sources of uncertainty that enter UrbanSim. Our method accounts for all the listed sources except one: uncertainty in model structure. A combination of our approach with other methods, such as Bayesian model averaging (Hoeting et al. 1999), could close this gap.

Acknowledgments

This research has been funded in part by National Science Foundation Grants EIA-0121326 and IIS-0534094.

Appendix

A Estimation

Let I, J, K denote the sample size of the input parameters (Step 1 of Section 3.3), the number of runs for the same input parameters (Step 2) and the dimension of the output space, respectively. Then, the quantities in Equations 6 and 7 can be estimated as follows:

- Estimation of μ_{ik}:

$$\hat{\mu}_{ik} = \frac{1}{J} \sum_{j=1}^{J} \Phi_{ijk} \tag{12}$$

- Estimation of σ_δ^2:

$$\hat{\sigma}_\delta^2 = \frac{1}{IJK} \sum_{ijk} (\Phi_{ijk} - \hat{\mu}_{ik})^2 \tag{13}$$

- Estimation of a:

$$\hat{a} = \frac{1}{IK} \sum_{i,k} (y_k - \hat{\mu}_{ik}) \,. \tag{14}$$

- Estimation of σ_i^2:

$$\hat{\sigma}_i^2 = \frac{1}{K} \sum_{k} (y_k - \hat{a} - \hat{\mu}_{ik})^2 \,. \tag{15}$$

B Building the posterior distribution

To build $p(y_k|\Theta_i)$, note that

$$(y_k - a - \hat{\mu}_{ik}) = (y_k - a - \mu_{ik}) - (\hat{\mu}_{ik} - \mu_{ik})$$

and so

$$E(y_k - a - \hat{\mu}_{ik}) = 0 \quad \text{and}$$

$$Var(y_k - a - \hat{\mu}_{ik}) = Var(y_k - a - \mu_{ik}) + Var(\hat{\mu}_{ik} - \mu_{ik}) = \sigma_i^2 + \frac{\sigma_\delta^2}{J}.$$

Then we get Equation (8).

References

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate 9*, 1518–1530.

Beven, K. (2000). *Rainfall-runoff modelling: The primer*. John Wiley & Sons, Ltd.

Beven, K. and A. Binley (1992a). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes 6*, 279–298.

Beven, K. and A. Binley (1992b). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes 6*, 279–298.

Beven, K. J., J. Freer, B. Hankin, and K. Schulz (2000). The use of generalized likelihood measures for uncertainty estimation in high order models of environmental systems. In W. J. Fitzgerald et al. (Ed.), *Nonlinear and Nonstationary Signal Processing*, pp. 115–151. Cambridge, U. K.: Cambridge University Press.

Christensen, S. (2003). A synthetic groundwater modelling study of the accuracy of GLUE uncertainty intervals. *Nordic Hydrology 35*(1), 45–59.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc. 74*, 829–836.

Currin, C., T. J. Mitchell, M. Morris, and D. Ylvisaker (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association 86*, 953–963.

Dubus, I. G., C. D. Brown, and S. Beulke (2003). Sources of uncertainty in pesticide fate modelling. *The Science of the Total Environment 317*, 53–72.

Flyvbjerg, B., M. K. S. Holm, and S. L. Buhl (2005). How (in)accurate are demand forecasts in public works projects? the case of transportation. *Journal of the American Planning Association 71*(2), 131–146.

Gel, Y., A. E. Raftery, and T. Gneiting (2004). Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method (with discussion). *Journal of the American Statistical Association 99*, 575–590.

Givens, G. H. and A. E. Raftery (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *Journal of the American Statistical Association 91*, 132–141.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science 14*, 382–417. [A corrected

version is available at
www.stat.washington.edu/www/research/online/hoeting1999.pdf.].

Keyfitz, N. (1981). The limits of population forecasting. *Population and Development Review 7*, 579–593.

Korving, H., J. M. van Noortwijk, P. H. A. J. M. van Gelder, and R. S. Parkhi (2003). Coping with uncertainty in sewer system rehabilitation. In Bedford and van Gelder (Eds.), *Safety and Reliability*. Swets & Zeitlinger, Lisse, ISBM 90 5809 551 7.

Krishnamurthy, S. and K. Kockelman (2002). Uncertainty propagation in an integrated land use-transport modeling framework: Output variation via urbansim. *Transportation Research Record* (1805), 128–135.

