Kobe University Repository : Kernel

PDF issue: 2025-12-05

Criteria for selecting model updating methods for better temporal transferability

Sanko, Nobuhiro

(Citation)
Transportmetrica A: Transport Science,16(3):1310-1332

(Issue Date)
2020-04-03

(Resource Type)
journal article

(Version)
Accepted Manuscript

(Rights)
This is an Accepted Manuscript of an article published by Taylor & Francis in [Transportmetrica A: Transport Science on 2020] available online: http://www.tandfonline.com/10.1080/23249935.2020.1746862

(URL)
https://hdl.handle.net/20.500.14094/90007003

KOBE UNIVERSITY

Criteria for selecting model updating methods for better temporal transferability

Nobuhiro Sanko

Graduate School of Business Administration, Kobe University, Japan

2-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan

E-mail: sanko@kobe-u.ac.jp

https://orcid.org/0000-0002-2159-1765

Abstract

When older and more recent datasets have large and small numbers of observations, respectively, then discrete choice modellers must decide whether to utilise both datasets with model updating (transfer scaling, joint context estimation, Bayesian updating, and combined transfer estimation) or only the more recent dataset. This study investigates the case when the data collection time points and the number of observations from each time point differ. Bootstrapping was applied to commuting mode choice models utilising datasets from Nagoya, Japan. The following criteria are proposed: (1) when the more recent time point has a large number of observations, use only the more recent data; (2) when the more recent time point has a smaller number of observations, use transfer scaling or joint context estimation based on the differences in the contexts of the two time points and the sample size from the older time point.

Keywords

Model updating; Data collection time points; Number of observations; Temporal transferability; Repeated cross-sectional data

1. Introduction

Discrete choice models have been applied to transport demand forecasting, and the success of forecasting relies on the temporal transferability of the models. The transferability of models is strongly influenced by the data used for estimation, and datasets with a large number of recent observations are ideal (Sanko 2017). Modellers, however, may not have access to ideal data. The present study addresses the case in which an older dataset has a large number of observations while a more recent dataset has a smaller number of observations. This is likely to happen in areas where large-scale surveys are conducted at infrequent intervals. For example, in Japanese metropolises, such surveys are typically conducted every ten years, so the most recent data may be ten years old. Modellers may collect new data, but conducting a large-scale survey by themselves is difficult, so their survey may have a limited number of observations.

When data from two time points is available, forecasting performance can be improved by utilising a model updating method, such as transfer scaling, joint context estimation, Bayesian updating, and combined transfer estimation.[1] The model updating methods utilise both a larger number of observations from the older time point and a smaller number of observations from the more recent time point. This method, however, raises significant concerns about the impacts of the combinations of data collection time points and the numbers of observations from each time point.

No studies have proposed clear criteria for choosing updating methods. Although studies have addressed this issue (e.g., Badoe and Miller (1995b) and Karasmaa and Pursula (1997)), their results are inconclusive and are not supported by any statistical tests. This is discussed in Section 2 of this paper. Furthermore, these studies showed that utilising the older and newer datasets together does little to improve forecasting performance even when the more recent time point has very few observations (e.g., 400 observations). While

[1] Other updating methods utilising aggregate data is not of interest in the present study. Interested readers refer to Ortúzar and Willumsen (2011, Section 9.5).

this issue has yet to be investigated, the present study questions the meaning of model updating methods themselves.

By combining the concepts of Sanko (2017) and the present study, the author has grasped the overall concept of using data from two time points. When data from two time points is available, modellers have three options: 1) use both the older and newer datasets, 2) use only the newer dataset, and 3) use only the older dataset. Sanko (2017) compared options 2 and 3 and found that option 2—use only the newer dataset—generally produces better forecasts, with and without statistical significance, than option 3—use only the older dataset. When the newer dataset has very few observations, then option 3—use only the older dataset—produces better forecasting without statistical significance. However, option 1—use both the older and newer datasets—might produce much better forecasts than either option 2 or 3, and there is no reason to rule out option 1 from the analysis. This study compares the first and second options: 1) use both the older and newer datasets, and 2) use only the newer dataset.[2] Since Sanko (2017) demonstrated that option 2 is preferred to option 3, if option 2 is preferred to option 1 in the present study, then the value of option 2—use only the newer dataset—is more strongly supported.

This study proposes criteria for selecting model updating methods by focusing on the data collection time points and the number of observations. This study considers the option 2 above—utilising only the more recent data—a special case of model updating in which the older dataset is not weighted. The study offers practical answers to the following questions.

1. Suppose that data from two points in time is available. Which model updating method should be used for forecasting?
2. Suppose that old data is available, but new data for modelling and forecasting will soon be collected. When the sample size from the older time point is known, how many

[2] Options 1 and 3 are not compared. Comparing options 1 and 2 has more value than comparing options 1 and 3, since Sanko (2017) showed that option 2 produced better forecasts than option 3. Both Sanko (2015, 2018) and the present study compare options 1 and 2. For option 1, Sanko (2015, 2018) considered transfer scaling, while the present study considers transfer scaling, joint context estimation, Bayesian updating, and combined transfer estimation.

observations should be collected for a specified updating method?

Although answering the above two research questions is an ultimate goal of the study, testing the research questions requires some operationalisations. This is the first study to compare and statistically test different model updating methods in terms of their forecasting performance, especially from the viewpoints of data collection time points and numbers of observations. The analysis is kept as simple as possible and utilises datasets available to the author. The operationalisations are noted from the two viewpoints of modelling and data.

First, three dimensions relating to modelling are operationalised, even though they may affect temporal transferability: 1) the underlying theory of travel behaviour, 2) the mathematical model structure, and 3) the empirical specification.[3] For the purpose of operationalising these three dimensions, this study assumes 1) a theory of utility maximisation, 2) linear-in-parameters multinomial logit models, and 3) a single model specification throughout the paper.[4]

Second, data availability allowed operationalising two dimensions of interests—data collection time points and the numbers of observations. This study utilises household travel survey data from Nagoya, Japan collected in 1971, 1981, 1991, and 2001. The present study analyses commuting-to-work mode choice behaviours, which are also analysed in previous studies by the author. The data from 1971, 1981, and 1991 are used for model estimation and

[3] Four levels of hierarchy that affect transferability are discussed in Sikder (2013). The top three levels are 1), 2), and 3) in the main text. Examples included in the three levels are 1) utility maximisation vs. lexicographic and trip-based vs. tour-based; 2) logit vs. nested logit; and 3) choice of explanatory variables, linear vs. non-linear formulation of explanatory variables, and consideration of heterogeneity. The present study's interest lies in 4) model parameter estimates. The two dimensions of interests—data collection time points and the numbers of observations—impact the model parameter estimates.

[4] Several studies have investigated the impact of model specifications on temporal transferability, with most concluding that models with more explanatory variables are more temporally transferable (Badoe and Miller, 1995b; Fox et al., 2014; Karasmaa and Pursula, 1997; Train, 1979). However, models estimated with more explanatory variables sometimes result in overfitting (Badoe and Miller, 1995a). The present study does not address the impact of the model specifications, and the same model specifications are utilised throughout the analysis. Different model specifications result in different levels of temporal transferability. However, if these specifications improved or worsened to the same extent in all of the updated models, then the impact of the model specifications would be cancelled out. If the model specifications improved or worsened to different degrees, then the specifications would affect the results of the study. This is a topic for future study.

updating, and the data from 2001 are used for validation. The number of observations usable for modelling differs across time points, but more than 10000 observations are available in all of the time points. Regarding the dimensions of the data collection time points, combinations of 1971 and 1981, 1971 and 1991, and 1981 and 1991 are compared. Regarding the dimensions of the numbers of observations, the author considers 12 values between 100 and 10000, and compares combinations of these 12 values from each time point.[5] Although the data availability restricts combinations relating to the two dimensions of interest, the dataset still enables the author to compare different combinations of data collection time points and number of observations. Furthermore, the data has some advantages. The survey was conducted by the same governmental bodies in a similar manner each year. Therefore, the datasets are appropriate for analysing data collection time points and sample sizes, since the quality of the data is believed to differ little across years. This data is utilised in a series of the author's studies, so direct comparison is possible. The data comes from the 1971–2001 period, and the most recent data from 2001 is more than 15 years old, which is less of a concern.

The criteria proposed in this paper are based on analyses of problems to which the above operationalisations are applied. This study utilises a bootstrapping technique to obtain more concrete insights with statistical meaning, but the quantitative results, such as the number of observations from each time point, may have meaning only in the contexts of operationalisation. The criteria for more general contexts, as defined in the research questions, must be inferred from the analysis. Therefore, the criteria proposed in the present study include qualitative rather than quantitative expressions.

This paper is organised as follows. Section 2 reviews the literature that compares model updating methods in terms of forecasting performance. Section 3 describes the datasets. Section 4 describes a multinomial logit model, various model updating methods, a

[5] Cleaning the data for estimation purposes produced more than 30000 usable observations from all of the time points. However, in other studies by the author and in the previous studies reported in Section 2, analysing up to 10000 observations contributed significantly to the results. Therefore, the present study follows the same method.

bootstrapping procedure, and hypothesis testing. Section 5 reports estimates and compares and statistically tests the forecasting performance of the models. Criteria for selecting model updating methods are proposed. Section 6 presents the concluding remarks.

