Bayesian Melding in Urban Simulations

Hana Ševčíková
Center for Urban Simulation and Policy Analysis
University of Washington, Seattle
UTIA Dept. of Pattern Recognition, Praha
www.stat.washington.edu/hana, www.urbansim.org

Joint work with Adrian Raftery and Paul Waddell

2nd International Workshop DAR, Třešť 2006

Outline

- Motivation

Outline

- Motivation
- Description of the computer model UrbanSim

Outline

- Motivation
- Description of the computer model UrbanSim
- Methodology of applying Bayesian melding to UrbanSim

Outline

- Motivation
- Description of the computer model UrbanSim
- Methodology of applying Bayesian melding to UrbanSim
- Results

Outline

- Motivation
- Description of the computer model UrbanSim
- Methodology of applying Bayesian melding to UrbanSim
- Results
- Summary

Motivation

- Target: Simulation models for land use, transportation and environmental planning

Motivation

- Target: Simulation models for land use, transportation and environmental planning
- Such models are subject to uncertainty

Motivation

- Target: Simulation models for land use, transportation and environmental planning
- Such models are subject to uncertainty
  - measurement errors in input data,

Motivation

- Target: Simulation models for land use, transportation and environmental planning
- Such models are subject to uncertainty
  - measurement errors in input data, systematic errors,

Motivation

- Target: Simulation models for land use, transportation and environmental planning
- Such models are subject to uncertainty
  - measurement errors in input data, systematic errors, model structure,

Motivation

- Target: Simulation models for land use, transportation and environmental planning
- Such models are subject to uncertainty
  - measurement errors in input data, systematic errors, model structure, input parameters,

Motivation

- Target: Simulation models for land use, transportation and environmental planning
- Such models are subject to uncertainty
  - measurement errors in input data, systematic errors, model structure, input parameters, stochasticity

Motivation

- Target: Simulation models for land use, transportation and environmental planning
- Such models are subject to uncertainty
  - measurement errors in input data, systematic errors, model structure, input parameters, stochasticity
- They usually provide point predictions.

Motivation

- Target: Simulation models for land use, transportation and environmental planning
- Such models are subject to uncertainty
  - measurement errors in input data, systematic errors, model structure, input parameters, stochasticity
- They usually provide point predictions.
- Need for expressing uncertainty about model output quantities of policy interest.

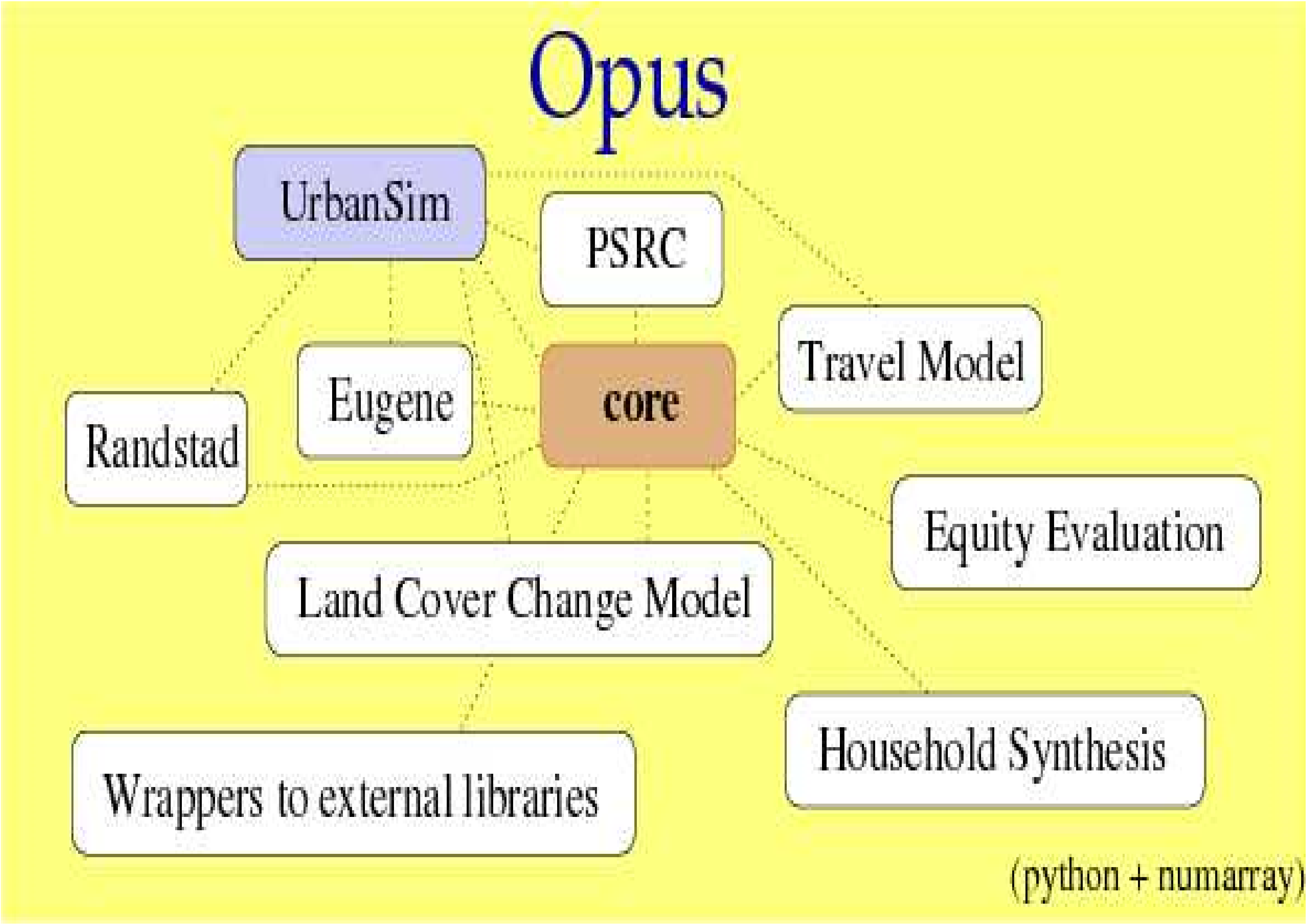
Motivation

- Target: Simulation models for land use, transportation and environmental planning
- Such models are subject to uncertainty
  - measurement errors in input data, systematic errors, model structure, input parameters, stochasticity
- They usually provide point predictions.
- Need for expressing uncertainty about model output quantities of policy interest.
- Multiple runs not satisfactory.

