DREAM
Danish Research Institute for
Economic Analysis and Modelling

SMILE

Dwelling choice transition probabilities

Joint vs sequential modelling

Maria Juul Hansen

Background paper
16 April 2026

www.dreamgruppen.dk

1 Background

The purpose of this note is to assess whether there is a difference in predictive accuracy between two approaches to estimating transition probabilities in two dimensions, here for the choice of dwelling type and dwelling usage when moving. The comparison examines the difference between (i) estimating the joint distribution as a single outcome (bundle or joint model over all combinations of (dwelling type, dwelling usage)), and (ii) estimation via sequential conditional models (dwelling type conditional on dwelling usage, then dwelling usage conditional on dwelling type, and vice versa). All models are trained in R using light gradient boosting machine (LightGBM)[1] on register data from Statistics Denmark at the individual level. The question is whether the ordering or factorization of the same joint distribution leads to systematically different predictions or errors when the models are evaluated in a comparable way. The dwelling type and usage categories in the data are listed in Tables 1.2 and 1.1.

Table 1.1: Dwelling types: Danish and English labels (support of $T_{i,t}$)

Danish	English
Ejerbolig	Owner-occupied
Almen	Social (non-profit) housing
Andel	Co-op housing
Udlejning offentlig	Public rental
Udlejning privat	Private rental

Note: Co-op housing (*andel*) in Denmark is a tenure form where residents hold a share in a housing association that confers the right to occupy a specific dwelling, rather than owning the unit outright as in full owner-occupation.

Table 1.2: Dwelling usages: Danish and English labels (support of $A_{i,t}$)

Danish	English
Stuehus	Main house (primary dwelling on an agricultural property)
Parcelhus	Detached house
Række-, kæde-, dobbelthus	Terraced/semi-detached/linked house
Etageboligbebyggelse	Apartment building
Kollegium	Student housing
Anden helårsbeboelse	Other permanent dwelling
Erhvervsbolig/erhvervsenhed	Commercial dwelling/unit
Døgninstitution	Residential institution (24-hour care)
Fritidshus	Holiday house

[1] R version 4.3.2. R package `lightgbm` version 4.3.0.

2 Theoretical framework

2.1 Notation

Dwelling type and usage are treated as random variables and denoted by uppercase letters $T_{i,t}$ and $A_{i,t}$, each carrying both an individual index i and a time index t. Let τ denote the indices for $T_{i,t}$ and a the indices for $A_{i,t}$. Observed outcomes in the data are written $\tau^*_{i,t}$ and $a^*_{i,t}$, i.e. the realised outcomes satisfy $T_{i,t} = \tau^*_{i,t}$ and $A_{i,t} = a^*_{i,t}$. Simulated draws are distinguished by the subscript sim, with the relevant individual and time index taken from context. Throughout, $P(\cdot)$ indicates the probability of an event. Finally, $\mathbf{X}_{i,t}$ is the feature vector for individual i at time t, i.e. covariates measured in period t when modelling post-move outcomes $(T_{i,t+1}, A_{i,t+1})$.

2.2 Model specifications

For a discrete joint distribution over two dimensions, the basic factorisation holds for each individual i and for all category indices τ and a. With outcomes $(T_{i,t+1}, A_{i,t+1})$ after the move and covariates $\mathbf{X}_{i,t}$[2],

$$\begin{aligned} P(T_{i,t+1} = \tau,\, A_{i,t+1} = a \mid \mathbf{X}_{i,t}) &= P(A_{i,t+1} = a \mid T_{i,t+1} = \tau,\, \mathbf{X}_{i,t})\, P(T_{i,t+1} = \tau \mid \mathbf{X}_{i,t}) \\ &= P(T_{i,t+1} = \tau \mid A_{i,t+1} = a,\, \mathbf{X}_{i,t})\, P(A_{i,t+1} = a \mid \mathbf{X}_{i,t}). \end{aligned} \tag{1}$$

From (1) it follows that modelling the choice of dwelling type and dwelling usage can be split into two sequential models where type is conditioned on usage and usage on type (or vice versa), or modelled jointly from the outset. The four models examined in this note are therefore Model 1, Model 2a, Model 2b, and Model 3a, Model 3b, see Table 2.1.

Table 2.1: Overview of model specifications

Model name	Model specification
Model 1	$P(T_{i,t+1} = \tau,\, A_{i,t+1} = a \mid \mathbf{X}_{i,t})$ for all (τ, a)
Model 2a	$P(A_{i,t+1} = a \mid \mathbf{X}_{i,t})$ for all a
Model 2b	$P(T_{i,t+1} = \tau \mid A_{i,t+1} = a,\, \mathbf{X}_{i,t})$ for all (τ, a)
Model 3a	$P(T_{i,t+1} = \tau \mid \mathbf{X}_{i,t})$ for all τ
Model 3b	$P(A_{i,t+1} = a \mid T_{i,t+1} = \tau,\, \mathbf{X}_{i,t})$ for all (τ, a)

[2] The full set of variables in $\mathbf{X}_{i,t}$ are: family type (single male, single female or couple), functions of highest completed education among adult household members, origin indicators for adult household members, labor market attachment indicators for adult household members, indicator for existence of children in the household, dwelling type, usage, size, and construction year for the dwelling in period t, indicators for current city size and municipality and indicator for municipality chosen in the next period $t + 1$.

2.3 Propagating errors

In Models 2b and 3b, estimation uses the observed value of the other dimension at $t+1$ as a regressor: Model 2b conditions on $A_{i,t+1} = a^*_{i,t+1}$, Model 3b on $T_{i,t+1} = \tau^*_{i,t+1}$. Models 2a and 3a condition only on $\mathbf{X}_{i,t}$. The joint model (Model 1) estimates probabilities $P(T_{i,t+1} = \tau, A_{i,t+1} = a \mid \mathbf{X}_{i,t})$ without conditioning on any outcomes from $t + 1$.

