HEC Montréal

Affiliée à l'université de Montréal

Mixed Effects Trees and Forests for Clustered Data

PAR

AHLEM HAJJEM

THÈSE

présentée en vue de l'obtention du grade

de Philosophiæ Doctor (Ph.D.) en administration

Spécialisation : Méthodes Quantitatives

SEPTEMBRE 2010

HEC Montréal

Affiliée à l'université de Montréal

Cette thèse est intitulée :

Mixed Effects Trees and Forests for Clustered Data

Présentée par Ahlem Hajjem

a été évaluée par un jury composé des personnes suivantes:

Marc Fredette

.....

Président rapporteur

François Bellavance

.....

Co-directeur de recherche

Denis Larocque

.....

Co-directeur de recherche

Cataldo Zuccaro

.....

Membre du jury

Thierry Duchesne

.....

Examinateur externe

Raf Jans

.....

Représentant du doyen de la FES

RÉSUMÉ

Les méthodes d'arbres, qui sont des outils populaires et appréciés d'analyse et d'exploitation de données, étaient à l'origine développées sous l'hypothèse de données indépendantes. Les travaux antérieurs qui ont adapté ces méthodes aux données corrélées sont basés sur l'approche multivariée des mesures répétées. L'objectif principal de cette thèse est d'adapter la méthode d'arbre standard aux données corrélées du fait de leur structure hiérarchique. Pour cela, nous avons suivi une approche par les effets mixtes. Cette approche est plus flexible en ce qui concerne les données puisque les observations corrélées sont perçues comme étant imbriquées à l'intérieur des groupes et non pas comme des vecteurs de réponses multiples.

Cette thèse est composée de trois articles. Dans le premier article, nous procédons à une extension de la méthode d'arbre de régression standard aux données hiérarchiques avec une variable de réponse continue. Nous proposons alors une méthode d'arbre nommée "mixed effects regression tree" (MERT). Dans le second article, nous procédons à une extension de la méthodologie MERT à d'autres types de réponses (réponses binaires, données de comptage, réponses catégorielles ordonnées, réponses multicatégorielles nominales). Pour cela, nous proposons une méthode d'arbre nommée "generalized mixed effects regression tree" (GMERT). Nous proposons dans le troisième article la méthode de forêt aléatoire à effets mixtes, nommée "mixed effects random forest" (MERF).

Les résultats des études de simulations menées dans les trois articles montrent qu'en présence de corrélation intra-groupe, les nouvelles méthodes d'arbres sont préférables à celles supposant l'indépendance des données.

Mots clés : Méthodes d'arbres, forêt aléatoire, données hiérarchiques, effets mixtes, algorithme d'espérance-maximisation (EM), quasi-vraisemblance pénalisée (PQL).

ABSTRACT

Tree based methods, which are very popular and appreciated data analysis tools, were firstly developed under the assumption of independent data. Previous works adapting them to correlated data are based on the multivariate repeated-measures approach. The main goal of this thesis is to extend standard tree methods to clustered and hence correlated data, using the mixed effects approach. This approach is more flexible in terms of data requirements because the correlated observations are viewed as nested within clusters rather than as vectors of multivariate repeated responses.

This thesis is composed of three articles. In the first paper, we propose the "mixed effects regression tree" (MERT) method. It is an extension of the standard regression tree method to the case of clustered data with continuously measured outcome. The second paper presents the generalized mixed effects regression tree (GMERT) method, which is an extension of MERT methodology to other types of outcomes (binary outcomes, counts data, ordered categorical outcomes, and multicategory nominal scale outcomes). We propose in the third paper the "mixed effects random forest" (MERF) method, which is an extension of the standard random forest method to the case of clustered data with continuously measured outcome.

The results of the simulations studies conducted in the three papers show that, when cluster-correlation is present, the new tree methods are preferable over the standard ones assuming independence of the data.

Keywords : Tree based methods, random forest, clustered data, mixed effects, expectation-maximization (EM) algorithm, penalized quasi-likelihood (PQL) algorithm.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

REMERCIEMENTS

Je remercie Dieu pour m'avoir donné la volonté et la patience pour accomplir ce travail.

Un grand merci à ma famille pour son soutien et encouragement.

Je voudrais aussi exprimer toute ma gratitude envers mes co-directeurs, Professeur François Bellavance et Professeur Denis Larocque, pour leur grande disponibilité, leur implication, et leurs conseils précieux.

Je remercie aussi tous les membres du jury pour avoir accepté de lire et commenter ce travail.

Mes remerciements s'adressent également à HEC Montréal, au Conseil de Recherche en Sciences Naturelles et en Génie du Canada (CRSNG), et au Fonds Québécois de la Recherche sur la Nature et les Technologies (FQRNT), pour leur support financier.

Merci infiniment !

INTRODUCTION GÉNÉRALE

Les méthodes d'arbres sont des techniques traditionnelles d'analyse et d'exploitation de données. Elles sont devenues populaires grâce à l'algorithme CART (classification and regression trees) de Breiman et al. (1984). Comparativement aux modèles de régression paramétriques, ces méthodes ont plusieurs avantages : Elles peuvent analyser facilement des grandes bases de données comprenant un nombre élevé de covariables, elles peuvent détecter de façon automatique les interactions potentielles entre ces dernières, et elles sont robustes face aux problèmes d'observations extrêmes et de colinéarité.

Les méthodes d'arbres supposent l'indépendance des données. Or, cette hypothèse n'est certainement pas satisfaite dans le cas de données hiérarchiques. Ces dernières sont souvent obtenues par un échantillonnage multiniveaux, où les observations sont imbriquées à l'intérieur d'unités de niveau supérieur (groupes). Elles sont communément présentes dans plusieurs champs de recherche (e.g., Raudenbush and Bryk, 2002 ; Goldstein, 2003 ; Fitzmaurice, Laird, and Ware, 2004). La structure hiérarchique de ces données implique que les observations provenant d'un même groupe sont souvent plus similaires entre elles que les observations provenant de groupes différents. Souvent, ces données comprennent deux types de covariables, celles décrivant l'observation au niveau hiérarchique inférieur et celles décrivant le groupe, et incluent deux sources de variations, intra- et inter- groupes. Des effets fixes mais aussi aléatoires servent à expliquer, au moins partiellement, ces deux sources de variabilité.

L'objectif principal de cette thèse est d'adapter les méthodes d'arbres standards aux données hiérarchiques, et ce en suivant une approche par les effets mixtes (fixes et aléatoires). Les travaux antérieurs (Segal, 1992 ; Zhang, 1998 ; Abdolell, Leblanc, Stephens, and Harrison, 2002 ; Lee, 2005) qui ont étendu les méthodes d'arbres dans le but d'accommoder la dépendance des données sont basés sur l'approche multivariée des mesures répétées. L'approche par les effets mixtes est plus flexible en termes de données parce que les observations corrélées sont perçues comme étant imbriquées à l'intérieur des groupes plutôt que comme des vecteurs de réponses multiples. Il y a un avantage à suivre cette approche puisqu'elle permet : 1) d'analyser des données où les groupes

sont de tailles inégales, 2) de considérer les covariables du niveau observation dans le processus d'embranchement, ce qui permet de séparer les observations provenant d'un même groupe dans des noeuds différents, et 3) d'inclure des effets aléatoires.

Trois articles font l'objet de cette thèse. Dans le premier article, nous proposons une extension des méthodes d'arbres standards aux données hiérarchiques avec une variable réponse continue. Nous avons nommé cette extension "mixed effects regression tree" (MERT). Nous l'avons implémenté en utilisant un algorithme d'arbre standard à l'intérieur du cadre bien connu de l'algorithme "espérance-maximisation" (EM). Nous l'avons aussi illustré en analysant des données sur les revenus du box-office de la première semaine des films présentés dans la province de Québec au Canada sur la période allant de 2001 à 2008. Les résultats de la simulation montrent que la performance prédictive de MERT est meilleure que celle de l'arbre de régression standard, en particulier lorsque les effets aléatoires sont importants.

Dans le deuxième article, nous proposons une extension de la méthodologie d'arbre de régression à effets mixtes (MERT), qui est conçue pour une réponse continue, à d'autres types de réponses (réponses binaires, données de comptage, réponses catégorielles ordonnées, réponses multicatégorielles nominales). Nous avons nommé cette extension "generalized mixed effects regression tree" (GMERT). Cette méthode utilise la quasi-vraisemblance pénalisée (PQL) pour l'estimation et l'algorithme espérance-maximisation (EM) pour la computation. Les résultats de l'étude de simulation menée pour le cas de réponse binaire montrent qu'en présence d'effets aléatoires la méthode GMERT a une performance prédictive nettement meilleure que celle de l'arbre de classification standard.

Par ailleurs, la performance prédictive d'un seul arbre peut souvent être améliorée au dépend de l'interprétabilité en utilisant un ensemble d'arbres. Le bagging et la forêt aléatoire en général (Breiman, 1996, 2001) sont des méthodes ensemblistes très connues et très puissantes dans le cas des arbres. Sur la base des conclusions des deux premiers articles, il est devenu clair que l'application directe de l'algorithme standard de forêt aléatoire aux données hiérarchiques impliquerait nécessairement une performance prédictive moins qu'optimale de la part de chaque arbre individuel à l'intérieur de la forêt. Ainsi, nous proposons dans le troisième article une méthode de forêt aléatoire à effets mixtes. Nous avons nommée cette méthode "mixed effects random forest" (MERF). Il s'agit d'une extension de la méthode standard de forêt aléatoire aux données hiérarchiques avec une

réponse continue. Nous l'avons implémentée en utilisant un algorithme standard de forêt aléatoire à l'intérieur de l'algorithme EM. Les résultats de la simulation menée dans cet article sont prometteurs et montrent que le gain sur le plan prédictif suite à l'utilisation de MERF à la place de la forêt standard augmente en fonction de l'importance des effets aléatoires.

ARTICLE I

MIXED EFFECTS REGRESSION TREES 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

1.1 Abstract

This paper presents an extension of the standard regression tree method to clustered data. Previous works extending tree methods to accommodate correlated data are mainly based on the multivariate repeated-measures approach. We propose a "mixed effects regression tree" method where the correlated observations are viewed as nested within clusters rather than as vectors of multivariate repeated responses. The proposed method can handle unbalanced clusters, allows observations within clusters to be splitted, and can incorporate random effects and observation-level covariates. We implemented the proposed method using a standard tree algorithm within the framework of the expectation-maximization (EM) algorithm. The simulation results show that the proposed regression tree method provide substantial improvements over standard trees when the random effects are non negligible. A real data example illustrates the proposed method.

Keywords : Tree based methods, clustered data, mixed effects, expectation-maximization (EM) algorithm.

1.2 Introduction

Clustered data, often obtained by multistage sampling with observations nested within higher-level units (clusters), is common throughout many areas of research (e.g., Raudenbush and Bryk, 2002 ; Goldstein, 2003 ; Fitzmaurice, Laird, and Ware, 2004). The data structure consists of individuals nested within groups. These data may include two types of covariates, observation-level and cluster-level covariates, and involve two sources of variation, within and between clusters. Usually, observations that belong to the same cluster tend to be more similar to each other than observations from different clusters. The focus of this paper is to extend the standard regression tree methods to clustered data and therefore take into account the correlation between observations within a cluster.

Tree based methods became popular with the CART (classification and regression trees) paradigm (Breiman, Friedman, Olshen, and Stone, 1984). They provide many advantages compared to parametric models : They can handle large data sets with many covariates, they are robust to outliers and collinearity problems, and they detect automatically potential interactions between covariates.

If a standard tree algorithm is directly applied to clustered data, any tree node could include observations belonging to different clusters, and the question of which summary response value should be attached to them arises, i.e. overall average response or cluster-specific average response within each node. Furthermore, the inclusion of the observation-level and cluster-level covariates as candidates in the splitting process is not always enough to ensure that the nested structure of the data is fully taken into account. Not considering the clustered aspect of the data in the splitting process constitutes an evident loss of likely valuable information. To that end, statistical models to analyze clustered data often imply an additional random-effect component in addition to the fixed-effect component. The larger the random effects, the harder it will be for a standard tree algorithm to find the right tree structure, which should affect negatively the prediction accuracy. This will be illustrated in the simulation study in Section 1.4.

To legitimize the application of standard tree methodology to clustered data, one could remove the random or cluster-specific component, and then apply a standard tree algorithm, such as CART, only to the fixed or population-averaged component. This constitutes the key point of the regression tree approach presented in this paper, named "mixed effects regression tree". It is an extension of standard regression trees to clustered data that can appropriately deal with random effects.

The proposed mixed effects regression tree method have the following characteristics :

1. It can handle clusters with different numbers of observations (unbalanced clusters).
2. It allows the inclusion of observation-level and cluster-level covariates in the splitting process, and consequently, observations from the same cluster can be separated into different nodes during the tree growing process.
3. It allows observation-level covariates to have random effects.

Previous extensions of tree based methods to accommodate the correlation structure induced by clustered data were developed for longitudinal settings (e.g., Segal, 1992 ; Zhang, 1998 ; Yu and Lambert, 1999 ; Abdolell, Leblanc, Stephens, and Harrison, 2002 ; Lee, 2005 ; Ghattas and Nerini, 2007). These extensions do not allow observations within a cluster (i.e.

repeated observations over time for a given subject) to be splitted into different nodes.

This paper presents and evaluates an extension of regression trees for clustered data. The remainder of this article is organized as follows : Section 1.3 describes the proposed mixed effects regression tree approach ; Section 1.4 presents a simulation study to evaluate the performance of the method ; Section 1.5 illustrates the application of the method with a real data set ; Section 1.6 discusses a number of related issues.

1.3 Mixed Effects Regression Tree Approach

Statistical model for clustered data typically include two components : A fixed or population-averaged and a random or cluster-specific component. The basic idea behind the proposed mixed effects regression tree is to dissociate the fixed from the random effects. We use a standard regression tree to model the fixed effects and a node-invariant linear structure at each terminal node of the tree to model the random effects. The method is implemented using a standard tree algorithm within the framework of the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977 ; McLachlan and Krishnan, 1997). More precisely, the linear estimation of the fixed component in the linear mixed effects (LME) model (Harville, 1976, 1977 ; Laird and Ware, 1982) is replaced by a standard regression tree algorithm. Let's first briefly review the LME model and the EM algorithm.

