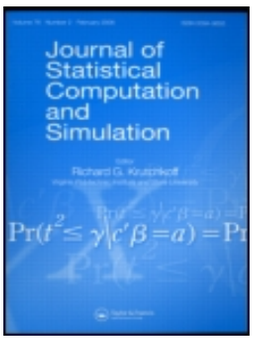

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Mixed-effects random forest for clustered data

Ahlem Hajjem, François Bellavance & Denis Larocque

To cite this article: Ahlem Hajjem, François Bellavance & Denis Larocque (2014) Mixed-effects random forest for clustered data, Journal of Statistical Computation and Simulation, 84:6, 1313-1328, DOI: 10.1080/00949655.2012.741599

To link to this article: https://doi.org/10.1080/00949655.2012.741599

Published online: 12 Nov 2012.

Journal of Statistical Computation and Simulation, 2014
Vol. 84, No. 6, 1313–1328, http://dx.doi.org/10.1080/00949655.2012.741599

Taylor & Francis
Taylor & Francis Group

Mixed-effects random forest for clustered data

Ahlem Hajjem*, François Bellavance and Denis Larocque

Department of Management Sciences, HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine, Montréal, QC, Canada H3T 2A7

(*Received 3 February 2012; final version received 16 October 2012*)

This paper presents an extension of the random forest (RF) method to the case of clustered data. The proposed 'mixed-effects random forest' (MERF) is implemented using a standard RF algorithm within the framework of the expectation–maximization algorithm. Simulation results show that the proposed MERF method provides substantial improvements over standard RF when the random effects are non-negligible. The use of the method is illustrated to predict the first-week box office revenues of movies.

Keywords: clustered data; mixed effects; regression tree; random forest

1. Introduction

Tree-based methods are appreciated by practitioners because they often provide reasonable and easy-to-interpret models even when a large number of covariates is present due to their ability to select the most significant predictors and to handle interactions automatically. However, the prediction performance of a single tree can often be improved, at the expense of interpretability, by using ensemble of trees. Random forests (RFs) [1] is a very powerful ensemble method for trees. Both empirical studies (e.g. [1,2]) and theoretical results [3] have demonstrated the power of this method and this is why it became an active area of research. Recent surveys about RFs and ensemble methods can be found in [4–6].

Using the mixed-effects approach, Hajjem *et al.* [7] extended the regression tree (RT) algorithms such as CART [8] to the case of clustered data for a continuous outcome. Simulation results showed that their mixed-effects regression tree (MERT) algorithm provides substantial improvements over standard trees when the random effects are non-negligible. The key idea of MERT is to dissociate the fixed from the random effects. It is basically an iterative call to a standard RT algorithm within the framework of the expectation–maximization (EM) algorithm [9,10]. At each iteration, a standard RT is built from the transformed response data where the current estimate of the random-effect component is removed from the original response. Sela and Simonoff [11] independently proposed a similar approach, called random effects expectation–maximization (RE–EM) trees.

*Corresponding author. Email: ahlem.hajjem@uqam.ca

A possible generalization of RFs to clustered data consists in replacing the RT within each iteration of the MERT algorithm with a forest of RTs. The goal of this paper is to introduce this proposed generalization of the RF method, called 'mixed-effects random forest' (MERF), and to investigate its performance with a simulation study. The predictive mean-squared error (PMSE) of MERF is compared with the PMSE of five alternative models, including the standard RF, under different key features such as the strength of both the total and the random effects and the level of dependence between the predictors. The main finding is that MERF is more appropriate than a standard RF for clustered data, particularly when the random effects are non-negligible.

The remainder of this article is organized as follows: Section 2 describes the proposed MERF approach; Section 3 presents a simulation study to evaluate the performance of MERF; Section 4 illustrates the application of the method with a real data set; and Section 5 gives some concluding remarks.

2. MERF approach

We define the MERF of RTs as follows:

$$\begin{aligned} y_i &= f(X_i) + Z_i b_i + \epsilon_i, \\ b_i &\sim N(0, D), \quad \epsilon_i \sim N(0, R_i), \quad i = 1, \ldots, n, \end{aligned} \tag{1}$$

where $y_i = [y_{i1}, \ldots, y_{in_i}]^{\mathrm{T}}$ is the $n_i \times 1$ vector of responses for the n_i observations in cluster i, $X_i = [x_{i1}, \ldots, x_{in_i}]^{\mathrm{T}}$ is the $n_i \times p$ matrix of fixed-effects covariates, $Z_i = [z_{i1}, \ldots, z_{in_i}]^{\{\mathrm{T}\}}$ is the $n_i \times q$ matrix of random-effects covariates, $b_i = (b_{i1}, \ldots, b_{iq})^{\mathrm{T}}$ is the $q \times 1$ unknown vector of random effects for cluster i, $\epsilon_i = [\epsilon_{i1}, \ldots, \epsilon_{in_i}]^{\mathrm{T}}$ is the $n_i \times 1$ vector of errors, and the unknown function $f(X_i)$ is estimated using a standard forest of RTs. The random part, $Z_i b_i$, is assumed linear. The total number of observations is $N = \sum_{i=1}^{n} n_i$. The covariance matrix of b_i is D, while R_i is the covariance matrix of ϵ_i.

We further assume that b_i and ϵ_i are independent and normally distributed and that the between-cluster observations are independent. Hence, the covariance matrix of the vector of observations y_i in cluster i is $V_i = \mathrm{Cov}(y_i) = Z_i D Z_i^{\mathrm{T}} + R_i$, and $V = \mathrm{Cov}(y) = \mathrm{diag}(V_1, \ldots, V_n)$, where $y = [y_1^{\mathrm{T}}, \ldots, y_n^{\mathrm{T}}]^{\mathrm{T}}$. We will also assume that the correlation is induced solely via the between-cluster variation, that is, R_i is diagonal ($R_i = \sigma^2 I_{n_i}, i = 1, \ldots, n$). This assumption is suitable for a large class of clustered data problems [12, p. 30]. Possible extensions to allow for other correlation structures are discussed in Section 5.

The MERF algorithm is the MERT algorithm [7] where the single RT structure used to estimate the fixed part of the model is replaced by a RF, that is, an ensemble of unpruned RTs.

The MERF algorithm is similar to the EM algorithm for the linear mixed-effects (LMEs) model, as described by Wu and Zhang [13, Section 2.2.5], and is as follows:

Step 0. Set $r = 0$. Let $\hat{b}_{i(0)} = 0$, $\hat{\sigma}^2_{(0)} = 1$, and $\hat{D}_{(0)} = I_q$.

Step 1. Set $r = r + 1$. Update $y^*_{i(r)}$, $\hat{f}(X_i)_{(r)}$, and $\hat{b}_{i(r)}$.

(i) $y^*_{i(r)} = y_i - Z_i \hat{b}_{i(r-1)}, i = 1, \ldots, n$.

(ii) Build a forest of trees using a standard RF algorithm with $y^*_{ij(r)}$ as the training set responses and x_{ij} as the corresponding training set of covariates, $i = 1, \ldots, n, j = 1, \ldots, n_i$. The bootstrap training samples to build the forest are simple random samples drawn with replacement from the training set $(y^*_{ij(r)}, x_{ij})$.

(iii) Obtain an estimate $\hat{f}(x_{ij})_{(r)}$ of $f(x_{ij})$ using only the subset of trees in the forest that are build with the bootstrap samples not containing observation j in cluster i, that is, the out-of-bag prediction of the RF; let $\hat{f}(X_i)_{(r)} = [\hat{f}(x_{i1})_{(r)}, \ldots, \hat{f}(x_{in_i})_{(r)}]^{\mathrm{T}}$.

(iv) $\hat{b}_{i(r)} = \hat{D}_{(r-1)} Z_i^{\mathrm{T}} \hat{V}_{i(r-1)}^{-1} (y_i - \hat{f}(X_i)_{(r)}), i = 1, \ldots, n$, where $\hat{V}_{i(r-1)} = Z_i \hat{D}_{(r-1)} Z_i^{\mathrm{T}} + \hat{\sigma}^2_{(r-1)} I_{n_i}$, $i = 1, \ldots, n$.