The crucial difference compared to simulations in SMILE is the chain of conditions: the next stage conditions on outcomes in $t+1$ that are themselves simulated, not observed. Inconsistencies in an early stage of the chain of estimations therefore propagate through the simulated chain of events. If, for example, the probability of rental housing is positively biased, too many households are assigned rental housing in the simulation. Even if $P(A_{i,t+1} = a \mid T_{i,t+1} = \tau, \mathbf{X}_{i,t})$ is well estimated across (τ, a), this alone can lead to overprediction of e.g. social housing, because social housing is relatively frequent among rentals.

Such issues have previously characterized the transition probabilities when they were estimated using conditional inference trees (CTREE) and therefore required some manual adjustments. This note asks whether a similar challenge can be detected with LGBM when simulating from R. If so, the bundle model would be expected to outperform the sequential models in terms of accuracy.

3 Basis for comparison

3.1 Data and hyperparameters

The evaluation keeps the data, the train–validation–test split, and the covariate specification $\mathbf{X}_{i,t}$ fixed across approaches. The data used for training and validation are a random sample of 80 % of the original population data of movers, while test data comprise the remaining 20 %[3]. The training/validation data is further split into training data (80 %) and validation data (20 %). The validation set is used to assess predictive performance during training when early stopping is applied.[4] Hyperparameters are chosen following Cursor's AI recommendations for a standard parameter set for a sequential model and a joint model, respectively, see Table 3.1. Because the joint model

[3]The population data consists of household members that are not considered children in the household by Statistics Denmark. In some cases children and young teenagers show up as adults in the household. This note has not taken account of this, which will be reflected in the graphs shown in coming sections, but results are not expected to be affected. The data are further restricted to individuals living in Denmark in both the beginning and end of the year, but with a change of address. Only one change of address is counted each year. Note that while individual-level data are used in this comparison exercise, data going into SMILE is aggregated at the family level to ensure all family members move simultaneously and to the same dwelling. In addition, dwelling choice events are simulated in SMILE conditional on reasons for moving (e.g. separate probabilities for people moving from the parental home or moving due to divorce or partnership formation). In the comparison underlying this note, conditioning is only as explained above, not on the reason for moving. Conclusions are not expected to be affected by this difference either.

[4]Early stopping ends training when a chosen metric does not improve for a specified number of rounds.

has 45 classes versus 9 for the usage model and 5 for the type model, hyperparameters are also adjusted to reflect this difference in complexity.

Table 3.1: LightGBM hyperparameters: bundle vs. sequential

Parameter	Bundle	Sequential
objective	multiclass	multiclass
metric	multi_logloss	multi_logloss
boosting	gbdt	gbdt
learning_rate	0.01	0.03
num_leaves	127	63
max_depth	8	7
min_data_in_leaf	500	300
min_sum_hessian_in_leaf	20	10
min_gain_to_split	0.01	0.01
feature_fraction	0.6	0.7
bagging_fraction	0.6	0.7
bagging_freq	1	1
lambda_l1	0.2	0.1
lambda_l2	5.0	3.0
max_bin	255	255
max_cat_to_onehot	6	6
cat_l2	15	10
cat_smooth	300	200
min_data_per_group	200	100
force_col_wise	TRUE	TRUE
nrounds	20 000	20 000
early_stopping_rounds	200	200

Note: Official LightGBM reference with a full description of each parameter: lightgbm.readthedocs.io/en/latest/Parameters.html. The R package uses the same parameter names.

3.2 Accuracy: Log-loss

As the basis for comparison we use multi log-loss on the test set, which has not been used in estimation, i.e. out-of-sample log-loss. Multi log-loss (MLL) is defined as

$$\text{MLL} = -\frac{1}{N}\sum_{i=1}^{N} \log \hat{P}\big(Y_{i,t+1} = y^{*}_{i,t+1} \mid \mathbf{X}_{i,t}\big), \tag{2}$$

where N is the number of observations in the test set, $Y_{i,t+1}$ is the (vector-valued or scalar) random outcome for individual i in period $t{+}1$ for the model at hand, and $y^{*}_{i,t+1}$ is its observed realisation. $\hat{P}(\cdot)$ is the predicted probability of the stated event given $\mathbf{X}_{i,t}$.

Ultimately the goal is to know how accurately the model predicts the pair $(T_{i,t+1}, A_{i,t+1})$. Only Model 1 models this joint event directly. However, combining Model 2a and 2b or Model 3a and 3b also yields predictions for both type and usage. The comparison therefore focuses on accuracy for predicting $(T_{i,t+1}, A_{i,t+1})$ for Model 1 vs. Model 2 $(2a{+}2b)$ vs. Model 3 $(3a{+}3b)$.

For Model 2a–3b, log-loss is only available for the ability to fit the conditional models. To compute log-loss evaluated at the probability of the *observed joint event* $\{T_{i,t+1} = \tau^*_{i,t+1}, A_{i,t+1} = a^*_{i,t+1}\}$, the basic probability factorization above is used. For example, log-loss from joint Model 1 can be written as a function of log-loss from sequential Model 2 (Model $2a$+$2b$):

$$
\begin{aligned}
\text{logloss}^{\text{Model 1}} &= -\frac{1}{N}\sum_{i=1}^{N} \log P\big(T_{i,t+1} = \tau^*_{i,t+1}, A_{i,t+1} = a^*_{i,t+1} \mid \mathbf{X}_{i,t}\big) \\
&= -\frac{1}{N}\sum_{i=1}^{N} \log\Big(P\big(T_{i,t+1} = \tau^*_{i,t+1} \mid A_{i,t+1} = a^*_{i,t+1}, \mathbf{X}_{i,t}\big) \\
&\qquad \cdot P\big(A_{i,t+1} = a^*_{i,t+1} \mid \mathbf{X}_{i,t}\big)\Big) \\
&= -\frac{1}{N}\sum_{i=1}^{N} \Big[\log P\big(T_{i,t+1} = \tau^*_{i,t+1} \mid A_{i,t+1} = a^*_{i,t+1}, \mathbf{X}_{i,t}\big) \\
&\qquad + \log P\big(A_{i,t+1} = a^*_{i,t+1} \mid \mathbf{X}_{i,t}\big)\Big] \\
&= \text{logloss}^{\text{Model } 2b} + \text{logloss}^{\text{Model } 2a},
\end{aligned}
\tag{3}
$$

where each $P(\cdot)$ is the probability of the event considered. A similar derivation applies to Model 3a and 3b. It follows from (3) that a comparable measure between log-loss from Model 1 and the sequential models can be obtained by summing log-loss from Model 2a and 2b or Model 3a and 3b, provided the probability functions are consistently estimated for all models on the same data.