2. Literature Review

No studies have statistically tested model updating methods in terms of their forecasting performance. Moreover, no studies have proposed criteria for selecting model updating methods. However, the following two studies provided some insights.

Karasmaa and Pursula (1997) utilised data from Helsinki, Finland, that had been collected in 1981 and 1988 to evaluate the transferability of mode and destination choice models. They estimated four updating models (transfer scaling, joint context estimation, Bayesian updating, and combined transfer estimation) utilising both a fixed number of observations (n = 2428) from 1981 and varied numbers of observations (n = 400, 800, 1600, 3200, and 4575) from 1988.[6] They also estimated models utilising only the various numbers of observations from 1988, which is considered a small sample model since the varied numbers of observations are sometimes smaller than the fixed number of observations. The updating models and the small sample model were then used to forecast behaviours for 1988. Note that the 1988 dataset was used for both model estimation and validation, since datasets from only two points in time were available. They examined two model specifications: LOS (leve-of-service) and LOS+SE (socio-economic). Note that the LOS and LOS+SE specifications include LOS and LOS+SE variables, respectively as explanatory variables. Results indicated that, common to both specifications, the small sample model and the combined transfer estimation produced almost identical forecasting performance, and their forecasting performance was always superior to that by transfer scaling, joint context estimation, and Bayesian updating.

[6] The fixed number of observations and the varied numbers of observations respectively correspond to the larger number of observations from the older time point and the smaller number of observations from the more recent time point in the present study. Their study includes a case in which the varied numbers of observations are larger than the fixed number of observations; this is not of interest in the present study.

Badoe and Miller (1995b) utilised data collected in 1964 and 1986 in Toronto, Canada, to evaluate the transferability of morning peak work trip mode choice models. The study follows exactly the same procedure as Karasmaa and Pursula (1997), but the fixed number of observations was n = 8066 from 1964 and the varied numbers of observations were n = 400, 800, 1600, 3200, 6400, and 32328 from 1986. They reached the same conclusions as Karasmaa and Pursula (1997), for both the LOS+SE specification and the LOS specification with $n \geq 1600$ from 1986. However, joint context estimation produced the highest forecasting performance, followed by transfer scaling, combined transfer estimation, and small sample model for the LOS specification with $n \leq 800$ from 1986.

These two studies have limited implications for the following reasons. First, two time points are inadequate for data use; data must be prepared from at least three time points, and data from one time point must be used solely for validation. Second, the results might be coincidental, since these two studies randomly, but only once, used a small number of observations. The present study is designed to overcome these limitations and provide criteria for selecting model updating methods focusing on the data collection time points and the numbers of observations. Specifically, the first reason above is addressed by utilising data from four points in time. The second reason is addressed by utilising a bootstrap technique, which already has been applied to studies of transferability. Karasmaa and Pursula (1997) randomly sampled the 1998 datasets 60 times and evaluated how the number of observations from 1998 and the updating method affected the stability of the estimated parameters. However, they did not apply the bootstrapped estimates to forecasting. In the context of spatial transferability, Karasmaa (2003) randomly sampled the dataset 100 times and evaluated how the number of observations and the updating method affected forecasting performance. However, Karasmaa's study did not use statistical tests.

3. Data

This study utilises the household travel survey data in the Nagoya metropolitan area of Japan.

The data is repeated cross-sectional data from 1971, 1981, 1991, and 2001. The data from 1971, 1981, and 1991 are used for modelling, while the data from 2001 is used for validation. The main contents of the survey have changed little over the years. Commuting-to-work mode choice models are estimated, and three transportation modes—rail, bus, and car—are considered. A full description of the datasets is found in Sanko (2014), but two points must be emphasised. First, travel cost is not considered, since allowances are provided by most companies for purchasing commuting passes or fuel. Second, transportation modal shares changed substantially between 1971 and 1981, but have changed less since then.[7]

4. Methodology

Since this is the first study to statistically test forecasting performances calculated by different model updating methods, simple multinomial logit models were employed. However, the methodology is applicable to other model formulations. This section presents multinomial logit models and model updating methods, followed by the bootstrapping procedure, and hypothesis testing utilising the bootstrap. The methodology is graphically depicted in Fig. 1. In the following explanation, *t*1 and *t*2 represent the older and more recent time points, respectively.

***** Fig. 1 *****

4.1. Multinomial Logit Models

Random utility models are assumed, and total utility is decomposed into a deterministic component and an error component. The deterministic component of individual *p*'s utility for alternative *i*, at *t* (*t* = *t*1 or *t*2 in the following explanations), V_{ip}^{t}, is expressed as Eq. (1).

[7] After cleaning the data for estimation purposes, the shares of travel modes among rail, bus, and car in 1971, 1981, 1991, and 2001 were 28%, 28%, 26%, and 25%, respectively, for rail, 21%, 9%, 5%, and 3%, respectively, for bus, and 51%, 63%, 68%, and 72%, respectively, for car.

$$V_{ip}^{t} = \alpha_{i}^{t} + \sum_{k} \beta_{ik}^{t} x_{ikp}^{t} \quad (1)$$

where x_{ikp}^{t} denotes the *k*-th explanatory variable for individual *p* for alternative *i* at *t*, and β_{ik}^{t} denotes its related parameter; α_{i}^{t} denotes an alternative-specific constant for alternative *i* at *t*; scale parameters are not explicitly included since they are fixed to unity values for identification, which is relaxed in Sections 4.2.1 and 4.2.2.

Assuming independent and identical Gumbel distributions for the error components, multinomial logit models are derived, where the probability of individual *p*'s choosing alternative *i* among alternative *j*'s in his/her choice set at *t*, P_{ip}^{t}, is expressed as:

$$P_{ip}^{t} = \frac{\exp\left(V_{ip}^{t}\right)}{\sum_{j} \exp\left(V_{jp}^{t}\right)} \quad (2)$$

The log-likelihood function, L^{t}, is defined as Eq. (3), and parameters are estimated by using the maximum likelihood method.

$$L^{t} = \sum_{p} \sum_{j} y_{jp}^{t} \ln\left(P_{jp}^{t}\right) \quad (3)$$

where y_{jp}^{t} denotes an indicator that takes a value of one if individual *p* chose alternative *j* at *t* and zero otherwise.

4.2. Model Updating Methods

Five model updating methods are examined. A model utilising only the more recent dataset also is considered as a special case of model updating that poses zero weight on the older dataset and unity weight on the more recent dataset. For expositional ease, the following

abbreviations are used: *cst* for transfer scaling, since it updates ConStanTs and a scale parameter; *jnt* for JoiNT context estimation; *bay* for BAYesian updating; *com* for COMbined transfer estimation; and *sma* for SMAll sample model from the more recent dataset.

4.2.1. Transfer Scaling (abbreviation: *cst*) (Atherton and Ben-Akiva, 1976)

Models are estimated as explained in Section 4.1 by utilising data at *t*1; the data from *t*2 is utilised to update the parameters. The probability of individual *p*'s choosing alternative *i* at *t*2, P_{ip}^{t2}, is expressed as:

$$P_{ip}^{t2} = \frac{\exp\left(\mu^{t2}\left(\alpha_i^{t2} + \sum_k \hat{\beta}_{ik}^{t1} x_{ikp}^{t2}\right)\right)}{\sum_j \exp\left(\mu^{t2}\left(\alpha_j^{t2} + \sum_k \hat{\beta}_{jk}^{t1} x_{jkp}^{t2}\right)\right)} \tag{4}$$

where, a hat (^) indicates an estimate.

By utilising $\mathbf{x}^{t2}$, $\mathbf{y}^{t2}$, and Eqs. (3) and (4), a scale parameter (μ^{t2}) and alternative-specific constants ($\boldsymbol{\alpha}^{t2}$) are estimated, while parameters related to explanatory variables ($\hat{\boldsymbol{\beta}}^{t1}$) are used as is.

The forecasting performance, log-likelihood on 2001, as shown in Eq. (3), is evaluated by applying $\hat{\boldsymbol{\beta}}^{t1}$, $\hat{\boldsymbol{\alpha}}^{t2}$, $\hat{\mu}^{t2}$, $\mathbf{x}^{2001}$, and $\mathbf{y}^{2001}$ to Eqs. (3) and (4).

4.2.2. Joint Context Estimation (abbreviation: *jnt*) (Badoe and Miller, 1995b)

Models are estimated by utilising data from both *t*1 and *t*2. The data from *t*1 is applied to Eqs. (1)–(3), while the data from *t*2 is applied to Eqs. (3) and (5). The probability of individual *p*'s choosing alternative *i* at *t*2, P_{ip}^{t2}, is expressed as:

$$P_{ip}^{t2} = \frac{\exp\left(\alpha_i^{t2} + \mu_{LOS}^{t2} \sum_{k'} \beta_{ik'}^{t2} x_{ik'p}^{t2} + \mu_{SE}^{t2} \sum_{k''} \beta_{ik''}^{t2} x_{ik''p}^{t2} \right)}{\sum_j \exp\left(\alpha_j^{t2} + \mu_{LOS}^{t2} \sum_{k'} \beta_{jk'}^{t2} x_{jk'p}^{t2} + \mu_{SE}^{t2} \sum_{k''} \beta_{jk''}^{t2} x_{jk''p}^{t2} \right)} \quad (5)$$

where, explanatory variables are categorised as LOS and SE variables; the *k*' and *k*" respectively are used to refer to LOS- and SE-related variables and parameters. Two different scale parameters, μ_{LOS}^{t2} and μ_{SE}^{t2}, are estimated for the LOS and SE components of the utility function, respectively, by following an approach adopted by Badoe and Wadhawan (2002). The following constraint is applied: $\boldsymbol{\beta}^{t1} = \boldsymbol{\beta}^{t2}$. The $\boldsymbol{\beta}^{t1} = \boldsymbol{\beta}^{t2}$, $\boldsymbol{\alpha}^{t1}$, $\boldsymbol{\alpha}^{t2}$, μ_{LOS}^{t2}, and μ_{SE}^{t2} are estimated by maximising the sum of log-likelihoods for the *t*1 and *t*2 datasets.