Step 2. Update $\hat{\sigma}^2_{(r)}$ and $\hat{D}_{(r)}$ using

$$\hat{\sigma}^2_{(r)} = N^{-1} \sum_{i=1}^{n} \{\hat{\epsilon}^{\mathrm{T}}_{i(r)} \hat{\epsilon}_{i(r)} + \hat{\sigma}^2_{(r-1)} [n_i - \hat{\sigma}^2_{(r-1)} \text{ trace } (\hat{V}_{i(r-1)})]\},$$

$$\hat{D}_{(r)} = n^{-1} \sum_{i=1}^{n} \{\hat{b}_{i(r)} \hat{b}^{\mathrm{T}}_{i(r)} + [\hat{D}_{(r-1)} - \hat{D}_{(r-1)} Z_i^{\mathrm{T}} \hat{V}^{-1}_{i(r-1)} Z_i \hat{D}_{(r-1)}]\},$$

where $\hat{\epsilon}_{i(r)} = y_i - \hat{f}(X_i)_{(r)} - Z_i \hat{b}_{i(r)}$.

Step 3. Keep iterating by repeating steps 1 and 2 until convergence.

In words, the algorithm starts at step 0 with default values for $\hat{b}_i$, $\hat{\sigma}^2$, and $\hat{D}$. At step 1, it first calculates the fixed part of the response variable, y_i^*, that is, the response variable from which we remove the current available value of the random part. Second, the algorithm takes bootstrap samples from the training set (y_{ij}^*, x_{ij}) to build a forest of trees. To minimize overfitting, the predicted fixed part $\hat{f}(x_{ij})$ for observation j from cluster i is obtained with the subset of trees in the forest that are build using the bootstrap samples not containing observation j from cluster i (i.e. out-of-bag prediction). Fourth, it computes $\hat{b}_i$ with the updated estimate of the random part of Equation (1). At step 2, it updates the variance components $\hat{\sigma}^2$ and $\hat{D}$ based on updated estimates of the residuals. The algorithm keeps iterating by repeating steps 1 and 2 until convergence.

The convergence of the algorithm is monitored by computing at each iteration the following generalized log-likelihood (GLL) criterion:

$$\begin{aligned} \text{GLL}\,(f, b_i|y) = \sum_{i=1}^{n} \{&[y_i - f(X_i) - Z_i b_i]^{\mathrm{T}} R_i^{-1} [y_i - f(X_i) - Z_i b_i] \\ &+ b_i^{\mathrm{T}} D^{-1} b_i + \log |D| + \log |R_i|\}. \end{aligned} \tag{2}$$

To predict the response for a new observation j that belongs to a cluster i among those used to fit the MERF model, we use both its corresponding population-averaged RF prediction, $\hat{f}(x_{ij})$, and the predicted random part corresponding to its cluster, $Z_i \hat{b}_i$. For a new observation that belongs to a cluster not included in the sample used to estimate the model parameters, we can only take the corresponding population-averaged RF prediction, $\hat{f}(x_{ij})$.

3. Simulation

We investigate the performance of the proposed MERF of RTs through a simulation study. We compare the PMSE of MERF and five alternative models, namely (1) the standard RF of RTs, (2) the MERT, (3) the standard RT, (4) the LME model, and (5) the linear model (LM).

We implemented the proposed MERF algorithm in R [14] using the package *randomForest* [15] to estimate the fixed component of Equation (1). The function *randomForest* implements Breiman's RF algorithm for classification and regression, based on Breiman and Cutler's original Fortran code. The default settings of the function *randomForest* are used, except for the parameter *ntree* (number of trees to grow within the forest) that we set to 300 instead of the default value of 500 to save overall computing time for the simulation. Note that this smaller number still ensures that every observation in the learning set gets predicted by about 100 trees in each iteration since

the out-of-bag set is formed by about $\frac{1}{3}$ of the original sample on average. RT and MERT models are also fitted with the default settings of the function *rpart* [16].

For MERF convergence, we suggest a minimum number of iterations to avoid early stopping and keep iterating until the absolute change in GLL is less than a given small value. For the simulation, preliminary testing suggested the following approach. We force a minimum of 100 iterations and keep iterating while the absolute change in GLL is not less than 1E–04 or we reach a maximum of 200 iterations. Convergence was reached before the 200th iteration in 98% of cases, with 137 iterations on average. The final MERF model is the one at the last iteration. For MERT, we force a minimum of 50 iterations and keep iterating while the absolute change in GLL is not less than 1E–04 or we reach a maximum of 200 iterations. Once the stopping criterion is met, we run an additional 50 iterations. The mixed-effects tree model chosen is the one corresponding to the last iteration where the number of leaves is equal to the modal value over the last 50 mixed-effects tree models [7]. Convergence was reached in 100% of cases, with 118 iterations on average. Computing time per iteration of MERF is on average 2.88 s which is 78% greater than the average time of MERT (1.62 s per iteration). The latter is slightly more than twice the computing time of one RF with 300 trees, which is 0.78 s on average. The simulations were performed on Opteron AMD 275 CPU 1.0 GHz processors.

3.1. *Simulation design*

The simulation design has a hierarchical structure of 100 unbalanced clusters and 5000 observations: 20 clusters with 10 observations, 20 with 30 observations, 20 with 50 observations, 20 with 70 observations, and 20 with 90 observations. The first 10% of the generated observations in each cluster form the training sample, and the other 90% are kept for the first test sample. Consequently, the RTs are built with 500 observations nested within 100 unbalanced clusters having 1, 3, 5, 7, or 9 observations. The remaining 4500 observations form the first test set. We call this test set 'known clusters'.

Using the same scheme, we generated another 4500 observations nested within 100 new clusters. They form the second test set called 'new clusters'. The first and the second test sets are used to evaluate two types of out-of-sample performance of MERF: its performance for new observations from clusters included in the training data set and its performance for new observations from clusters not included in the training data set.

Table 1. DGPs for the simulation study.

DGP	ρ	PTEV[a]	PREV[b]	σ^2_{Fixed}	m	σ^2_b	ICC[c]
1	0.0	90	10	8.1	0.8	0.9	47.4
2			30	6.3	0.7	2.7	73.0
3			50	4.5	0.6	4.5	81.8
4		60	10	1.4	0.3	0.2	13.0
5			30	1.1	0.3	0.5	31.0
6			50	0.8	0.2	0.8	42.9
7	0.4	90	10	8.1	0.7	0.9	47.4
8			30	6.3	0.6	2.7	73.0
9			50	4.5	0.5	4.5	81.8
10		60	10	1.4	0.3	0.2	13.0
11			30	1.1	0.3	0.5	31.0
12			50	0.8	0.2	0.8	42.9

[a]PTEV $= ((\sigma^2_{\text{Fixed}} + \sigma^2_b)/(\sigma^2_{\text{Fixed}} + \sigma^2_b + \sigma^2_\varepsilon)) \times 100$.
[b]PREV $= (\sigma^2_b/(\sigma^2_{\text{Fixed}} + \sigma^2_b)) \times 100$.
[c]Intracluster Correlation $= (\sigma^2_b/(\sigma^2_b + \sigma^2_\varepsilon)) \times 100$.

Table 2. Results of the PMSE on the first test set (known clusters) of MERF, RF, MERT, RT, LME, and LM models based on 100 simulation runs.