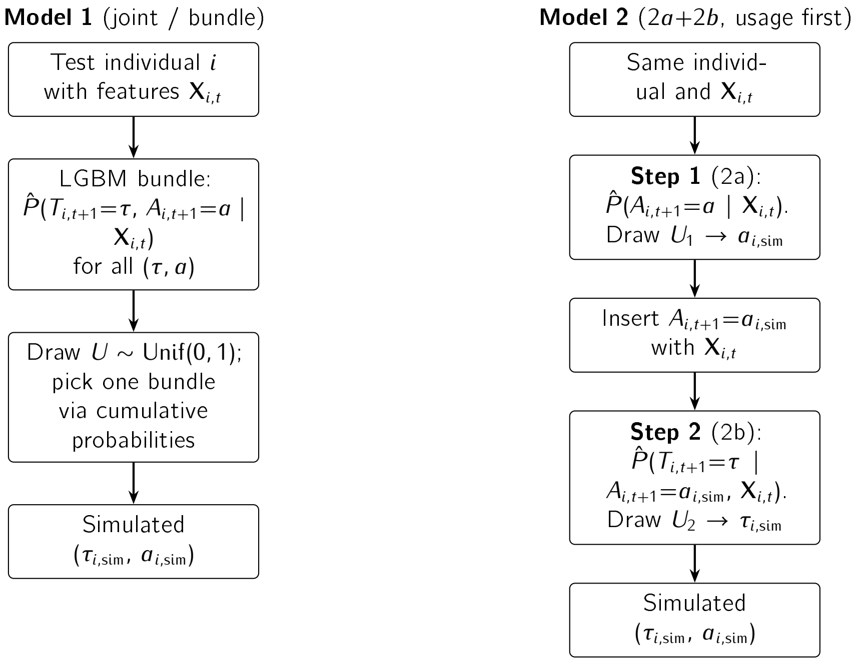
3.3 Accuracy: Visualizations

As a supplement to assessing accuracy, visualizations based on both model predictions and simulations from Model 1, Model 2 ($2a$+$2b$), and Model 3 ($3a$+$3b$) are used. Type and usage are simulated for each individual in the test data. The simulated joint distribution of $(T_{i,t+1}, A_{i,t+1})$ is compared graphically to the corresponding empirical distribution. This makes it possible to examine whether the simulation from SMILE might show discrepancies between data and simulation for specific parts of the distribution.

3.3.1 Simulation

Simulated classes are assigned by inverse transform sampling: for each individual i, $U \sim \text{Unif}(0,1)$ is drawn, and an outcome is chosen from the cumulative predicted probabilities. Figure 3.1 shows the order of steps for Model 1 and Model 2. For Model 3 the sequential logic is repeated in reverse order: first type is drawn with Model 3a from estimated $\hat{P}(T_{i,t+1} = \tau \mid \mathbf{X}_{i,t})$, then usage conditional on the simulated type with Model 3b from estimated $\hat{P}(A_{i,t+1} = a \mid T_{i,t+1} = \tau_{i,\text{sim}}, \mathbf{X}_{i,t})$.

Figure 3.1: Schematic procedure for one test individual. All steps are repeated independently for each individual. With multiple replications the whole chain is repeated with new draws.



4 Results

Model 1 converges slowly[5], so a random 50 % subsample is taken from the original population data from which the training, validation and test sets are constructed. A 50 % sample corresponds to roughly 2.8 million observations. This already exceeds the number of observations previously used to train CTREE models for usage given a default moving reason (about 2.1 million), and far exceeds what was used to train CTREE models for type given a default moving reason (just under 0.85 million). The size of the dataset used for the LGBM models is therefore large enough to support conclusions for the comparison even with a 50 % sample.

4.1 Log-loss

For the 50 % subsample and the model settings used, log-loss lies between 1.77 and 1.78 across Model 1, 2, and 3, see Table 4.1. On the basis of log-loss there is thus no difference in accuracy across models that would clearly favour one model over another.

4.2 Marginal distributions

We have compared marginal conditional distributions for each dwelling type and usage over age, conditional on factor-variable levels in $\mathbf{X}_{i,t}$. For the model curves the starting

[5] For a 100 % sample the model had still not converged after 20 000 training rounds and two days.

Table 4.1: Runtimes and multi log-loss on test for a 50 % sample

Model	Training (min)	Prediction, test (min)	Pred. per 10^5 obs. (min)	Multi log-loss
Model 1	428.6	36.4	5.13	1.7706
Model 2a	62.6	3.7	0.52	—
Model 2b	71.6	4.6	0.65	—
Model 3a	62.4	3.8	0.54	—
Model 3b	70.7	4.8	0.67	—
Model 2 (2a+2b), total	134.2	8.3	1.17	1.7773
Model 3 (3a+3b), total	133.1	8.6	1.21	1.7784

Note: Training data: $N_{\text{train}} = 2\,839\,439$ observations. Test data: $N_{\text{test}} = 709\,860$ observations. Reported runtimes are from a single run. The column *Pred. per* 10^5 is prediction time scaled to 100 000 test observations. For Model 2 and 3 combined, training and prediction time are the sum of the sub-models (2a+2b and 3a+3b respectively).

point is the model's *predicted* probabilities on the test data: each individual i in the test set has a probability vector for the post-move outcomes $(T_{i,t+1}, A_{i,t+1})$, and the average predicted probability for each class τ or a is plotted within each cell defined by factor level and age. For a given age and a fixed factor level f, let $\mathcal{I}_{f,\text{age}}$ denote the set of test observations in that cell.