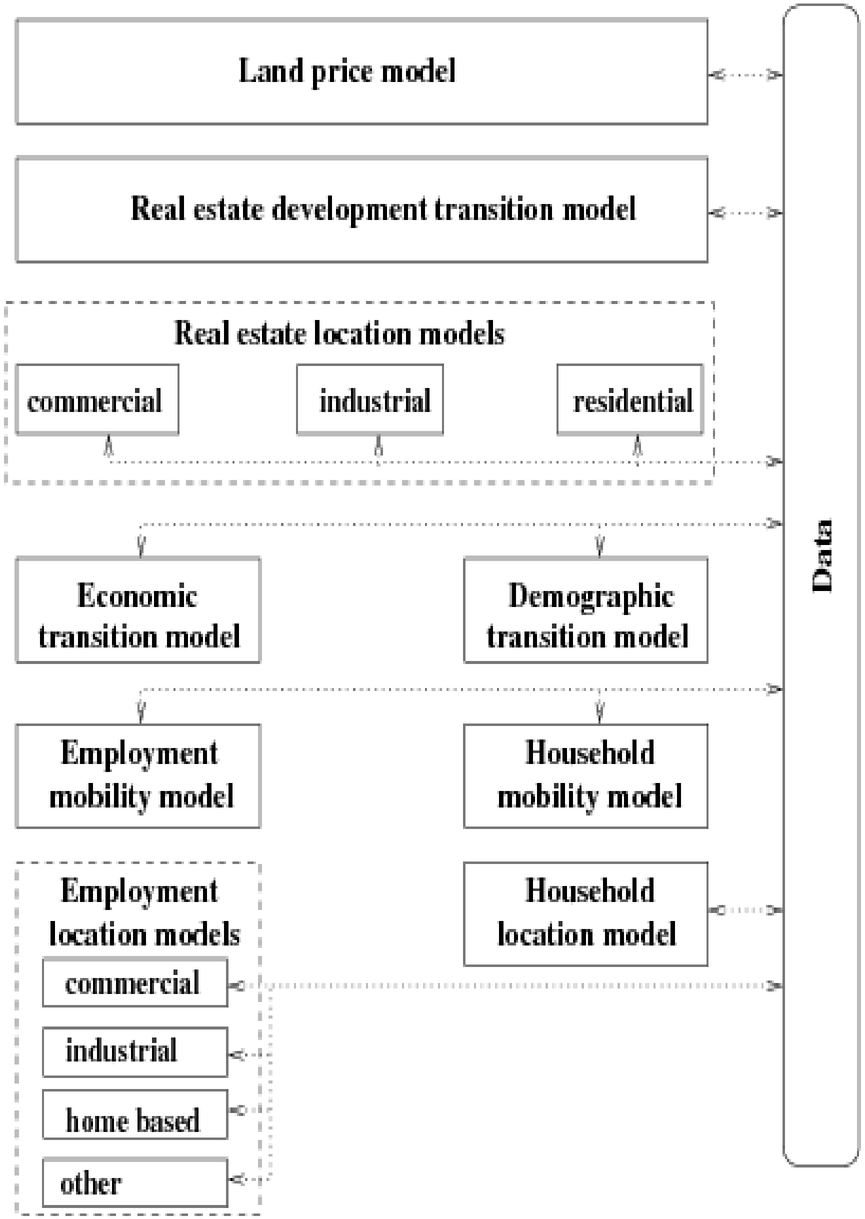
Motivation

- Target: Simulation models for land use, transportation and environmental planning
- Such models are subject to uncertainty
  - measurement errors in input data, systematic errors, model structure, input parameters, stochasticity
- They usually provide point predictions.
- Need for expressing uncertainty about model output quantities of policy interest.
- Multiple runs not satisfactory.
- Method: Bayesian Melding

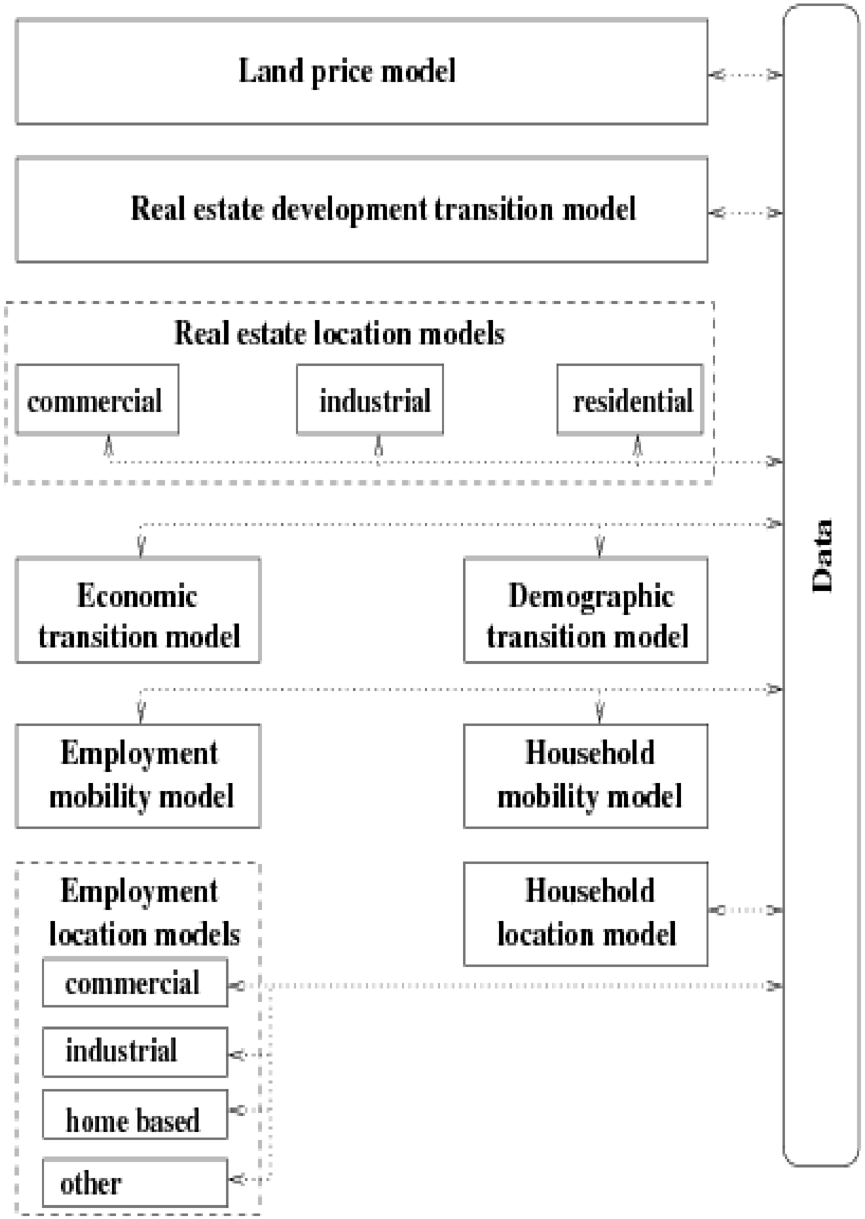
Open Platform for Urban Simulation



UrbanSim


Land price model
Real estate development transition model
Real estate location models
commercial
industrial
residential
Economic transition model
Demographic transition model
Employment mobility model
Household mobility model
Employment location models
commercial
industrial
home based
other
Household location model
Data

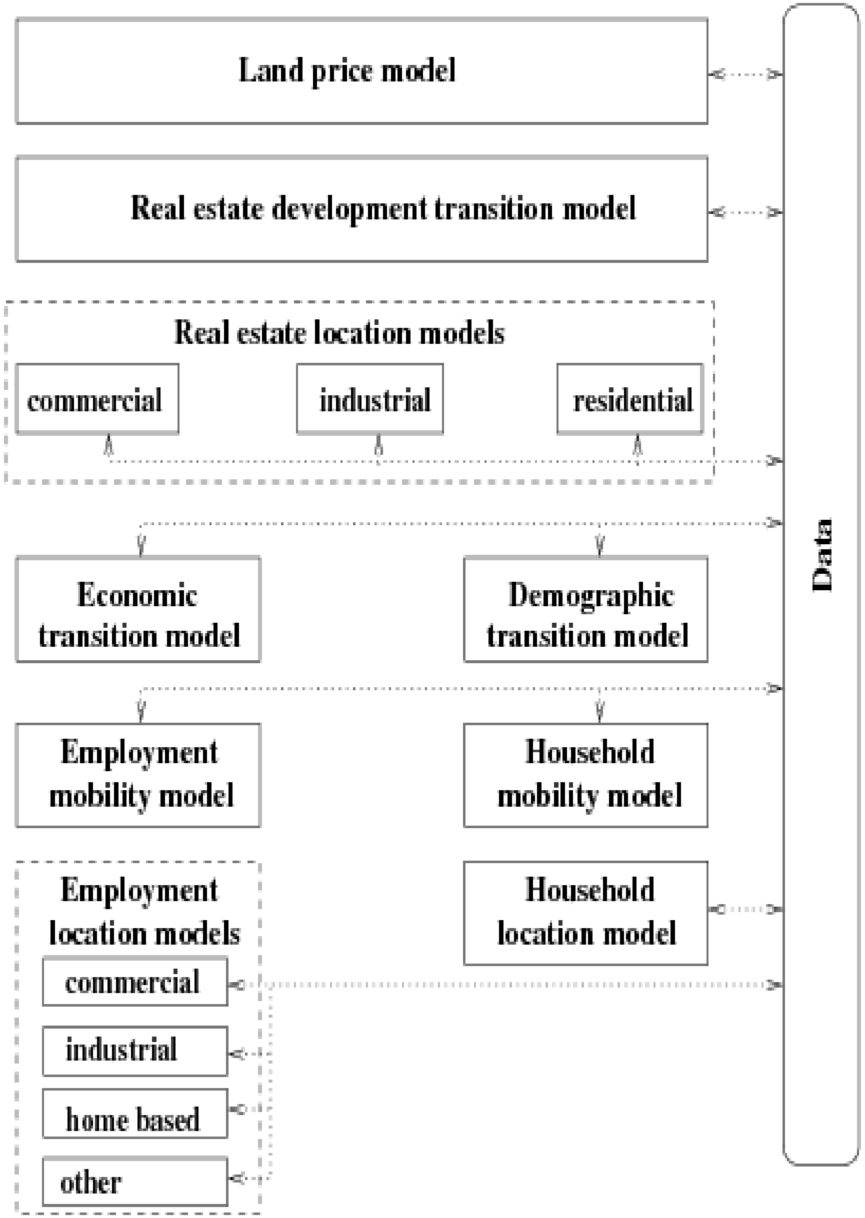
UrbanSim



Uncertainty:

75 submodels (51 stochastic)