1.3.1 EM Algorithm for the Linear Mixed Effects Model

The LME model is generally written in the following form :

$$\begin{gathered} y_i = X_i\beta + Z_i b_i + \epsilon_i, \\ b_i \sim N_q(0, D), \epsilon_i \sim N_{n_i}(0, R_i), \\ i = 1, ..., n, \end{gathered} \tag{1.1}$$

where $y_i = [y_{i1}, ..., y_{in_i}]^T$ is the $n_i \times 1$ vector of responses for the n_i observations in cluster i, $X_i = [x_{i1}, ..., x_{in_i}]^T$ is the $n_i \times p$ matrix of fixed-effects covariates, $Z_i = [z_{i1}, ..., z_{in_i}]^T$ is the

$n_i \times q$ matrix of random-effects covariates, $\epsilon_i = [\epsilon_{i1}, ..., \epsilon_{in_i}]^T$ is the $n_i \times 1$ vector of errors, $b_i = (b_{i1}, ..., b_{iq})^T$ is the $q \times 1$ unknown vector of random effects for cluster i, and β is the $p \times 1$ unknown vector of parameters for the fixed effects. 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. The usual LME model also assumes that b_i and ϵ_i are independent and normally distributed and that the between-clusters observations are independent. Hence, the covariance matrix of the vector of observations y_i in cluster i is $V_i = Cov(y_i) = Z_i D Z_i^T + R_i$, and $V = Cov(y) = diag(V_1, \ldots, V_n)$, where $y = [y_1^T, ..., y_n^T]^T$. We will further assume that the correlation is induced solely via the between-clusters variation, that is, R_i is diagonal ($R_i = \sigma^2 I_{n_i}, i = 1, ..., n$). This assumption is suitable for a large class of clustered data problems (Raudenbush and Bryk, 2002, page 30).

The parameters in LME models can be estimated by the method of maximum likelihood (ML) implemented with the EM algorithm. This algorithm addresses the problem of maximizing the likelihood by considering it like a missing data problem. More precisely, the y_i are the observed data and the b_i are the missing data. Thus, the complete data are (y_i, b_i), $i = 1, ..., n$, while β, σ^2, and D are the parameters to be estimated. The general technique is to calculate the expected values of the missing objects, given current parameter estimates (expectation step), and then to use those expected values to update the parameter estimates (maximization step). These two steps are repeated until convergence.

The major cycle for the ML-based EM-algorithm, as described in §2.2.5 of Wu and Zhang (2006), is as follows :

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

Step 1. Set $r = r + 1$. Update $\hat{\beta}_{(r)}$ and $\hat{b}_{i(r)}$

$$\hat{\beta}_{(r)} = \left(\sum_{i=1}^{n} X_i^T \hat{V}^{-1}_{i(r-1)} X_i \right)^{-1} \left(\sum_{i=1}^{n} X_i^T \hat{V}^{-1}_{i(r-1)} y_i \right),$$
$$\hat{b}_{i(r)} = \hat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} \left(y_i - X_i \hat{\beta}_{(r)} \right), i = 1, ..., n,$$

where $\hat{V}_{i(r-1)} = Z_i \hat{D}_{(r-1)} Z_i^T + \hat{\sigma}^2_{(r-1)} I_{n_i}, i = 1, ..., n.$

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

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

where $\hat{\epsilon}_{i(r)} = y_i - X_i \hat{\beta}_{(r)} - Z_i \hat{b}_{i(r)}$, $N = \sum_{i=1}^{n} n_i$.

Step 3. Repeat steps 1 and 2 until convergence.

1.3.2 EM Algorithm for the Mixed Effects Regression Trees

The proposed mixed effects regression tree model is :

$$\begin{gathered} y_i = f(X_i) + Z_i b_i + \epsilon_i, \\ b_i \sim N_q(0, D), \epsilon_i \sim N_{n_i}(0, R_i), \\ i = 1, ..., n, \end{gathered} \tag{1.2}$$

where all quantities are defined as in Section 1.3.1 except that the linear fixed part $X_i\beta$ in (1.1) is replaced by the function $f(X_i)$ that will be estimated with a standard tree based model. The random part, $Z_i b_i$, is still assumed linear.

The mixed effects tree algorithm is the ML-based EM-algorithm in which we replace the linear structure used to estimate the fixed part of the model by a standard tree structure. The algorithm 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) $y^*_{i(r)} = y_i - Z_i \hat{b}_{i(r-1)}, i = 1, ..., n,$

ii) Let $\hat{f}(X_i)_{(r)}$ be an estimate of $f(X_i)$ obtained from a standard tree algorithm with $y^*_{i(r)}$ as responses and X_i, $i = 1, \ldots, n$, as covariates. Note that the tree is built

as usual using all N individual observations as inputs along with their covariate vectors,

iii) $\hat{b}_{i(r)} = \hat{D}_{(r-1)} Z_i^T \hat{V}_{i(r-1)}^{-1} \left(y_i - \hat{f}(X_i)_{(r)} \right), i = 1, ..., n,$

where $\hat{V}_{i(r-1)} = Z_i \hat{D}_{(r-1)} Z_i^T + \hat{\sigma}^2_{(r-1)} I_{n_i}, i = 1, ..., n.$

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

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

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

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

Step 3. Repeat 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^*, i.e., the response variable from which we remove the current available value of the random part. Second, it estimates the fixed component $\hat{f}(X_i)$ using a standard tree algorithm with y_i^* as responses and X_i as covariates. Third, it updates $\hat{b}_i$. At step 2, it updates the variance components $\hat{\sigma}^2$ and $\hat{D}$ based on the residuals after the estimated fixed component $\hat{f}(X_i)$ is removed from the raw data y_i It 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 :

$$\begin{aligned} GLL(f, b_i|y) = \sum_{i=1}^{n} \{ & [y_i - f(X_i) - Z_i b_i]^T R_i^{-1} [y_i - f(X_i) - Z_i b_i] \\ & + b_i^T D^{-1} b_i + \log |D| + \log |R_i| \}. \end{aligned} \tag{1.3}$$

At each iteration, a single large tree is built and a subtree is selected using a pruning and cross-validation method. Doing so introduces instability over the iteration process. Indeed, a small change in the updated data (i.e., $y^*_{i(r)}$) could produce a selected subtree with a

different number of leaves (terminal nodes). In order to give insight about the behavior of GLL, Figure 1.1 shows the iteration process for one data set in one simulation run from the simulation study described in more details in the next section. The GLL decreases sharply at the beginning and stabilizes around iteration 40, but its value jumps once in a while from iteration 50 to 200 (Figure 1.1d). These jumps occur when there is a change in the number of leaves of the tree (Figure 1.1a). We also observe these jumps in the estimated variance parameters (Figure 1.1c) and in the mean squared errors (Figure 1.1b). This is mainly due to the instability associated with the choice of a single subtree at each iteration. All subtree structures in this simulation run are exactly the same except that those with only three terminal nodes do not have the split on the variable X_2 (see Figure 1.1).

Insert Figure 1.1 about here

In practice, we suggest the following method to stop the iteration process and select a final subtree model. First, we impose a minimum number of iterations to avoid early stopping (e.g. 50), then we keep iterating until the absolute change in GLL is less than a given small value (e.g. 1E-06). Once the stopping criterion is reached, we let the process continue for an additional pre-determined number of iterations (e.g. 50 in Figure 1.1). We then find the most frequent (modal value) number of leaves for the selected subtrees in the sequence of additional iterations. The final subtree model chosen is the one corresponding to the last iteration where the number of leaves is equal to the modal value. In the example presented in Figure 1.1, the subtree model selected is the one in the very last iteration, a tree with four leaves since it is the most frequent number of leaves in the 50 additional iterations after the GLL stabilizes.

This algorithm is similar in terms of computational complexity to bagged trees (Breiman, 1996). While the latter uses bootstrap replicates of the learning data set, the proposed

algorithm iteratively computes updated data sets in terms of the response variable (i.e., $y^*_{i(r)}$). Both algorithms fit a standard regression tree to each one of the modified data sets. This process entails no additional challenge in terms of computational complexity if it uses one of the available and efficient implementation of a standard regression tree algorithm. Updating the learning data set at each iteration in the proposed algorithm for mixed effects regression trees is not too demanding since we have closed form expressions for the estimators of the random effects b_i and of the variance components σ^2 and D. Note however that the number of bootstrap samples is arbitrarily fixed in advance in the bagging algorithm, while the number of iterations depends on the speed of convergence of the proposed EM algorithm for mixed effects regression trees. Many factors may affect this convergence (e.g. : sample size, initial values, instability of standard regression trees). The main disadvantage of the EM algorithm is that it may require a large number of iterations before reaching the stopping criteria.

To predict the response for a new observation that belongs to a cluster among those used to fit the mixed effects regression model, we use both its corresponding population-averaged tree prediction and the predicted random part corresponding to its cluster. 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 tree prediction.

There exist a number of other nonlinear or nonparametric methods to model the fixed part $f(X_i)$ and/or the random part $Z_i b_i$ in (1.2) (e.g., Davidian and Giltinan, 1995 ; Zhang and Davidian, 2004 ; Zhang, 1997 ; Wu and Zhang, 2006). These alternatives may be more suitable in some applications. Tree methods are however attractive because they propose easily interpretable models and are able, through their automatic detection of possible significant interactions between covariates, to represent complex relationships.

1.4 Simulation

In this section, we investigate the performance of the mixed effects regression trees in comparison to standard trees. The proposed method was implemented in R (R Development Core Team, 2007) using the function *rpart* (Therneau and Atkinson, 1997). This

function implements cost-complexity pruning based on cross-validation after an initial large tree is grown. The default settings of *rpart* are used ; the largest tree is grown and pruned automatically using the 1-SE rule of Breiman and al. (1984).

Within the mixed tree approach, we force the first 50 iterations, then we keep iterating while the absolute change in GLL is not less than 1E-06 or we reach a maximum of 1000 iterations. Once the stopping criterion is met, we run an additional 50 iterations. The mixed 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 tree models.

To compare the performance of the standard and mixed effects regression tree methods, we evaluate both their ability to find the true tree structure used to generate the data, and their predictive accuracy measured by the predictive mean squared error (PMSE). In addition, we look at how well are estimated the variance-covariance components at the observation-level (σ^2) and at the cluster-level (D) with the mixed effects regression tree approach.

1.4.1 Simulation Design

The simulation design used has a hierarchical structure of 100 clusters with 55 observations generated in each cluster. The first five observations in each cluster form the training sample, and the other 50 observations are left for the test sample. Consequently, the trees are built with 500 observations (100 clusters of 5 observations). Three random variables, X_1, X_2, and X_3, are first generated independently with a uniform distribution in the interval $[0, 10]$; they serve as predictors. The response variable y is generated based on the following fixed tree rules along with the random components :

Leaf 1. If $x_{1ij} \leq 5$ and $x_{2ij} \leq 5$ then $y_{ij} = \mu_1 + z_{ij}^T b_i + \epsilon_{ij}$,

Leaf 2. if $x_{1ij} \leq 5$ and $x_{2ij} > 5$ then $y_{ij} = \mu_2 + z_{ij}^T b_i + \epsilon_{ij}$,

Leaf 3. if $x_{1ij} > 5$ and $x_{3ij} \leq 5$ then $y_{ij} = \mu_3 + z_{ij}^T b_i + \epsilon_{ij}$,

Leaf 4. if $x_{1ij} > 5$ and $x_{3ij} > 5$ then $y_{ij} = \mu_4 + z_{ij}^T b_i + \epsilon_{ij}$,

where b_i and ϵ_i are generated according to $N(0, D)$ and $N(0, I)$ respectively, for $i = 1, ..., 100$ and $j = 1, ..., 55$. Each observation j in cluster i falls into only one of the four terminal nodes with mean response value equal to μ_1, μ_2, μ_3, or μ_4 respectively (see Figure 1.2).

Insert Figure 1.2 about here

Insert Table 1.I about here

We consider 14 different data generating processes (DGP), summarized in Table 1.I. Two different scenarios are selected for the fixed components. In the first scenario, the means of the four terminal nodes are widely spread with $\mu_1 = -20$, $\mu_2 = -10$, $\mu_3 = 10$ and $\mu_4 = 20$, while in the second scenario, they are closer with $\mu_1 = 10$, $\mu_2 = 11$, $\mu_3 = 12$ and $\mu_4 = 13$. The random components are generated based on the following three different scenarios :

1. No random effects (NRE), i.e. $D = 0$.
2. Random intercept (RI), i.e. $z_{ij} = 1$ for $i = 1, ..., 100$, and $j = 1, ..., 55$, and $D = d_{11} > 0$.
3. Random intercept and covariate (RIC) which is a RI with a linear random effect for X_1. More precisely, $z_{ij} = [1, x_{1ij}]$ for $i = 1, ..., 100$, $j = 1, ..., 55$, and $D = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}$, $d_{11} > 0$ and $d_{22} > 0$.

In all cases, the within-cluster variance σ^2 is set to 1. An equivalent alternative would be to fix the terminal nodes means while varying the σ^2 value so that large fixed effects coincide with small values for σ^2 and small fixed effects coincide with large values for σ^2. We consider two levels for the between-clusters covariance matrix D. In the RI case, we

use $D = d_{11} = 0.25$ and 0.5 which are equivalent to an intra-cluster correlation coefficient of 0.20 and 0.33 respectively. In the RIC case, we have two additional conditions based on the value of the correlation between the random components, $d_{12}/\sqrt{d_{11} + d_{22}} = 0$ and $d_{12}/\sqrt{d_{11} + d_{22}} = 0.5$; in the first correlation scenario, $d_{11} = d_{22} = 0.25$, and in the second $d_{11} = d_{22} = 0.5$.

We adjusted three models for each DGP scenario : 1) a standard (STD) tree model, 2) a random intercept (RI) tree model, and 3) a random intercept and covariate (RIC) tree model. The true model is the one corresponding to the DGP used to generate the data. Overall, we built 42 regression tree models (14 scenarios $\times$ 3 models). The simulation results are obtained by means of 100 runs.