For the *bundle* model (Model 1), joint probabilities $\hat{P}(T_{i,t+1} = \tau,\, A_{i,t+1} = a \mid \mathbf{X}_{i,t})$ are predicted for all pairs (τ, a). Marginal predicted curves conditional on factor level f and age are therefore obtained by:

$$\begin{aligned}
\widehat{P}_1^{\text{pred}}(T_{i,t+1} = \tau \mid i \in \mathcal{I}_{f,\text{age}}) &= \frac{1}{|\mathcal{I}_{f,\text{age}}|} \sum_{i \in \mathcal{I}_{f,\text{age}}} \sum_{a} \hat{P}(T_{i,t+1} = \tau,\, A_{i,t+1} = a \mid \mathbf{X}_{i,t}), \\
\widehat{P}_1^{\text{pred}}(A_{i,t+1} = a \mid i \in \mathcal{I}_{f,\text{age}}) &= \frac{1}{|\mathcal{I}_{f,\text{age}}|} \sum_{i \in \mathcal{I}_{f,\text{age}}} \sum_{\tau} \hat{P}(T_{i,t+1} = \tau,\, A_{i,t+1} = a \mid \mathbf{X}_{i,t}).
\end{aligned} \tag{4}$$

For the *sequential* models (Model 2 and 3), marginal type and usage curves use the marginal predictions from the relevant sub-model in each case. Let $\widehat{\pi}_{i,m}^{\text{type}}(T_{i,t+1} = \tau)$ and $\widehat{\pi}_{i,m}^{\text{use}}(A_{i,t+1} = a)$ denote those predicted probabilities for individual i under sequential specification $m \in \{2, 3\}$. Then

$$\begin{aligned}
\widehat{P}_m^{\text{pred}}(T_{i,t+1} = \tau \mid i \in \mathcal{I}_{f,\text{age}}) &= \frac{1}{|\mathcal{I}_{f,\text{age}}|} \sum_{i \in \mathcal{I}_{f,\text{age}}} \widehat{\pi}_{i,m}^{\text{type}}(T_{i,t+1} = \tau), \\
\widehat{P}_m^{\text{pred}}(A_{i,t+1} = a \mid i \in \mathcal{I}_{f,\text{age}}) &= \frac{1}{|\mathcal{I}_{f,\text{age}}|} \sum_{i \in \mathcal{I}_{f,\text{age}}} \widehat{\pi}_{i,m}^{\text{use}}(A_{i,t+1} = a).
\end{aligned} \tag{5}$$

For Model 2, $\widehat{\pi}_{i,2}^{\text{type}}$ comes from Model 2b and $\widehat{\pi}_{i,2}^{\text{use}}$ from Model 2a, for Model 3, $\widehat{\pi}_{i,3}^{\text{type}}$ comes from Model 3a and $\widehat{\pi}_{i,3}^{\text{use}}$ from Model 3b.

The reference curve *Data* is based on *observed* class shares in the test set within each cell:

$$\begin{aligned}\widehat{P}^{\text{data}}(T_{i,t+1} = \tau \mid i \in \mathcal{I}_{f,\text{age}}) &= \frac{1}{|\mathcal{I}_{f,\text{age}}|} \sum_{i \in \mathcal{I}_{f,\text{age}}} \mathbf{1}\{\tau^*_{i,t+1} = \tau\}, \\ \widehat{P}^{\text{data}}(A_{i,t+1} = a \mid i \in \mathcal{I}_{f,\text{age}}) &= \frac{1}{|\mathcal{I}_{f,\text{age}}|} \sum_{i \in \mathcal{I}_{f,\text{age}}} \mathbf{1}\{a^*_{i,t+1} = a\}.\end{aligned} \tag{6}$$

There is one graph for each factor level and each class which results in 460 graphs in total. We therefore only show those that do not condition on a factor level, see Figures 4.1 and 4.2. Hence, the presented graphs correspond to the weighted average over factor levels of each of the probabilities in (5)–(6). The conclusion from these graphs is that there are no clear systematic differences between Model 1, 2a, and 3b in their ability to predict usage, or between Model 1, 2b, and 3a in their ability to predict type. The same holds if conditioning on factor levels.

4.3 Joint distributions

From the simulated data, we have computed the share of simulated individuals in each dwelling bundle class over age, conditional on factor-variable levels in $\mathbf{X}_{i,t}$. A bundle class b labels one joint outcome (τ, a), i.e. the event $\{T_{i,t+1} = \tau, A_{i,t+1} = a\}$. This is therefore the joint distribution implied by the simulated data.

The calculation follows the same principle as for marginal data shares. For each model m the curves show simulated shares in each cell for a given factor level and age, pooling over R independent replications.

For a given age and fixed factor level f, the same cell $\mathcal{I}_{f,\text{age}}$ is used as in the section on marginal conditional distributions. Let $r = 1, \dots, R$ denote the replication index. For individual i, model m, and replication r, let $B^{\text{sim}}_{i,r,m}$ denote the *random* bundle-class label from the simulation, with realised value $b^{\text{sim}}_{i,r,m}$. The model curve for factor level f and the given age is the empirical share of draws in each class b,

$$\widehat{P}^{\text{sim}}_m(B^{\text{sim}}_{i,r,m} = b \mid i \in \mathcal{I}_{f,\text{age}}) = \frac{1}{\sum_{r=1}^{R} |\mathcal{I}_{f,\text{age}}|} \sum_{r=1}^{R} \sum_{i \in \mathcal{I}_{f,\text{age}}} \mathbf{1}\{b^{\text{sim}}_{i,r,m} = b\}, \tag{7}$$