Levy, J., M. K. Clayton, and G. Chesters (1998). Using an approximation of the three-point Gauss-Hermite quadrature formula for model prediction and quantification of uncertainty. *Hydrogeology Journal 6*, 457–468.

Levy, J. and E. E. Ludy (2000). Uncertainty quantification for delineation of wellhead protection areas using the Gauss-Hermite quadrature approach. *Ground Water 38*, 63–75.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics 21*, 239–245.

Morgan, M. G. and M. Henrion (1990). *Uncertainty : a guide to dealing with uncertainty in quantitative risk and policy analysis*. New York: Cambridge University Press.

Neuman, S. P. (2003). Maximum likelihood bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment 17*, 291–305.

Poole, D. and A. E. Raftery (2000). Inference for deterministic simulation models: the Bayesian melding approach. *Journal of the American Statistical Association 95*(452), 1244–1255.

Raftery, A. E., G. H. Givens, and J. E. Zeh (1995). Inference from a deterministic population dynamics model for bowhead whales. *Journal of the American Statistical Association 90*(430), 402–416.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review 133*, 1155–1174.

Raftery, A. E. and J. E. Zeh (1993). Estimation of bowhead whale, *balaena mysticetus*, population size (with discussion). In C. Gatsonis et al. (Ed.), *Case Studies in Bayesian Statistics*, Number 83 in Lecture Notes in Statistics, pp. 163–240. Springer-Verlag.

Refsgaard, J. C. and H. J. Henriksen (2004). Modelling guidelines – terminology and guiding principles. *Advances in Water Resources 27*, 71–82.

Regan, H. M., H. R. Akcakaya, S. Ferson, K. V. Root, S. Carroll, and L. R. Ginzburg (2003). Treatments of uncertainty and variability in ecological risk assessment of single-species populations. *Human and Ecological Risk Assessment 9*(4), 889–906.

Rubin, D. B. (1987). Comment. *Journal of the American Statistical Association 82*, 543–546.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989). Design and analysis of computer experiments (with discussion). *Statistical Science 4*, 409–435.

Smith, S. K. and T. Sincich (1988). Stability over time in the distribution of population forecast errors. *Demography 25*, 461–474.

Stoto, M. (1983). The accuracy of population forecasts. *Journal of the American Statistical Association 78*, 13–20.

Talagrand, O., R. Vautard, and B. Strauss (1997). Evaluation of probabilistic prediction systems. In *Proceedings, ECMWF Workshop on Predictability*, pp. 1–25. ECMWF. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, U.K.].

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press.

van Asselt, M. B. A. and J. Rotmans (2002). Uncertainty in integrated assessment modelling – from positivism to pluralism. *Climatic Change 54*(1-2), 75–105.

Waddell, P. A. (2002). UrbanSim: Modeling urban development for land use, transportation and environmental planning. *Journal of the American Planning Association 68*(3), 297–314.

Waddell, P. A., B. J. L. Berry, and I. Hoch (1993). Residential property values in multinodal urban area: New evidence on the implicit price of location. *Journal of Real Estate Finance and Economics 7*, 117–141.

Waddell, P. A., A. Borning, M. Noth, N. Freier, M. Becke, and G. Ulfarsson (2003). Microsimulation of urban development and location choices: Design and implementation of UrbanSim. *Networks and Spatial Economics 3*(1), 43–67.

List of Tables

Table 1: Procedure type, type of estimation of input parameters, number of input parameters, and whether or not random numbers (RNs) are used, for each model in UrbanSim

Model	Procedure	Estim. of input par.	#par.	RNs
Accessibility	deterministic	–	–	no
Economic/Demographic transition	algorithm based on joint prob. distr. of jobs/households	external sources	17	yes
Employment/Household mobility	random sampling	observed rates	79	yes
Employment/Household location choice	multinomial logit	multinomial logit	332	yes
Land price	hedonic regression	hedonic regression	49	no
Development	multinomial logit	multinomial logit	615	yes

Table 2: Observed data in 2000 and predictions for 2000 made in 1978 and 1988 for the Eugene-Springfield area.