1.4.2 Simulation Results

Firstly, we evaluate the performance of the approaches in terms of recovering the right tree structure. Here, an estimated tree is considered to be right if it has the same structure as the model generating the data, i.e. if its first split is on X_1, then the left side of the tree splits on X_2, while the right side of the tree splits on X_3, and the number of terminal nodes equals four (Figure 1.2). We do not consider the cut-off values for the splits in assessing the true structure of the tree.

The results are presented in Table 1.II. In all scenarios where the means of the terminal nodes are very different (i.e. large fixed effect : DGPs 1, 3, 4, 7, 8, 11, and 12), both the proposed approach (RI and RIC tree) and the standard tree algorithm succeed in finding the right tree structure. However, when the difference between the means of the terminal nodes is small, the higher the intra-cluster correlation is the harder it is for all methods to find the right tree structure (see DGPs 5 vs 6, 9 vs 10, and 13 vs 14). In all of these cases however, RIC tree results are closer to the true data partition compared to partitions obtained from the RI tree or the standard tree. For DGPs 9, 10, 13 and 14, the standard tree has never identified the right tree structure, while the RIC tree approach does best with recovery rates of 64%, 60%, 68%, and 67%, respectively.

Insert Table 1.II about here

The performance of the methods is also judged based on their predictive accuracy measured by the predictive mean squared error :

$$PMSE = \frac{\sum_{i=1}^{100} \sum_{j=1}^{50} (y_{ij} - \hat{y}_{ij})^2}{5000},$$

where $\hat{y}_{ij}$ is the predicted response for observation j in cluster i in the test set. Recall that the trees are built with 100 clusters of 5 observations each but the PMSE is computed on 5000 observations in the test set (50 observations in each cluster). The average, median, minimum, maximum and standard deviation of PMSE over the 100 runs were calculated, and the results are presented in Table 1.II.

All three methods have exactly the same average performance when the data are uncorrelated (DGPs 1 and 2). But in all cases with a random component (DGPs 3 to 14), the proposed mixed effects approach does better than the standard tree algorithm even with the wrong specification of the random component part. Again, the higher the intra-cluster correlation the more difficult it is for the standard tree to predict accurately the response variable, but not for the mixed effects approach which handles appropriately this correlation. The improvement of the new approach over the standard tree algorithm is often large, especially when a random covariate effect is present (DGPs 7 to 14). For example, in DGP 14, the RIC tree has an average PMSE of 1.42 compared to 21.6 for the standard tree.

Insert Table 1.III about here

Table 1.III gives the summary statistics of the estimated variance at the observation-level. If we compare the estimated value of σ^2 to its true value of 1 we can conclude that the proposed mixed effects approach is very efficient even when the random structure is over-specified, i.e. the RIC tree always estimates σ^2 correctly. However, in cases where the fitted model is a RI tree while the true model is a RIC tree, the mixed effects approach seems to retrieve some of the cluster-level variance of the omitted random component in the estimation of the observation-level variation σ^2. The higher the variance components of D the more important is the inflation of the estimated σ^2.

Insert Table 1.IV about here

Table 1.IV gives the summary statistics of the estimated variance-covariance components at the cluster-level. First, under-specification of the random structure seems to be harmful while over-specification is not. The estimates of d_{11} are inflated in cases where the fitted model is a RI tree while the true model is a RIC tree; the higher the magnitude of the intra-cluster correlation the more important is the inflation of the d_{11} estimates. Second, in the in-depth analysis of the simulation run under DGP 6 (Figure 1.1), we observe that the MSE improves until about iteration 40 (Figure 1.1b), which is the point in the iteration process where good estimates of the variance components are reached. Notice also that the tree at the first iteration corresponds to a standard tree. It has only three leaves with a PMSE equal to 1.65 while the final RI tree model selected recovers the true tree structure with four leaves and has a PMSE equal to 1.25.

1.5 Data Example

In this section, we illustrate the proposed tree method using a real data set on first-week box office revenues of movies presented in the province of Quebec in Canada from

2001 to 2008. The unit of analysis is a screen showing the new movie during its first week of release. The importance of the first-week revenues is well-known in the industry. Typically, it represents about 25 % of the total box office of a general public film (Simonoff and Sparrow, 2000). The total number of observations (screens) is 60175. This data includes information on 2656 movies and each movie is treated as a cluster. These clusters are highly unbalanced with an average size of 22.7 screens per movie ($minimum = 1$; $first\ quartile = 1$; $median = 8$; $third\ quartile = 47$; $maximum = 93$).

1.5.1 Description of Observation and Cluster Level Covariates

We have three covariates at the screen-level (observation-level) and eight at the movie-level (cluster-level). The three screen-level covariates are : (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-entreprise ; 5-Famous Players ; 6-Cinémas R.G.F.M. ; 7-Cinémas Fortune ; 8-AMC).

The eight movie-level covariates are : (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 between 70 to 227 minutes, (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-Thirteen years old ; 3-Sixteen years old ; 4-Eighteen years 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, commonly used as a proxy for the marketing effort.

Using a learning sub-sample of 30018 screens within the 2656 movies, we fitted the following three models : 1) a standard regression tree (SRT) model, 2) a random intercept regression tree (RIRT) model, and 3) a random intercept linear regression (RILR) model. As

commonly done in box office prediction studies, we model the log transform of the first-week box office revenues since it has a distribution highly skewed to the right. We also took the logarithm of the covariate *Size* to lessen its asymmetry and improve the fit of the RILR model. Note that the latter asymmetry has no effect for the SRT and RIRT models but affects the linear mixed effects model.

1.5.2 Results

All covariates are statistically significant in the RILR model (results not shown), but only eight covariates (*Size*, *Region*, *Theater*, *Language*, *length*, *Month*, *rating*) are retained in the SRT model and only four (*Size*, *Region*, *Theater*, *Language*) are retained by the algorithm in the RIRT model. The SRT structure is larger than the RIRT structure, i.e., the standard regression tree has 44 leaves while the random intercept regression tree has 28 leaves. However, the RIRT is not a subtree of the SRT ; the first splits of the two trees are identical, but their second splits use different partitions based on the same movie-level covariate *Region* (i.e. $Region = 2; 4; 5; 6$ vs. $Region = 2; 4; 6$, respectively). Figures 1.3 and 1.4 show the first three levels of the fitted SRT and RIRT, respectively.

Insert Figure 1.3 about here

Insert Figure 1.4 about here

The RIRT model has the smallest in-sample MSE (0.44). The MSE of the SRT model and of the RILR model are 0.86 and 0.54, respectively. Thus, in-sample, the RIRT reduces

the MSE of the SRT model by 48.93% and reduces the MSE of the RILR model by 18.30%. Using the test sub-sample of 30157 screens within 1920 movies, the RIRT model also has the best predictive performance ; its PMSE is 0.53 while the PMSE of the SRT and RILR models are 0.90 and 0.62, respectively. Thus, the RIRT reduces the PMSE of the SRT model by 41.63% and reduces the PMSE of the RILR model by 14.94%.

1.6 Discussion

Statistical models for clustered data typically include two components : A fixed or population-averaged and a random or cluster-specific component. If these two components have an underlying linear and additive structure, and if the normality assumption is reasonable, the LME models are appropriate. If the linear assumption is too restrictive, other structures may be more suitable to represent the true underlying relationship between the covariates and the response variable.

There exist a number of nonlinear and/or nonparametric methods that are based on the mixed effects modeling approach and that have relaxed partially or completely the linearity or normality assumptions of LME models. We mention for example, the nonlinear mixed effects models (Davidian and Giltinan, 1995), the generalized additive mixed effects model (Zhang and Davidian, 2004), and the multivariate adaptive splines for the analysis of longitudinal data (Zhang, 1997). These methods may be more suitable to represent the underlying true relationship with the dependent variable in some applications.

The proposed mixed effects regression tree method relaxes the linearity assumption of the fixed component of LME models. As for the standard regression tree, this method is attractive because it proposes easily interpretable models that can be graphically displayed which make them easily understandable by non statisticians, and is able, through its automatic detection of possible significant interactions between covariates, to represent complex relationships.

Others have extended tree methods to clustered data, but mainly in the context of

longitudinal studies. Segal (1992) extended the regression tree methodology to repeated measures and longitudinal data by modifying the split function to accommodate multiple responses. He developed several split functions based either on deviations around clusters subgroup mean vectors or on two-sample statistics measuring clusters subgroup separation. One of his objectives was the identification of clusters subgroups, i.e., subgroups of growth curves. Hence, all the observations in a cluster end up in the same terminal node and describe the growth curve corresponding to that terminal node. Zhang (1998) treated the multivariate binary response case in a similar setting. Lee (2005) suggested a tree-based method that can analyze any type of multiple responses. His tree algorithm fits a marginal regression tree at each node using the generalized estimating equations, then separates clusters into two subgroups based on the sign of their Pearson's residual average. By using a likelihood ratio test statistic from a mixed model as the splitting criterion, Abdolell *et al.* (2002) were able to lift the requirements that subjects have an equal number of repeated observations. Others extended and applied these multivariate tree approaches to functional data, i.e. data where the response is a high-dimensional vector. The basic idea is to reduce the dimensionality then fit a multivariate tree to the reduced multivariate response (e.g. Yu and Lambert, 1999 ; Ghattas and Nerini, 2007).

All the latter extensions of tree based methods to handle correlation induced by the data structure do not allow observation-level covariates to be candidates in the splitting process and, consequently, all repeated observations from a given subject remain together during the tree building process and can not be splitted across different nodes. This is different from the method proposed here which can split observations within clusters since observation-level covariates are candidates in the splitting process. Moreover, the proposed tree method can appropriately deal with the possible random effects of observation-level covariates.

Although the focus in this paper is on the most common form of clustered data, i.e. individuals nested within groups, the proposed mixed effects regression tree approach can be applied to analyze longitudinal data. Indeed, we can adjust a tree growth model

where the time period and other time-varying covariates, as well as baseline measures (e.g., characteristics of the subject's background, or of an experimental treatment) are used as candidates in the splitting process. However, the proposed tree algorithm assumes that the correlation structure is solely induced via the between-cluster variation. For data sets with a short time series, Bryk and Raudenbush (1987) noted that this assumption is often most practical and unlikely to distort the results. For other circumstances, one needs to adapt the EM algorithm to generalize the approach to alternative covariance structures. To this end, Jennrich and Schluchter (1986) described an hybrid EM scoring algorithm that could be used to adapt the EM algorithm presented in Section 1.3.2 for the mixed effects regression tree model in order to allow alternative within-subject covariance structures.

The main drawback of standard regression trees is their instability, i.e. a slight change in the training sample can lead to a radically different tree model. One solution to improve the predictive accuracy of trees is the use of ensemble methods such as bagging (Breiman, 1996) and forest of trees (Breiman, 2001). This observation applies also to mixed effects regression trees and the proposed method should be a good candidate for ensemble algorithms.

1.7 Conclusion

We proposed a simple approach to extend the standard regression tree methods to clustered data. Simulation results showed that, as it is the case in the parametric framework, improper handling of the correlation induced by clustered data may result in the true relationship between variables not being identified by a standard tree algorithm. The mixed effects regression trees can be used as a modeling tool in their own right, or as an exploratory tool for finding better predictive models. Past studies (e.g., Kuhnert, Do, and McClure, 2000) suggested that usual tree model could be used as a precursor to a parametric model. This is also true for mixed effects regression tree models that can be used as a precursor to a parametric mixed effects model. The standard tree methodology has some advantages in comparison to parametric simple regression modeling approach (e.g., handling of large data sets with many variables, handling outliers and collinearity problems, etc.), and all of these

advantages carry over naturally to the mixed effects regression tree methodology.

In the light of the simulation results and the example, the proposed mixed effects regression tree approach seems to be more appropriate for clustered data than standard tree procedures, particularly when the random effects are non negligible. This method is appropriate for clustered data where the outcome is continuous. Extending it to other kind of outcomes, e.g. binary, would be important for practitioners. Also, further investigations about the robustness of the method when its main assumptions are seriously violated (i.e. the fixed component is non piecewise constant, the random component is non linear, the fixed and random components are non additive, the errors are non normal) and when the tree structure is more complex than the one used in the simulation study, remain to be done.

An R program implementing the mixed effect regression tree procedure is available from the first author.

1.8 References

Abdollel, M., LeBlanc, M., Stephens, D. and Harrison, R. V. (2002). Binary partitioning for continuous longitudinal data : Categorizing a prognostic variable. *Statistics in Medicine,* 21, 3395-3409.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees.* Wadsworth International Group. Belmont, California.

Breiman, L. (1996). Bagging predictors. *Machine Learning,* 24, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning,* 45, 5-32.

Bryk, A. S., and Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin,* 101, 147-158.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Mixed Effects Models for Repeated Measurement Data.* Chapman and Hall.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,* 39, 1-38.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied longitidunal analysis.* New York : Wiley.

Ghattas, B., and Nerini, D. (2007). Classifying densities using functional regression trees : Applications in oceanology. *Computational Statistics & Data Analysis,* 51, 4984-4993.

Goldstein, H. (2003). *Multilevel statistical models (3rd Edition).* Arnold, London.

Harville, D. A. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *Annals of Statistics,* 4, 384-395.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association,* 72, 320-38.

Jennrich, R. I., and Schluchter, M. D. (1986). Unbalanced Repeated-Measures with Structured Covariance Matrices. *Biometrics,* 42, 805-820.

Kuhnert, P. M., Do, K.-A., and McClure, R. (2000). Combining nonparametric models with logistic regression : An application to motor vehicle injury data. *Computational Statistics and Data Analysis,* 34, 371-386.

Laird, N. M. and Ware J. H. (1982). Random-effects models for longitudinal data. *Biometrics,* 38, 963-974.

Lee, S. K. (2005). On Generalized multivariate decision tree by using GEE. *Computational Statistics & Data Analysis,* 49, 1105-1119.

McLachlan G. J. and Krishman T. (1997). *The EM algorithm and extensions.* Wiley. New

York.

R Development Team (2007). *R : A Language and environment for statistical computing*. R Foundation for Statistical Computing : www.R-project.org.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models : Applications and data analysis method (2nd Edition).* Sage. Newbury Park, CA.

Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association,* 87, 407-418.