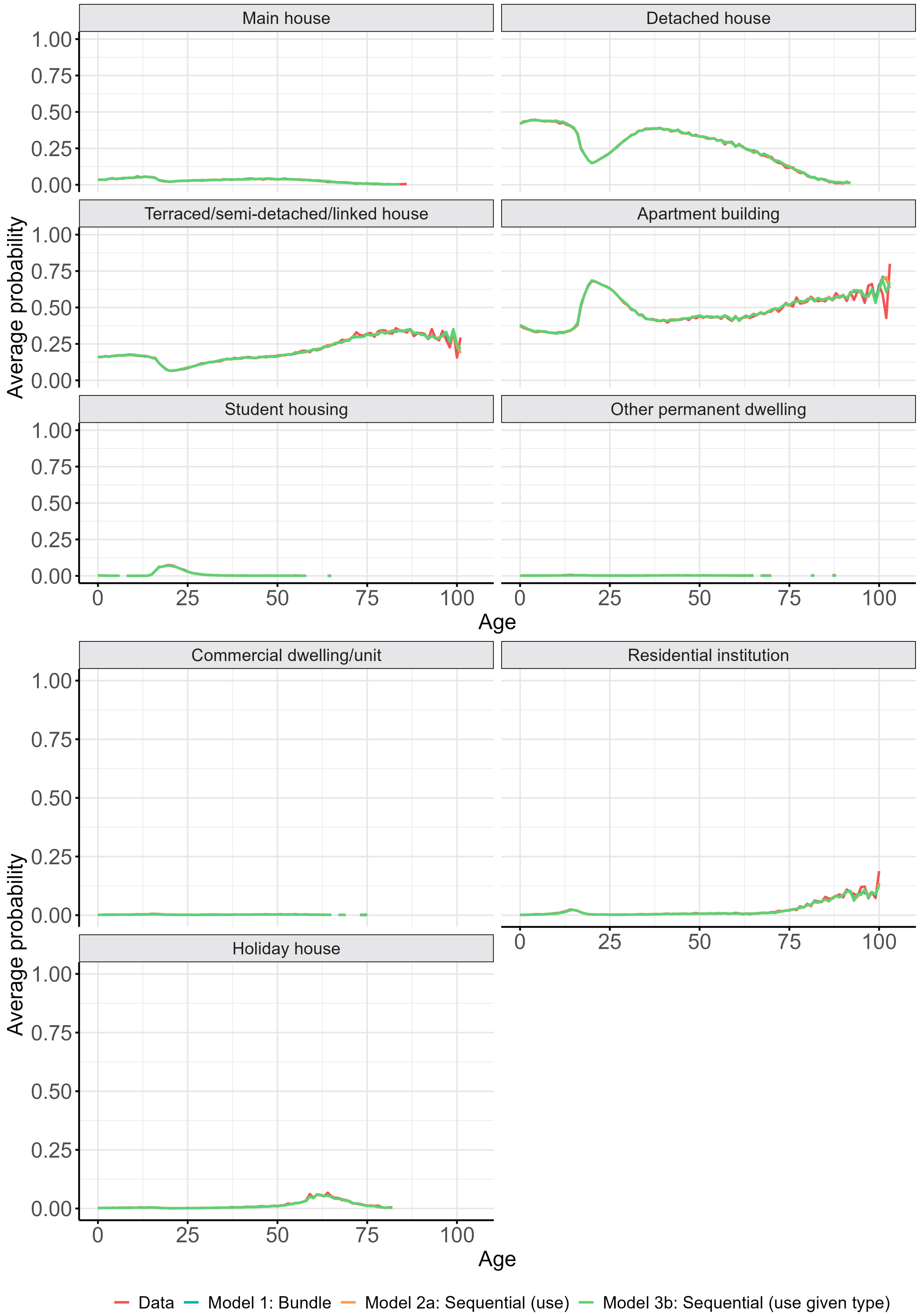
This amounts to pooling all R draws into one empirical distribution over bundle classes within the cell. As explained in Section 2.3 and Section 3.3.1, the difference from using model predictions directly as in Section 4.2 is that here the object is simulated $(T_{i,t+1}, A_{i,t+1})$ under Model 2 or 3, so intermediate draws condition on simulated $A_{i,t+1}$ (Model 2b) or simulated $T_{i,t+1}$ (Model 3b).

Let $B_{i,t+1}$ denote the random bundle-class label for individual i at $t{+}1$ (joint outcome for $(T_{i,t+1}, A_{i,t+1})$). The reference curve *Data* uses observed realisations $b^*_{i,t+1}$, corresponding to $(\tau^*_{i,t+1}, a^*_{i,t+1})$:

$$\widehat{P}^{\text{data}}(B_{i,t+1} = b \mid i \in \mathcal{I}_{f,\text{age}}) = \frac{1}{|\mathcal{I}_{f,\text{age}}|} \sum_{i \in \mathcal{I}_{f,\text{age}}} \mathbf{1}\{b^*_{i,t+1} = b\}. \tag{8}$$

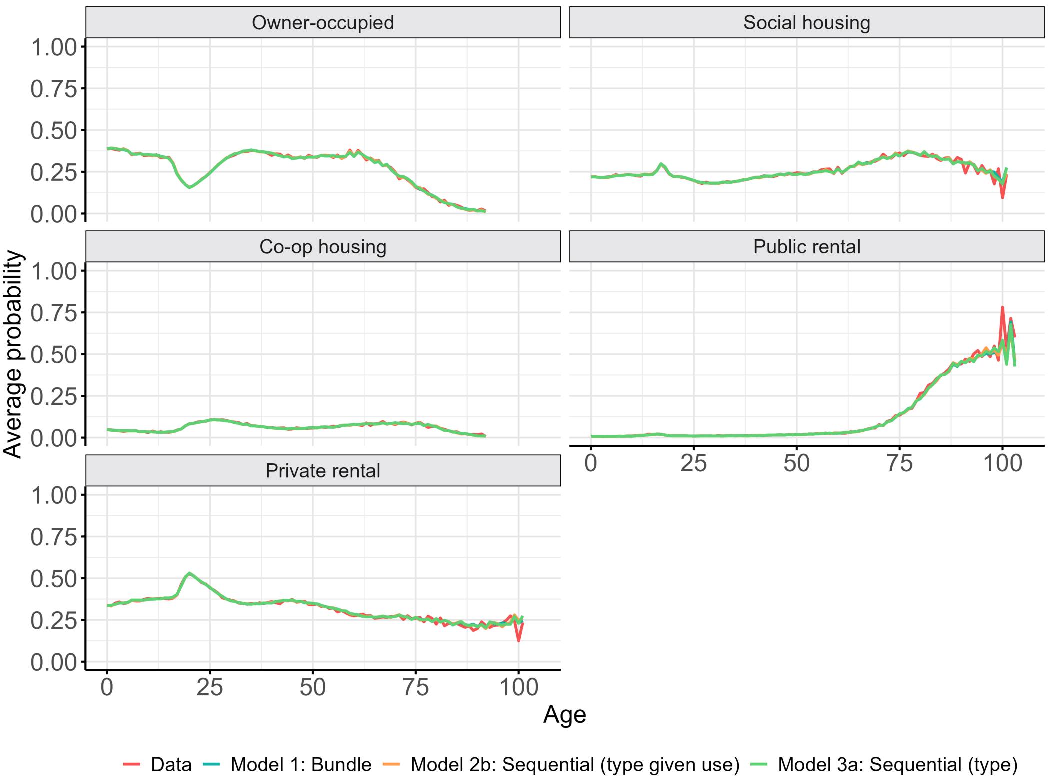
As for the marginal distributions, there are too many graphs to show all of them (more than 1.300). We therefore only show those that do not condition on a factor level, see

Figure 4.1: Marginal distributions of dwelling *usage* (out-of-sample): Models 1, 2a, and 3b.