UrbanSim



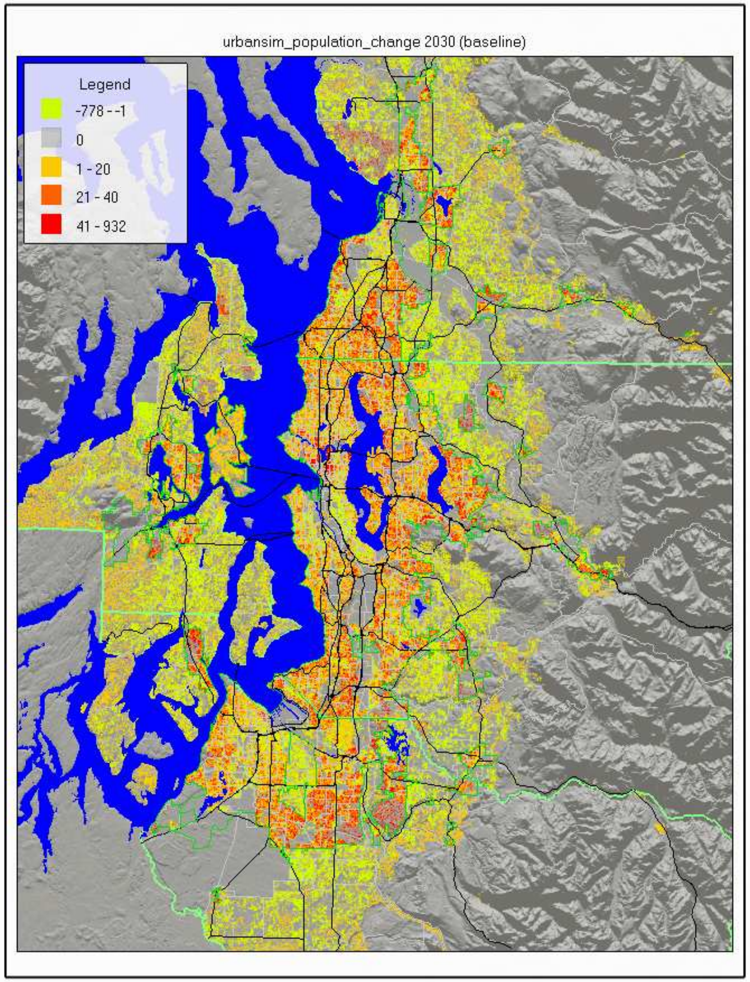
Uncertainty:

75 submodels (51 stochastic)

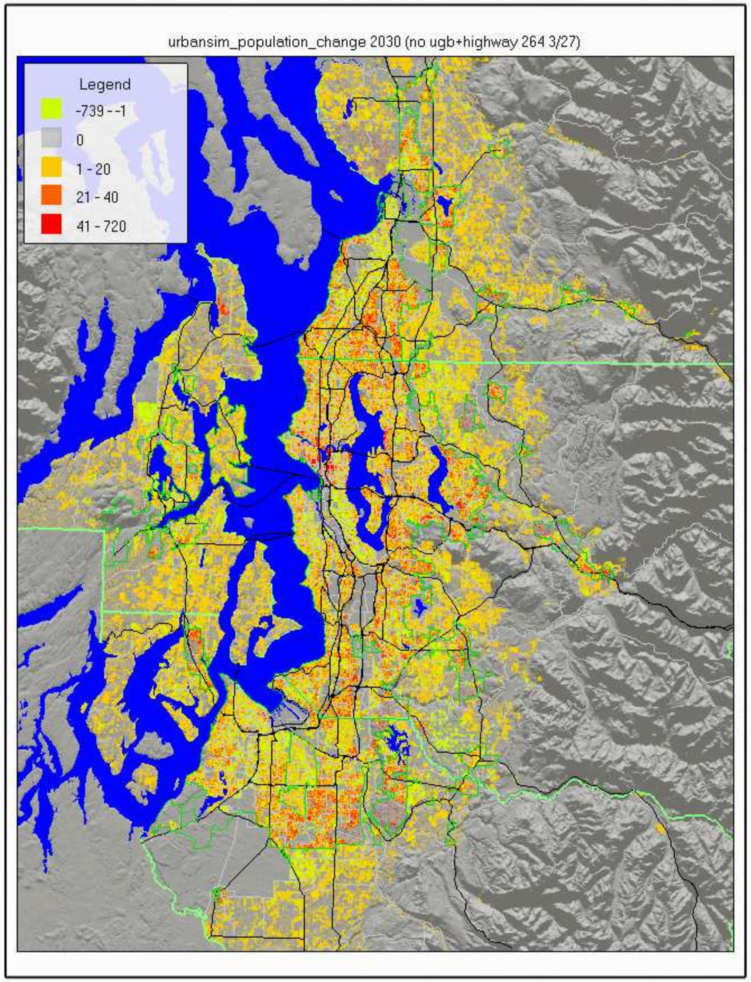
972 input par. for PSRC appl.

Population change in Puget Sound area (30 years prediction)

baseline



no UGB + highway



Bayesian Melding notation

Θ collection of model inputs

Bayesian Melding notation

Θ	collection of model inputs
$q(\Theta)$	prior probability distribution of inputs

Bayesian Melding notation

Θ	collection of model inputs
$q(\Theta)$	prior probability distribution of inputs
Φ	collection of model outputs about which we have information

Bayesian Melding notation

Θ	collection of model inputs
$q(\Theta)$	prior probability distribution of inputs
Φ	collection of model outputs about which we have information
Ψ	quantities of policy interest

Bayesian Melding notation

Θ	collection of model inputs
$q(\Theta)$	prior probability distribution of inputs
Φ	collection of model outputs about which we have information
Ψ	quantities of policy interest
y	observed data

Bayesian Melding notation

Θ	collection of model inputs
$q(\Theta)$	prior probability distribution of inputs
Φ	collection of model outputs about which we have information
Ψ	quantities of policy interest
y	observed data

$$\Phi = M_{\Phi}(\Theta)$$

Bayesian Melding notation

Θ	collection of model inputs
$q(\Theta)$	prior probability distribution of inputs
Φ	collection of model outputs about which we have information
Ψ	quantities of policy interest
y	observed data

$$\Phi = M_\Phi(\Theta)$$

$$\Psi = M_\Psi(\Theta, \Phi) = M_\Psi(\Theta, M_\Phi(\Theta))$$

Bayesian Melding notation

Θ	collection of model inputs
$q(\Theta)$	prior probability distribution of inputs
Φ	collection of model outputs about which we have information
Ψ	quantities of policy interest
y	observed data

$$\Phi = M_\Phi(\Theta)$$

$$\Psi = M_\Psi(\Theta, \Phi) = M_\Psi(\Theta, M_\Phi(\Theta))$$

$$L(\Phi) = \mathrm{Prob}(y|\Phi)$$

Bayesian Melding notation

Θ	collection of model inputs
$q(\Theta)$	prior probability distribution of inputs
Φ	collection of model outputs about which we have information
Ψ	quantities of policy interest
y	observed data

$$\Phi = M_\Phi(\Theta)$$

$$\Psi = M_\Psi(\Theta, \Phi) = M_\Psi(\Theta, M_\Phi(\Theta))$$

$$L(\Phi) = \mathsf{Prob}(y|\Phi)$$

$$L(\Theta) = \mathsf{Prob}(y|M_\Phi(\Theta))$$

Bayesian Melding notation

Θ	collection of model inputs
$q(\Theta)$	prior probability distribution of inputs
Φ	collection of model outputs about which we have information
Ψ	quantities of policy interest
y	observed data

$$\Phi = M_\Phi(\Theta)$$

$$\Psi = M_\Psi(\Theta, \Phi) = M_\Psi(\Theta, M_\Phi(\Theta))$$

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

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

$$\pi(\Theta) \propto q(\Theta)L(\Theta)$$

Bayesian Melding notation

Θ	collection of model inputs
$q(\Theta)$	prior probability distribution of inputs
Φ	collection of model outputs about which we have information
Ψ	quantities of policy interest
y	observed data