A forecasting performance, log-likelihood on 2001 as shown in Eq. (3), is evaluated by applying $\hat{\boldsymbol{\beta}}^{t1} = \hat{\boldsymbol{\beta}}^{t2}$, $\hat{\boldsymbol{\alpha}}^{t2}$, $\hat{\mu}_{LOS}^{t2}$, $\hat{\mu}_{SE}^{t2}$, $\mathbf{x}^{2001}$, and $\mathbf{y}^{2001}$ to Eqs. (3) and (5).

4.2.3. Bayesian Updating (abbreviation: *bay*) (Atherton and Ben-Akiva, 1976)

Models are estimated as shown in Section 4.1 by utilising data from *t*1; and estimated parameters and variance-covariance matrix are denoted as $\begin{pmatrix} \hat{\boldsymbol{\alpha}}^{t1} \\ \hat{\boldsymbol{\beta}}^{t1} \end{pmatrix}$ and $\boldsymbol{\Sigma}^{t1}$, respectively. The same applies to data from *t*2, and the $\begin{pmatrix} \hat{\boldsymbol{\alpha}}^{t2} \\ \hat{\boldsymbol{\beta}}^{t2} \end{pmatrix}$ and $\boldsymbol{\Sigma}^{t2}$ are defined similarly.

Parameters updated are expressed as:

$$\begin{pmatrix} \boldsymbol{\alpha}^{up} \\ \boldsymbol{\beta}^{up} \end{pmatrix} = \left(\left(\boldsymbol{\Sigma}^{t1}\right)^{-1} + \left(\boldsymbol{\Sigma}^{t2}\right)^{-1} \right)^{-1} \left(\left(\boldsymbol{\Sigma}^{t1}\right)^{-1} \begin{pmatrix} \hat{\boldsymbol{\alpha}}^{t1} \\ \hat{\boldsymbol{\beta}}^{t1} \end{pmatrix} + \left(\boldsymbol{\Sigma}^{t2}\right)^{-1} \begin{pmatrix} \hat{\boldsymbol{\alpha}}^{t2} \\ \hat{\boldsymbol{\beta}}^{t2} \end{pmatrix} \right) \quad (6)$$

A forecasting performance, log-likelihood on 2001 as shown in Eq. (3), is evaluated by

applying $\begin{pmatrix}\hat{\boldsymbol{\alpha}}^{up}\\ \hat{\boldsymbol{\beta}}^{up}\end{pmatrix}$, $\mathbf{x}^{2001}$, and $\mathbf{y}^{2001}$ to Eqs. (1)–(3).

4.2.4. Combined Transfer Estimation (abbreviation: *com*) (Ben-Akiva and Bolduc, 1987)

Parameters updated are expressed as:

$$\begin{pmatrix}\boldsymbol{\alpha}^{up}\\ \boldsymbol{\beta}^{up}\end{pmatrix}=\left(\left(\boldsymbol{\Sigma}^{t1}+\boldsymbol{\Delta}\boldsymbol{\Delta}'\right)^{-1}+\left(\boldsymbol{\Sigma}^{t2}\right)^{-1}\right)^{-1}\left(\left(\boldsymbol{\Sigma}^{t1}+\boldsymbol{\Delta}\boldsymbol{\Delta}'\right)^{-1}\begin{pmatrix}\hat{\boldsymbol{\alpha}}^{t1}\\ \hat{\boldsymbol{\beta}}^{t1}\end{pmatrix}+\left(\boldsymbol{\Sigma}^{t2}\right)^{-1}\begin{pmatrix}\hat{\boldsymbol{\alpha}}^{t2}\\ \hat{\boldsymbol{\beta}}^{t2}\end{pmatrix}\right) \quad (7)$$

where, $\begin{pmatrix}\hat{\boldsymbol{\alpha}}^{t1}\\ \hat{\boldsymbol{\beta}}^{t1}\end{pmatrix}$, $\boldsymbol{\Sigma}^{t1}$, $\begin{pmatrix}\hat{\boldsymbol{\alpha}}^{t2}\\ \hat{\boldsymbol{\beta}}^{t2}\end{pmatrix}$, and $\boldsymbol{\Sigma}^{t2}$ are defined in Section 4.2.3, and $\boldsymbol{\Delta}=\begin{pmatrix}\hat{\boldsymbol{\alpha}}^{t2}-\hat{\boldsymbol{\alpha}}^{t1}\\ \hat{\boldsymbol{\beta}}^{t2}-\hat{\boldsymbol{\beta}}^{t1}\end{pmatrix}$.

If $\boldsymbol{\Delta}=\mathbf{0}$, then this is identical to Bayesian updating. A forecasting performance, log-likelihood on 2001 as shown in Eq. (3), is evaluated by applying $\begin{pmatrix}\hat{\boldsymbol{\alpha}}^{up}\\ \hat{\boldsymbol{\beta}}^{up}\end{pmatrix}$, $\mathbf{x}^{2001}$, and $\mathbf{y}^{2001}$ to Eqs. (1)–(3).

4.2.5. Small Sample Model (abbreviation: *sma*)

Models are estimated as shown in Section 4.1 by utilising data from *t*2. A forecasting performance, log-likelihood on 2001 as shown in Eq. (3), is evaluated by applying $\hat{\boldsymbol{\beta}}^{t2}$, $\hat{\boldsymbol{\alpha}}^{t2}$, $\mathbf{x}^{2001}$, and $\mathbf{y}^{2001}$ to Eqs. (1)–(3).

4.3. Bootstrapping

Bootstrapping, which was first proposed by Efron and Tibshirani (1993), was applied in this study as follows. First, 10000 commuting trips were randomly selected from each of three time points (1971, 1981, and 1991). From these 10000 observations, a smaller number of observations was chosen for analysis. The same 10000 observations were chosen from each time point in order to avoid the impact on forecasting performance that might occur if different

numbers of observations from each time point were used. 10000 commuting trips also were selected randomly from the 2001 dataset that was used to evaluate forecasting performance.

Three notations—*y*, *n*, and *b*—are defined as follows.

- *y*: the year when the data was collected (1971, 1981, and 1991).
- *n*: the number of observations. This study examined 12 values for *n* (100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, and 10000).
- *b*: a bootstrap repetition (*b* = 1, 2, . . ., 1000).

From each *y*, *n* observations were randomly drawn 1000 times, with replacement from 10000 commuting trips already selected from each year. (Note that for each *b*-th draw from the same *y*, the large *n* observations will contain all of the records included in the small *n* observations.) In total, 36000 (= 3 *y*'s × 12 *n*'s × 1000 *b*'s) datasets were generated.

This study examines combinations of data collection time points and sample sizes; further notations—y_1, y_2, n_1, and n_2—are defined below.

- y_1 and y_2 denote older and more recent time points, respectively.
- n_1 and n_2 denote the numbers of observations from the older and the more recent time points, respectively.

The y_1 and y_2 are chosen from the above defined three *y*'s, while n_1 and n_2 are chosen from the above defined 12 *n*'s. However, two inequalities—$y_1 < y_2$ and $n_1 \geq n_2$—must be satisfied, meaning that y_1 is older than y_2 and n_1 is larger than or equal to n_2. Realisable combinations of y_1, y_2, n_1, and n_2 are 3 × 78 = 234, which is a multiplication of three (=3 × 2/2) combinations relating to the choices of y_1 and y_2 by 78 (= 12 × 13/2) combinations relating to the choices of n_1 and n_2. For each y_1, y_2, n_1, n_2, and *b*, the procedure described in Section 4.2 is repeated. This is shown below.

(a1) Preparation: estimation by utilising n_1 observations from y_1;

(a2) Preparation: estimation by utilising n_2 observations from y_2;

(b1) Transfer scaling: updating alternative-specific constants and a scale parameter of the

model estimated in step (a1) by utilising n_2 observations from y_2;

(b2) Joint context estimation: estimation by utilising both n_1 and n_2 observations from y_1 and y_2, respectively;

(b3) Bayesian updating: updating parameters by utilising estimates in steps (a1) and (a2);

(b4) Combined transfer estimation: updating parameters by utilising estimates in steps (a1) and (a2);

(b5) Small sample model: the estimates in step (a2) are used as is; and

(c) Forecasting behaviours for 2001 by models developed in (b1)–(b5).

Four estimations are required (in steps (a1), (a2), (b1), and (b2)). Hence, in total, 4 × 1000 (*b*'s) × 234 (combinations of the y_1, y_2, n_1, and n_2) = 936000 estimations are required. However, 936000 estimations are not actually required, since, for example, estimates utilising data from 1971 are used in cases of both (y_1, y_2) = (1971, 1981) and (1971, 1991).