Note: Model 1 follows the usage row of (4), Model 2a and 3b the usage row of (5) and the data line the usage row of (6). All formulas only condition on age here. All cells with less than 3 observations have been set to NA to comply with anonymization requirements from Statistics Denmark.

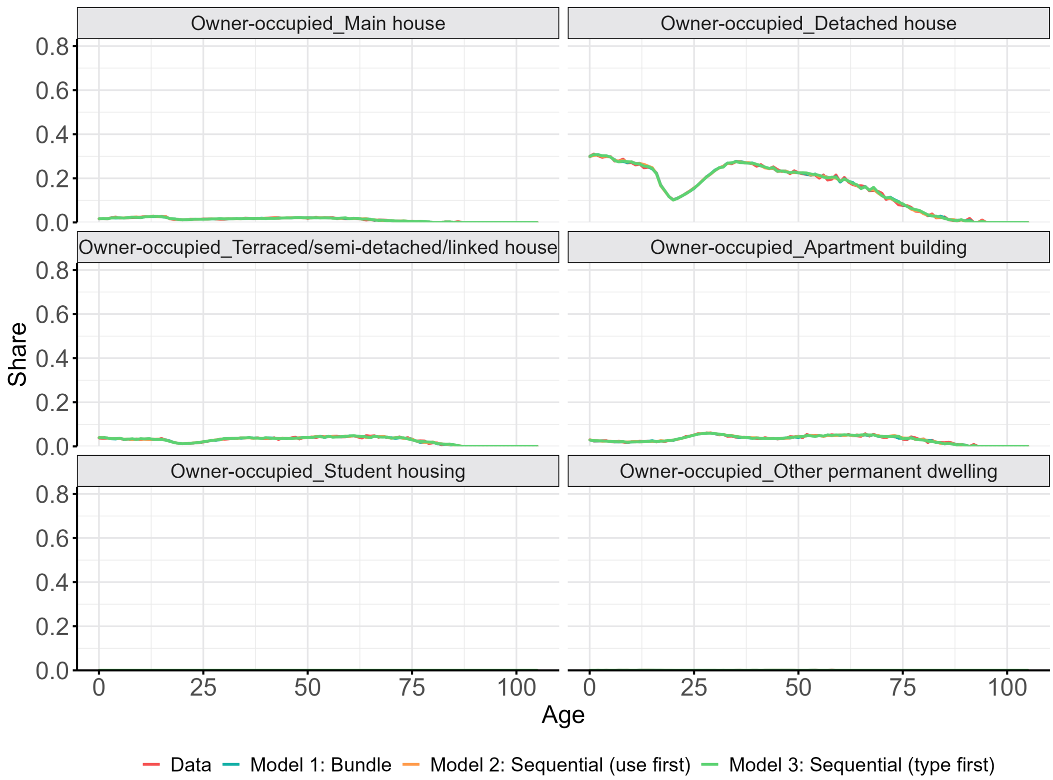
Figure 4.2: Marginal distributions of dwelling *type* (out-of-sample): Models 1, 3a, and 2b.



Note: Model 1 follows the type row of (4), Model 2b and 3a the type row of (5) and the data line the type row of (6). All formulas only condition on age here. All cells with less than 3 observations have been set to NA to comply with anonymization requirements from Statistics Denmark.

Figures 4.3–4.10. Consequently, the presented graphs are the weighted average over factor levels of the probabilities in (7)–(8). The conclusion from the graphs is that at bundle level there are still no clear systematic differences between Model 1, Model 2 ($2a$+$2b$), and Model 3 ($3a$+$3b$). This also holds if conditioning on factor levels.

Figure 4.3: Simulated housing-bundle shares over age (out-of-sample): Model 1, Model 2 ($2a$+$2b$), and Model 3 ($3a$+$3b$). Part 1 of 8.