	MERF					RF					MERT				
DGP	Avg	Med.	Min.	Max.	Std	Avg	Med.	Min.	Max.	Std	Avg	Med.	Min.	Max.	Std
PMSE															
1	4.11	4.05	3.43	5.08	0.35	4.55	4.50	3.82	5.54	0.37	5.66	5.58	4.21	8.40	0.81
2	3.66	3.66	3.06	4.36	0.23	5.96	5.86	4.94	7.50	0.56	4.80	4.72	3.88	6.46	0.52
3	3.03	3.01	2.64	3.77	0.21	7.46	7.47	5.72	10.62	0.82	4.00	3.93	2.88	5.61	0.49
4	1.63	1.63	1.40	1.83	0.07	1.67	1.67	1.46	1.84	0.07	1.96	1.93	1.69	2.30	0.13
5	1.59	1.59	1.47	1.80	0.07	1.90	1.89	1.69	2.21	0.10	1.87	1.87	1.61	2.28	0.12
6	1.53	1.54	1.37	1.67	0.06	2.12	2.11	1.86	2.45	0.14	1.78	1.78	1.54	2.09	0.10
7	3.23	3.20	2.87	3.79	0.19	3.77	3.75	3.29	4.45	0.26	4.45	4.41	3.69	6.28	0.47
8	2.92	2.92	2.58	3.43	0.18	5.35	5.30	4.40	6.93	0.57	3.93	3.88	3.10	5.15	0.42
9	2.49	2.48	2.09	2.90	0.16	6.81	6.70	5.25	8.93	0.82	3.29	3.28	2.61	4.27	0.33
10	1.47	1.46	1.36	1.64	0.05	1.52	1.52	1.41	1.67	0.05	1.78	1.76	1.51	2.17	0.12
11	1.47	1.47	1.36	1.64	0.06	1.78	1.77	1.57	2.02	0.10	1.74	1.73	1.56	2.01	0.11
12	1.43	1.43	1.31	1.55	0.05	2.04	2.04	1.67	2.48	0.15	1.66	1.65	1.47	1.87	0.08
RD[a] *in PMSE*															
1						9.78	10.00	−2.10	17.76	4.20	26.54	26.92	7.73	44.35	8.70
2						38.22	37.32	26.71	50.01	5.35	22.97	25.17	4.57	41.12	7.57
3						58.96	59.79	46.33	72.13	4.97	23.55	23.58	2.48	40.76	7.50
4						2.56	2.64	−2.27	6.76	1.68	16.39	16.37	5.59	29.17	5.09
5						16.04	15.62	8.83	22.94	3.32	14.63	14.64	−0.86	24.95	5.02
6						27.75	27.95	15.85	37.43	4.83	14.00	14.01	4.92	22.94	3.82
7						17.04	16.74	3.77	32.85	5.52	32.67	30.94	6.80	80.53	12.25
8						44.91	44.81	32.65	57.82	5.13	25.02	25.77	8.37	38.87	6.74
9						63.02	62.94	50.08	72.65	4.74	23.82	24.07	7.53	39.72	6.70
10						3.32	3.47	−1.16	7.93	1.93	17.10	17.15	6.11	28.60	4.71
11						16.92	16.73	7.30	25.37	3.86	15.08	14.72	6.19	24.77	4.02
12						29.50	29.84	15.02	40.14	4.85	13.49	13.47	3.33	19.59	3.66
	RT					LME					LM				
DGP	Avg	Med.	Min.	Max.	Std	Avg	Med.	Min.	Max.	Std	Avg	Med.	Min.	Max.	Std
PMSE															
1	6.10	6.02	4.73	8.32	0.77	7.12	7.07	6.40	8.25	0.33	7.52	7.50	6.81	8.63	0.33
2	7.37	7.32	5.59	9.92	0.89	6.06	6.05	5.46	6.79	0.24	8.07	7.99	7.20	9.39	0.53
3	8.82	8.86	6.44	12.16	1.04	4.74	4.73	4.14	5.29	0.21	8.76	8.76	7.33	11.66	0.74
4	2.01	2.00	1.74	2.40	0.14	2.06	2.05	1.89	2.25	0.07	2.10	2.10	1.91	2.31	0.07
5	2.23	2.23	1.94	2.53	0.13	1.92	1.92	1.77	2.08	0.06	2.19	2.20	1.98	2.41	0.09
6	2.41	2.41	2.11	2.74	0.14	1.74	1.74	1.62	1.86	0.06	2.29	2.29	2.05	2.64	0.13
7	5.01	4.98	4.13	6.37	0.48	5.26	5.25	4.68	5.71	0.20	5.74	5.73	5.24	6.27	0.23
8	6.60	6.49	5.22	8.56	0.71	4.52	4.49	3.96	5.04	0.19	6.72	6.67	5.82	8.03	0.52
9	8.03	7.93	5.89	10.72	0.96	3.61	3.58	3.21	4.02	0.16	7.62	7.55	6.15	9.76	0.76
10	1.83	1.82	1.57	2.22	0.12	1.74	1.74	1.63	1.85	0.05	1.79	1.79	1.67	1.90	0.05
11	2.05	2.03	1.77	2.35	0.14	1.68	1.68	1.53	1.81	0.05	1.96	1.95	1.76	2.16	0.09
12	2.27	2.29	1.77	2.66	0.17	1.56	1.56	1.40	1.68	0.05	2.13	2.14	1.78	2.58	0.14
RD[a] *in PMSE*															
1	32.01	32.96	10.89	48.23	7.93	42.38	42.44	30.88	49.02	3.27	45.42	45.51	33.35	51.71	3.33
2	49.73	49.81	36.46	62.17	5.61	39.57	39.74	31.78	46.09	2.90	54.48	54.76	46.11	61.87	3.38
3	65.21	65.63	54.42	75.67	4.39	36.19	36.14	26.74	43.28	2.83	65.20	65.79	55.59	74.62	3.55
4	18.40	18.31	5.07	29.57	5.32	20.74	21.10	12.92	27.07	2.55	22.27	22.73	15.19	29.99	2.64
5	28.47	28.81	15.53	35.81	3.93	16.97	17.14	9.44	20.56	2.30	27.34	26.81	22.62	33.58	2.88
6	36.48	36.57	23.15	46.52	4.02	12.01	12.09	4.80	18.05	2.27	33.17	32.99	24.25	42.35	4.04
7	40.08	39.94	22.76	64.37	8.75	40.89	41.50	27.76	53.92	5.23	47.92	47.23	38.17	57.20	3.58
8	55.36	55.24	43.65	65.00	4.07	35.36	35.55	28.57	41.87	3.10	56.35	56.01	48.98	64.74	3.31
9	68.63	68.81	57.00	76.48	4.02	31.13	31.07	23.33	39.03	2.93	67.08	67.39	56.88	74.98	3.68
10	19.48	19.39	6.91	32.31	4.86	15.58	15.62	10.56	19.89	2.05	17.89	17.71	12.12	23.78	2.28
11	28.07	27.94	18.21	37.24	4.39	12.37	12.65	7.13	17.17	2.12	24.76	25.09	14.23	31.56	3.25
12	36.75	37.16	20.03	47.90	4.79	8.42	8.65	1.63	13.69	2.27	32.58	32.73	20.11	44.14	4.43

[a]RD in PMSE $= ((\text{PMSE}_{\text{Alternative}} - \text{PMSE}_{\text{MERF}})/\text{PMSE}_{\text{Alternative}}) \times 100$.

Table 3. Results of the PMSE on the second test set (new clusters) of MERF. RF, MERT, RT, LME, and LM models based on 100 simulation runs.