	Observed data in 2000	Prediction from 1978	Prediction from 1988
Population	322 977	379 500	331 600
Employment	158 271	173 400	NA
Wage & salaried employment	143 900	NA	129 200

Table 3: Estimates required for computing the weights. Estimates for $\hat{\sigma}^2_{(\cdot)}$ and $v_{(\cdot)}$, respectively, denote the mean of $\hat{\sigma}^2_i$ and v_i, respectively, across all i.

	$\hat{a}$	$\hat{\sigma}^2_\delta$	$\hat{\sigma}^2_{(\cdot)}$	$v_{(\cdot)}$
Estimate	-0.156	0.072	15.035	15.071
SD across i	–	–	2.323	2.323

Table 4: The seven highest weights and the smallest weight.

i	64	13	12	76	68	88	23	33(min)
w_i	0.8058	0.0883	0.0370	0.0217	0.0122	0.0116	0.0068	$4 \cdot 10^{-45}$

Table 5: Coverage of for the 90% confidence interval. Missed cases give the number of observations that fall outside of the confidence interval. The total number of observations is 265.

method	missed cases	coverage
Bayesian melding	31	0.88
multiple runs	163	0.38

Table 6: Comparison of coverage from Table 5 in simulations with different number of runs.

simulation size	100×2		1000×3	
method	missed cases	coverage	missed cases	coverage
Bayesian melding	31	0.883	29	0.890
multiple runs	163	0.385	165	0.375

List of Figures

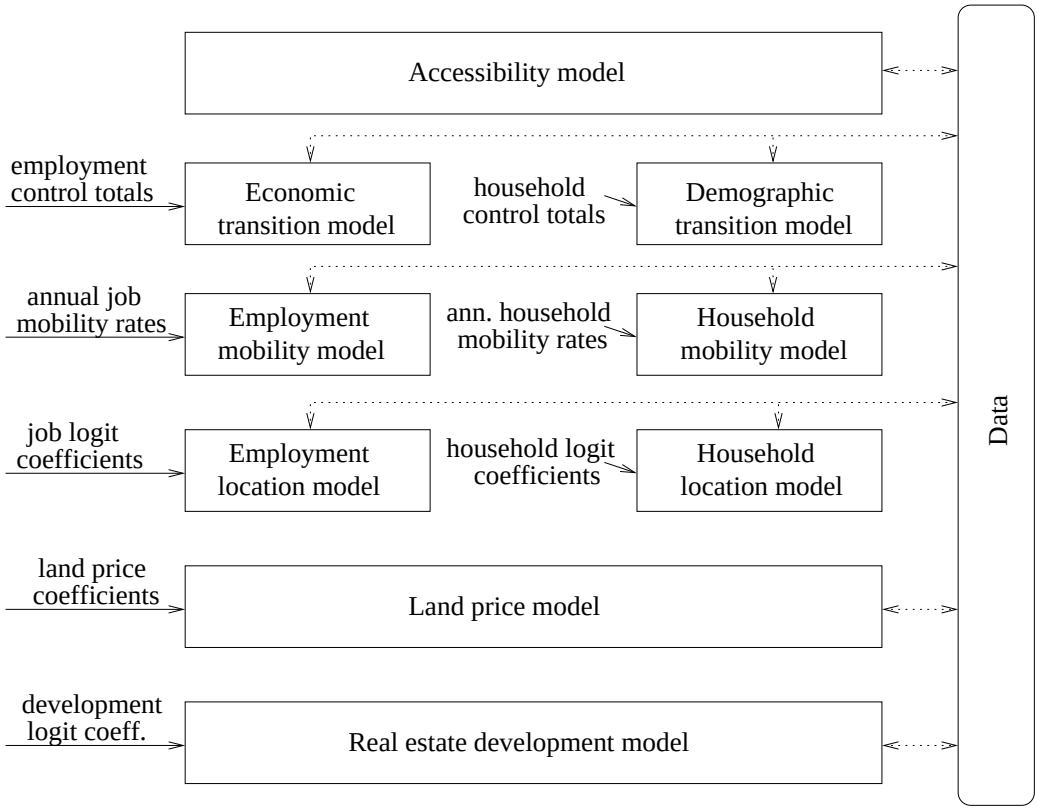

Figure 1: UrbanSim models in order of their runs (from top to bottom). The solid arrows show external input parameters. The dashed arrows mark the data flow: data are input to the different models, and these in turn modify the data.