Simonoff, J. S. and Sparrow, I. R. (2000). Predicting movie grosses : Winners and losers, blockbusters and sleepers. *Chance,* 13(3), 15-24.

Therneau, T. M. and Atkinson, E. J. (1997). *An introduction to recursive partitioning using the rpart routines.* Technical Report 61, Department of Health Science Research, Mayo Clinic, Rochester.

Yu, Y. and Lambert, D. (1999). Fitting Trees to Functional Data : With an Application to Time-of-day Patterns. *Journal of Computational and Graphical Statistics,* 8, 749-762.

Wu, H. and Zhang, J. T. (2006). *Nonparametric regression methods for longitidunal data analysis : Mixed-effects modeling approaches.* Wiley. New York.

Zhang, H., (1998). Classification trees for multiple binary responses. *Journal of the American Statistical Association,* 93, 180-193.

Zhang, H., (1997). Multivariate Adaptive Splines for Analysis of Longitudinal Data. *Journal of Computational and Graphical Statistics,* 6, 74 - 91.

Zhang, D. and Davidian, M. (2004). Likelihood and conditional likelihood inference for generalized additive mixed models for clustered data. *Journal of Multivariate Analysis,* 91, 90-106.

Table 1.I Data generating processes (DGP) for the simulation study.

DGP	Data Structure								
	Fixed Component					Random Component			
	Effect	μ_1	μ_2	μ_3	μ_4	Structure	d_{11}	d_{22}	d_{12}
1	Large	-20	-10	10	20	No random effect	0.00	0.00	0.00
2	Small	10	11	12	13				
3	Large	-20	-10	10	20	Random intercept	0.25	0.00	0.00
4							0.50	0.00	0.00
5	Small	10	11	12	13		0.25	0.00	0.00
6							0.50	0.00	0.00
7	Large	-20	-10	10	20	Random intercept and covariate X_1 with 0 correlation	0.25	0.25	0.00
8							0.50	0.50	0.00
9	Small	10	11	12	13		0.25	0.25	0.00
10							0.50	0.50	0.00
11	Large	-20	-10	10	20	Random intercept and covariate X_1 with 0.5 correlation	0.25	0.25	0.125
12							0.50	0.50	0.25
13	Small	10	11	12	13		0.25	0.25	0.125
14							0.50	0.50	0.25

Table 1.II Results of the 100 simulation runs in terms of recovering the right tree structure and the predictive mean square error (PMSE).

DGP	Fixed effect	Random effect	Fitted tree model*	% of trees with the right tree structure	PMSE				
					Avg.	Med.	Min	Max	Std
1	Large	No random effect	STD	100	2.14	1.95	1.04	6.10	0.97
			RI	100	2.14	1.95	1.04	6.10	0.97
			RIC	100	2.15	1.96	1.04	6.10	0.97
2	Small		STD	95	1.04	1.03	0.96	1.21	0.04
			RI	97	1.04	1.03	0.96	1.21	0.04
			RIC	97	1.04	1.04	0.96	1.21	0.04
3	Large	Random intercept	STD	100	2.43	2.09	1.26	5.49	1.01
			RI	100	2.29	1.96	1.14	5.38	1.01
			RIC	100	2.29	1.96	1.14	5.38	1.01
4			STD	100	2.61	2.37	1.39	5.95	0.91
			RI	100	2.24	1.94	1.11	5.53	0.91
			RIC	100	2.25	1.94	1.11	5.53	0.91
5	Small		STD	77	1.31	1.30	1.18	1.52	0.07
			RI	91	1.16	1.15	1.07	1.33	0.05
			RIC	91	1.17	1.16	1.08	1.33	0.05
6			STD	60	1.58	1.59	1.35	1.82	0.10
			RI	86	1.20	1.18	1.08	1.37	0.06
			RIC	88	1.20	1.19	1.08	1.37	0.06
7	Large	Random intercept and covariate with 0 correlation	STD	100	10.95	10.99	7.62	14.96	1.62
			RI	100	4.90	4.70	3.25	7.91	1.02
			RIC	100	2.48	2.22	1.30	5.68	0.94
8			STD	100	19.49	19.13	13.15	28.68	2.69
			RI	100	7.44	7.08	5.00	13.98	1.42
			RIC	100	2.69	2.41	1.32	8.17	1.25
9	Small		STD	0	10.28	9.95	7.10	14.58	1.45
			RI	6	3.93	3.91	3.05	4.96	0.38
			RIC	64	1.41	1.40	1.23	1.61	0.10
10			STD	0	18.90	18.65	14.30	26.44	2.59
			RI	0	6.46	6.31	4.90	9.99	0.91
			RIC	60	1.46	1.46	1.25	1.82	0.11
11	Large	Random intercept and covariate with 0.5 correlation	STD	100	12.25	11.85	8.65	18.47	2.15
			RI	100	4.96	4.59	3.40	10.10	1.45
			RIC	100	2.57	2.11	1.30	7.28	1.33
12			STD	100	21.52	21.19	15.45	30.98	2.85
			RI	100	7.10	6.91	5.21	12.22	1.11
			RIC	100	2.34	2.06	1.27	6.76	0.90
13	Small		STD	0	11.75	11.47	9.04	17.92	1.70
			RI	5	4.01	4.00	2.82	5.50	0.41
			RIC	68	1.39	1.38	1.21	1.72	0.10
14			STD	0	21.60	21.51	15.80	28.85	2.91
			RI	1	6.45	6.40	4.96	8.59	0.79
			RIC	67	1.42	1.41	1.20	1.77	0.11

* STD : Standard tree model ; RI : Random intercept tree model ; RIC : Random intercept and covariate tree model

Table 1.III Results of the 100 simulation runs for the estimation of the observation-level variance (the true value is $\sigma^2 = 1$).

DGP	Fixed effect	Random effect	Fitted tree model*	$\hat{\sigma}^2$ Avg.	Med.	Min	Max	Std
1	Large	No random effect	RI	0.98	0.98	0.74	1.14	0.07
			RIC	0.96	0.95	0.73	1.13	0.07
2	Small		RI	0.98	0.98	0.81	1.16	0.08
			RIC	0.96	0.97	0.80	1.15	0.08
3	Large	Random intercept	RI	0.99	1.00	0.83	1.15	0.08
			RIC	0.98	0.98	0.79	1.15	0.08
4			RI	0.99	0.99	0.83	1.16	0.07
			RIC	0.98	0.97	0.80	1.15	0.07
5	Small		RI	0.98	0.98	0.83	1.14	0.07
			RIC	0.96	0.97	0.80	1.14	0.07
6			RI	1.00	1.00	0.84	1.25	0.08
			RIC	0.99	0.99	0.81	1.25	0.08
7	Large	Random intercept and covariate with 0 correlation	RI	3.09	3.09	2.29	4.46	0.36
			RIC	0.98	0.99	0.82	1.21	0.09
8			RI	5.11	5.07	3.42	7.29	0.71
			RIC	1.01	1.00	0.79	1.38	0.09
9	Small		RI	3.29	3.25	2.34	4.49	0.39
			RIC	1.03	1.02	0.83	1.33	0.10
10			RI	5.32	5.11	3.85	8.76	0.82
			RIC	1.02	1.01	0.82	1.31	0.10
11	Large	Random intercept and covariate with 0.5 correlation	RI	3.07	3.02	2.35	4.10	0.38
			RIC	1.00	1.00	0.79	1.18	0.08
12			RI	5.14	5.07	3.65	7.35	0.77
			RIC	1.00	1.00	0.82	1.25	0.09
13	Small		RI	3.28	3.24	2.23	5.22	0.42
			RIC	1.00	1.01	0.78	1.21	0.09
14			RI	5.30	5.16	4.03	6.91	0.78
			RIC	1.01	1.02	0.83	1.24	0.09

* STD : Standard tree model ; RI : Random intercept tree model ; RIC : Random intercept and covariate tree model

Table 1.IV Results of the 100 simulation runs for the estimation of the cluster-level variance-covariance components.

DGP	Fixed effect	Random effect	Fitted tree model*	$\hat{d}_{11}$ True value	Avg.	Med.	Min	Max	Std	$\hat{d}_{22}$ True value	Avg.	Med.	Min	Max	Std	$\hat{d}_{12}$ True value	Avg.	Med.	Min	Max	Std
1	Large	No random effect	STD	-	0.01	0.00	0.00	0.08	0.01	-	-	-	-	-	-	-	-	-	-	-	-
			RI	-	0.01	0.00	0.00	0.08	0.02	-	-	-	-	-	-	-	-	-	-	-	-
			RIC	-	0.06	0.02	0.00	0.29	0.07	-	0.00	0.00	0.00	0.01	0.00	-	-0.01	0.00	-0.05	0.00	0.01
2	Small		STD	-	0.01	0.00	0.00	0.08	0.02	-	-	-	-	-	-	-	-	-	-	-	-
			RI	-	0.05	0.02	0.00	0.25	0.06	-	0.00	0.00	0.00	0.01	0.00	-	-0.01	0.00	-0.05	0.00	0.01
3	Large	Random intercept	STD	0.25	0.24	0.24	0.10	0.41	0.06	-	-	-	-	-	-	-	-	-	-	-	-
			RI	0.25	0.29	0.28	0.03	0.63	0.13	-	0.00	0.00	0.00	0.01	0.00	-	-0.01	-0.01	-0.05	0.01	0.02
4			STD	0.50	0.50	0.49	0.27	1.03	0.12	-	-	-	-	-	-	-	-	-	-	-	-
			RI	0.50	0.52	0.51	0.21	1.00	0.18	-	0.00	0.00	0.00	0.01	0.00	-	-0.01	0.00	-0.08	0.03	0.02
5	Small		STD	0.25	0.24	0.24	0.08	0.38	0.06	-	-	-	-	-	-	-	-	-	-	-	-
			RI	0.25	0.29	0.27	0.08	0.65	0.13	-	0.00	0.00	0.00	0.01	0.00	-	-0.01	-0.01	-0.06	0.01	0.02
6			STD	0.50	0.48	0.47	0.26	0.81	0.10	-	-	-	-	-	-	-	-	-	-	-	-
			RI	0.50	0.48	0.44	0.13	0.94	0.16	-	0.00	0.00	0.00	0.01	0.00	-	0.00	0.00	-0.07	0.02	0.02
7	Large	Random intercept and covariate with 0 correlation	STD	0.25	6.66	6.63	3.88	10.06	1.14	0.25	-	-	-	-	-	0.00	-	-	-	-	-
			RI	0.25	0.24	0.24	0.01	0.65	0.14	0.25	0.26	0.26	0.16	0.34	0.03	0.00	0.00	0.00	-0.13	0.14	0.06
8			STD	0.50	12.81	12.68	7.87	19.19	2.20	0.50	-	-	-	-	-	0.00	-	-	-	-	-
			RI	0.50	0.49	0.50	0.01	0.99	0.19	0.50	0.50	0.50	0.32	0.69	0.07	0.00	-0.01	0.00	-0.20	0.23	0.09
9	Small		STD	0.25	6.49	6.31	4.26	11.57	1.25	0.25	-	-	-	-	-	0.00	-	-	-	-	-
			RI	0.25	0.22	0.22	0.01	0.77	0.17	0.25	0.25	0.25	0.16	0.38	0.04	0.00	0.01	0.00	-0.13	0.14	0.06
10			STD	0.50	12.83	12.58	7.39	19.80	2.31	0.50	-	-	-	-	-	0.00	-	-	-	-	-
			RI	0.50	0.46	0.48	0.01	1.24	0.25	0.50	0.49	0.49	0.33	0.76	0.08	0.00	0.00	0.01	-0.32	0.23	0.11
11	Large	Random intercept and covariate with 0 correlation	STD	0.25	7.66	7.41	4.85	11.32	1.34	0.25	-	-	-	-	-	0.125	-	-	-	-	-
			RI	0.25	0.25	0.24	0.00	0.85	0.17	0.25	0.25	0.25	0.15	0.36	0.04	0.125	0.12	0.13	0.00	0.28	0.05
12			STD	0.50	15.00	14.82	9.28	24.30	2.47	0.50	-	-	-	-	-	0.25	-	-	-	-	-
			RI	0.50	0.47	0.48	0.07	0.92	0.21	0.50	0.49	0.48	0.33	0.72	0.07	0.25	0.24	0.24	-0.11	0.43	0.09
13	Small		STD	0.25	7.86	7.67	5.41	11.26	1.39	0.25	-	-	-	-	-	0.125	-	-	-	-	-
			RI	0.25	0.24	0.20	0.03	0.66	0.15	0.25	0.25	0.25	0.17	0.39	0.04	0.125	0.14	0.14	-0.02	0.27	0.05
14			STD	0.50	15.27	15.30	10.18	21.16	2.44	0.50	-	-	-	-	-	0.25	-	-	-	-	-
			RI	0.50	0.47	0.46	0.05	0.94	0.20	0.50	0.49	0.48	0.36	0.69	0.08	0.25	0.26	0.26	0.06	0.45	0.09

* STD : Standard tree model ; RI : Random intercept tree model ; RIC : Random intercept and covariate tree model