A forecasting performance, defined by a log-likelihood on the 2001 dataset as explained in Section 4.2, is termed *Lm* (y_1, y_2, n_1, n_2, *b*) for model updating method *m* (=*cst*, *jnt*, *bay*, *com*, and *sma*) by utilising the *b*-th repetition of n_1 and n_2 observations from y_1 and y_2, respectively. Since the *sma* utilises only n_2 observations from y_2, this can be termed explicitly as *Lsma* (•, y_2, •, n_2, *b*).[8]

4.4. Hypothesis testing

This section proposes a test comparing the forecasting performance of two model updating methods: *m*1 and *m*2. The forecasting performance is determined from log-likelihood values on the 2001 dataset when parameters estimated by the two models are applied to that dataset. This is equivalent to comparing two models estimated using the 2001 dataset, where the number of free parameters is zero since all parameter values are fixed to those calculated by the transferred models. Since the two models are not nested, a standard likelihood ratio

[8] Log-likelihood values are used for evaluation, since they are used for calculating other measures to assess transferability.

test is not applicable. Therefore, the author takes the following approach. (See, for example, Ette (1996).)

Null and alternative hypotheses are described by comparing two population parameters: the two forecasting performances produced by models $m1$ and $m2$. The present study evaluates forecasting performance using the log-likelihood on the 2001 dataset; the null hypothesis is that the two models have the same log-likelihood on the 2001 dataset, while the alternative hypothesis is that the two models do not have the same log-likelihood on the 2001 dataset, i.e., one or the other model has better forecasting performance.

This test is conducted by evaluating the differences of the log-likelihood values on 2001 dataset between two models. The present study approximates its sample distribution by bootstrapping. Specifically, the x_b (b = 1, 2, …, 1000) defined in Eq. (8) is calculated for each combination of y_1, y_2, n_1, and n_2 that satisfy $y_1 < y_2$ and $n_1 \geq n_2$.

$$x_b = Lm1\ (y_1, y_2, n_1, n_2, b) - Lm2\ (y_1, y_2, n_1, n_2, b) \tag{8}$$

Note that the same b is used for both L's and that the x_b is defined only when both L's are calculated. The calculation of x_b's is unsuccessful for some b's, which is likely to happen when n_1 and/or n_2 are small. If updating method $m1$ produces better forecasts, then x_b is more positive.

A 95 percent confidence interval of differences of the log-likelihood values on the 2001 dataset between the two models is expressed as Formula (9), where $x_{b\,(0.025)}$ and $x_{b\,(0.975)}$ represent 2.5 and 97.5 percentiles of the x_b, respectively.

$$(x_{b\,(0.025)},\ x_{b\,(0.975)}) \tag{9}$$

The null hypothesis is rejected at a five percent level of significance if $x_{b\,(0.025)} > 0.0$ or $x_{b\,(0.975)} < 0.0$.

5. Results and Discussion

This section presents estimates, the characteristics of the forecasting performance, and the results of hypothesis testing. Criteria for selecting model updating methods are proposed.

5.1. Estimates

The following dummy variables are defined: male (1 for male, 0 for female), 20 years old or older (1 if 20 years old or older, 0 if younger), 65 years old or older (1 if 65 years old or older, 0 if younger), and Nagoya (1 if origin and/or destination of the trip are in Nagoya City, 0 if not). Descriptive statistics for the variables included in the mode choice models are fully interpreted in Sanko (2014), but one point must be restated. When the number of observations is small, then extremely large shares of 20 years old or older and extremely small shares of 65 years old or older, together with the smaller share of bus users presented in Section 3, will result in poor estimates.[9]

Table 1 shows the journey-to-work multinomial logit mode choice model estimates for (y_1, y_2) = (1971, 1981) when 10000 observations are used from both time points (for the sake of brevity, the estimates for the other (y_1, y_2) are not presented). Notations of (a)'s and (b)'s correspond to those used in Section 4.3. Although the data from 2001 is used solely for validation, a model using 2001 data was estimated and presented as a reference. The results reported here are the best and are based on examining many combinations of variables. Car ownership was not taken into account as an explanatory variable, since it is very much related to mode choice (or car choice) and the two are to some extent endogenous. Throughout the present paper, the model specification reported here is utilised.[10] For the

[9] In 1971, 1981, 1991, and 2001, percentages of 20 years old or older are 94.1%, 96.2%, 97.1%, and 98.8%, respectively; percentages of 65 years old or older are 1.5%, 1.6%, 2.1%, and 3.1%, respectively.

[10] Anonymous reviewers expressed concerns about too small a number of explanatory variables in the models. One of reviewers asked if the author adopted a set of variables currently/previously used in practice in Japan. Actually, the models are developed by the author and independent from any models used in practice. Of course, the present study intends to answer research questions which will provide practical implications, but the present

reason stated in Section 3, travel cost is not included in the models.[11] Estimates of the models are fully interpreted in Sanko (2014). The forecasting performance for behaviours in 2001 (see the row labelled 'Log-likelihood on 2001 data') is, from the best, *sma*, *com*, *jnt*, *cst*, and *bay*.[12] The same order of forecasting performance was seen for the other combinations of (y_1, y_2).

***** Table 1 *****

5.2. Forecasting performance

This section examines the characteristics of forecasting performance. Only the results for the case of (y_1, y_2) = (1971, 1981) are presented. The other cases of (y_1, y_2) are noted when necessary.

Table 2 consists of 12 panels expressed as combinations of (a)–(d) and (1)–(3) notations. The (a), (b), (c), and (d) notations refer to the model updating methods *cst*, *jnt*, *bay*, and *com*, respectively. Averages and standard deviations of forecasting performance, the log-likelihood on 2001 data as defined by Lm (y_1 = 1971, y_2 = 1981, n_1, n_2, b), are depicted in panels (1) and (2), respectively. Although each updating model was estimated 1000 times for each combination of n_1 and n_2, some of the 1000 times resulted in poor estimates, especially

---

study is based on academic interests. While reviewing models developed by others provides much insight into choosing explanatory variables, relying on them too much should be avoided, since this makes the present research a post-project evaluation. For the purpose of the present study, the author chose models that he considered the best. Another reviewer raised a question relating to sensitivity to the model specification, which is beyond the scope of the present study. See footnotes 3 and 4 for more details. There are infinite model specifications, none of which is perfect, meaning that no sensitivity analysis can persuade all researchers and practitioners. However, the impact of the model specifications on temporal transferability receives huge attention, so this direction is a topic for future study.

[11] The author's approach of not considering the travel cost is empirically justified by Sanko et al. (2013). They estimated commuting mode choice models between car and public transportation for the Nagoya metropolitan area. They found that the car cost parameter was not estimated significantly and that the public transportation cost parameter was estimated with the wrong sign and was excluded from the model.

[12] Mean absolute error, which calculates absolute differences between predicted and actual shares for each mode and sums up the absolute differences, sometimes is used as an accuracy measure. Although not presented in the study, the mean absolute error produced the same order of model superiority for all combinations of data collection time points.

when the sample sizes were small. The averages and standard deviations are calculated after excluding the poor estimates. The remaining number of bootstrap repetitions, after excluding the problematic estimates, also is presented in panels (3) of Table 2. Results for *m* = *sma* are not presented in separate panels, but they are depicted in all panels as a reference. In each panel, the horizontal axis is the number of observations from the more recent time point of 1981 (n_2).

***** Table 2 *****

In each panel, 12 lines are drawn for 12 n_1's; a line for *m* = *sma* also is drawn as a reference. The lines are self-explanatory, since they are drawn only when $n_1 \geq n_2$ is satisfied. Only the line for n_1 = 100 terminates at n_2 = 100, only the line for n_1 = 200 terminates at n_2 = 200, and so forth. Two lines terminate at n_2 = 10000: the lines for n_1 = 10000 and *m* = *sma*; dashed lines are used for *m* = *sma*.

In panels (a1), (b1), (c1), and (d1), the lines for *cst*, *jnt*, *bay*, *com*, and *sma* rise towards the right. When the sample size from the older time point is the same, then the larger the sample size from the more recent time point, the better the forecasting. On the other hand, when the sample size from the more recent time point is the same, the larger the sample size from the older time point, the better the forecasting for *cst*. This is because a large n_1 improves the transferability of the parameters related to explanatory variables. The same is true for *jnt* when n_2 is small, but not when n_2 is large. For *jnt*, the impacts of a large n_1 are twofold: (i) a large n_1 makes estimates more accurate; and (ii) a constraint of $\boldsymbol{\beta}^{t1} = \boldsymbol{\beta}^{t2}$ together with a large n_1 make the parameter estimates being influenced more by the older data, which is less transferable. The (i) and (ii) outperform for a small and large n_2, respectively. For *bay*, the larger the sample size from the older time point, the worse the forecast when the sample size from the more recent time point is the same. This is because parameters with a large n_1 are estimated more precisely, and updated parameters weigh

more on the parameters from the older time point, which is less transferable. For *com*, n_1 does not affect the lines very much. The lines for *sma* sometimes are drawn above those for the other updating methods. This means that a use of only the more recent data produces better forecasting than using both datasets. Cases of other combinations of the older and more recent time points are noted: *bay* sometimes has better forecasts than *sma* when n_2 is small in the case of (y_1, y_2) = (1981, 1991).