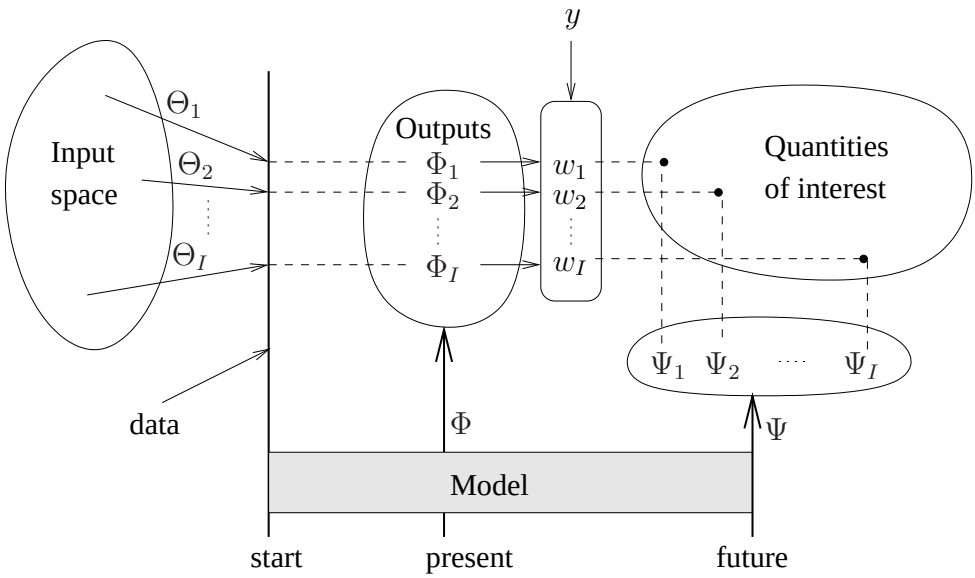

Figure 2: Illustration of the Bayesian melding method. The uncertain model inputs, Θ are assumed to refer to the starting time of the simulation, and the outputs, Φ and the data relevant to the outputs, y, are observed at the "present" time, while the quantities of interest, Ψ, refer to the future. The quantities Θ_i, Φ_i and Ψ_i refer to the i-th simulated values of inputs, outputs and quantities of interest respectively.

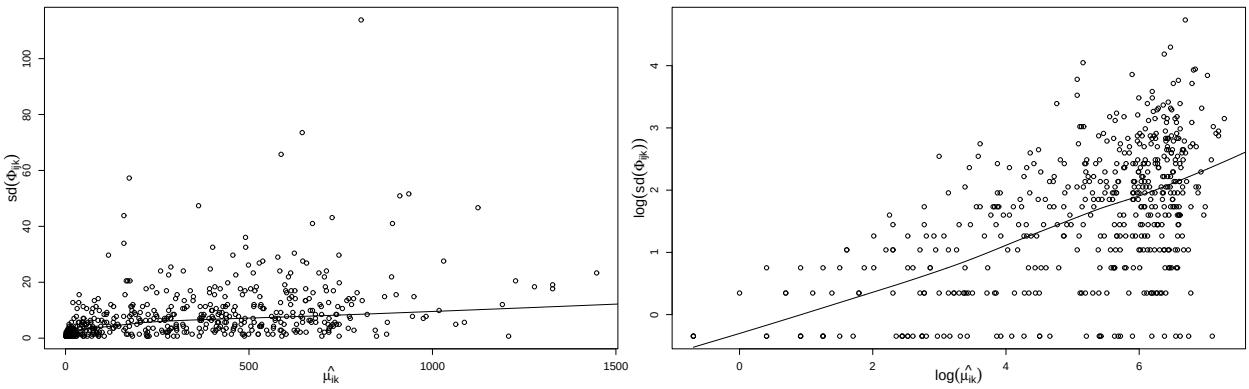

Figure 3: Scatter plot of $\hat{\mu}_{ik}$ vs. standard deviation of Φ_{ijk} and the corresponding lowess curve (raw data in the left panel, data on the log-log scale in the right panel). The plots contain 500 randomly selected points from $I \cdot K = 29\,500$ points in total (excluding points where either $\hat{\mu}_{ik}$ or $\mathrm{sd}(\Phi_{ijk})$ is equal zero). The lowess curve is based on all 29 500 points.