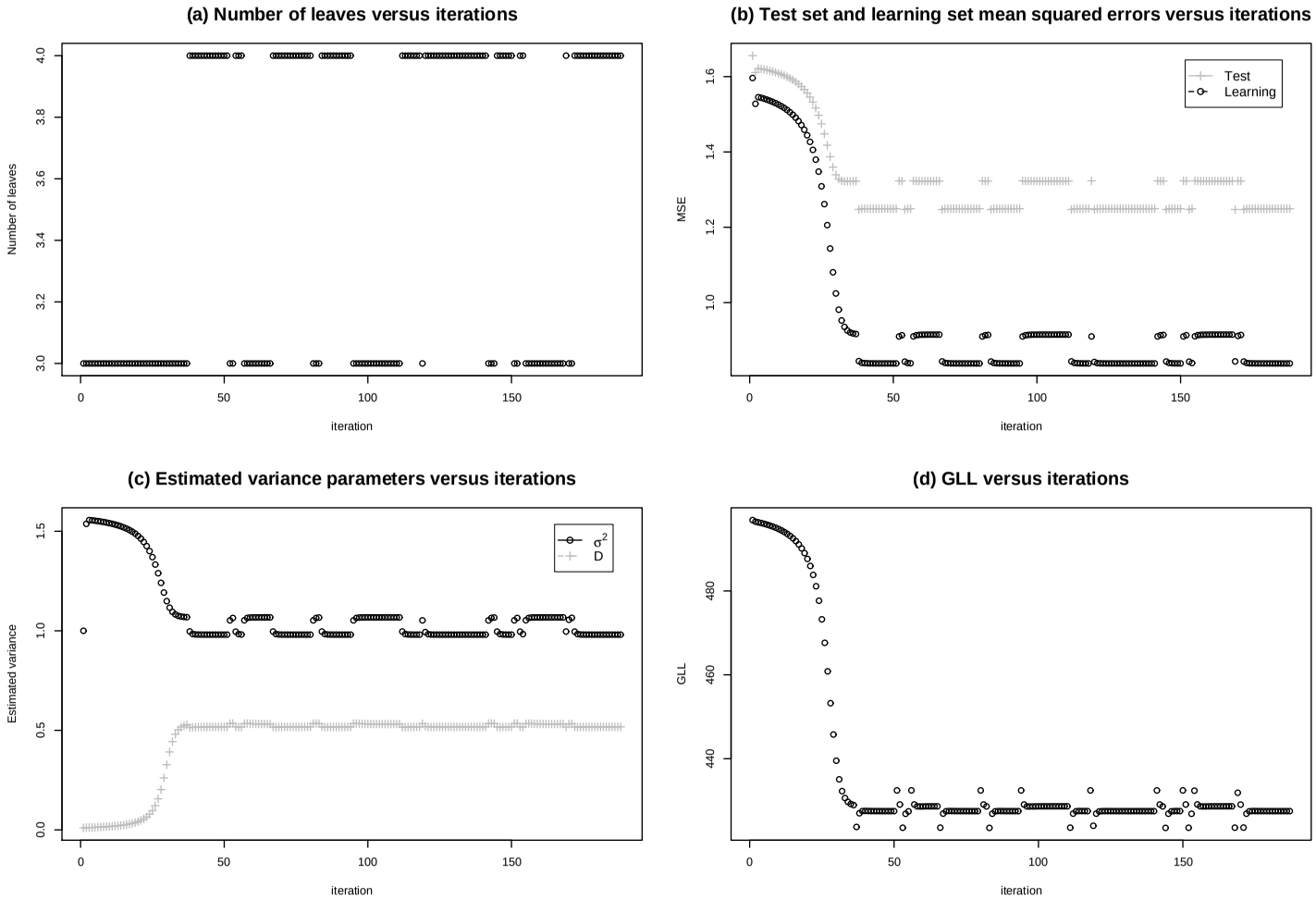

Figure 1.1 Behavior of different key elements of the mixed effects regression tree algorithm through the iteration process for fitting the random intercept model to one sample from DGP 6, i.e. small fixed effect with a random intercept structure with $D = d_{11} = 0.5$ and $\sigma^2 = 1$

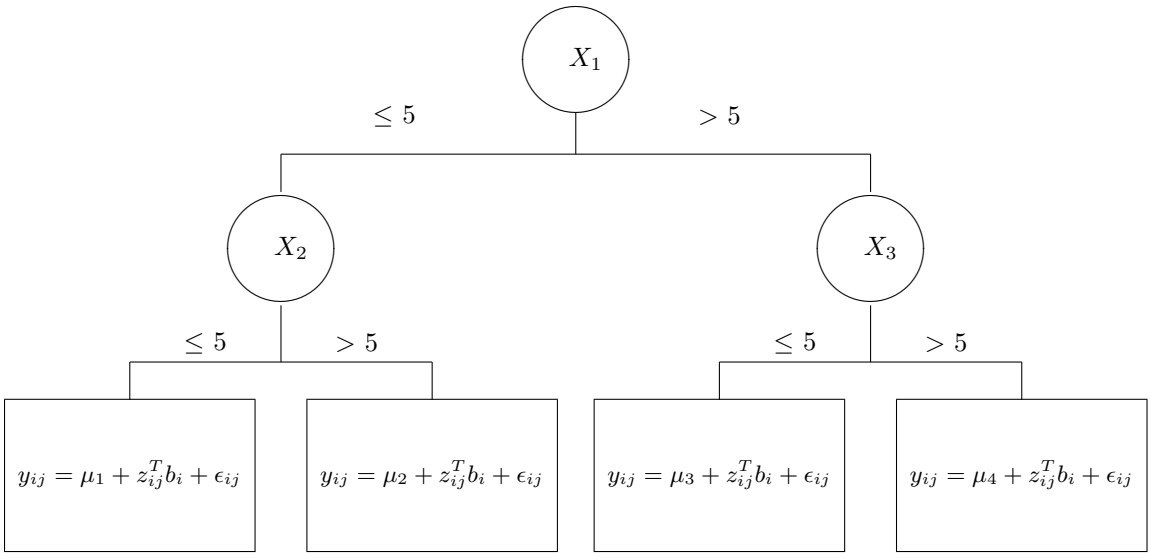

Figure 1.2 Mixed effects regression tree structure used for the simulation study.

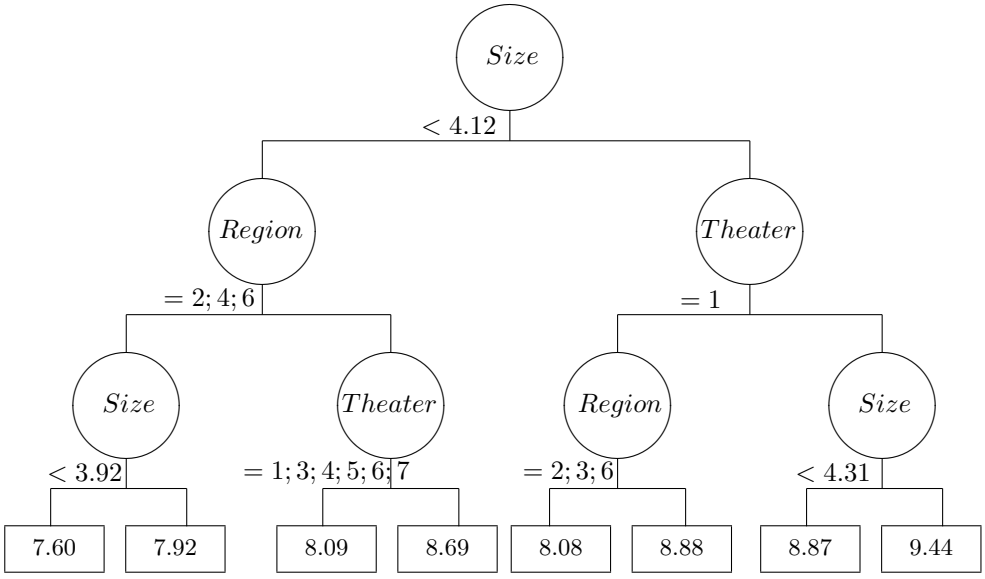

Figure 1.3 The first three levels of the standard regression tree for the data example on first-week box office revenues (on the log scale). When the condition below a node is true then go to the left node, otherwise go to the right node. The complete tree has 44 leaves.

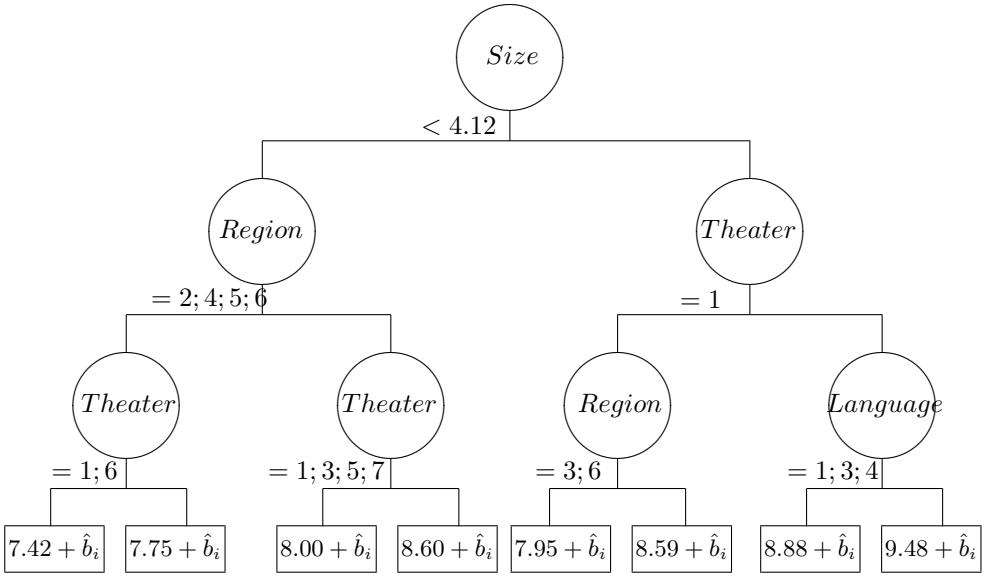

Figure 1.4 The first three levels of the random intercept regression tree for the data example on first-week box office revenues (on the log scale). When the condition below a node is true then go to the left node, otherwise go to the right node. The complete tree has 28 leaves.

ARTICLE II

GENERALIZED MIXED EFFECTS REGRESSION TREES

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

2.1 Abstract

This paper presents the generalized mixed effects regression tree (GMERT) method, an extension of the mixed effects regression tree (MERT) methodology designed for continuous outcomes to other types of outcomes (e.g., binary outcomes, counts data, ordered categorical outcomes, and multicategory nominal scale outcomes). This extension uses the penalized quasi-likelihood (PQL) method for the estimation and the expectation-maximization (EM) algorithm for the computation. The simulation results in the binary response case show that, when random effects are present, the proposed generalized mixed effects regression tree method provides substantial improvements over standard classification trees.

Keywords : Tree based methods, clustered data, mixed effects, penalized quasi-likelihood (PQL) algorithm, expectation-maximization (EM) algorithm.

2.2 Introduction

Tree based methods are a classic data mining technique. These methods became popular with the CART (classification and regression tree) algorithm (Breiman, Friedman, Olshen, and Stone, 1984). They have many advantages compared to parametric methods. For instance, they are able to detect automatically possible significant interactions between covariates, and they propose easily interpretable models that can be graphically displayed. However, when the data are clustered (i.e., observations nested within clusters) with covariates at the observation- and at the cluster-level, the standard tree algorithm is no longer appropriate. A number of extensions of standard tree methods to the case of clustered data were proposed in the literature.

Segal (1992) extended the regression tree methodology to repeated measures and longitudinal data (i.e., repeated observations nested within subjects) by modifying the split function to accommodate multiple responses. All the observations in a cluster end up in the same terminal node and describe the growth curve corresponding to that terminal node. Zhang (1998) proposed two splitting criteria for the case of multiple binary responses. These extensions require that subjects have an equal number of repeated observations. By using a likelihood ratio test statistic from a mixed model as the splitting criterion, Abdolell, Leblanc,

Stephens, and Harrison (2002) were able to lift this requirement. Lee (2005) suggested a tree-based method that can analyze continuous or discrete multiple responses. His tree algorithm fits a marginal regression tree at each node using generalized estimating equations, then separates clusters into two subgroups based on the sign of their Pearson's residual average.

All the above extensions of tree based methods to handle the correlation induced by the data structure (i.e., repeated observations nested within subjects) do not allow observation-level (i.e., time-varying) covariates to be candidates in the splitting process and, consequently, 1) no random or subject-specific effect of these covariates is allowed, and 2) all repeated observations from a given subject remain together during the tree building process and can not be splitted across different nodes. Hajjem, Bellavance, and Larocque (2008) proposed a mixed effects regression tree (MERT) method. It is an extension of the standard regression tree method to the case of clustered data where individuals are nested within groups. In contrast to the above extensions, this tree method can appropriately deal with the possible random effects of observation-level covariates and can split observations within clusters since observation-level covariates are candidates in the splitting process. Moreover, it does not require that the clusters have an equal number of observations. However, MERT was designed for a continuous response.

Following the logic of the generalized linear mixed models (GLMMs) (e.g., Breslow and Clayton, 1993), and adjusting for some new issues that arise in tree modeling framework, we propose a tree based method, named "generalized mixed effects regression tree" (GMERT), which is suitable for other types of outcomes (e.g., binary outcomes, counts data, ordered categorical outcomes, and multicategory nominal scale outcomes). Basically, the GMERT algorithm is a repeated call to a weighted MERT algorithm. The proposed GMERT method can handle unbalanced clusters, allows observations within clusters to be splitted, and can incorporate random effects and observation-level covariates.

This paper presents and evaluates the proposed generalized mixed effects regression tree method. The remainder of this article is organized as follows : Section 2.3 describes the

proposed approach ; Section 2.4 presents a simulation study to evaluate its performance of the method ; Section 2.5 discusses a number of related issues.

2.3 Generalized Mixed Effects Regression Tree

The basic idea behind the proposed generalized mixed effects regression tree method is to replace the linear structure used to model the fixed effects component in the GLMM's linear predictor with a regression tree structure, while the random component is still represented using a linear structure as in GLMMs. For the estimation of the GMERT model, we use the penalized quasi-likelihood (PQL) method (Breslow and Clayton, 1993), and for the computation we use the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977 ; McLachlan and Krishman 1997). Let's first review the key components of GLMM and the PQL algorithm.