	MERF					RF					MERT				
DGP	Avg	Med.	Min.	Max.	Std	Avg	Med.	Min.	Max.	Std	Avg	Med.	Min.	Max.	Std
PMSE															
1	4.61	4.58	3.69	5.40	0.32	4.57	4.54	3.65	5.31	0.32	6.03	5.93	4.33	8.27	0.72
2	5.83	5.81	4.76	7.28	0.48	5.94	5.93	4.78	7.29	0.51	6.87	6.84	5.39	8.41	0.59
3	7.23	7.18	5.89	9.47	0.70	7.50	7.46	5.93	9.79	0.73	8.07	8.02	6.41	10.24	0.82
4	1.68	1.69	1.52	1.87	0.07	1.68	1.68	1.52	1.88	0.07	2.00	1.99	1.67	2.40	0.14
5	1.89	1.87	1.66	2.12	0.10	1.90	1.90	1.67	2.15	0.10	2.15	2.14	1.86	2.53	0.13
6	2.10	2.11	1.78	2.46	0.12	2.14	2.15	1.74	2.51	0.13	2.34	2.34	2.00	2.72	0.14
7	3.83	3.84	3.33	4.31	0.24	3.78	3.82	3.25	4.29	0.25	4.97	4.92	4.09	7.00	0.44
8	5.22	5.24	4.34	6.51	0.41	5.31	5.27	4.37	6.59	0.43	6.12	6.08	4.97	7.48	0.50
9	6.74	6.71	4.87	8.61	0.72	6.97	6.93	5.08	9.15	0.74	7.42	7.42	5.57	9.31	0.77
10	1.53	1.52	1.42	1.69	0.06	1.53	1.52	1.42	1.66	0.05	1.84	1.82	1.61	2.09	0.10
11	1.78	1.78	1.56	2.07	0.09	1.80	1.79	1.59	2.08	0.09	2.03	2.03	1.80	2.43	0.13
12	2.02	2.00	1.76	2.42	0.13	2.06	2.04	1.76	2.43	0.13	2.23	2.21	1.99	2.64	0.14
RD[a] *in PMSE*															
1						−0.89	−0.71	−6.22	2.52	1.56	22.79	22.61	5.86	43.12	8.31
2						1.77	1.60	−5.31	8.56	2.46	14.96	14.94	1.42	31.36	5.66
3						3.55	3.51	−4.51	11.94	2.77	10.30	10.01	1.83	21.19	4.02
4						−0.19	−0.20	−3.06	1.81	0.88	15.52	15.65	2.70	27.90	4.95
5						0.87	0.95	−3.53	3.45	1.30	12.25	12.66	2.17	20.19	4.05
6						1.85	1.93	−2.88	4.92	1.52	10.02	10.16	3.91	16.10	2.56
7						−1.13	−1.17	−6.72	4.30	1.99	22.67	21.95	4.95	45.71	6.41
8						1.69	1.28	−3.68	10.60	2.50	14.46	14.31	1.87	23.43	4.83
9						3.18	3.12	−5.22	9.03	2.29	9.04	8.94	0.47	19.51	3.64
10						−0.18	−0.17	−2.17	2.26	0.78	16.35	16.21	5.68	27.38	4.36
11						0.93	0.93	−1.51	3.57	1.16	12.20	12.01	4.73	22.45	3.34
12						1.81	1.76	−1.95	5.45	1.29	9.23	9.26	2.22	14.76	2.68
	RT					LME					LM				
	Avg	Med.	Min.	Max.	Std	Avg	Med.	Min.	Max.	Std	Avg	Med.	Min.	Max.	Std
PMSE															
1	6.12	5.99	4.51	7.85	0.71	7.53	7.52	6.74	8.48	0.36	7.53	7.54	6.75	8.46	0.36
2	7.34	7.31	5.82	9.07	0.77	8.02	8.00	7.09	9.19	0.48	8.05	8.02	7.08	9.20	0.49
3	8.86	8.81	7.04	11.48	0.91	8.72	8.70	7.29	11.09	0.72	8.79	8.78	7.28	11.16	0.73
4	2.01	2.00	1.76	2.40	0.15	2.10	2.11	1.94	2.33	0.07	2.10	2.11	1.94	2.33	0.07
5	2.24	2.22	1.94	2.55	0.13	2.19	2.20	1.95	2.42	0.09	2.20	2.20	1.95	2.43	0.10
6	2.43	2.43	2.00	2.91	0.16	2.30	2.29	1.93	2.66	0.12	2.31	2.31	1.94	2.67	0.12
7	5.04	4.96	4.10	6.23	0.46	5.74	5.73	5.24	6.32	0.22	5.75	5.74	5.21	6.35	0.21
8	6.56	6.55	5.34	8.07	0.54	6.64	6.63	5.75	7.91	0.41	6.69	6.64	5.77	7.90	0.42
9	8.19	8.10	6.18	10.95	0.93	7.70	7.71	6.07	9.59	0.70	7.79	7.75	6.09	9.63	0.72
10	1.84	1.83	1.60	2.16	0.10	1.80	1.79	1.68	1.97	0.06	1.80	1.80	1.68	1.96	0.06
11	2.08	2.07	1.77	2.44	0.13	1.98	1.97	1.76	2.19	0.08	1.98	1.98	1.78	2.19	0.08
12	2.30	2.30	2.00	2.87	0.15	2.14	2.11	1.84	2.49	0.12	2.15	2.12	1.85	2.50	0.13
RD[a] *in PMSE*															
1	23.87	24.16	−2.83	43.15	8.71	38.79	39.09	33.05	45.52	2.77	38.83	38.99	32.82	45.48	2.76
2	20.15	19.53	5.04	37.26	6.87	27.37	27.27	19.58	34.19	2.60	27.66	27.72	20.07	34.45	2.63
3	18.23	19.09	5.63	29.05	5.07	17.15	16.98	11.36	21.69	2.00	17.85	17.79	9.52	23.78	2.31
4	15.95	15.55	4.84	27.90	5.33	19.99	19.79	13.68	24.77	2.26	19.97	19.81	13.58	24.44	2.28
5	15.67	15.77	3.82	24.45	3.91	13.99	14.13	8.05	18.12	2.01	14.19	14.37	8.17	18.70	2.04
6	13.47	13.37	4.08	18.95	3.29	8.41	8.28	4.33	12.47	1.60	9.02	8.99	3.40	13.83	1.77
7	23.65	23.52	5.84	38.52	6.58	33.38	33.26	27.51	41.31	2.79	33.49	33.37	27.96	41.82	2.81
8	20.30	20.10	8.49	33.36	4.76	21.43	21.32	16.59	27.87	2.42	21.93	21.79	17.12	28.59	2.59
9	17.54	17.52	6.72	31.26	4.44	12.58	12.62	4.21	19.85	2.52	13.49	13.41	6.20	20.68	2.46
10	16.62	16.04	6.90	31.33	4.69	14.99	15.06	9.94	21.24	2.03	15.05	15.10	9.96	21.24	2.03
11	14.10	13.99	5.73	22.47	3.82	9.95	10.16	4.90	13.70	1.88	10.13	10.27	5.22	14.29	1.94
12	11.91	11.70	4.00	21.89	3.33	5.46	5.48	−0.16	9.02	1.58	5.99	5.97	2.62	8.99	1.65

[a]RD in PMSE $= ((\text{PMSE}_{\text{Alternative}} - \text{PMSE}_{\text{MERF}})/\text{PMSE}_{\text{Alternative}}) \times 100$.

The data generating process (DGP) is as follows. Nine random variables are first generated from a multivariate normal distribution $(X_1, \ldots, X_9) \sim N(0, \Sigma)$ with Σ chosen such that all variables have unit variance and are correlated with $\sigma_{k,k'} = \rho$ for $k \neq k' \leq 9$. Then, the continuous response variable y is generated according to the following non-LM, using only the first three random variables:

$$\begin{aligned} y_{ij} &= m \times g(x_{ij}) + b_i + \varepsilon_{ij}, \\ g(x_{ij}) &= 2x_{1ij} + x_{2ij}^2 + 4(x_{3ij} > 0) + 2\log|x_{1ij}|x_{3ij}, \\ b_i &\sim N(0, \sigma_b^2), \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \ \ i = 1, \ldots, 100, j = 1, \ldots, n_i, \end{aligned} \tag{3}$$

where $m \times g(x_{ij})$ represents the response fixed part, with a non-linear form and a variance $\sigma^2_{\text{Fixed}} = m^2\sigma_g^2$. The parameter m simply serves as a tuning parameter to control the magnitude of σ^2_{Fixed} in the simulation design.

The proportion of total-effects variability (PTEV) of the model in Equation (3) is given by

$$\text{PTEV} = \frac{\sigma^2_{\text{Fixed}} + \sigma_b^2}{\sigma^2_{\text{Fixed}} + \sigma_b^2 + \sigma_\varepsilon^2} \times 100, \tag{4}$$