$$\Phi = M_\Phi(\Theta)$$

$$\Psi = M_\Psi(\Theta, \Phi) = M_\Psi(\Theta, M_\Phi(\Theta))$$

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

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

$$\pi(\Theta) \propto q(\Theta)L(\Theta)$$

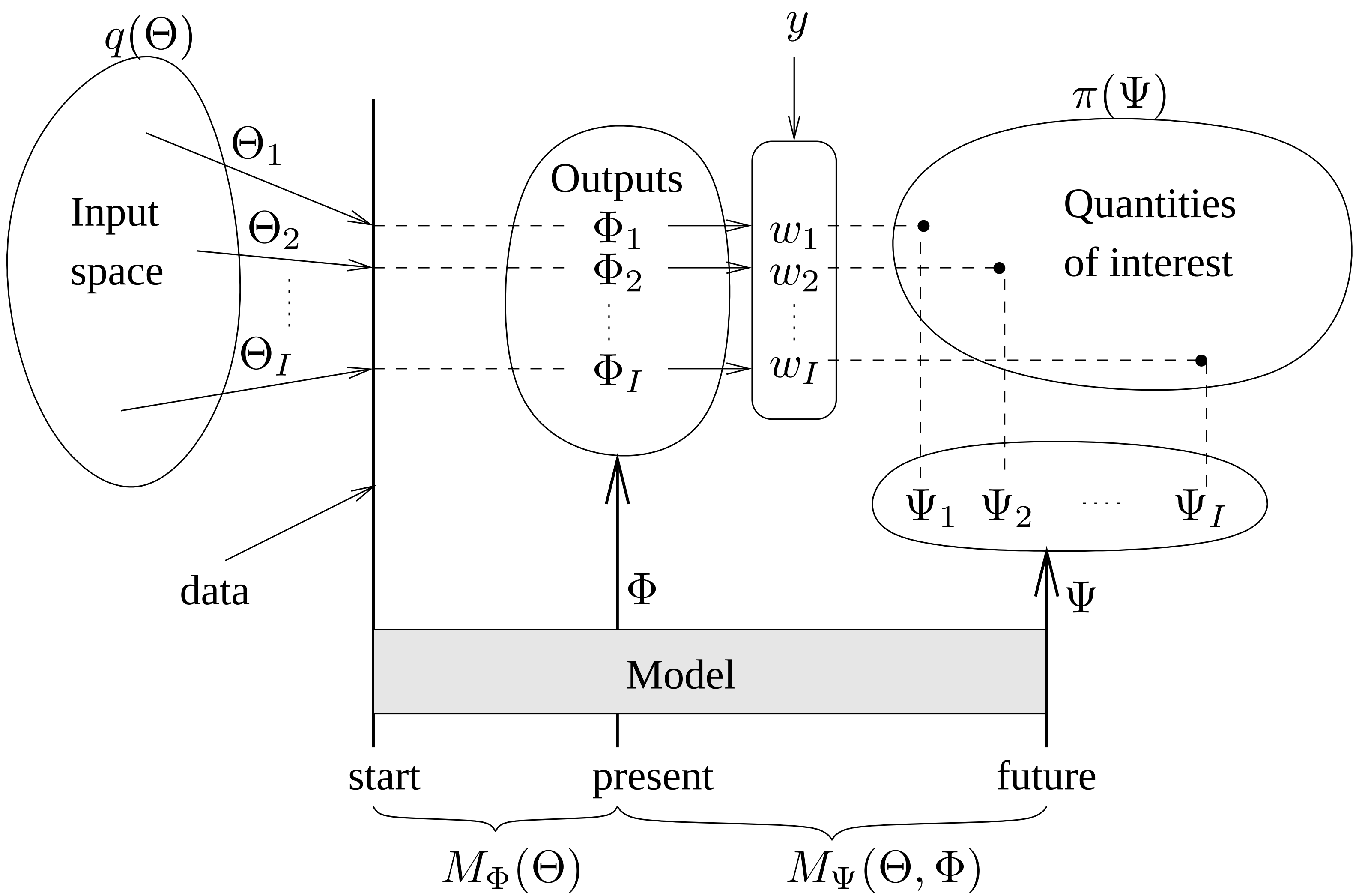
yields $\pi(\Psi)$

Bayesian Melding

For a deterministic model, based on Importance Sampling:

Bayesian Melding

For a deterministic model, based on Importance Sampling:



Computing the Posterior Distribution

For a stochastic model:

Computing the Posterior Distribution

For a stochastic model:

1. Draw a sample $\{\Theta_1, \ldots, \Theta_I\}$ from $q(\Theta)$.

Computing the Posterior Distribution

For a stochastic model:

1. Draw a sample $\{\Theta_1, \ldots, \Theta_I\}$ from $q(\Theta)$.
2. For each Θ_i, run the model J times with different seeds to obtain $\Phi_{ij}, j = 1, \ldots, J$.

Computing the Posterior Distribution

For a stochastic model:

1. Draw a sample $\{\Theta_1, \ldots, \Theta_I\}$ from $q(\Theta)$.
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 $\bar{\Phi}_i = \frac{1}{J}\sum_{j=1}^{J} \Phi_{ij}$.

Computing the Posterior Distribution

For a stochastic model:

1. Draw a sample $\{\Theta_1, \ldots, \Theta_I\}$ from $q(\Theta)$.
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 $\bar{\Phi}_i = \frac{1}{J}\sum_{j=1}^{J} \Phi_{ij}$.
   Result: An approximate posterior distribution of inputs with values $\{\Theta_1, \ldots, \Theta_I\}$ and probabilities proportional to $\{w_1, \ldots, w_I\}$.

Computing the Posterior Distribution

For a stochastic model:

1. Draw a sample $\{\Theta_1, \ldots, \Theta_I\}$ from $q(\Theta)$.
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 $\bar{\Phi}_i = \frac{1}{J}\sum_{j=1}^{J} \Phi_{ij}$.
   Result: An approximate posterior distribution of inputs with values $\{\Theta_1, \ldots, \Theta_I\}$ and probabilities proportional to $\{w_1, \ldots, w_I\}$.
4. Distributions of Φ and Ψ are finite mixtures:

$$\pi(\Phi) = \sum_{i=1}^{I} w_i \, p(\Phi|\Theta_i)$$

Computing the Posterior Distribution

For a stochastic model:

1. Draw a sample $\{\Theta_1, \ldots, \Theta_I\}$ from $q(\Theta)$.
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 $\bar{\Phi}_i = \frac{1}{J}\sum_{j=1}^{J} \Phi_{ij}$.
   Result: An approximate posterior distribution of inputs with values $\{\Theta_1, \ldots, \Theta_I\}$ and probabilities proportional to $\{w_1, \ldots, w_I\}$.
4. Distributions of Φ and Ψ are finite mixtures:

$$\pi(\Phi) = \sum_{i=1}^{I} w_i \, p(\Phi|\Theta_i) \qquad \pi(\Psi) = \sum_{i=1}^{I} w_i \, p(\Psi|\Theta_i)$$

Application to UrbanSim

- Test database: Eugene Springfield, OR

Application to UrbanSim

- Test database: Eugene Springfield, OR
- Baseyear data available for 1980.

Application to UrbanSim

- Test database: Eugene Springfield, OR
- Baseyear data available for 1980.
- We have observed the number of households in each zone in 1994, $y_1, \ldots, y_K$ $(K = 295)$.

Application to UrbanSim

- Test database: Eugene Springfield, OR
- Baseyear data available for 1980.
- We have observed the number of households in each zone in 1994, $y_1, \ldots, y_K$ $(K = 295)$.
- Prediction for 2000, $\Psi_1, \ldots, \Psi_K$.

Application to UrbanSim

- Test database: Eugene Springfield, OR
- Baseyear data available for 1980.
- We have observed the number of households in each zone in 1994, $y_1, \ldots, y_K$ $(K = 295)$.
- Prediction for 2000, $\Psi_1, \ldots, \Psi_K$.
- start: 1980, present: 1994, future: 2000.

Likelihood and posterior distribution

We are interested in

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

Likelihood and posterior distribution

We are interested in

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

based on

$$\Phi_{ijk} = \mu_{ik} + \delta_{ijk}, \text{ where } \delta_{ijk} \overset{iid}{\sim} N(0, \sigma_\delta^2)$$

Likelihood and posterior distribution

We are interested in

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

based on

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

Likelihood and posterior distribution

We are interested in

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

based on

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

Estimation of μ_{ik}, σ_δ^2, σ_i^2, and a done by approximate maximum likelihood.

Likelihood and posterior distribution (cont.)

$$y_k|\Theta_i \sim N(\hat{a} + \hat{\mu}_{ik}, v_i)$$

Likelihood and posterior distribution (cont.)

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

Likelihood and posterior distribution (cont.)

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

$$w_i \propto p(y|\Theta_i) =$$

Likelihood and posterior distribution (cont.)

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

$$w_i \propto p(y|\Theta_i) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi v_i}} \exp\left[-\frac{1/2(y_k - \hat{a} - \hat{\mu}_{ik})^2}{v_i}\right]$$

Likelihood and posterior distribution (cont.)