2.3.1 PQL Algorithm for the Generalized Linear Mixed Models

Let $y_i = [y_{i1}, ..., y_{in_i}]^T$ denote the $n_i \times 1$ vector of responses for the n_i observations in cluster i. Let $X_i = [x_{i1}, ..., x_{in_i}]^T$ denote the $n_i \times p$ matrix of fixed-effects covariates, and $Z_i = [z_{i1}, ..., z_{in_i}]^T$ denote the $n_i \times q$ matrix of random-effects covariates. Let b_i denote the $q \times 1$ unknown vector of random effects for cluster i. Then, conditional on the b_i, the GLMM assumes that the response vector y_i follows a distribution from the exponential family (McCullagh and Nelder, 1989) with density $f(y_i|b_i, \beta)$ where β is common for all the clusters and is the $p \times 1$ unknown vector of parameters for the fixed effects. The total number of observations is $N = \sum_{i=1}^{n} n_i$. Let $\mu_i = E(y_i|b_i)$ and $Cov(y_i|b_i) = \sigma^2 v_i(\mu_i)$, where σ^2 is a dispersion parameter that may or may not be known and $v_i(\mu_i) = diag[v(\mu_{i1}), ..., v(\mu_{in_i})]$ with $v(.)$ being a known variance function. This formulation implies that the correlation is completely induced via between-clusters variation , i.e. given b_i, the observations are assumed independent. This assumption is suitable for a wide range of applications (Breslow and Clayton, 1993). Let $\eta_i = g(\mu_i)$ where $g(\mu_i) = [g(\mu_{i1}), ..., g(\mu_{in_i})]^T$ with $g(.)$ being a known link function. The GLMM is often written in the following form (see for example,

§2.4.1 of Wu and Zhang, 2006) :

$$\eta_i = X_i\beta + Z_i b_i, \tag{2.1}$$
$$b_i \sim N_q(0, D), i = 1, ..., n,$$

where D is the variance-covariance matrix of random effects. Estimation of the parameters in GLMM is not as simple as for the linear mixed effects (LME) model (Harville, 1976). When the errors at the observation-level are non normally distributed and the random effects at the cluster-level are assumed multivariate normal, the integration needed to obtain the likelihood is not available in closed form (e.g., Raudenbush and Bryk, 2002, page 456). An approximation via the linearization, known as the penalized quasi likelihood (PQL) approach, was developed and implemented in a number of mixed effects modeling softwares such as the *glmmPQL* function (Venables and Ripley, 2002) of R (R Development Core Team, 2007), HLM6 (Raudenbush, Bryk, Cheong, Congdon, and du Toit, 2004), and SAS *GLIMMIX* procedure (SAS Institute Inc., 2008). This method linearizes the non linear response variable y_i with a first-order Taylor series expansion. The resulting pseudo-response variable $y_{li} = g(\mu_i) + (y_i - \mu_i)g^{'}(\mu_i)$, where $g'(.)$ is the first derivative of $g(.)$ ($g'(\mu_i) = v_i^{-1}(\mu_i)$ for the canonical link function), follows approximately a normal distribution. Hence, the integration is available in a closed form, and the maximization of the likelihood can be done using available estimation and computation algorithms, such as the method of maximum likelihood (ML) implemented within the EM algorithm framework (ML-based EM algorithm). The resulting LME pseudo-model is defined as follows :

$$y_{li} = X_i\beta + Z_i b_i + e_i \tag{2.2}$$

where b_i and e_i are assumed independent and normally distributed and the between clusters observations are assumed independent. Consequently, based on the above pseudo-model, we have $V = Cov(y_l) = \text{diag}(Cov(y_{li}), ..., Cov(y_{ln}))$, where $y_l = [y_{l1}^T, ..., y_{ln}^T]^T$, the covariance matrix of within-cluster observations vector y_{li} for the i^{th} cluster is $V_i = Cov(y_{li}) = Z_i D Z_i^T + R_i$ where $R_i = diag[\sigma^2 v_{ij} g'(\mu_{ij})^2]$ with $v_{ij} = Var(y_{ij}|b_i)$ and σ^2 is a dispersion parameter

which can be estimated from the usual residual sum of squares or fixed to 1 if the assumed distribution does not have a scale parameter and no under- or over-dispersion parameter is to be estimated.

Using the weights $W_i = diag(w_{ij})$ with $w_{ij} = (v_{ij}g'(\mu_{ij})^2)^{-1}$ and $w_{ij} = v_{ij}$ for the canonical link function, we derive the following weighted LME pseudo-model

$$W_i^{\frac{1}{2}} y_{li} = W_i^{\frac{1}{2}} X_i \beta + W_i^{\frac{1}{2}} Z_i b_i + W_i^{\frac{1}{2}} e_i \tag{2.3}$$

with $Cov(W_i^{\frac{1}{2}} y_{li}) = W_i^{\frac{1}{2}} Z_i D (W_i^{\frac{1}{2}} Z)_i^T + \sigma^2 I_{n_i}$. This weighted LME pseudo-model can be fitted using the ML-based EM algorithm.

The PQL algorithm is detailed below.

MACRO STEP 0. Set $M = 0$. Given initial estimates of the mean values, $\hat{\mu}_{ij}^{(0)}$, $j = 1, ..., n_i$, fit a weighted LME pseudo-model using the linearized pseudo responses, $y_{li}^{(0)}$, and the weights, $W_i^{(0)} = diag(w_{ij}^{(0)})$.

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

Micro Step 1. Set $m = m + 1$. Update $\hat{\beta}_{(m)}$ and $\hat{b}_{i(m)}$

$$\hat{\beta}_{(m)} = \left(\sum_{i=1}^{n} (W_i^{\frac{1}{2}(M)} X_i)^T \hat{V}^{-1}_{i(m-1)} W_i^{\frac{1}{2}(M)} X_i \right)^{-1} \left(\sum_{i=1}^{n} (W_i^{\frac{1}{2}(M)} X_i)^T \hat{V}^{-1}_{i(m-1)} W_i^{\frac{1}{2}(M)} y_{li}^{(M)} \right),$$
$$\hat{b}_{i(m)} = \hat{D}_{(m-1)} (W_i^{\frac{1}{2}(M)} Z_i)^T \hat{V}^{-1}_{i(m-1)} \left(W_i^{\frac{1}{2}(M)} y_{li}^{(M)} - W_i^{\frac{1}{2}(M)} X_i \hat{\beta}_{(m)} \right),$$

where $\hat{V}_{i(m-1)} = W_i^{\frac{1}{2}(M)} Z_i \hat{D}_{(m-1)} (W_i^{\frac{1}{2}(M)} Z_i)^T + \hat{\sigma}^2_{(m-1)} I_{n_i}, i = 1, ..., n.$

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

$$\hat{\sigma}^2_{(m)} = N^{-1} \sum_{i=1}^{n} \left\{ \hat{\epsilon}^T_{i(m)} \hat{\epsilon}_{i(m)} + \hat{\sigma}^2_{(m-1)} [n_i - \hat{\sigma}^2_{(m-1)} \text{trace}(\hat{V}_{i(m-1)})] \right\},$$
$$\hat{D}_{(m)} = n^{-1} \sum_{i=1}^{n} \left\{ \hat{b}_{i(m)} \hat{b}^T_{i(m)} + [\hat{D}_{(m-1)} - \hat{D}_{(m-1)} (W_i^{\frac{1}{2}(M)} Z_i)^T \hat{V}^{-1}_{i(m-1)} W_i^{\frac{1}{2}(M)} Z_i \hat{D}_{(m-1)}] \right\},$$

where $\hat{\epsilon}_{i(m)} = W_i^{\frac{1}{2}(M)} y_{li}^{(M)} - W_i^{\frac{1}{2}(M)} X_i \hat{\beta}_{(m)} - W_i^{\frac{1}{2}(M)} Z_i \hat{b}_{i(m)}$.

Micro Step 3. Repeat steps 1 and 2 until convergence in terms of the generalized log-likelihood value :

$$GLL(\beta, b_i|y) = \sum_{i=1}^{n} \{\hat{\epsilon}_{i(m)}^T (\hat{\sigma}_{(m)}^2 I_{n_i})^{-1} \hat{\epsilon}_{i(m)} + \hat{b}_{i(m)}^T \hat{D}_{(m)}^{-1} \hat{b}_{i(m)} + \log|\hat{D}_{(m)}| + \log|\hat{\sigma}_{(m)}^2 I_{n_i}|\}.$$

MACRO STEP 1. Set $M = M + 1$. Set $\hat{\eta}_i^{(M-1)} = \hat{y}_{li}^{(M-1)} = X_i\hat{\beta} + Z_i\hat{b}_i$, where $\hat{\beta}$ and $\hat{b}_i$ are the estimated values at the micro level convergence of the previous macro iteration. Set $\hat{\mu}_i^{(M)} = g^{-1}(\hat{\eta}_i^{(M-1)})$, and fit a new weighted LME pseudo-model using the updated $y_{li}^{(M)}$ and $W_i^{(M)}$, i.e. repeat the micro steps 0 to 3 using as initial values for $\hat{\sigma}^2$, and $\hat{D}$, their micro-level convergence values in the previous macro iteration.

MACRO STEP 2. Repeat macro step 1 until convergence of $\hat{\eta}_i$.

The PQL algorithm is a doubly iterative process (i.e., micro iterations within macro iterations). At each macro iteration, the linearized response variable and the weights are updated. The micro iterations represent the iterative fitting process of a standard LME model where the current linearized response variable and weights values serve as the response variable and the weights, respectively.

2.3.2 PQL Algorithm for the Generalized Mixed Effects Regression Trees

The proposed generalized mixed effects regression tree (GMERT) model can be written as :

$$\eta_i = f(X_i) + Z_i b_i, \tag{2.4}$$
$$b_i \sim N_q(0, D), i = 1, ..., n,$$

where all quantities are defined as in Section 2.3.1 except that the linear fixed part $X_i\beta$ in (2.1) is replaced by the function $f(X_i)$ that will be estimated with a standard regression tree model. The random part, $Z_i b_i$, is still assumed linear.

Following the PQL approach used to estimate the GLMM, we can derive a MERT

pseudo-model from the above GMERT model, exactly as the LME pseudo-model derived from the GLMM. More precisely, a first-order Taylor-series expansion yields the linearized response variable, $y_{li} = g(\mu_i) + (y_i - \mu_i)g^{'}(\mu_i)$, and the MERT pseudo-model is defined as follows :

$$y_{li} = f(X_i) + Z_i b_i + e_i. \tag{2.5}$$

The GMERT algorithm is basically the PQL algorithm used to fit GLMMs where the weighted LME pseudo-model is replaced by a weighted MERT pseudo-model. Consequently, the fixed-part $f(X_i)$ is estimated with a standard regression tree model while the random part, $Z_i b_i$, is still estimated using a linear structure. The GMERT algorithm is detailed below.

MACRO STEP 0. Set $M = 0$. Given initial estimates of the mean values, $\hat{\mu}_{ij}^{(0)}$, $j = 1, ..., n_i$, fit a weighted MERT pseudo-model using the linearized pseudo responses, $y_{li}^{(0)}$, and the weights, $W_i^{(0)} = diag(w_{ij}^{(0)})$.

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

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

i) $y^*_{li(m)} = y_{li}^{(M)} - Z_i \hat{b}_{i(m-1)}$,

ii) Let $\hat{f}_{(m)}(X_i)$ an estimate of $f(X_i)$ obtained from a standard regression tree algorithm with $y^*_{li(m)}$ as responses, X_i as covariates, and W_i as weights, $i = 1, \ldots, n$. Note that the tree is built as usual using all N observations as inputs along with their covariate vectors but with the specified weights (see the appendix in Section 2.7 for details),

iii) $\hat{b}_{i(m)} = \hat{D}_{(m-1)}(W_i^{\frac{1}{2}(M)} Z_i)^T \hat{V}^{-1}_{i(m-1)} \left(W_i^{\frac{1}{2}(M)} y_{li}^M - W_i^{\frac{1}{2}(M)} \hat{f}_{(m)}(X_i) \right)$,

where $\hat{V}_{i(m-1)} = W_i^{\frac{1}{2}(M)} Z_i \hat{D}_{(m-1)} (W_i^{\frac{1}{2}(M)} Z_i)^T + \hat{\sigma}^2_{(m-1)} I_{n_i}, i = 1, ..., n.$

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

$$\hat{\sigma}^2_{(m)} = N^{-1}\sum_{i=1}^{n}\left\{\hat{\epsilon}^T_{i(m)}\hat{\epsilon}_{i(m)} + \hat{\sigma}^2_{(m-1)}[n_i - \hat{\sigma}^2_{(m-1)}\text{trace}(\hat{V}_{i(m-1)})]\right\},$$

$$\hat{D}_{(m)} = n^{-1}\sum_{i=1}^{n}\left\{\hat{b}_{i(m)}\hat{b}^T_{i(m)} + [\hat{D}_{(m-1)} - \hat{D}_{(m-1)}(W_i^{\frac{1}{2}(M)}Z_i)^T\hat{V}^{-1}_{i(m-1)}W_i^{\frac{1}{2}(M)}Z_i\hat{D}_{(m-1)}]\right\},$$

where $\hat{\epsilon}_{i(m)} = W_i^{\frac{1}{2}(M)}y_{li}^{(M)} - W_i^{\frac{1}{2}(M)}\hat{f}_{(m)}(X_i) - W_i^{\frac{1}{2}(M)}Z_i\hat{b}_{i(m)}$.

Micro Step 3. Repeat steps 1 and 2 until convergence in terms of the generalized log-likelihood value :

$$GLL(f, b_i|y) = \sum_{i=1}^{n}\{\hat{\epsilon}^T_{i(m)}(\hat{\sigma}^2_{(m)}I_{n_i})^{-1}\hat{\epsilon}_{i(m)} + \hat{b}^T_{i(m)}\hat{D}^{-1}_{(m)}\hat{b}_{i(m)} + \log|\hat{D}_{(m)}| + \log|\hat{\sigma}^2_{(m)}I_{n_i}|\}.$$

MACRO STEP 1. Set $M = M + 1$. Set $\hat{\eta}_i^{(M-1)} = \hat{y}_{li}^{(M-1)} = \hat{f}(X_i) + Z_i\hat{b}_i$, where $\hat{f}$ and $\hat{b}_i$ equal their estimated values at the micro level convergence of the previous macro iteration. Set $\hat{\mu}_i^{(M)} = g^{-1}(\hat{\eta}_i^{(M-1)})$ and fit a new weighted MERT pseudo-model using the updated $y_{li}^{(M)}$ and $W_i^{(M)}$, i.e., repeat the micro steps 0 to 3 using as initial values for $\hat{b}_i$, $\hat{\sigma}^2$, and $\hat{D}$, their micro-level convergence values in the previous macro iteration.

MACRO STEP 2. Repeat macro step 1 until convergence of $\hat{\eta}_i$.