In panels (a2), (b2), (c2), and (d2), the lines for *cst*, *jnt*, *bay*, *com*, and *sma* generally fall towards the right. When the sample size from the older time point is the same, the larger the sample size from the more recent time point, the smaller the standard deviations of the forecasting performance. (Some exceptions were found in panel (b2), which might be the limitations of bootstrapping with 1000 repetitions.) On the other hand, in general, when the sample size from the more recent time point is the same, the larger the sample size from the older time point, the smaller the variance.

In panels (a3), (b3), (c3), and (d3), the lines for *cst*, *jnt*, *bay*, *com*, and *sma* generally rise towards the right. When the sample size from the older time point is the same, the larger the sample size from the more recent time point, the larger the percentage of 1000 estimations that are free of poor estimations. On the other hand, the lines with larger n_1 generally appear above those with smaller n_1. This relates to the model specifications. As shown in Table 1, the models have alternative-specific constants in two of three alternatives. Therefore, the dataset must include at least one choice from each alternative. Dummy variables face a similar constraint. The above requirement often is not satisfied when the number of observations is small.[13]

The characteristics of forecasting performance presented in Table 1 summarises cases where utilising data from two time points has merits, as follows:

[13] The two requirements of (a) alternative-specific constants and (b) dummy variables apply differently to five updating methods. For *sma*, both requirements apply to data from y_2. For *bay* and *com*, both requirements apply to data from y_1 and y_2 separately. For *cst*, both requirements apply to data from y_1, and (a) applies to data from y_2. For *jnt*, (a) applies to data from y_1 and y_2 separately, but (b) applies to pooled data from y_1 and y_2. Therefore, the remaining numbers of repetitions are larger in *cst* and *jnt* than *sma* but smaller in *bay* and *com* than *sma* for the dataset used in the present study.

- The *cst* and *jnt* are more appropriate than *sma* when n_1 and n_2 are larger and smaller, respectively, where the forecasting performance is better on average (panels (a1) and (b1)); its standard deviation is smaller (panels (a2) and (b2)); the remaining *b*'s is larger (panels (a3) and (b3)).
- The *bay* is more appropriate than *sma* when n_2 is small for (y_1, y_2) = (1981, 1991), where the forecasting performance is better on average; its standard deviation is smaller; however, the remaining *b*'s is smaller.

5.3. Results of hypothesis testing

Model updating methods are tested pairwise with respect to their forecasting performance; the results are presented in Table 3. For ease of presentation, the models are reordered as *sma*, *com*, *jnt*, *cst*, and *bay*. All test results are based on x_b, which is calculated when the forecasting performances from both updating methods are available. Note that, although panels (a3), (b3), (c3), and (d3) in Table 2 present the number of *b*'s remaining after excluding problematic estimates for each model updating method, the test results here consider two updating methods at the same time. Only results for the case of (y_1, y_2) = (1971, 1981) are presented. The other cases of (y_1, y_2) are noted in the main text when necessary.

***** Table 3 *****

In each panel, the horizontal and vertical axes represent the number of observations from the older time point of 1971 and the more recent time point of 1981, respectively. Parts shaded in grey in the upper left of each panel represent areas where the sample size from the more recent time point is greater than that from the older time point. Therefore, the present study focuses on the lower right of each panel. The cells are shaded in black if the updating method shown in rows labelled (1)–(5) produced statistically significantly better forecasts at five percent levels of significance than that shown in columns labelled (a)–(e). For example,

the crossing of the '(1) *sma*' row and the '(c) *jnt*' column means that the *sma* produced statistically significantly better forecasts than the *jnt* when n_1 = 2000 and n_2 = 900–2000, and when n_1 = 10000 and n_2 = 800–10000.

Interestingly, the panels containing cells shaded in black appear only in the upper right of Table 3. Therefore, if one updating method produces statistically significantly better forecasts than another updating method for at least one combination of n_1 and n_2, then the latter updating method never produces statistically significantly better forecasts than the former updating method for any combination of n_1 and n_2. Therefore, the statistical tests produced a clear ranking of the updating methods from best to worst: *sma*, *com*, *jnt*, *cst*, and *bay*. The same order was found for (y_1, y_2) = (1971, 1991) and (1981, 1991).

5.4. Criteria for selecting model updating methods

Updating methods utilising data from two points in time never produced statistically significantly better forecasts than *sma*, although they sometimes produced better forecasts on average, as shown in Table 2. This section proposes criteria for selecting model updating methods using the results both with and without statistical tests.

Updating methods that most frequently produced the highest forecasting performance are presented in Fig. 2.[14] An updating method producing the highest forecasting performance on average can be chosen intuitively as the best method. However, averages have at least two drawbacks. First, they are sensitive to outliers. Second, averages depend on how poor estimates are handled. The averages in Table 2 are calculated after excluding poor estimates and are unsuitable for comparing two models, where one model produces poor estimates more frequently than the other model, but produces a higher average in forecasting performances after excluding poor estimates. Methods that most frequently produce the highest forecasting performance do not have these drawbacks. The results for all three combinations of y_1 and y_2 are presented.

[14] Out of 1000 bootstrap repetitions, repetitions where all five updating methods produced poor estimates are removed from the analysis. In the repetitions remained, the updating method producing the highest forecasting performance most frequently is reported.

***** Fig. 2 *****

Summarised below are cases where updating methods other than *sma* are chosen.

- For (y_1, y_2) = (1971, 1981) and (1971, 1991), when n_2 is small, *jnt* is chosen.
- For (y_1, y_2) = (1981, 1991), when n_2 is small; *jnt* and *cst* are chosen when n_1 is small and large, respectively.

As discussed in Section 3, the shares of travel modes have changed substantially between 1971 and 1981, but have changed less since then. The differences in contexts are the largest in (1971, 1991) followed by (1971, 1981) and (1981, 1991). Similarities in contexts between 1981 and 1991 make parameters other than alternative-specific constants transferable, so *cst* most frequently produces the best forecasting.

The results presented in Fig. 2 are consistent with the studies by Karasmaa and Pursula and Badoe and Miller (see Section 2), although the present study sets aside data from 2001 only for validation purposes, which differs from their studies. In their studies, using data from the older time point does little to improving forecasting performance when there are at least 400 observations from the more recent time point. Fig. 2 shows that the small sample model most frequently produced the best forecasting performance when the sample sizes of the more recent data are at least 300, 300–400, and 400–500 in panels (a), (b), and (c), respectively. Although previous studies have not examined cases where the sample size from the more recent data was less than 400, the present study thoroughly investigated cases where the sample size from the more recent data was 100–1000 with an interval of 100 and reached similar conclusions. In Badoe and Miller, joint context estimation produced the highest forecasting performance, followed by transfer scaling, combined transfer estimation, and small sample model, when the only LOS variables are included as explanatory variables and there were 400 or 800 observations from the more recent time point. Fig. 2 also shows

cases where joint context estimation and transfer scaling most frequently produced the highest forecasting performance. Irrespective of the number of observations from the older time point, the minimum sample size from the more recent time point where the small sample model most frequently produces the highest forecasting performance was surprisingly stable (300, 300–400, and 400–500 in panels (a), (b), and (c), respectively). This can be useful for finding quantitative insights, such as the specific number of observations required, irrespective of the number of observations from the older time point.

The criteria for selecting model updating methods are summarised in Fig. 3. Although the author found a clue to identifying a quantitative recommendation in the previous paragraph, the present study analyses contexts with many assumptions and the author infers implications that can be applicable to more general contexts. Therefore, the recommendation uses a qualitative rather than quantitative expression, which may have meaning only in the context of the present study. When the number of observations from the more recent time point is large, using only the more recent data is recommended. When the sample size from the more recent time point is small, using both older and more recent data may improve forecasting. When the difference in the contexts of the two time points is large, then joint context estimation is recommended. When the difference in the contexts of the two time points is small, then either joint context estimation or transfer scaling can be used; if the sample size from the older time point is small, then joint context estimation is recommended, but if the sample size from the older time point is large, then transfer scaling is recommended. Bayesian updating and combined transfer estimation are not recommended for any combination of data collection time points and sample sizes.

***** Fig. 3 *****

6. Conclusions

This study compares and statistically tests various model updating methods in terms of their

forecasting performance and proposes criteria for selecting model updating methods. The updating methods examined are transfer scaling, joint context estimation, Bayesian updating, combined transfer estimation, and small sample model. The former four methods utilise data from two time points, while the last method utilises data from only the more recent time point. This study considers the small sample model as a special case of model updating method, which poses zero weight on the older data. The data collection time points and the numbers of observations are two dimensions of interests in the present study, and the statistical tests and criteria are examined from these two aspects. This study examines one empirical case utilising repeated cross-sectional data: commuting mode choice behaviours in Nagoya, Japan. Using a bootstrapping technique, the study evaluated the effects of combining data collection time points (two time points from 1971, 1981, and 1991) and combining the numbers of observations from the two time points (12 different numbers of observations ranging from 100 to 10000) on forecasting performance for 2001 data.

Although not tested statistically, forecasting performance using the bootstrap technique has the following characteristics.

- When the number of observations from the older time point is the same, the larger the number of observations from the more recent time point, the better the forecast for any of the model updating methods.
- For transfer scaling, when the number of observations from the more recent time point is the same, then the larger the number of observations from the older time point, the better the forecast.
- For joint context estimation, when the number of observations from the more recent time point is the same and small, then the larger the number of observations from the older time point, the better the forecast. When the number of observations from the more recent time point is the same and large, then the larger the number of observations from the older time point, the worse the forecast.
- For Bayesian updating, when the number of observations from the more recent time point

is the same, then the larger the number of observations from the older time point, the worse the forecast.