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

$$w_i \propto p(y|\Theta_i) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi v_i}} \exp\left[-\frac{1/2(y_k - \hat{a} - \hat{\mu}_{ik})^2}{v_i}\right]$$

$$\pi(\Psi_k) = \sum_{i=1}^{I} w_i N(\hat{a} b_a + \frac{1}{J}\sum_{j=1}^{J}\Psi_{ijk}, v_i b_v)\,, \quad k = 1, \ldots, K$$

Prior on inputs

- parameters estimated by multinomial logistic regression or linear regression:

Prior on inputs

- parameters estimated by multinomial logistic regression or linear regression:

$$MVN(\hat{\Theta}, \mathsf{diag}\left(\mathsf{SE}(\hat{\Theta})^2\right))$$

Prior on inputs

- parameters estimated by multinomial logistic regression or linear regression:

$$MVN(\hat{\Theta}, \mathsf{diag}\left(\mathsf{SE}(\hat{\Theta})^2\right))$$

- mobility rates r:

Prior on inputs

- parameters estimated by multinomial logistic regression or linear regression:

$$MVN(\hat{\Theta}, \mathsf{diag}\left(\mathsf{SE}(\hat{\Theta})^2\right))$$

- mobility rates r:

$$N(\hat{r}, (\frac{\hat{r}(1-\hat{r})}{n})^2) \quad \text{truncated at zero}$$

Prior on inputs

- parameters estimated by multinomial logistic regression or linear regression:

$$MVN(\hat{\Theta}, \text{diag}\left(\text{SE}(\hat{\Theta})^2\right))$$

- mobility rates r:

$$N(\hat{r}, (\frac{\hat{r}(1-\hat{r})}{n})^2) \quad \text{truncated at zero}$$

- control totals:

Prior on inputs

- parameters estimated by multinomial logistic regression or linear regression:

$$MVN(\hat{\Theta}, \mathsf{diag}\left(\mathsf{SE}(\hat{\Theta})^2\right))$$

- mobility rates r:

$$N(\hat{r}, (\frac{\hat{r}(1-\hat{r})}{n})^2) \quad \text{truncated at zero}$$

- control totals:

$$N(\hat{c}, 20V(\hat{c})), \; V \text{ is estimated variance per year}$$

Simulation

- Run the model for 100 different outputs, 2 $\times$ with different seed $\Rightarrow I = 100,\ J = 2$ (results confirmed with $I = 1000,\ J = 3$)

Simulation

- Run the model for 100 different outputs, 2 $\times$ with different seed $\Rightarrow I = 100$, $J = 2$ (results confirmed with $I = 1000$, $J = 3$)
- Variances σ_δ^2 and σ_i^2 stabilized by applying the sqrt transform.

Simulation

- Run the model for 100 different outputs, 2 $\times$ with different seed $\Rightarrow I = 100,\ J = 2$ (results confirmed with $I = 1000,\ J = 3$)
- Variances σ_δ^2 and σ_i^2 stabilized by applying the sqrt transform.
- From the results Φ_{ijk} we estimated $\hat{a}$, $\hat{\sigma}_\delta^2$, and $\hat{\sigma}_i^2$

Simulation

- Run the model for 100 different outputs, 2 $\times$ with different seed $\Rightarrow I = 100,\ J = 2$ (results confirmed with $I = 1000,\ J = 3$)
- Variances σ_δ^2 and σ_i^2 stabilized by applying the sqrt transform.
- From the results Φ_{ijk} we estimated $\hat{a}$, $\hat{\sigma}_\delta^2$, and $\hat{\sigma}_i^2$
- Weights:

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	$4 \cdot 10^{-45}$

Simulation

- Run the model for 100 different outputs, 2 $\times$ with different seed $\Rightarrow I = 100,\ J = 2$ (results confirmed with $I = 1000,\ J = 3$)
- Variances σ_δ^2 and σ_i^2 stabilized by applying the sqrt transform.
- From the results Φ_{ijk} we estimated $\hat{a}$, $\hat{\sigma}_\delta^2$, and $\hat{\sigma}_i^2$
- Weights:

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	$4 \cdot 10^{-45}$

$R = 0.93$

Simulation

- Run the model for 100 different outputs, 2 $\times$ with different seed $\Rightarrow I = 100,\ J = 2$ (results confirmed with $I = 1000,\ J = 3$)
- Variances σ_δ^2 and σ_i^2 stabilized by applying the sqrt transform.
- From the results Φ_{ijk} we estimated $\hat{a}$, $\hat{\sigma}_\delta^2$, and $\hat{\sigma}_i^2$
- Weights:

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	$4 \cdot 10^{-45}$

$R = 0.93$ (for $i = 64$)

$R = 0.84$ (for $i = 33$)

Simulation

- Run the model for 100 different outputs, 2 $\times$ with different seed $\Rightarrow I = 100,\ J = 2$ (results confirmed with $I = 1000,\ J = 3$)
- Variances σ_δ^2 and σ_i^2 stabilized by applying the sqrt transform.
- From the results Φ_{ijk} we estimated $\hat{a}$, $\hat{\sigma}_\delta^2$, and $\hat{\sigma}_i^2$
- Weights:

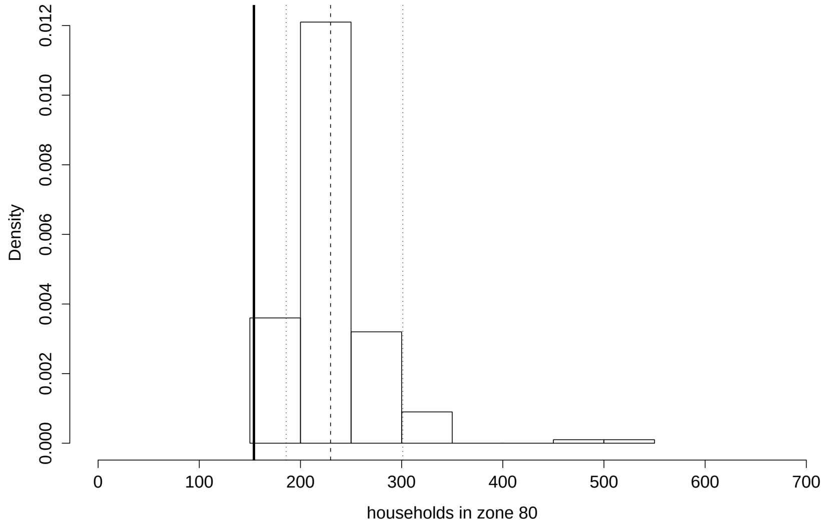
i	64	13	12	76	68	88	23	33(min)
w_i	0.8058	0.0883	0.0370	0.0217	0.0122	0.0116	0.0068	$4 \cdot 10^{-45}$

$R = 0.93$ (for $i = 64$) $\qquad R = 0.84$ (for $i = 33$)

- propagation factors b_a and b_v set to $20/14$ $\left(\frac{2000-1980}{1994-1980}\right)$