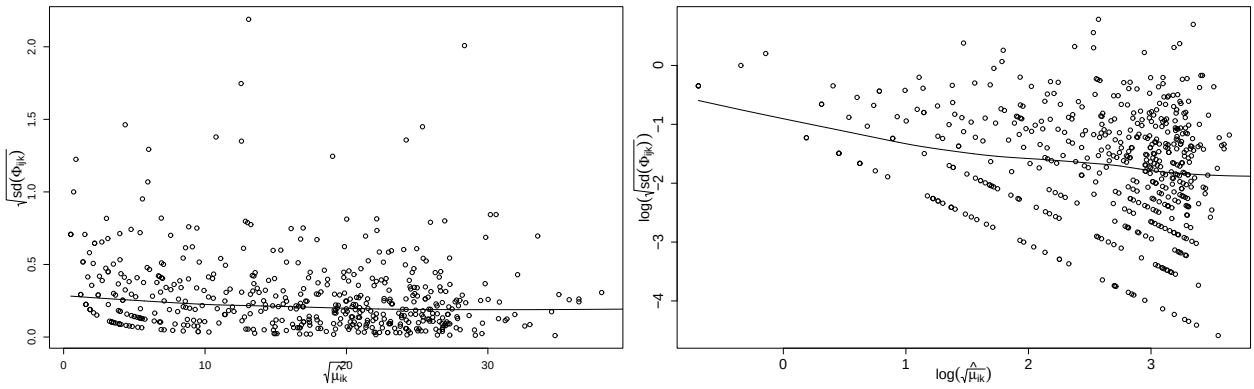

Figure 4: Data from Figure 3 on the square root scale.

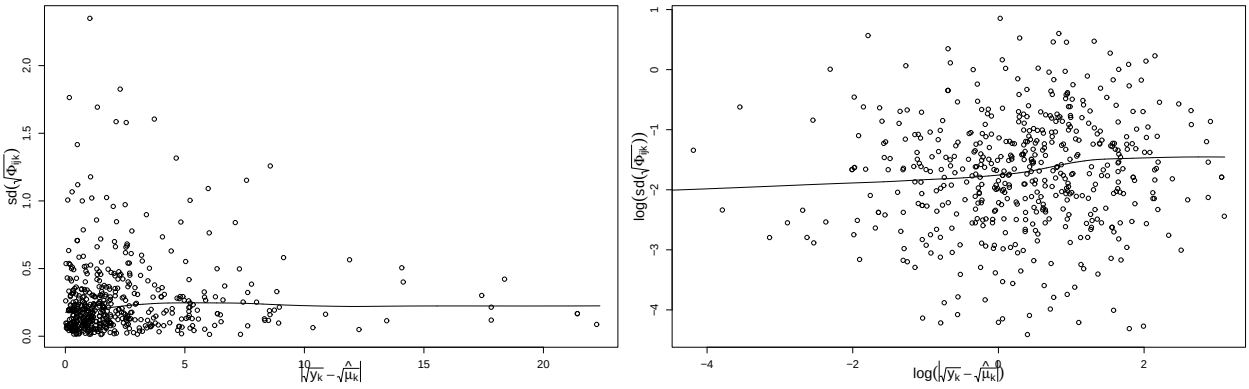

Figure 5: Absolute error of the transformed data, $|\sqrt{y_k} - \sqrt{\hat{\mu}_{ik}}|$, vs. standard deviation of $\sqrt{\Phi_{ijk}}$ for $t = 1994$ and the corresponding lowess curve (raw data in the left panel, data on the log-log scale in the right panel). As in Figure 3, the plots contain 500 randomly selected points, whereas the lowess curve is based on all points.