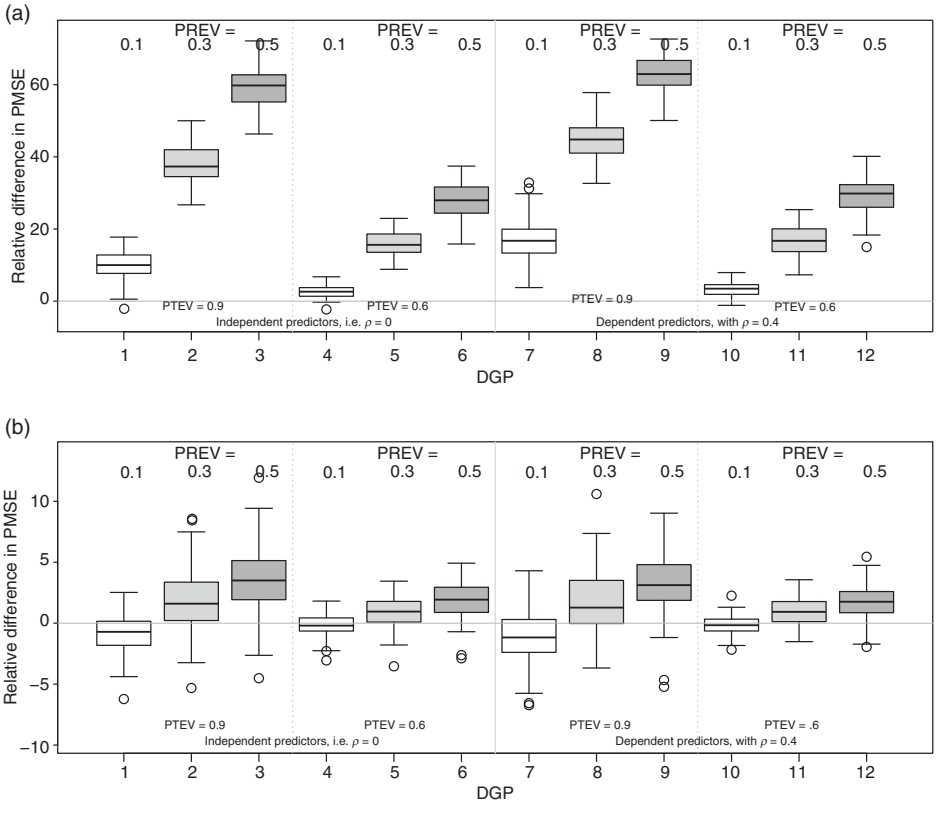

Figure 1. Distribution over the 100 simulation runs of the RD in PMSE, between MERF and RF, for (a) known clusters and (b) new clusters.

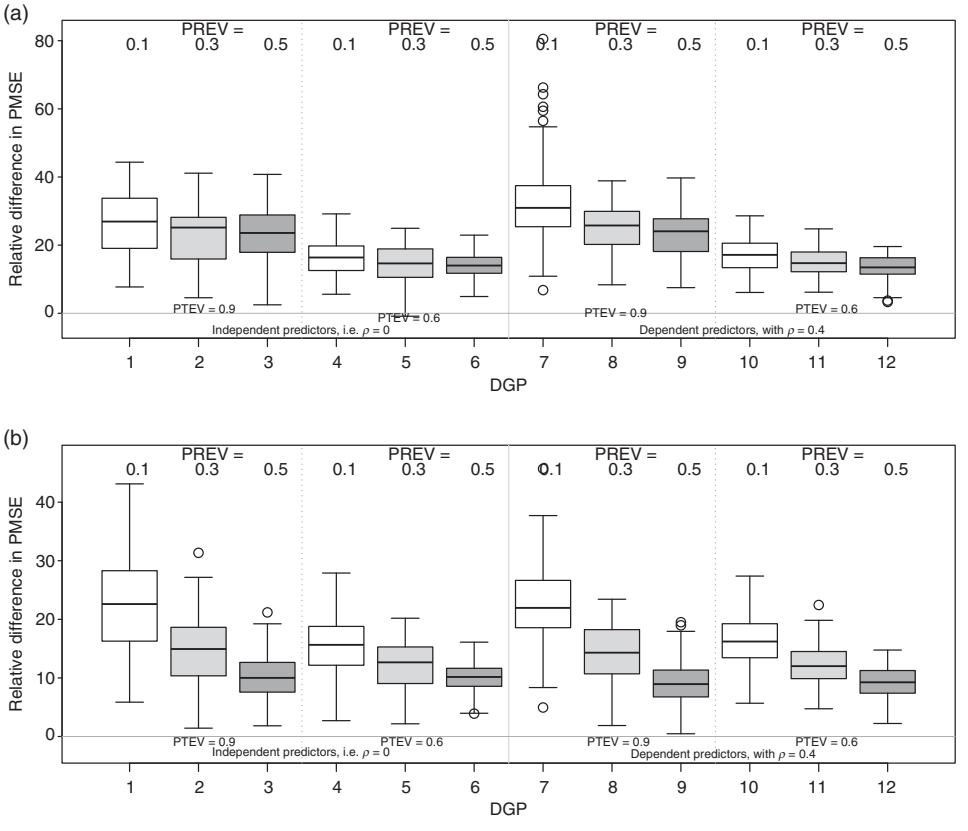

Figure 2. Distribution over the 100 simulation runs of the RD in PMSE, between MERF and MERT, for (a) known clusters and (b) new clusters.

and the proportion of random-effects variability (PREV) over total-effects variability is defined by

$$\text{PREV} = \frac{\sigma_b^2}{\sigma_{\text{Fixed}}^2 + \sigma_b^2} \times 100. \tag{5}$$

We consider 12 different DGPs, summarized in Table 1. In all cases, the within cluster variance σ_ε^2 is fixed at 1. We selected the values of 0 and 0.4 for ρ, 90% and 60% for PTEV (i.e. small and large noise), and 10%, 30%, and 50% for PREV (i.e. small, moderate, and large random effects).

Note that σ_g^2 depends only on the value of ρ. To estimate this variance, we conducted for each value of ρ a simulation where $g(x_{ij})$ was generated one million times. The observed variance was $\sigma_g^2 = 12.49$ when $\rho = 0$, and $\sigma_g^2 = 15.94$ when $\rho = 0.4$. We used these values of σ_g^2 in Equations (4) and (5) to obtain the values of m and σ_b^2 for each DGP in Table 1.

The simulation results are obtained by means of 100 runs.

3.2. *Simulation results*

Upper parts of Tables 2 and 3 present, for each DGP, the summary statistics of the PMSEs of the six fitted models on the first (known clusters) and the second (new clusters) test sets.

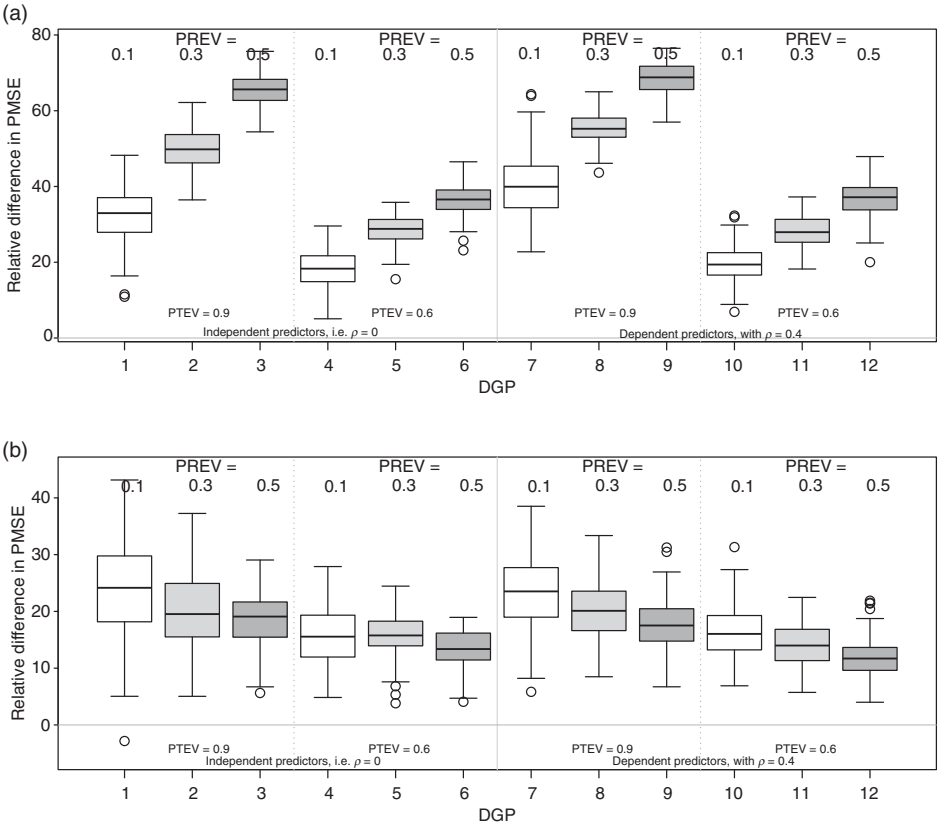

Figure 3. Distribution over the 100 simulation runs of the RD in PMSE, between MERF and RT, for (a) known clusters and (b) new clusters.

The two PMSEs are computed as

$$\text{PMSE} = \frac{\sum_{i=1}^{100}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_{ij})^2}{4500},$$

with $\hat{y}_{ij}$ being the predicted value of the jth observation in the ith cluster in the test set considered.