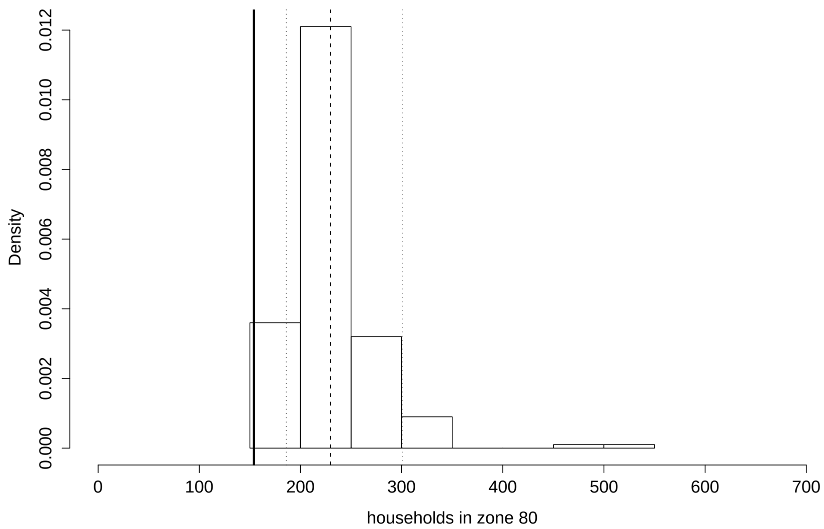
Results

Multiple runs

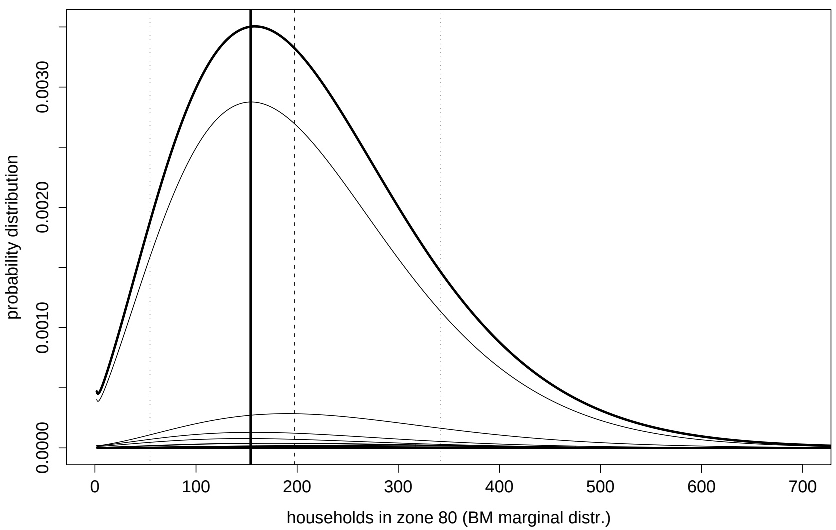


Results

Multiple runs

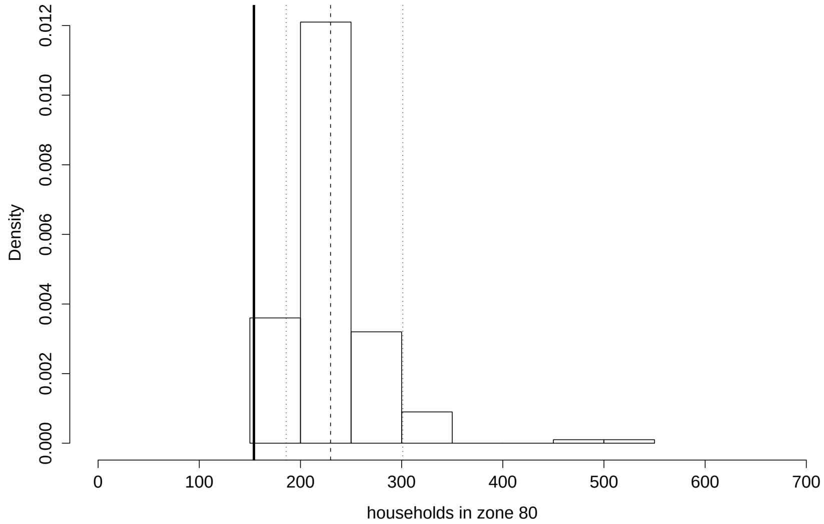


Bayesian melding

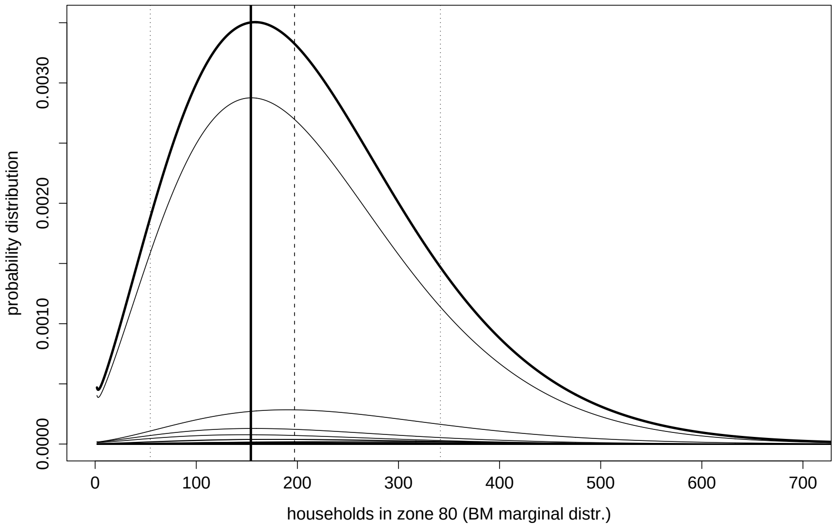


Results

Multiple runs



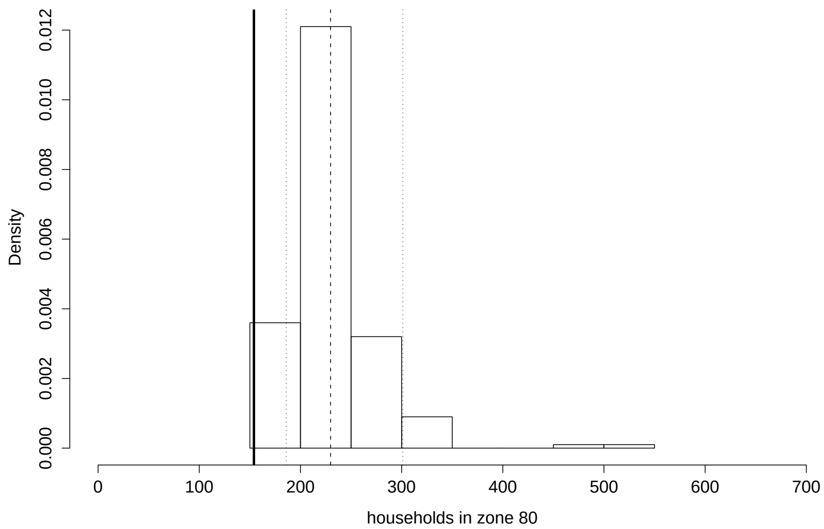
Bayesian melding



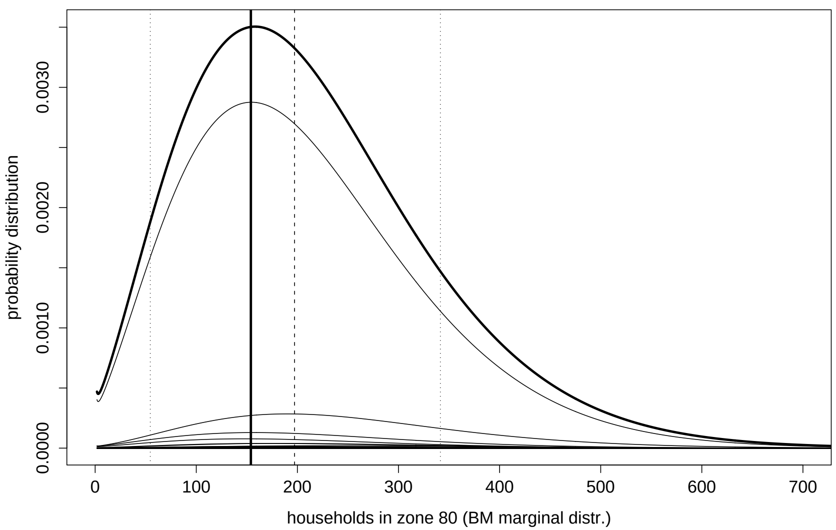
method	missed cases	coverage

Results

Multiple runs

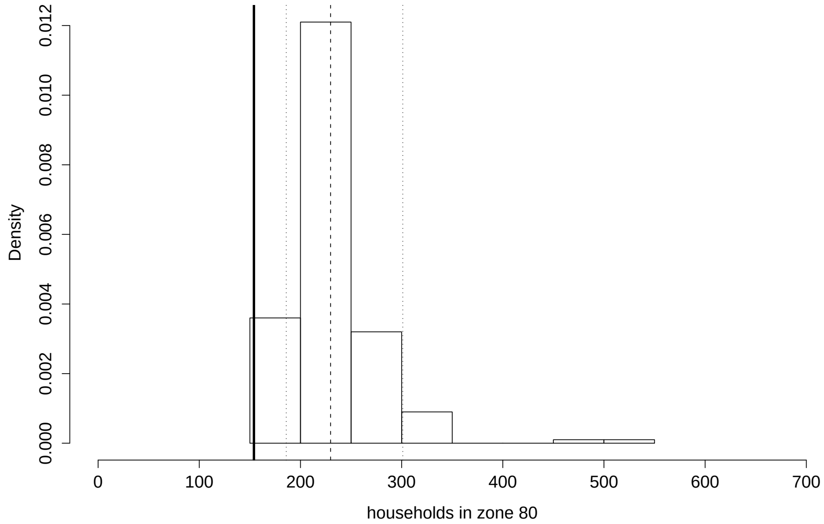
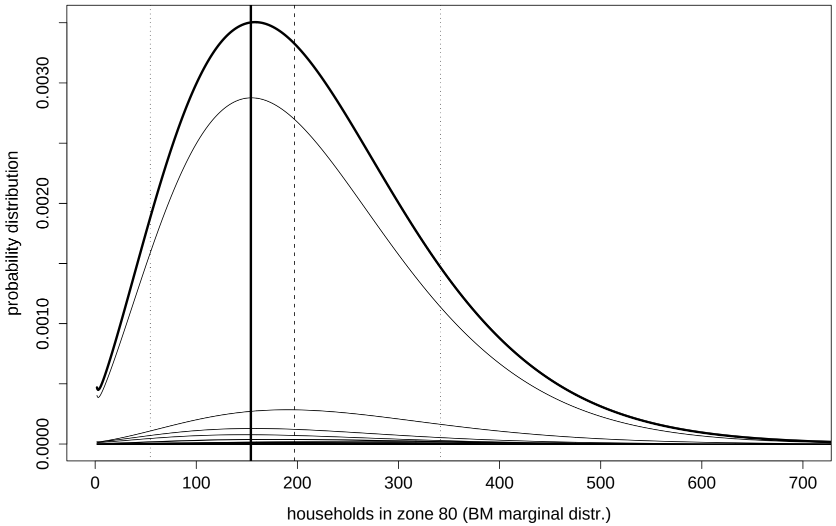