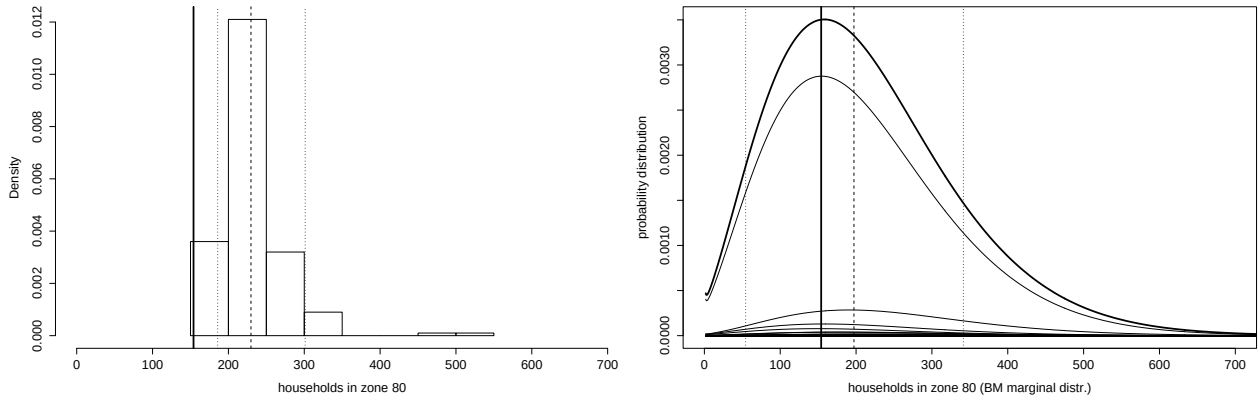

Figure 6: Results of zone 80. Right panel – histogram of multiple runs. Left panel – Bayesian melding probability distribution. In each plot, the vertical solid line marks the true observation (= 154), the vertical dashed line marks the mean and the vertical dotted lines mark the 90% confidence interval.

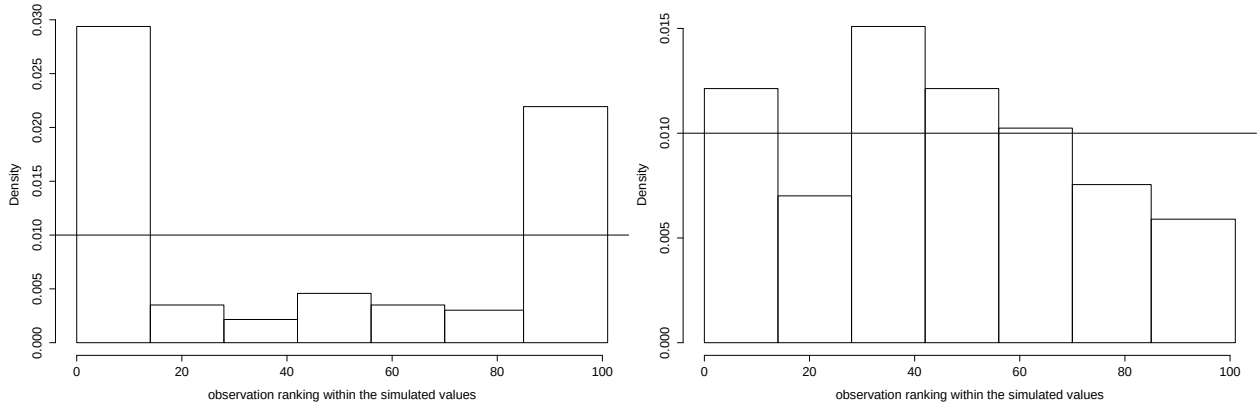

Figure 7: Verification rank histogram for the output from multiple runs (left panel) and from the Bayesian melding procedure (right panel). The closer the histogram is to being uniform, the better calibrated the corresponding method is.