Lower parts of Tables 2 and 3 present the summary statistics of the relative differences (RDs) in PMSE, between each alternative model and MERF.

The RDs are computed as

$$\text{RD} = \frac{\text{PMSE}_{\text{Alternative}} - \text{PMSE}_{\text{MERF}}}{\text{PMSE}_{\text{Alternative}}} \times 100.$$

Box-plots of the distribution of these RDs are given in Figures 1–5. In order to provide a quick way to compare all of the methods at once, Figure 6 shows the average of all RDs.

The primary interest of this simulation study is the comparison of MERF and RF (Figure 1). Looking first at the upper plot (a) of Figure 1, giving the results for predicting observations in known clusters, the main finding is that for a given value of PTEV and ρ, the benefit of MERF over RF increases greatly as PREV increases. This can be seen by looking at the progression of RD between DGP 1, 2, and 3 (PTEV $= 0.9$ and $\rho = 0$), between DGP 4, 5, and 6 (PTEV $= 0.6$ and $\rho = 0$), between DGP 7, 8, and 9 (PTEV $= 0.9$ and $\rho = 0.4$) and finally by looking at the

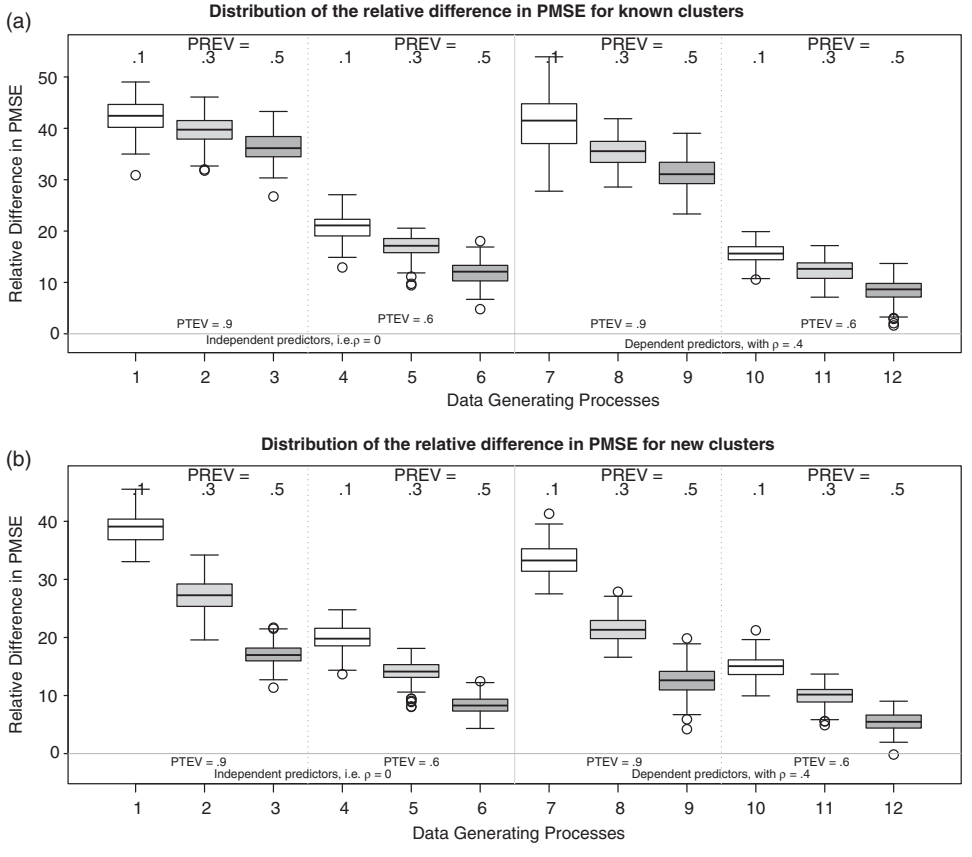

Figure 4. Distribution over the 100 simulation runs of the RD in PMSE, between MERF and LME, for (a) known clusters and (b) new clusters.

progression of RD between DGP 10, 11, and 12 (PTEV $= 0.6$ and $\rho = 0.4$). The largest median RD increase even reaches 62.9% (for DGP 9). This was expected as MERF, unlike RF, incorporates the cluster random effect into the final prediction. Thus, the more important the random effect is, the more gain in performance is achieved. The lower plot (b) of Figure 1 gives the results when we predict observations in new clusters that were not present in the training sample. This time, MERF uses only the forest prediction since the random effect is not available. However, since random effects are used during the training stage, the structure of the MERF trees can be different than those of the RF trees. The results show that MERF is still better than RF in most cases, but the improvement is modest compared with the known cluster case. The largest median RD increase is 3.55% (for DGP 3). Once again, the tendency is that, for a given value of PTEV and ρ, the benefit of MERF over RF increases as PREV increases. This suggests that, when the random effect is large, MERF can use it at the training stage to make some adjustments to the tree structures that will help the prediction of observations in new clusters. But it also shows that most of the benefit of MERF over RF is mainly due to the fact that the random effect is included in the prediction and this is why the improvement is much more important in the known cluster case. Inversely, it also shows that RF is not able to compensate appropriately for the omission of the random effects.

Looking at the other results, we observe that, except for a few cases (DGP 5 in Table 2 and DGPs 1 and 10 in Table 3), there is always some improvement (i.e. a positive minimum RD) of

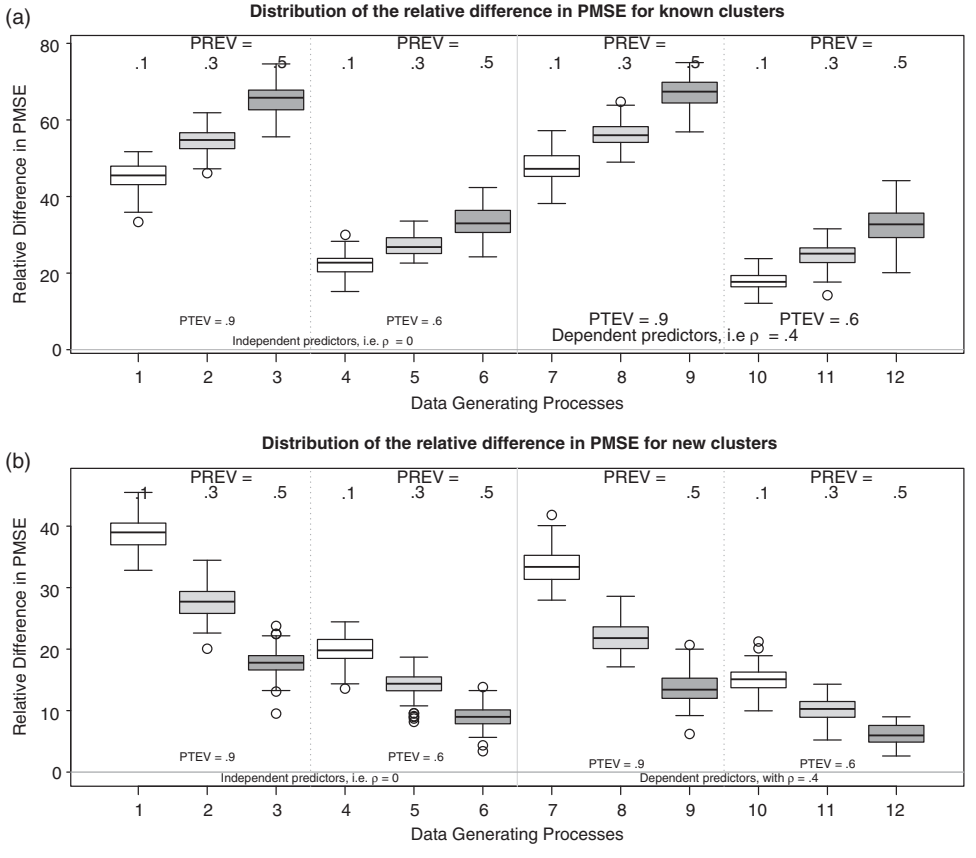

Figure 5. Distribution over the 100 simulation runs of the RD in PMSE, between MERF and LM, for (a) known clusters and (b) new clusters.