The GMERT model can be used to get the predicted response for a new observation that belongs to a cluster among those used to fit this model as well as for a new observation that belongs to a cluster not included in the sample used to fit this model. To predict the response for a new observation that belongs to a cluster among those used to fit the generalized mixed effects regression tree model, we use both its corresponding fixed component prediction and the predicted random part corresponding to its cluster. This is a cluster-specific estimate. For a new observation that belongs to a cluster not included in the sample used to estimate the model parameters, we can only use its corresponding fixed component prediction (i.e., the random part is set to 0).

2.3.3 GMERT Model in the Binary Response Case

For clustered data with a binary response variable, i.e., $y_{ij} = \mu_{ij} + \varepsilon_{ij}$ with $E(\varepsilon_{ij}) = 0$ and $Var(\varepsilon_{ij}) = \sigma^2 v_{ij} = \sigma^2 \mu_{ij}(1 - \mu_{ij})$, the commonly used parametric model is the mixed effects logistic regression model with the logit link function, namely

$$\eta_{ij} = g(\mu_{ij}) = logit(\mu_{ij}) = \ln[\frac{\mu_{ij}}{1 - \mu_{ij}}] = x_{ij}^T \beta + z_{ij}^T b_i. \tag{2.6}$$

The conditional expectation $\mu_{ij} = E(y_{ij}|b_i, x_{ij}) = P(y_{ij} = 1|b_i, x_{ij})$ is the conditional probability of success given the random effects and covariate values. This model can also be written as follows :

$$P(y_{ij} = 1|b_i, x_{ij}) = g^{-1}(\eta_{ij}), \tag{2.7}$$

where $g^{-1}(\eta_{ij}) = \frac{1}{1+\exp(-\eta_{ij})}$ is the logistic cumulative distribution function.

The GMERT model in the binary response case (i.e., mixed effects classification tree) and its corresponding MERT pseudo-model are respectively defined as follows :

$$\eta_{ij} = \ln[\frac{\mu_{ij}}{1 - \mu_{ij}}] = f(x_{ij}) + z_{ij}^T b_i, \tag{2.8}$$

$$y_{lij} = \eta_{ij} + e_{ij}, \tag{2.9}$$

where $e_{ij} = (y_{ij} - \mu_{ij})g'(\mu_{ij})$, $g'(\mu_{ij}) = [\mu_{ij}(1 - \mu_{ij})]^{-1}$, and $Var(e_{ij}) = \sigma^2[\mu_{ij}(1 - \mu_{ij})]^{-1}$. The weights to be used in the GMERT algorithm are $w_{ij} = \mu_{ij}(1 - \mu_{ij})$.

The GMERT model can be used to get a predicted probability of success for a new observation that belongs to a cluster among those used to fit this model or for a new observation that belongs to a cluster not included in the sample used to fit this model. If the new observation j belongs to a cluster i in the first category, then its predicted probability of success $\hat{\mu}_{ij}$ equals $\frac{1}{1+\exp(-\hat{f}(x_{ij})-z_{ij}^T \hat{b}_i)}$, where $\hat{f}(x_{ij})$ is its predicted fixed component that

results from the fixed tree rules and $z_{ij}^T \hat{b}_i$ is its predicted random part corresponding to its cluster. However, if the new observation j belongs to a cluster i in the second category, then its predicted probability $\hat{\mu}_{ij}$ equals $\frac{1}{1+\exp(-\hat{f}(x_{ij}))}$ (i.e., the random part is set to 0).

2.4 Simulation

In this section, we investigate the performance of the GMERT method for binary outcomes in comparison to standard classification trees. The proposed GMERT method was implemented in R by means of a repeated call to the MERT algorithm. The latter uses the function *rpart* (Therneau and Atkinson, 1997). This function implements cost-complexity pruning based on cross-validation after an initial large tree is grown. In order to ensure that initial trees are sufficiently large, we set the complexity parameter to zero (i.e., $cp = 0$ means that any split that does not decrease at all the overall lack of fit is also attempted). Though there is a clear waste of computing time when not pruning off splits that are clearly not worthwhile, doing so ensure that the two methods to be compared were given equal chance to fit the data. In addition, we fixed the value of other parameters to reasonable (i.e., given the true tree model and the generated data sample to be used) and equal levels. That is, we set to five the maximum depth of any node of the final tree (i.e., $maxdepth = 5$), to 50 the minimum number of observations that must exist in a node in order for a split to be attempted (i.e., $minsplit = 50$), and to 10 the minimum number of observations in any terminal node (i.e., $minbucket = 10$). The largest tree is grown then pruned automatically based on minimum ten-folds cross-validated error.

For GMERT models, we used the following schema to stop the macro-micro iteration process and select a final model. Within each macro iteration, we follow the MERT algorithm convergence process. More precisely, we first impose a minimum of 50 micro iterations to avoid early stopping, then we keep iterating until either the absolute change in the generalized log-likelihood, GLL, is less than 1E-06 or we reach a maximum of 200 micro iterations. Once the stopping criterion is reached, we let the process continue for an additional 50 micro iterations. We then find the most frequent (modal value) number of leaves for the selected

subtrees in the sequence of additional iterations. The final subtree model chosen at the micro iteration level, is the one corresponding to the last micro iteration where the number of leaves is equal to the modal value.

At the macro iteration level, we keep iterating until either the absolute change in $\hat{\eta}_i$ is less than 1E-10 or we reach the maximum of 15 macro iterations. Once the stopping criterion is reached, we let the process continue for an additional 5 macro iterations. We then find the most frequent (modal value) number of leaves for the selected micro iteration subtrees in the sequence of additional macro iterations. The final GMERT model chosen is the one corresponding to the last macro iteration where the number of leaves is equal to the modal value.

To compare the performance of standard and mixed effects classification trees, we evaluate their predictive accuracy measured by the predictive mean absolute deviation in terms of the estimated probability (PMAD) and the predictive misclassification rate (PMCR).

2.4.1 Simulation Design

The simulation design used has a hierarchical structure of 100 clusters with 60 observations each. The first ten observations in each cluster form the training sample, and the other 50 observations are left for the test sample. Consequently, the trees are built from 1000 observations (100 clusters of 10 observations). Eight random variables, X_1 to X_8, independent and uniformly distributed in the interval $[0, 10]$ are generated. Only the first five are used predictors. The conditional or cluster-specific probabilities of success, μ_{ij}, are generated based on the following fixed tree rules along with the random component :

Leaf 1. If $x_{1ij} \leq 5$ and $x_{2ij} \leq 5$ then $\mu_{ij} = g^{-1}(g(\varphi^1) + z_{ij}^T b_i)$,

Leaf 2. If $x_{1ij} \leq 5$ and $x_{2ij} > 5$ and $x_{4ij} \leq 5$ then $\mu_{ij} = g^{-1}(g(\varphi^2) + z_{ij}^T b_i)$,

Leaf 3. If $x_{1ij} \leq 5$ and $x_{2ij} > 5$ and $x_{4ij} > 5$ then $\mu_{ij} = g^{-1}(g(\varphi^3) + z_{ij}^T b_i)$,

Leaf 4. If $x_{1ij} > 5$ and $x_{3ij} \leq 5$ and $x_{5ij} \leq 5$ then $\mu_{ij} = g^{-1}(g(\varphi^4) + z_{ij}^T b_i)$,

Leaf 5. If $x_{1ij} > 5$ and $x_{3ij} \leq 5$ and $x_{5ij} > 5$ then $\mu_{ij} = g^{-1}(g(\varphi^5) + z_{ij}^T b_i)$,

Leaf 6. If $x_{1ij} > 5$ and $x_{3ij} > 5$ then $\mu_{ij} = g^{-1}(g(\varphi^6) + z_{ij}^T b_i)$,

where $g()$ is the logit link function, φ^1 to φ^6 are the typical probabilities of success (i.e., probability of success when the random effects b_i equal zero), and $b_i \sim N(0, D)$, for $i = 1, ..., 100$, $j = 1, ..., 60$ (see figure 2.1). Each observation j in cluster i falls into only one of the six terminal nodes with a typical probability equal to φ^1, ..., φ^6 respectively. The binary response values y_{ij} are generated according to a Bernoulli distribution using the *rbinom* function of R with the *size* parameter fixed to one (i.e. one trial) and the *prob* parameter fixed to μ_{ij} (i.e. the generated conditional probability of success).

Insert Figure 2.1 about here

We consider 10 different data generating processes (DGP), summarized in Table 2.I. Two different scenarios are selected for the fixed components. In the large fixed effects scenario, the probabilities are chosen so that when there is no random effect, the standard classification tree is able to recover the true number of leaves most of the time (i.e., about 95% of times). In the small fixed effects scenario, the probabilities are chosen so that when there is no random effect, the standard classification tree is much less able to recover the true number of leaves (i.e., about 55% of times only).

Insert Table 2.I about here

The random components are generated based on the following three different scenarios :

1. No random effects (NRE), i.e. $D = 0$.

2. Random intercept (RI), i.e. $z_{ij} = 1$ for $i = 1, ..., 100$, and $j = 1, ..., 60$, and $D = d_{11} > 0$.
3. Random intercept and covariate (RIC) which is a RI with a linear random effect for X_1. More precisely, $z_{ij} = [1, x_{1ij}]$ for $i = 1, ..., 100$, $j = 1, ..., 60$, and $D = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}$, $d_{11} > 0$, $d_{22} > 0$, and $d_{12} = d_{21} = 0$.

Within each fixed effects scenario with random effects, we consider two levels (low and high) for the between-clusters covariance matrix D. More precisely, we consider that the random effect is small (large) when it results in about 10% (30%) of the observations' classes being shifted from 1 to 0 or vice versa.

We adjust three models for each DGP scenario : 1) a standard (STD) classification tree model, 2) a random intercept (RI) classification tree model, and 3) a random intercept and covariate (RIC) classification tree model. The true model corresponds to the DGP used to generate the data. In addition, using the *glmmPQL* function of R, we fitted for each DGP scenario a parametric mixed effects logistic regression model (MElog) that uses the true model leaves' indicators as predictors and the true random effects structure. Clearly, this model is not a real competitor since it is not possible in practice to specify this parametric structure without knowing the true underlying data generating process. The MElog model only serves as a benchmark for comparing the performance of the GMERT model. Overall, we built 40 models (10 scenarios $\times$ 4 models). The simulation results are obtained by means of 100 runs.

2.4.2 Simulation Results

Firstly, the performance of the methods is judged based on their predictive accuracy on the test set as measured by : 1) the predictive mean absolute deviation (PMAD) in terms of the estimated probability, and 2) the predictive misclassification rate (PMCR), i.e.,

$$PMAD = \frac{\sum_{i=1}^{100} \sum_{j=1}^{50} |\mu_{ij} - \hat{\mu}_{ij}|}{5000},$$

$$PMCR = \frac{\sum_{i=1}^{100}\sum_{j=1}^{50}|y_{ij} - \hat{y}_{ij}|}{5000},$$

where $\hat{\mu}_{ij}$ and $\hat{y}_{ij}$ are, respectively, the predicted probability and the predicted class of observation j in cluster i in the test data set. Secondly, we compare the performance of the GMERT approach to the MElog benchmark results.

The misclassification rate depends to some extent on the classification strategy and the cutpoint value used to classify the observations, in particular, when the data has a nested structure with clusters having different sizes in the training and the test data sets. The adopted strategy consists in these steps :

1. Sort the distinct predicted probabilities of the observations in the training set (there are, at most, number of clusters × number of terminal nodes distinct probabilities),
2. Classify the observations in the training set using in turn each one of these distinct predicted probabilities as a cutpoint ; classify as class 1 each observation in the training set that has a predicted probability equal to or higher than the cutpoint value,
3. Compute the proportions of class 1 that result from each one of the above cutpoint values,
4. Find the predicted probability among those in step 1 that yields the closest proportion of class 1 to the actual proportion of class 1 in the training set, and
5. Use this predicted probability as the cutpoint value in order to classify the observations in the test set, i.e., classify as class 1 each observation in the test set that has a predicted probability equal to or higher than this cutpoint value.

The average, median, minimum, maximum and standard deviation of the PMAD (columns 5 to 9) and the PMCR (columns 10 to 14) over the 100 runs were calculated and are presented in Table 2.II.

Insert Table 2.II about here

In terms of predictive accuracy (PMAD and PMCR), we note that when random effects are present (DGPs 3 to 10), the mixed effects classification tree does better than the standard classification tree even with a wrong specification of the random component part. The highest difference in terms of PMAD and PMCR is observed when both the fixed and the random effects are somewhat large (i.e., 21.65% and 17.2% in DGP 4, and 20.23% and 17.06% in DGP 8, respectively). The lowest difference in terms of PMAD and PMCR is observed when both the fixed and the random effects are somewhat small (i.e., 1.85% and 0.33% in DGP 5, and 1.82% and 0.46% in DGP 9, respectively). In addition, when there is no random effect (DGPs 1 and 2), the standard classification tree algorithm does slightly better in terms of PMAD and PMCR than the proposed GMERT approach with the highest difference being less than 2% (i.e., 1.08% and 1.15% in DGP 1, and 1.38% and 1.90% in DGP 2, respectively).

The difference in predictive accuracy (PMAD and PMCR) between the benchmark model MElog and the GMERT model reaches a minimum when the fixed effects are large while the random effects are small (DGPs 3 and 7), and a maximum when both the fixed and the random effects are small (DGPs 5 and 9). In terms of PMAD, this difference equals 0.80% and 1.13% in DGPs 3 and 7 respectively, and 2.50% and 2.25% in DGPs 5 and 9 respectively. In terms of PMCR, this difference equals 0.69% and 1.02% in DGPs 3 and 7 respectively, and 2.50% and 2.34% in DGPs 5 and 9 respectively. When there is no random effects, the difference in predictive accuracy between the benchmark model MElog and the GMERT model is, as anticipated, higher when the fixed effects are small (the PMAD difference in DGP 2 is 2.63 times the PMAD difference in DGP 1, and the PMCR difference in DGP 2 is 2.56 times the PMCR difference in DGP 1).

2.5 Discussion

Earlier extensions of standard tree methods to the case of correlated data (Segal, 1992; Zhang, 1998; Abdolell, Leblanc, Stephens, and Harrison, 2002; Lee, 2005) do not allow observation-level covariates to be candidates in the splitting process and, consequently, 1) no random or cluster-specific effect is allowed, and 2) all repeated observations from a given subject remain together during the tree building process and can not be splitted across different nodes. The MERT method (Hajjem et al., 2008) and the GMERT method proposed in this paper can appropriately deal with the possible random effects of observation-level covariates since these covariates are candidates in the splitting process. As a consequence, the observations within clusters may be splitted.