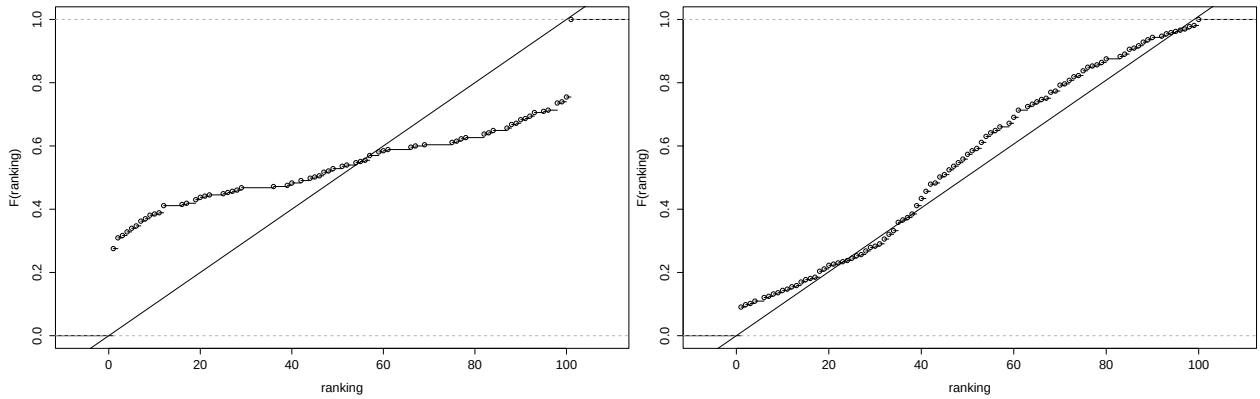

Figure 8: CDF for the output from multiple runs (left panel) and from the Bayesian melding procedure (right panel). The closer the CDF is to the diagonal line of equality shown on the plot, the better calibrated the corresponding method is.

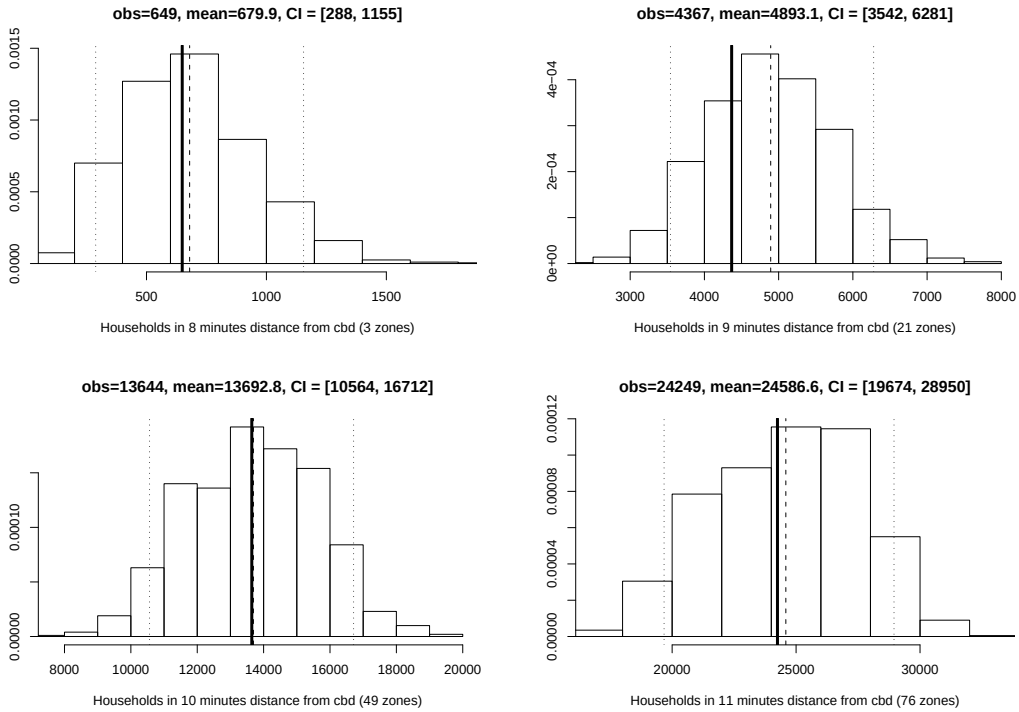

Figure 9: Histogram of the simulated posterior distributions for four aggregated quantities of interest: the numbers of households living within 8, 9, 10, and 11 minutes of the central business district. The true observations are marked by bold vertical lines.

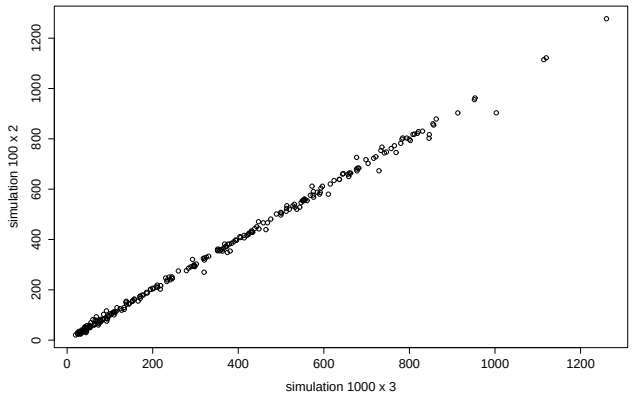

Figure 10: For each zone, mean value of 100 values simulated from the posterior distribution that was derived for simulation size 1000×3 (x-axis) and 100×2 (y-axis), respectively.