The forecasting performance of the model updating methods was statistically tested pairwise. The model updating method ranking is clear based on forecasting performance. The small sample model was ranked the highest, followed by combined transfer estimation, joint context estimation, transfer scaling, and Bayesian updating. A higher ranked method produced statistically significantly better forecasts than a lower ranked method in some combinations of numbers of observations from two points in time; however, a lower ranked method never produced statistically significantly better forecasts than a higher ranked method in any combinations of numbers of observations from two points in time.

Although updating methods utilising data from two points in time never produced statistically significantly better forecasts than the small sample model, they sometimes produced better forecasts than the small sample model without statistical significance. To make the best use of the results with and without statistical tests, the present study examines which model updating method most frequently produces the highest forecasting performance and proposes criteria for selecting model updating methods. The criteria are listed below.

- When the number of observations from the more recent time point is sufficiently large, use the small sample model.
- When the number of observations from the more recent time point is insufficiently large, the following criteria apply. If the difference in the contexts between the two time points is large, then use joint context estimation. If the difference in the contexts between the two time points is small, then use joint context estimation when the number of observations from the older time point is small and transfer scaling when the number of observations from the older time point is large.
- Irrespective of the number of observations from the older time point, the minimum sample size from the more recent time point where the small sample model most frequently produces the highest forecasting performance was surprisingly stable.

Combining results from Sanko (2017) and the present study, the use of only the more recent data outperforms both the use of only the older data as well as the use of both datasets. This supports the value of recent data. Although the results under many assumptions in the present study must be interpreted with care, there is a case in which a use of 500 observations from the more recent data produces better forecasting than the use of both 10000 and 500 observations from the older and more recent time points, respectively. These are not magic numbers applicable to any context, but they do suggest that smaller surveys conducted frequently rather than larger surveys conducted infrequently is a topic worth investigating.

Other topics must be addressed in a future study. First, this study focuses on data collection time points and numbers of observations. Other dimensions might affect transferability and so should be investigated (see footnote 3). Second, this study examines three combinations of older and recent time points, but the data comes from the same metropolitan area and the target forecast year is 2001 throughout the three combinations. Also, only mode choice models are examined. Future studies might analyse other contexts—e.g., the use of data from other metropolises, and other choices (such as destination choices) using the same Nagoya data. Third, the data used in the present study comes from 1971–2001, and the most recent data was already more than 15 years old. Therefore, future studies might look at more recent data. The main aim of the present study is to compare the forecasting performance of model updating methods, and the present study utilises available datasets. Since comparing updating methods itself is novel, investigating more recent travel behaviours was less of a concern. However, some ‘reverse trends,’ such as a decrease in car ownership, has been observed. It is worth investigating whether the conclusions in the present study are robust in the context of a reverse trend. The author believes that the impact of sample size on forecasting performance is an important research topic, especially when transport surveys must be conducted under limited budgets.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 25380564, 16K03931, and 19H01538. The author acknowledges the use of data provided by the Chubu Regional Bureau, Japan's Ministry of Land, Infrastructure, Transport and Tourism, and the NUTREND (Nagoya University TRansport and ENvironment Dynamics) Research Group. This paper is based on presentations at the 95th Annual Meeting of the Transportation Research Board in Washington, D.C., U.S.A., in January 2016 and the 52nd Infrastructure Planning Conference of the Japan Society of Civil Engineers, Akita, Japan, in November 2015.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Atherton, T.J., Ben-Akiva, M.E., 1976. Transferability and updating of disaggregate travel demand models. Transportation Research Record 610, 12–18.

Badoe, D.A., Miller, E.J., 1995a. Analysis of the temporal transferability of disaggregate work trip mode choice models. Transportation Research Record 1493, 1–11.

Badoe, D.A., Miller, E.J., 1995b. Comparison of alternative methods for updating disaggregate logit mode choice models. Transportation Research Record 1493, 90–100.

Badoe, D.A., Wadhawan, B., 2002. Jointly estimated cross-sectional mode choice models: specification and forecast performance. ASCE Journal of Transportation Engineering 128 (3), 259–269.

Ben-Akiva, M., Bolduc, D., 1987. Approaches to model transferability and updating: the combined transfer estimator. Transportation Research Record 1139, 1–7.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Ette, E.I., 1996. Comparing non-hierarchical models: application to non-linear mixed effects

modeling. Computers in Biology and Medicine 26 (6), 505–512.

Fox, J., Daly, A., Hess, S., Miller, E., 2014. Temporal transferability of models of mode-destination choice for the Greater Toronto and Hamilton Area. The Journal of Transport and Land Use 7 (2), 41–62.

Karasmaa, N., 2003. The Transferability of Travel Demand Models: An Analysis of Transfer Methods, Data Quality and Model Estimation. Ph.D. Thesis, Helsinki University of Technology, Transportation Engineering. Publication 106. Espoo, Finland.

Karasmaa, N., Pursula, M., 1997. Empirical studies of transferability of Helsinki metropolitan area travel forecasting models. Transportation Research Record 1607, 38–44.

Ortúzar, J.D., Willumsen, L.G., 2011. Modelling Transport. 4th edn. John Wiley and Sons, Chichester.

Sanko, N., 2014. Travel demand forecasts improved by using cross-sectional data from multiple time points. Transportation 41 (4), 673–695.

Sanko, N., 2015. Should small samples from recent time point be used with older data? applicability of updating models by transfer scaling. Presented at the hEART 2015 - 4th Symposium of the European Association for Research in Transportation, Copenhagen, Denmark.

Sanko, N., 2017. Temporal transferability: trade-off between data newness and the number of observations for forecasting travel demand. Transportation 44 (6), 1403–1420.

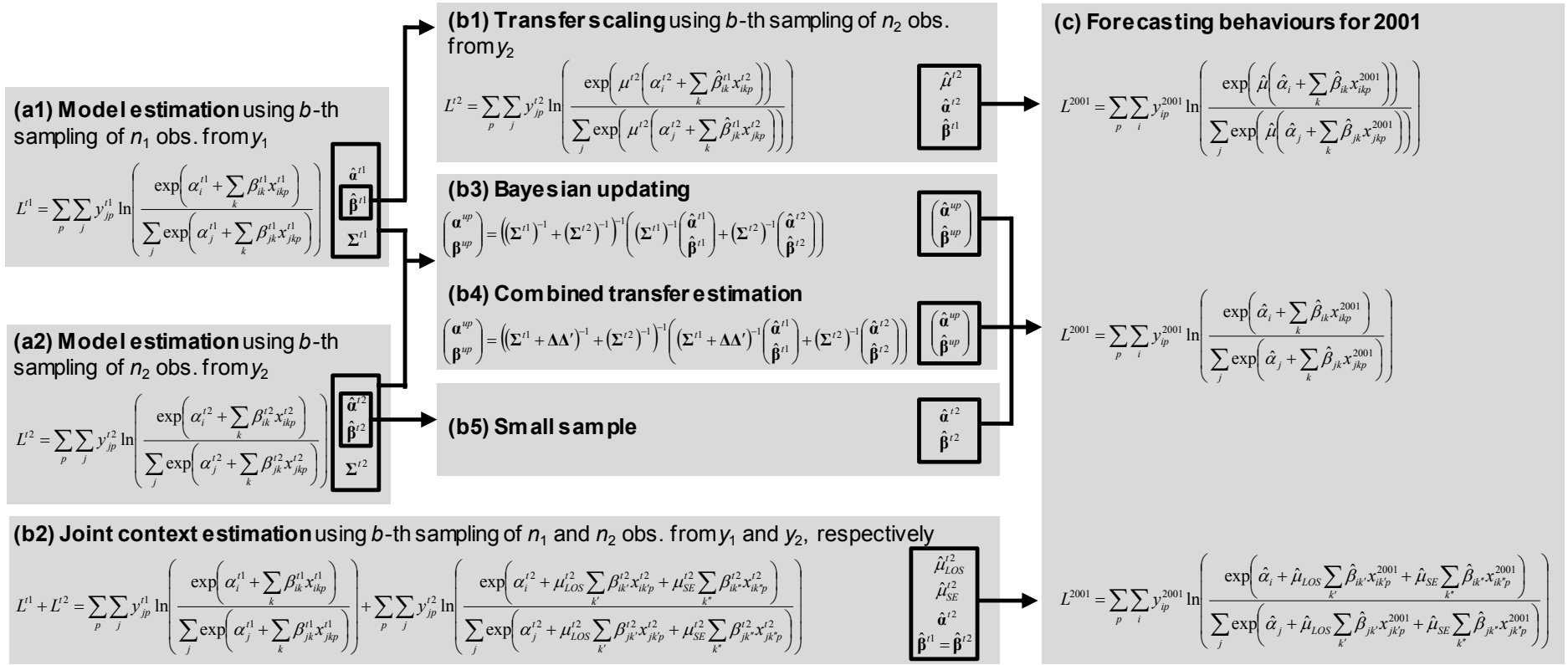
Sanko, N., 2018. Transfer scaling in travel demand forecasting models: its usefulness based on data collection time points and the numbers of observations. Journal of Japan Society of Civil Engineers Ser. D3 (Infrastructure Planning and Management) 74 (1), 21–34. (in Japanese)

Sanko, N., Morikawa, T., Kurauchi, S., 2013. Mode choice models' ability to express intention to change travel behaviour considering non-compensatory rules and latent variables. IATSS Research 36 (2), 129–138.