Bayesian melding



method	missed cases	coverage
multiple runs	163	0.39

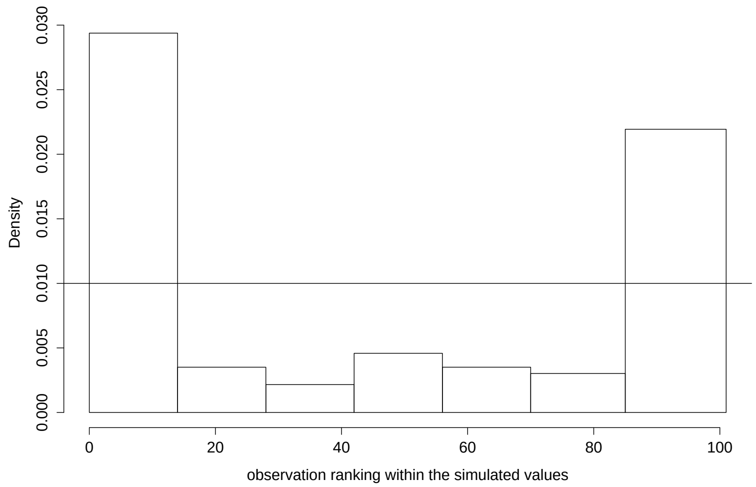
Results

Multiple runs	Bayesian melding
	

method	missed cases	coverage
multiple runs	163	0.39
Bayesian melding	31	0.88

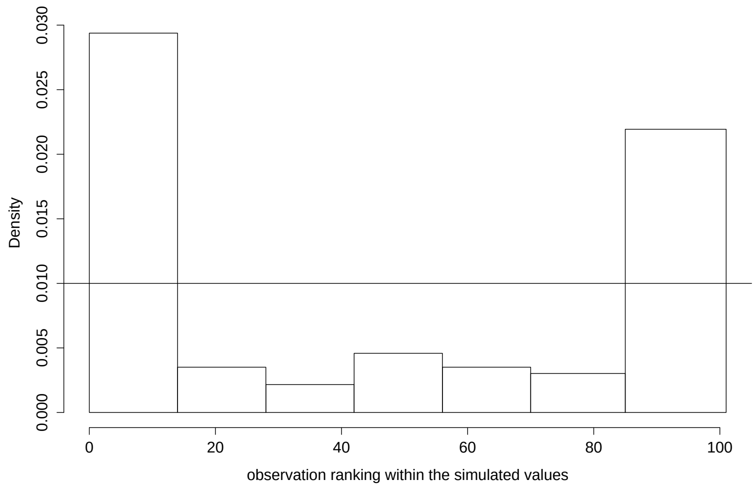
Verification rank histogram

Multiple runs

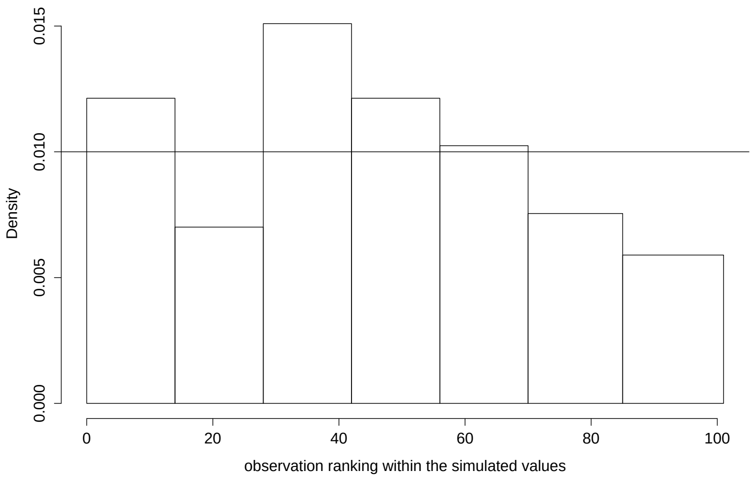


Verification rank histogram

Multiple runs

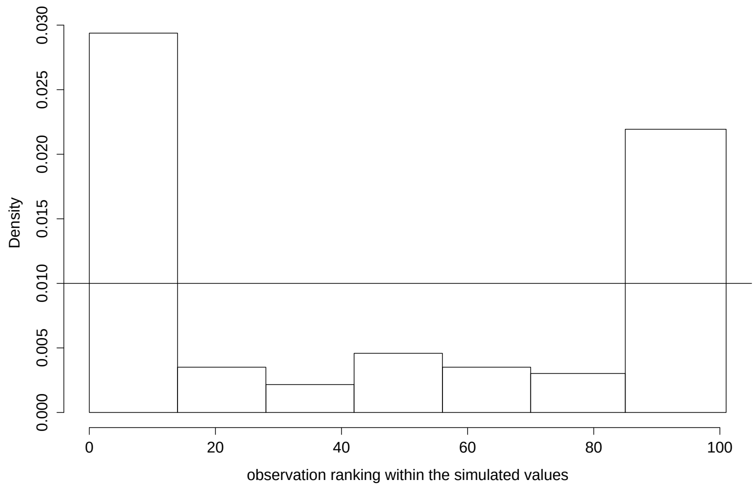


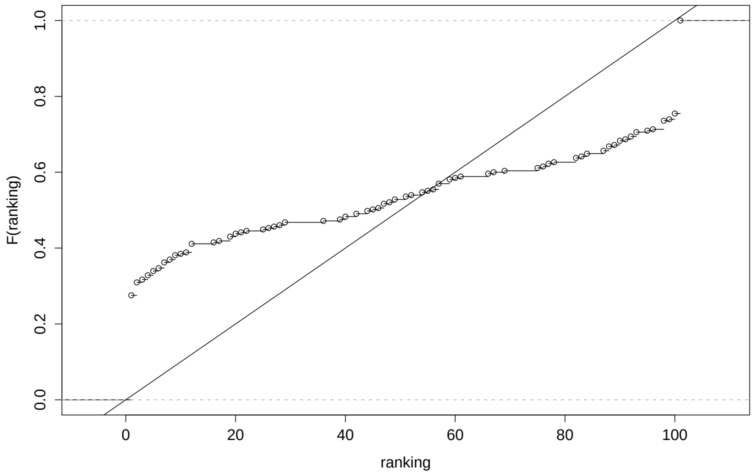
Bayesian melding



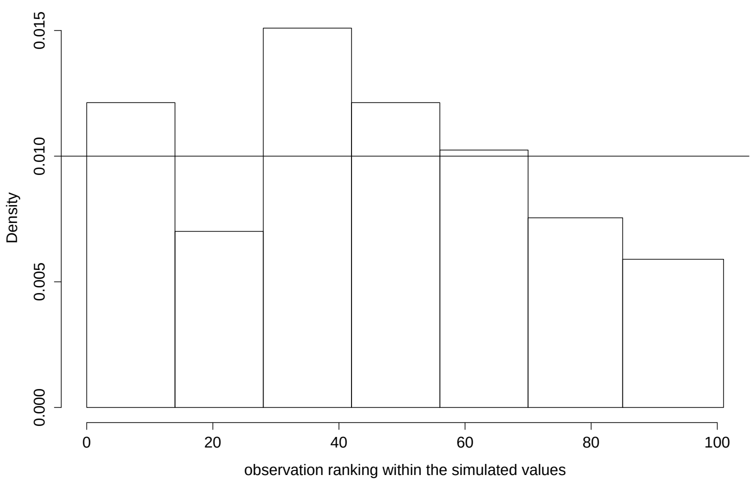
Verification rank histogram

Multiple runs



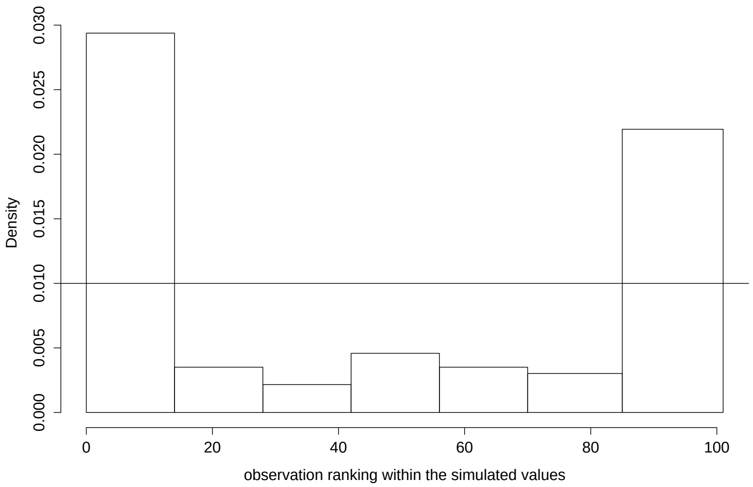


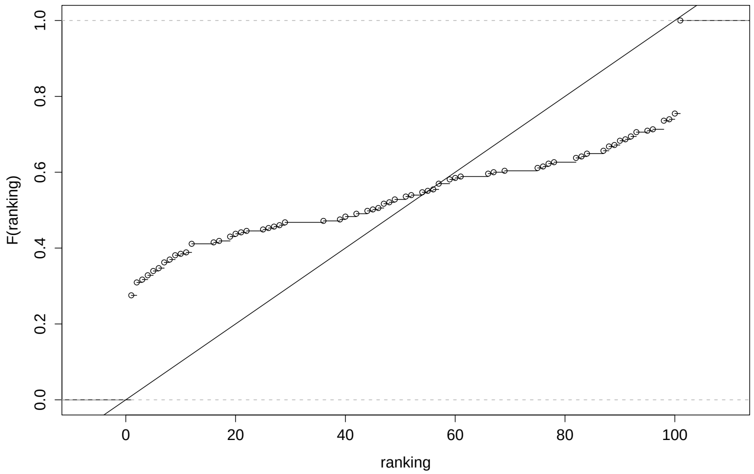
Bayesian melding



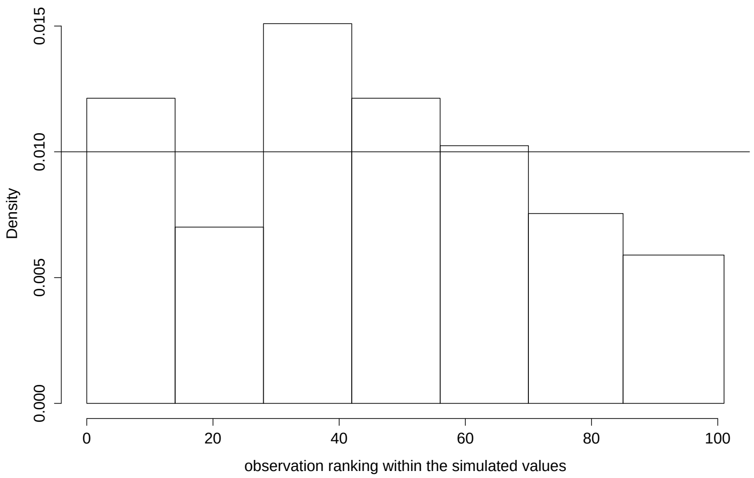
Verification rank histogram

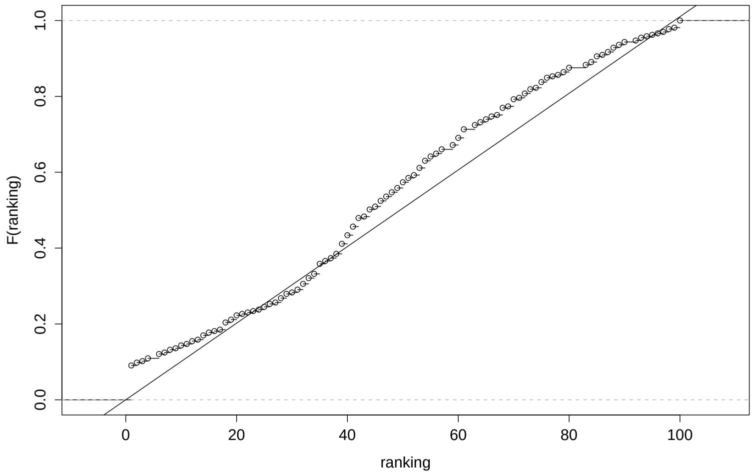
Multiple runs





Bayesian melding





Simulation size

simulation size	100×2		1000×3	
method	missed cases	coverage	missed cases	coverage

Simulation size

simulation size	100×2		1000×3	
method	missed cases	coverage	missed cases	coverage
Bayesian melding	31	0.88		

Simulation size

simulation size	100×2		1000×3	
method	missed cases	coverage	missed cases	coverage
Bayesian melding	31	0.88	29	0.89

Simulation size

simulation size	100×2		1000×3	
method	missed cases	coverage	missed cases	coverage
Bayesian melding	31	0.88	29	0.89
multiple runs	163	0.39		

Simulation size

simulation size	100×2		1000×3	
method	missed cases	coverage	missed cases	coverage
Bayesian melding	31	0.88	29	0.89
multiple runs	163	0.39	165	0.38

Simulation size