MERF over the other four alternative models: MERT, RT, LME, and LM (Tables 2 and 3 and Figures 2–5). In all cases, MERF did on average better (i.e. positive average and median RD) than these four alternative models. Indeed, even in cases where the random effects are relatively small (i.e. PREV = 10% in DGPs 1, 4, 7, and 10), the MERF median RD improvement over these four alternative models vary between 15.62% and 47.23% for new observations in known clusters, and between 15.06% and 39.09% for new observations in new clusters.

The most pronounced improvements of MERF over any alternative model appear in settings with relatively small noise (i.e. PTEV = 90% in DGPs 1, 2, 3 and 7, 8, 9). In addition, while the most pronounced improvements of MERF over models without random effects (i.e. RF, RT, and LM) appear in settings with large random effects (i.e. PREV = 50% in DGPs 3 and 9), the average improvement between MERF and the other mixed-effects models (i.e. LME, and MERT) is higher in settings with small random effects (i.e. PREV = 10% in DGPs 1 and 7). This is expected since the alternative mixed-effects models take into account the dependence of the data and estimate the random effects, as MERF do. Hence, when a considerable proportion of the response variability is explained by the random effects, the gap between their performance and that of MERF gets smaller, even when we predict new observations from new clusters.

In comparison with the MERF improvement over LME, the MERF improvement over MERT seems to be relatively less affected by the PREV (Figures 2 and 4). One additional and interesting point to notice is the relatively huge variability of the improvement over MERT in comparison

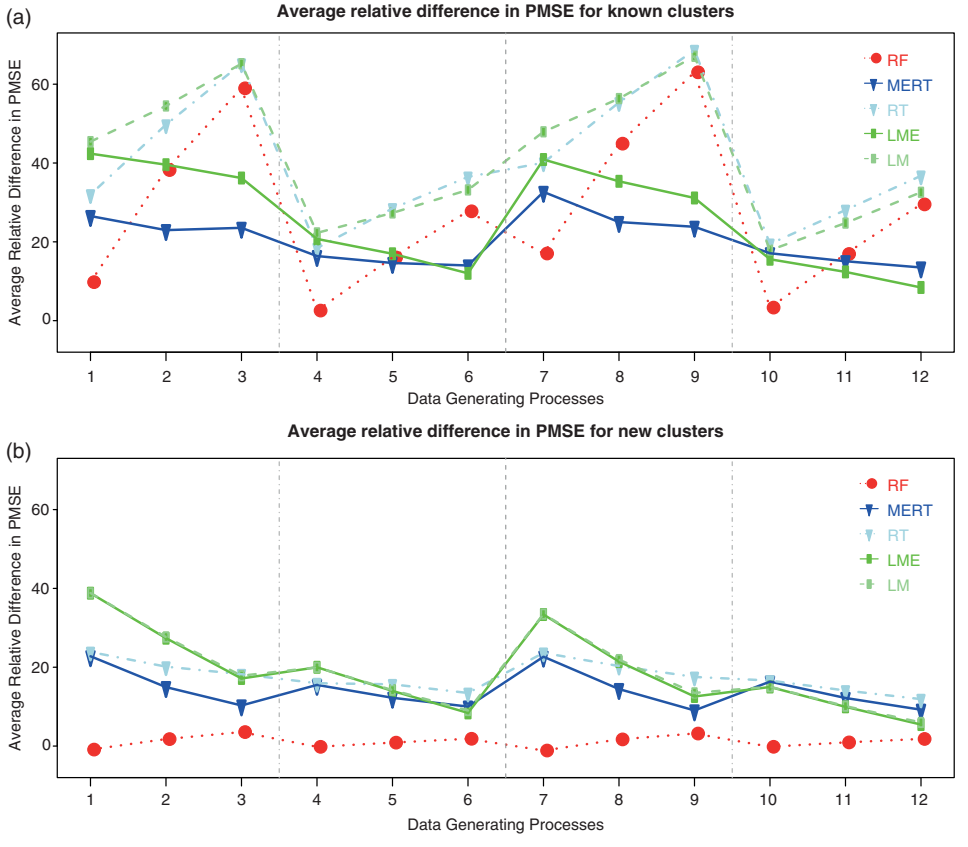

Figure 6. Average RD in PMSE, between MERF and each one of the alternative models: RF, MERT, RT, LME, and LM, for (a) known clusters and (b) new clusters.

with that over LME; the standard deviations of the improvement over MERT are more than twice than those of the improvement over LME (lower parts of Tables 2 and 3 and Figures 2 and 4).

Finally, there is no clear effect of the correlation between the predictors on the relative improvement of MERF over the alternative models.

4. Data example

We illustrate the proposed RF method using the same data set as in [7]. The data set consists of first-week box office revenues of 60,175 screens nested within 2656 new movies presented in the province of Québec in Canada from 2001 to 2008. On average, there are 22.7 screens per movie (minimum = 1; first quartile = 1; median = 8; third quartile = 47; maximum = 93). Each movie is treated as a cluster.

There are three screen-level covariates: (1) *Language* (1, French Version; 2, Original English Version; 3, Original French Version; 4, Original Version with Subtitles), (2) *Region* (1, Montréal; 2, Montérégie; 3, Québec City; 4, Laurentides; 5, Lanaudière; 6, Others), and (3) *Theater* owner (1, Independent; 2, Cinéplex; 3, Guzzo; 4, Ciné-enterprise; 5, Famous Players; 6, Cinémas R.G.F.M.; 7, Cinémas Fortune; 8, AMC).

In addition, there are eight movie-level covariates: (1) movie critics' *rating*, an ordinal covariate taking on values from 1 (the best) to 7 (the worst), (2) movie *length*, a continuous covariate ranging

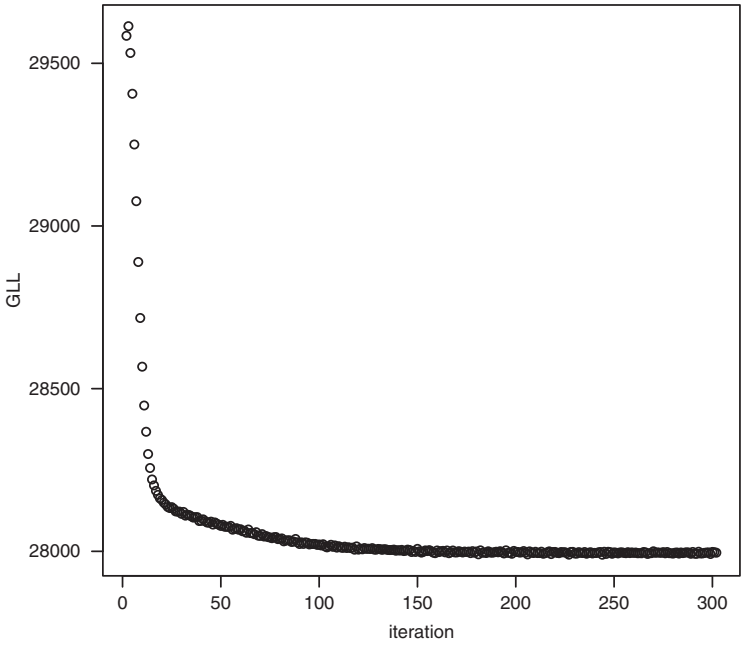

Figure 7. Behaviour of the GLL through the iteration process for fitting the MERF model for the example on first-week box office revenues.

Table 4. Results for the first-week box office revenues example.

		PMSE	Estimated ICC
RF	MERF	0.47	0.54
	RF	0.60	–
Single tree	MERT	0.53	0.51
	RT	0.90	–
Linear model	LME	0.62	0.42
	LM	1.00	–

Notes: The training data set has 60,175 observations (screens) nested within 2656 clusters (movies). The predictive mean-squared error (PMSE) is estimated with a hold-out sample of 30,157 observations nested within 1920 clusters.

between 70 and 227 min, (3) movie *genre* (1, Comedy; 2, Drama; 3, Thriller; 4, Action/Adventure; 5, Science Fiction; 6, Cartoons; 7, Others), (4) *Visa*, the assigned movie classification (1, General; 2, 13-year old; 3, 16-year old; 4, 18-year old), (5) *Month* of movie release, (6) Movie *distributer* (1, Vivafilm; 2, Sony; 3, Warner; 4, Fox; 5, Universal; 6, Paramount; 7, Disney; 8, Christal Films; 9, Films Séville; 10, DreamWorks; 11, MGM; 12, TVA Films; 13, Equinoxe; 14, Others), (7) *Country* of origin (1, USA; 2, Québec; 3, France; 4, Rest of Canada; 5, Other countries), and (8) *Size*, total number of screens for a movie in its first week (this is a common measure that approximates the marketing effort).