Sikder, S, 2013. Spatial Transferability of Activity-Based Travel Forecasting Models. Ph.D.

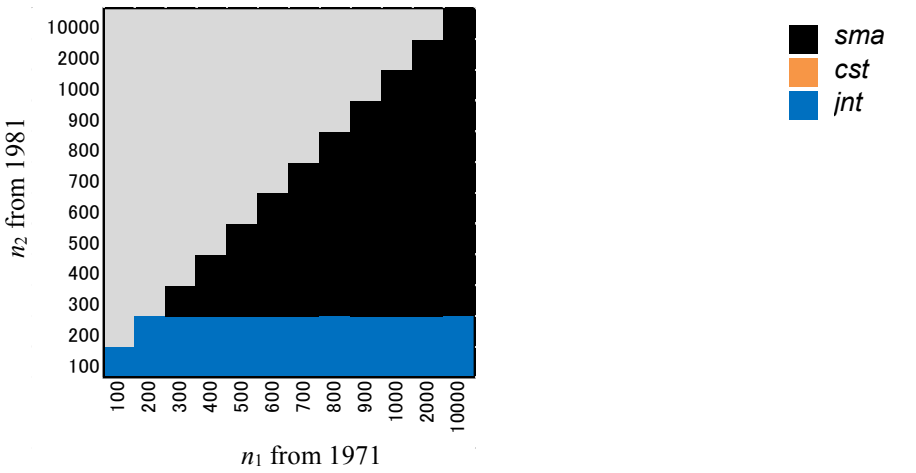
Dissertation, University of South Florida, Tampa, U.S.A.

Train, K.E., 1979. A comparison of the predictive ability of mode choice models with various levels of complexity. Transportation Research Part A 13 (1), 11–16.

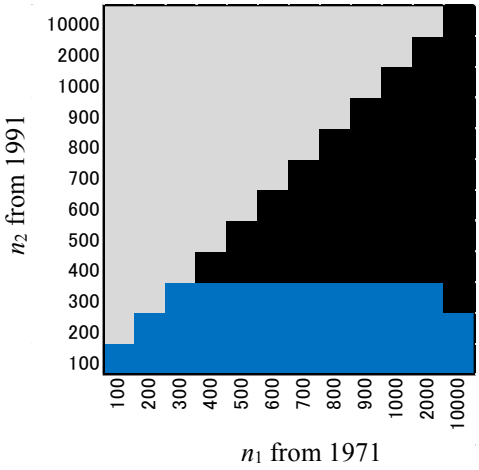


Note: The (a)–(c) notations correspond to those in Section 4.3. The procedure shown here is repeated for each combination of y_1, y_2, n_1, n_2, and b.

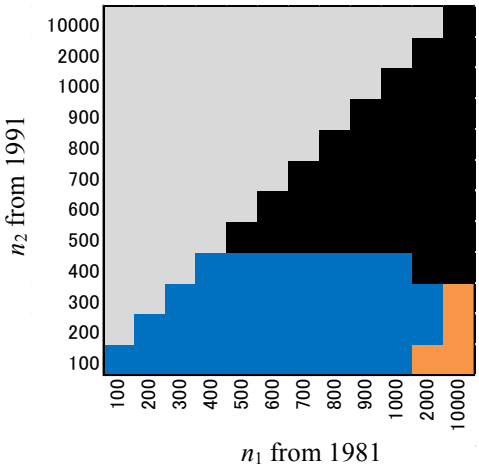
Fig. 1 Research methodology.



(a) y_1=1971 and y_2=1981



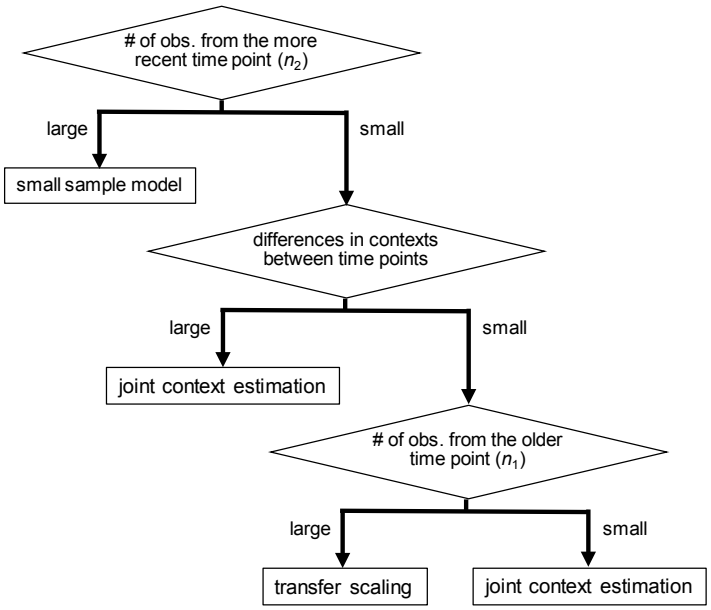
(b) y_1=1971 and y_2=1991



(c) y_1=1981 and y_2=1991

Note: Grey cells indicate $n_1 < n_2$, which is not of interest in the present study. Other coloured cells indicate the three updating methods that most frequently produce the highest forecasting performance: black for *sma*; orange for *cst*; and blue for *jnt*. Neither *bay* nor *com* produced the highest forecasting performance for any cell.

Fig. 2 Updating methods that most frequently produce the highest forecasting performance.



Note: The criteria utilise qualitative expression, i.e. ‘large’ and ‘small’, rather than quantitative expression. Bayesian updating and combined transfer estimation are not recommended in any combinations of data collection time points and sample sizes.

Fig. 3 Criteria for selecting model updating methods.

Table 1 Estimates

Variables	Ref. 1971[a]	(b1) *cst*	(b2) *jnt*	(b3) *bay*	(b4) *com*	(b5) *sma*	Ref. 2001[a]
Constant (B)	0.127*	--	*0.177**	-0.0698	-0.391	--	-1.03*
Constant (C)	-1.15*	--	*-0.768**	-0.864	-0.646	--	0.560*
Travel time [hr]	-0.606*	-0.606	-0.576*	-1.09	-1.81	-1.81*	-2.60*
Male dummy (R)	0.577*	0.577	0.634*	0.540	0.787	0.787*	0.511*
Male dummy (C)	1.97*	1.97	1.84*	1.84	2.17	2.17*	1.38*
20 years old or older dummy (C)	0.900*	0.900	0.768*	0.884	0.764	0.764*	0.511*
65 years old or older dummy (B)	1.91*	1.91	1.47*	1.50	1.37	1.37*	0.561*
Nagoya dummy (C)	-1.12*	-1.12	-1.29*	-1.45	-1.77	-1.77*	-2.21*
Constant (B) at 1981	--	-0.441*	-0.439*	--	--	-0.392*	--
Constant (C) at 1981	--	-0.986*	-1.04*	--	--	-0.645*	--
Scale parameter	--	1.27**	--	--	--	--	--
Scale parameter for LOS	--	--	3.34**	--	--	--	--
Scale parameter for SE	--	--	1.24**	--	--	--	--
N	10000	10000	20000	n.a.	n.a.	10000	10000
L(**β**)	-7776.86	-6138.88	-13791.50	n.a.	n.a.	-5985.02	-4716.28
L(**0**)	-8948.26	-8593.88	-17542.13	n.a.	n.a.	-8593.88	-8159.63
Adj. rho-squared	0.130	0.285	0.213	n.a.	n.a.	0.303	0.421
Log-likelihood on 2001 data	-6521.95	-5672.04	-5317.23	-5776.43	-5225.85	-5225.15	n.a.