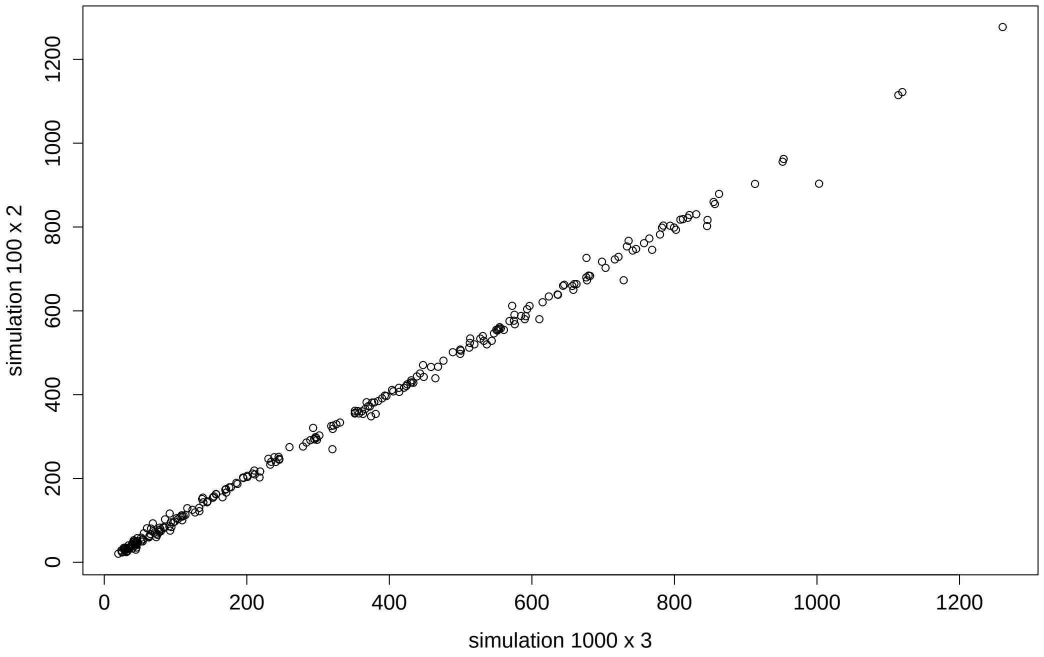
simulation size	100×2		1000×3	
method	missed cases	coverage	missed cases	coverage
Bayesian melding	31	0.88	29	0.89
multiple runs	163	0.39	165	0.38

Simulation size

simulation size	100×2		1000×3	
method	missed cases	coverage	missed cases	coverage
Bayesian melding	31	0.88	29	0.89
multiple runs	163	0.39	165	0.38

Simulation size

simulation size	100×2		1000×3	
method	missed cases	coverage	missed cases	coverage
Bayesian melding	31	0.88	29	0.89
multiple runs	163	0.39	165	0.38



Summary

- Goal is to assess uncertainty in the output from urban simulation models.

Summary

- Goal is to assess uncertainty in the output from urban simulation models.
- Bayesian melding extended for application to a stochastic simulation model.

Summary

- Goal is to assess uncertainty in the output from urban simulation models.
- Bayesian melding extended for application to a stochastic simulation model.
- Experiment on Eugene, OR, 1980-2000:

Summary

- Goal is to assess uncertainty in the output from urban simulation models.
- Bayesian melding extended for application to a stochastic simulation model.
- Experiment on Eugene, OR, 1980-2000:
  - Simple multiple runs underestimated uncertainty.

Summary

- Goal is to assess uncertainty in the output from urban simulation models.
- Bayesian melding extended for application to a stochastic simulation model.
- Experiment on Eugene, OR, 1980-2000:
  - Simple multiple runs underestimated uncertainty.
  - Bayesian melding provided well calibrated results.