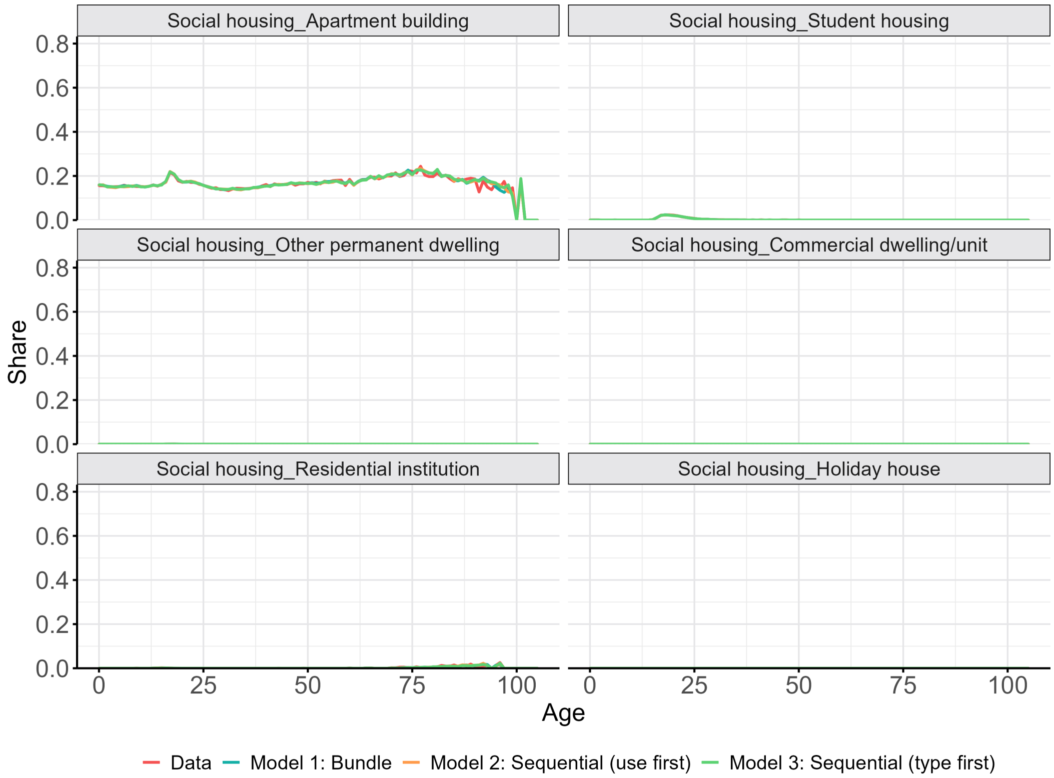


Note: Model curves follow (7) and the *Data* line follows (8), but without conditioning on f. All cells with less than 3 observations in the test data have been set to NA to comply with anonymization requirements from Statistics Denmark.

Figure 4.4: Simulated housing-bundle shares over age (out-of-sample). Part 2 of 8 *(continued)*.

Note: Same construction and reference curves as in Figure 4.3.

Figure 4.5: Simulated housing-bundle shares over age (out-of-sample). Part 3 of 8 *(continued)*.

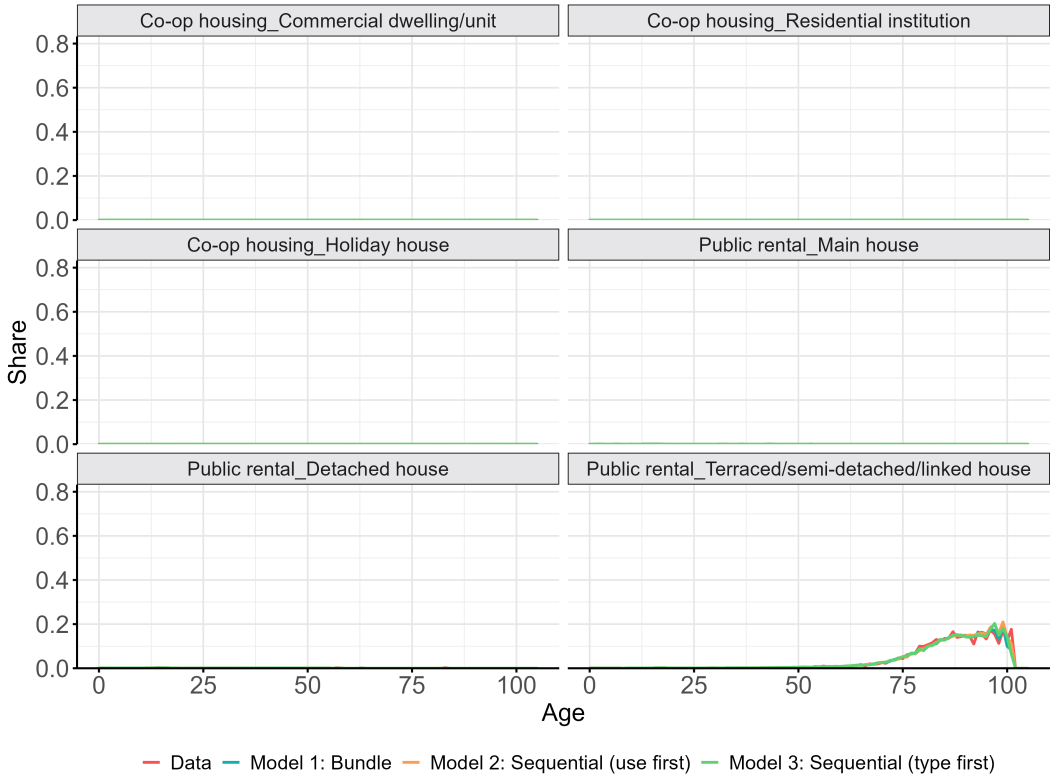


Note: Same construction and reference curves as in Figure 4.3.

Figure 4.6: Simulated housing-bundle shares over age (out-of-sample). Part 4 of 8 *(continued)*.

Note: Same construction and reference curves as in Figure 4.3.

Figure 4.7: Simulated housing-bundle shares over age (out-of-sample): Model 1, Model 2 ($2a+2b$), and Model 3 ($3a+3b$). Part 5 of 8.