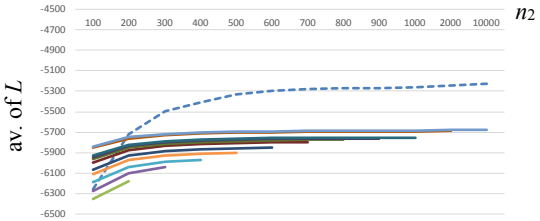
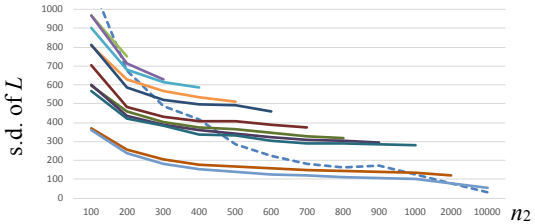
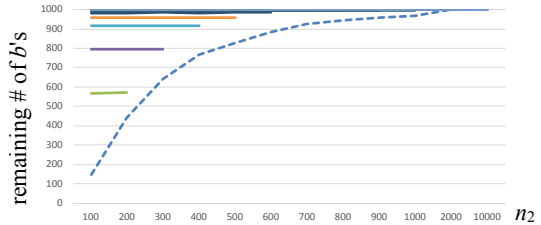
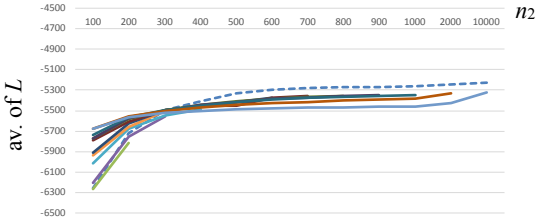
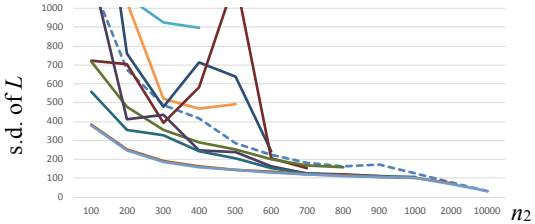
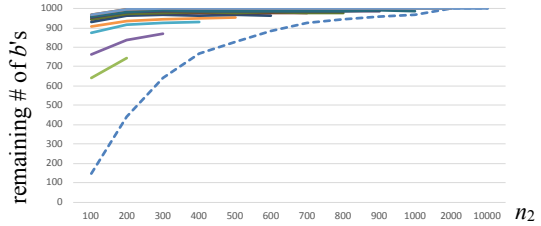
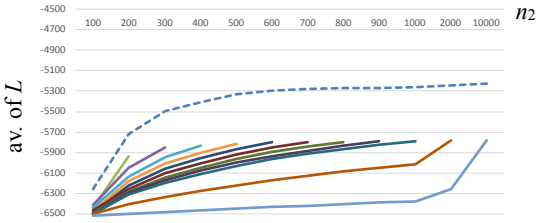
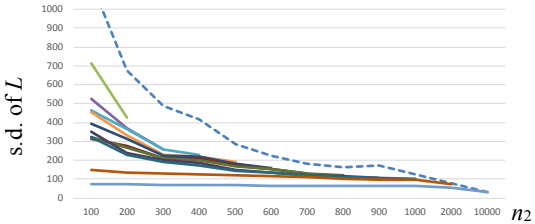
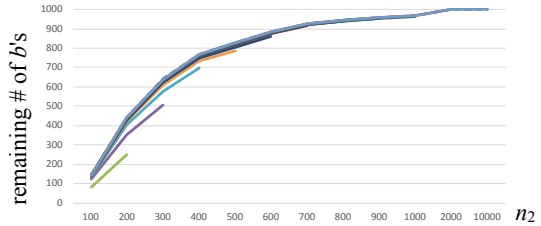
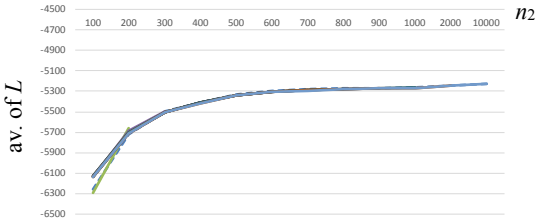
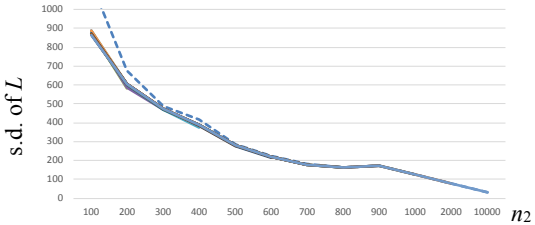
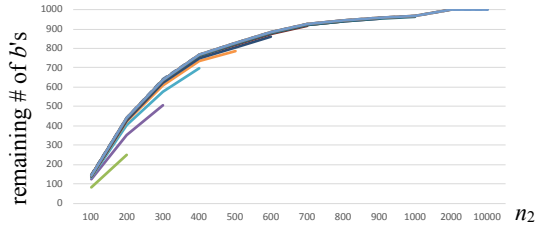
Note 1: (R), (B), and (C) notations refer to alternative-specific variables for rail, bus, and car, respectively. Variables without notations are generic.

Note 2: Notations of (a)'s and (b)'s correspond to those used in Section 4.3. The 1971 model is estimated in (a1) and the (b5) *sma* model is estimated in (a2). Italicised figures in models (b2) are not used for forecasting.

[a] Models presented as references and not used for forecasting. The 1971 model is included, since it is required to produce models of *cst*, *bay*, and *com*. The 2001 model is included, since estimates closer to this model will produce higher log-likelihood on 2001 data. Note that the parameters for the *cst* and *jnt* models must be adjusted by scale parameters before they can be compared to those in the 2001 model.

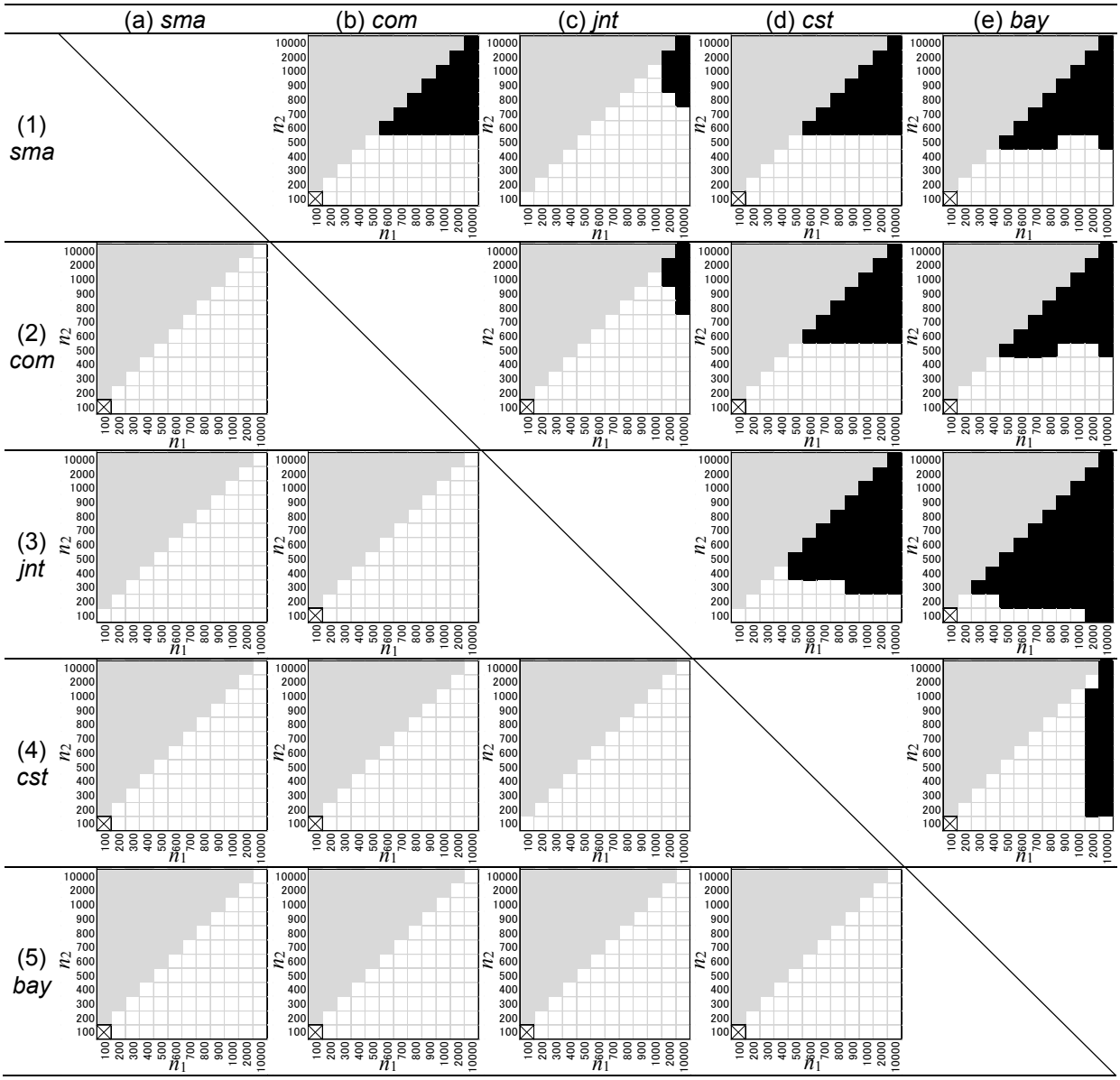
* and ** notations refer to estimated parameters. Note that all parameters indicated with * and ** notations differ from zero and one at a five percent level of significance, respectively. Parameters without these notations are not estimated, but are borrowed from model (a1) or calculated by utilising models of (a1) and (a2).

Table 2 Characteristics of forecasting performance using 1971/1981 datasets.

	(1) Averages	(2) Standard deviations	(3) Remaining # of repetitions
(a) *cst*			
(b) *jnt*			
(c) *bay*			
(d) *com*			

Note: The intervals for the horizontal axis, representing the number of observations from 1981 (n_2), are not equal. Lines for both n_1 = 10000 and m = *sma* are drawn in a domain of n_2 = 100–10000, but dashed lines indicate *sma*. The lines for m = *sma* are identical in each column, e.g., panels (a1), (b1), (c1), and (d1).

Table 3 Statistical tests comparing model updating methods using 1971/1981 datasets.



Note: In each panel, the horizontal and vertical axes represent the number of observations from 1971 (n_1) and 1981 (n_2), respectively. Grey cells indicate $n_1 < n_2$, which is not of interest in the present study. Black cells indicate that models noted as (1)–(5) produced statistically significantly better forecasting than those noted as (a)–(e). Cells that are crossed out indicate that fewer than 40 x_b's were successfully calculated and were unsuitable to produce 2.5 and 97.5 percentiles. Therefore, they are excluded from the analysis.