To allow comparison with the results of Hajjem *et al.* [7], we study the same relationship between the log transform of the first-week box office revenues and the 11 covariates, and we use the same learning sub-sample (30,018 screens within 2656 movies) and test sub-sample (30,157 screens within 1920 movies). We fit a standard RF model and a random intercept random forest (MERF) model. The results for the four other models (MERT, RT, LME and LM) were already given in [7] and are reproduced here for completeness. The number of trees to grow within each

forest is set to 300, and the number of variables randomly sampled as candidates at each split is set to 3. We set to 20 the minimum size of terminal nodes. For MERF convergence, we fixed the total number of iterations to 300. Figure 7 shows that the GLL stabilizes around iteration 150 well before reaching the last iteration (the absolute relative change in GLL equals 5E–5 at the last iteration, and its average over the last 150 iterations equals 1E–4).

Table 4 presents the out-of-sample performance of the six models.

MERF has the best predictive performance among all the alternative models evaluated; its PMSE is 0.47 while the PMSE of RF and MERT models are 0.60 and 0.53, respectively. Thus, MERF reduces the PMSE of the RF model by 21.49% and of the MERT model by 10.75%. It is noteworthy that in this example, the standard RF has a better performance than the LMEs model even if it neglects the clustering structure. The estimated intraclass correlation (ICC) for the three models involving a random intercept varies between 0.42 and 0.54, indicating the need to account for the clustered nature of the data.

5. Concluding remarks

The focus in this paper is on one specific and very common form of clustered data consisting of individuals nested within groups. Longitudinal data are another common form where each individual forms its own group. The proposed mixed-effects forest approach can also be applied to analyse longitudinal data. Indeed, we can adjust a forest of trees where the time period and other time-varying covariates, as well as baseline measures (e.g. individual characteristics or experimental treatments) are used as candidates in the splitting process. The proposed algorithm, however, assumes that the correlation structure is solely induced via the between-cluster variation. For data sets with a short time series, Bryk and Raudenbush [17] noted that this assumption is often most practical and unlikely to distort the results. However, if another covariance structure is needed, the proposed method would require some modifications. One possible avenue would be to generalize the EM algorithm to estimate alternative covariance structures. To this end, Jennrich and Schluchter [18] described an hybrid EM scoring algorithm. Alternatively, within the Sela and Simonoff [11] RE–EM framework, it is quite straightforward to estimate more general correlation structures since it is based on the *lme* function of the R *nlme* package. Building a forest of RE–EM trees with a specific correlation structures would then be one possibility.

One key feature of the RF approach is the need to resample the observations. With independent observations, using the standard bootstrap by resampling the individual observations works perfectly. However, things are not straightforward with clustered data. One key assumption of the approach proposed in this paper is that the random effects totally explain the intracluster correlation. Hence, the observations are independent once the random effects have been removed. This allows the use of standard bootstrap resampling after removing the random effects from the responses (see Step 1(ii) of the algorithm). The simulation results and the results for the real data set showed that this approach seems reasonable. A possibility for future work would be to investigate the robustness of the proposed approach when the intracluster correlation is not entirely explained by the random effects.

Another entirely different approach would be to build directly a forest of MERTs. With this approach, a bootstrap sample would be required for each individual MERT. However, since the original observations are possibly correlated, taking a standard bootstrap sample may not be the best choice. Bootstrapping directly clustered data can be done in different ways [19]. The three following strategies are possible: (1) resampling individual observations (observation bootstrap), (2) resampling entire clusters (cluster bootstrap), and (3) resampling of clusters and then of observations within them (two-stage bootstrap). Karpievitch *et al.* [20] proposed the RF++

method which performs cluster-based bootstrapping (i.e. strategy 2) to create learning data for single trees in a standard RF predictor. Adler *et al.* [21] found that resampling of clusters and then sampling one observation from them (strategy 3) is better compared with sampling entire clusters (strategy 2) since it further reduces similarity between single trees. One possibility for future work would be to investigate these strategies and compare them to the approach proposed in this paper. It is, however, important to note that the methods proposed in the latter two papers do not provide predictions of the random effects. They are basically adjusting the sampling method for clustering but do not incorporate random effects in the predictions. This is in contrast with the proposed approach which predicts the random effects and uses them in the final predictions of the response.

The objective of this paper is to propose a method to build a forest of trees with clustered data and to evaluate its performance. The results of the simulation study and of the real data set are promising. This new approach could lead the way for future research on ensemble methods for clustered data.

An R program implementing MERF is available from the first author.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by Le Fonds québécois de la recherche sur la nature et les technologies (FQRNT). The authors thank a reviewer for constructive and pertinent comments. They want to thank the Carmelle and Rémi Marcoux Chair in Arts Management for providing the movie box office data used in the example, Renaud Legoux for interesting discussions, and Mohamed Jendoubi for preparing the data set.

References

[1] L. Breiman, *Random forests*, Mach. Learn. 45 (2001), pp. 5–32.
[2] M. Hamza and D. Larocque, *An empirical comparison of ensemble methods based on classification trees*, J. Statist. Comput. Simul. 75 (2005), pp. 629–643.
[3] G. Biau, L. Devroye, and G. Lugosi, *Consistency of random forests and other averaging classifiers*, J. Mach. Learn. Res. 9 (2008), pp. 2015–2033.
[4] L. Rokach, *Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography*, Comput. Stat. Data Anal. 53 (2009), pp. 4046–4072.
[5] D.S. Siroky, *Navigating random forests and related advances in algorithmic modeling*, Stat. Surveys 3 (2009), pp. 147–163.
[6] A. Verikas, A. Gelzinis, and M. Bacauskiene, *Mining data with random forests: A survey and results of new tests*, Pattern Recognit. 44 (2011), pp. 330–349.
[7] A. Hajjem, F. Bellavance, and D. Larocque, *Mixed effects regression trees for clustered data*, Stat. Probab. Lett. 81 (2011), pp. 451–459.
[8] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1984.
[9] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. R. Statist. Soc. Ser. B 39 (1977), pp. 1–38.
[10] G.J. McLachlan and T. Krishman, *The EM Algorithm and Extensions*, Wiley, New York, 1997.
[11] R.J. Sela and J.S. Simonoff, *RE-EM trees: A data mining approach for longitudinal and clustered data*, Mach. Learn. 86 (2012), pp. 169–207.
[12] S.W. Raudenbush and A.S. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Method*, 2nd ed., Sage, Newbury Park, CA, 2002.
[13] H. Wu and J.T. Zhang, *Nonparametric regression methods for longitudinal data analysis: Mixed-effects modeling approaches*, Wiley, New York, 2006.
[14] R Development Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2007. Available at www.R-project.org.
[15] A. Liaw and M. Wiener, *Classification and regression by randomForest*, R News 2 (2002), pp. 18–22.
[16] T.M. Therneau and E.J. Atkinson, *An introduction to recursive partitioning using the rpart routines*, Tech. Rep. 61, Department of Health Science Research, Mayo Clinic, Rochester, 1997.
[17] A.S. Bryk and S.W. Raudenbush, *Application of hierarchical linear models to assessing change*, Psychol. Bull. 101 (1987), pp. 147–158.

[18] R.I. Jennrich and M.D. Schluchter, *Unbalanced repeated-measures with structured covariance matrices*, Biometrics 42 (1986), pp. 805–820.
[19] C.A. Field and A.H. Welsh, *Bootstrapping clustered data*, J. R. Statist. Soc. Ser. B 69 (2007), pp. 369–390.
[20] Y.V. Karpievitch, E.G. Hill, A.P. Leclerc, A.R. Dabney, and J.S. Almeida, *An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++*, PLoS ONE 4(9) (2009), Article no. e7087.
[21] W. Adler, S. Potapov, and B. Lausen, *Classification of repeated measurements data using tree-based ensemble methods*, Computational Statistics 26 (2011), pp. 355–369.