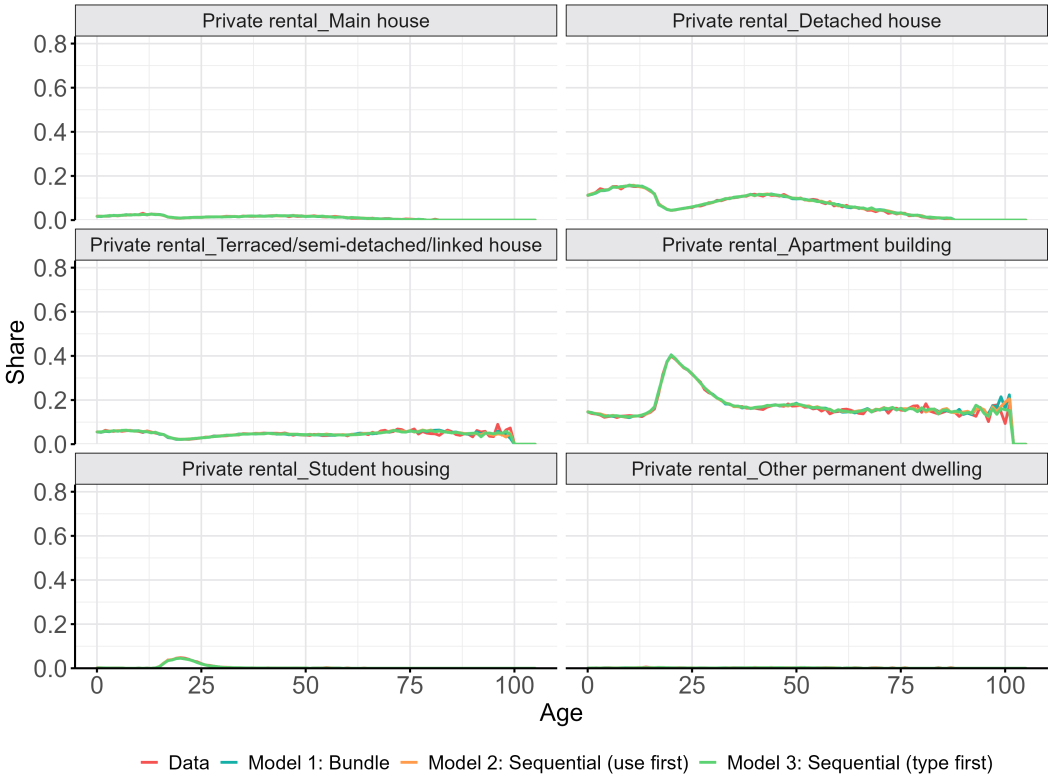


Note: Same construction and reference curves as in Figure 4.3.

Figure 4.8: Simulated housing-bundle shares over age (out-of-sample). Part 6 of 8 *(continued)*.

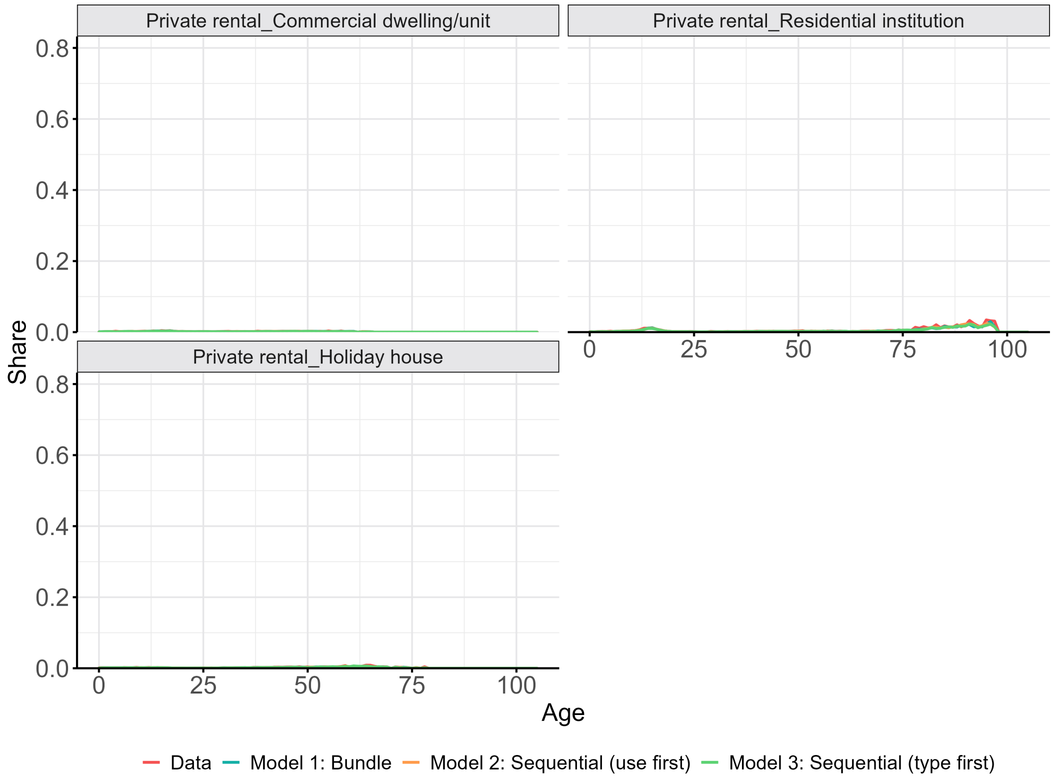
Note: Same construction and reference curves as in Figure 4.3.

Figure 4.9: Simulated housing-bundle shares over age (out-of-sample). Part 7 of 8 *(continued)*.



Note: Same construction and reference curves as in Figure 4.3.

Figure 4.10: Simulated housing-bundle shares over age (out-of-sample). Part 8 of 8 *(continued)*.



Note: Same construction and reference curves as in Figure 4.3.

5 Conclusion

With the present results showing near-identical log-loss (1.77–1.78) and visually aligned predictions across models, there is no evidence that the choice between joint bundle estimation and sequential models affects predictive accuracy in the setup tested.

By contrast, training and prediction times for the 50 % sample used ($N_{\text{train}} = 2\,839\,439$, $N_{\text{test}} = 709\,860$, cf. Table 4.1) differ markedly: training Model 1 takes about 429 minutes versus about 134 minutes for Model 2 ($2a$+$2b$) and about 133 minutes for Model 3 ($3a$+$3b$). Prediction on the full test set takes about 36 minutes for Model 1 versus about 8 minutes for Model 2 and about 9 minutes for Model 3. Thus the joint model spends roughly 4–5 times longer on prediction than each of the sequential models, which markedly worsens the bottleneck in SMILE. There is little difference in training or prediction time between Model 2 and 3, so the ordering of events in the sequential models is not decisive for runtimes.

Although the results above clearly show that the joint model has no sharp advantage over the sequential models, these tests cannot fully rule out that the joint model could be preferable with more than two events. A previous analysis restricted to visual comparisons considered four events: (dwelling type, dwelling usage, dwelling floor area, city size). The model then had to predict roughly 400 classes after removing classes representing less than 0.01 % of the data. For that model too, there were no clear differences in the visual predictions. In the comparison in this note, a 10 % sample was also used in addition to the 50 % sample, again without any clear differences in the visual predictions. For that reason, results for four events on a 10 % sample are expected to be representative of a larger sample as well. Moreover, we also expect the results to be valid when considering different reasons for moving in SMILE as well as when using data aggregated at the family rather than individual level. Overall, the assessment is therefore that joint models have no clear advantage over sequential ones in terms of predictive accuracy.