The GMERT model is a cluster-specific or conditional model which yields cluster-specific or conditional means estimates, $\mu_i = g^{-1}(\hat{f}(X_i) + Z_i\hat{b}_i)$, and not population-averaged or marginal means estimates, $E(\mu_i)$.

Although the simulation study focused on the binary response case, the GMERT method can be tailored to other types of response variables (e.g., counts data, ordered categorical outcomes, and multicategory nominal outcomes). Similarly to GLM, GMERT method transforms the expected outcome using an appropriate link function according to the type of the response variable and then equates it to a tree function of the fixed effects along with a linear function of the random effects. Future research would be to look for a tree structure representation for the random component as well, which may be more suitable in more complex problems.

2.6 Conclusion

In the present paper, we extended the mixed effect regression tree approach to other types of outcomes (e.g., binary outcomes, counts data, ordered categorical outcomes, and multicategory nominal scale outcomes).

The simulation results in the binary case show substantial improvements of the predictive accuracy over the standard classification tree, whenever random effects are present. However, the main limit of tree based method, including the one proposed here, is their instability. Ensemble methods such as bagging (Breiman, 1996) and forest of trees (Breiman, 2001) can greatly improve the predictive performance of trees. Hence, further improvement of the predictive accuracy of the GMERT method could be achieved if we use it as the base learner in an ensemble algorithms. This remains for future work.

An R program implementing the generalized mixed effects regression tree procedure is available from the first author.

2.7 Appendix : Weighted Standard Regression Tree Within GMERT Algorithm

Here we clarify how the weights intervene in the standard regression tree fitted at each micro iteration within the GMERT algorithm. At any micro iteration within a given macro iteration, the standard regression tree uses the corresponding $y^*_{li} = y_i - Z_i\hat{b}_i$ as the dependent variable and X_i as the covariates, along with the weights $W_i = diag(w_{ij})$, with $i = 1, ..., n$ and $j = 1, ..., n_i$.

Let T be the fitted standard regression tree, and let t be one of its nodes. Node t contains a subset of $N_t < N$ observations that belong to a subset of $n_t \leq n$ clusters with pseudo-responses $y^*_{li_tj_t}$, $i_t = 1, ..., n_t$ and $j_t = 1, ..., n_{i_t}$. Then, given the weights $w_{i_tj_t}$ of observation j_t in cluster i_t in node t, we have :

- The summary statistic to be attached to node t corresponds to its weighted response average $\bar{y}^*_{lt} = \frac{\sum_{i_t=1}^{n_t}\sum_{j_t=1}^{n_{i_t}} w_{i_tj_t} y^*_{li_tj_t}}{\sum_{i_t=1}^{n_t}\sum_{j_t=1}^{n_{i_t}} w_{i_tj_t}}$. This corresponds to the fitted value $\hat{y}^*_{lt} = \hat{f}(X_{i_t})$ when t is a terminal node.
- The error of node t equals its weighted sums of squares or corrected deviance DEV_t, with $DEV_t = \sum_{i_t=1}^{n_t}\sum_{j_t=1}^{n_{i_t}} w_{i_tj_t}(y^*_{li_tj_t} - \bar{y}^*_{lt})^2$.
- The splitting criterion is the improvement or the percent change in the weighted sums of squares for a given split of node t into two nodes t_l and t_r, i.e., $Improve = 1 - \frac{(DEV_{t_l} + DEV_{t_r})}{DEV_t}$.
- The cross-validated relative error corresponding to a given complexity parameter value for the tree T is defined as follows : $xerror = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}(y^*_{lij} - \hat{y}^*_{l(-ij)})^2}{DEV_{root}}$, with $\hat{y}^*_{l(-ij)}$ being the predicted value for observation j in cluster i, from the standard regression tree model that is fitted without this observation.

2.8 References

Abdollel, M., LeBlanc, M., Stephens, D. and Harrison, R. V. (2002). Binary partitioning for continuous longitudinal data : Categorizing a prognostic variable. *Statistics in Medicine,* 21, 3395-3409.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees.* Wadsworth International Group. Belmont, California.

Breiman, L. (1996). Bagging predictors. *Machine Learning,* 24, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning,* 45, 5-32.

Breslow, N.and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association,* 88, 9-25.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,* 39, 1-38.

Goldstein, H. and Rasbash, J. R. (1996). Improved Approximations for Multilevel Models with Binary Responses. *Journal of the Royal Statistical Society,* 159, 505-513.

Hajjem, A., Bellavance, F., and Larocque, D. (2008). Mixed Effects Regression Trees for Clustered Data. Submitted. *Les Cahiers du GERAD,* G-2008-57.

Harville, D. A. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *Annals of Statistics,* 4, 384-395.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics,* 38, 963-974.

Lee, S. K. (2005). On Generalized multivariate decision tree by using GEE. *Computational Statistics & Data Analysis,* 49, 1105-1119.

McLachlan, G. J. and Krishman, T. (1997). *The EM algorithm and extensions.* New York :Wiley.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models (2nd Edition).* Chapman & Hall/CRC. London.

R Development Team (2007). *R : A Language and environment for statistical computing.* R Foundation for Statistical Computing : www.R-project.org.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models : Applications and data analysis method (2nd Edition).* Sage. Newbury Park, CA.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., and du Toit, M. (2004). *HLM 6 : Hierarchical Linear & Nonlinear Modeling.* Scientific Software International, Inc.

SAS Institute Inc. (2008). *SAS/STAT 9.2 User's Guide : The GLIMMIX Procedure (Book*

Excerpt). Cary, NC : SAS Institute Inc.

Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association,* 87, 407-418.

Therneau, T. M. and Atkinson, E. J. (1997). *An introduction to recursive partitioning using the rpart routines.* Technical Report 61, Department of Health Science Research, Mayo Clinic, Rochester.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S (Fourth edition).* New York : Springer-Verlag.

Wu, H. and Zhang, J. T. (2006). *Nonparametric regression methods for longitidunal data analysis : Mixed-effects modeling approaches.* Wiley. New York.

Zhang, H. (1998). Classification trees for multiple binary responses. *Journal of the American Statistical Association,* 93, 180-193.

Table 2.I Data generating processes (DGP) for the simulation study.

DGP	Data Structure									
	Fixed Component							Random Component		
	Effect	φ^1	φ^2	φ^3	φ^4	φ^5	φ^6	Structure	d_{11}	d_{22}
1	Large	0.10	0.20	0.80	0.20	0.80	0.90	No random effect	0.00	0.00
2	Small	0.20	0.40	0.70	0.30	0.60	0.80			
3	Large	0.10	0.20	0.80	0.20	0.80	0.90	Random intercept	4.00	0.00
4									10.00	0.00
5	Small	0.20	0.40	0.70	0.30	0.60	0.80		0.50	0.00
6									4.00	0.00
7	Large	0.10	0.20	0.80	0.20	0.80	0.90	Random intercept and covariate	2.00	0.05
8									5.00	0.25
9	Small	0.20	0.40	0.70	0.30	0.60	0.80		0.25	0.01
10									2.00	0.05

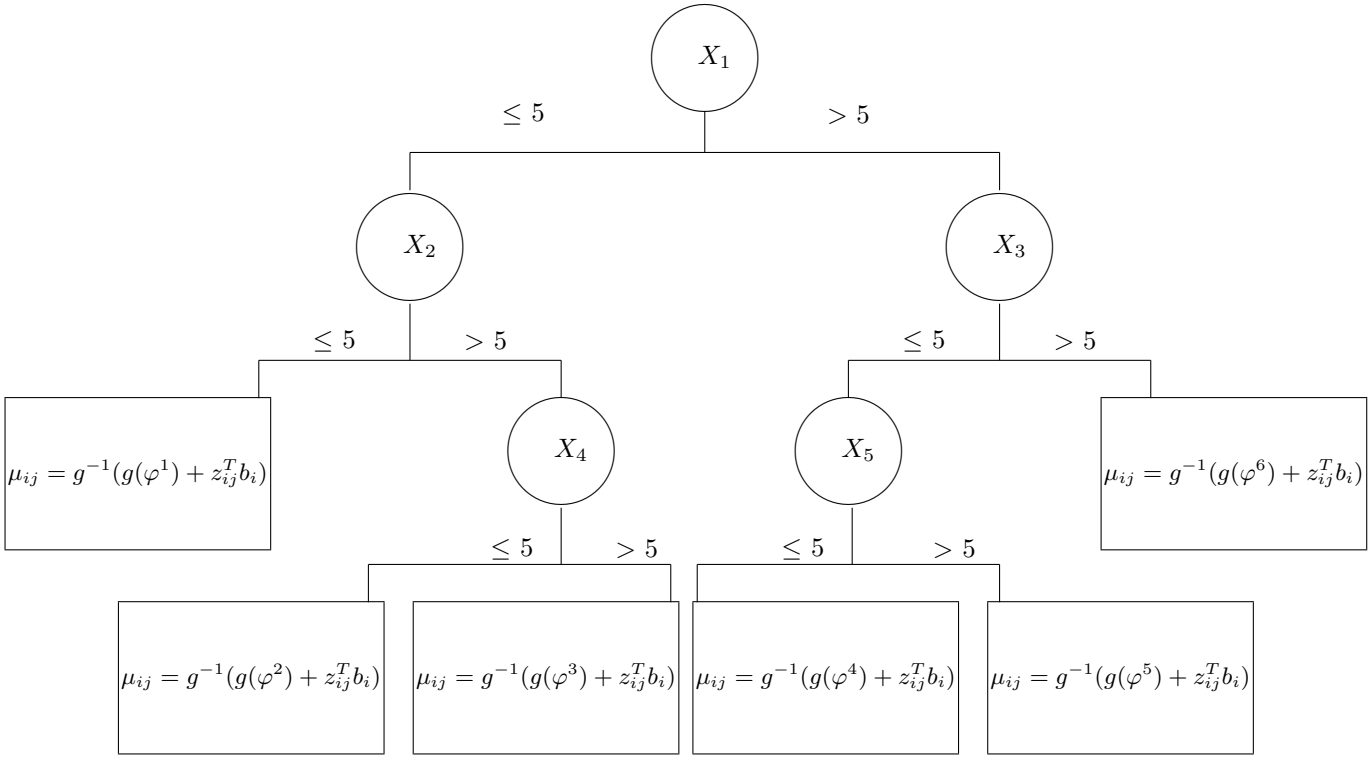

Figure 2.1 Generalized mixed effects tree structure used for the simulation study, with $g(.)$ being the logit link function and $g(.)^{-1}$ the inverse-logit or logistic function.

Table 2.II Results of the 100 simulation runs in terms of the predictive probability mean absolute deviation (PMAD) and the predictive misclassification rate (PMCR).

DGP	Fixed effect	Random effect	Fitted model*	PMAD (%)					PMCR (%)				
				Avg.	Med.	Min	Max	Std	Avg.	Med.	Min	Max	Std
1	Large	No random effect	STD	3.09	3.05	1.48	6.38	0.97	15.71	15.67	13.92	18.20	0.79
			RI	3.86	3.66	1.28	8.85	1.46	16.86	16.58	14.54	21.44	1.52
			RIC	4.17	3.98	1.31	8.85	1.49	16.85	16.60	14.52	21.78	1.55
			MElog	2.48	2.36	0.78	4.80	0.87	15.49	15.39	13.86	17.62	0.71
2	Small		STD	4.97	4.64	1.73	11.98	1.89	29.33	28.94	26.94	34.68	1.63
			RI	6.35	5.95	2.23	13.36	2.81	31.23	30.41	26.74	38.82	2.83
			RIC	6.32	5.82	2.43	12.52	2.68	31.00	30.18	26.66	38.68	2.70
			MElog	2.73	2.72	0.86	5.34	0.82	27.72	27.76	26.22	29.12	0.68
3	Large	Random intercept	STD	21.70	21.48	17.44	26.50	1.68	26.49	26.23	21.90	30.90	1.81
			RI	9.20	9.12	7.10	12.13	0.99	19.82	19.78	17.36	22.18	1.11
			RIC	9.69	9.58	7.10	14.87	1.20	20.08	20.02	17.82	22.86	1.16
			MElog	8.40	8.48	6.26	9.94	0.62	19.13	19.13	16.56	21.24	0.87
4			STD	30.24	29.97	25.29	35.50	1.98	33.65	33.23	28.92	41.14	2.58
			RI	8.59	8.52	6.80	11.42	0.84	16.45	16.46	12.20	20.16	1.16
			RIC	9.37	9.27	7.28	13.61	1.05	16.93	16.85	14.50	20.06	1.15
			MElog	7.59	7.57	6.06	9.14	0.65	15.69	15.73	11.82	18.34	1.07
5	Small		STD	12.56	12.36	10.40	15.97	1.30	31.70	31.36	29.06	36.44	1.67
			RI	10.71	10.54	7.81	15.44	1.58	31.37	31.17	28.14	36.58	1.62
			RIC	10.79	10.69	7.86	15.43	1.53	31.38	31.12	28.46	36.72	1.58
			MElog	8.21	8.17	6.61	9.89	0.60	28.87	28.85	27.46	30.64	0.64
6			STD	26.77	26.80	21.53	30.53	1.47	39.32	39.21	34.96	46.10	2.35
			RI	11.20	11.08	8.91	14.73	1.10	24.00	24.03	20.66	30.40	1.43
			RIC	11.40	11.20	9.32	14.66	1.02	24.09	24.06	20.94	30.30	1.42
			MElog	9.01	8.94	7.65	10.95	0.67	22.56	22.49	19.64	26.62	1.23
7	Large	Random intercept and covariate	STD	20.37	20.48	16.33	23.62	1.24	25.31	25.34	21.76	28.30	1.21
			RI	10.86	10.74	9.25	13.83	0.88	20.87	20.85	18.50	23.32	0.93
			RIC	10.58	10.47	8.43	14.14	0.98	20.83	20.79	18.02	23.54	1.02
			MElog	9.61	9.53	8.10	12.49	0.70	20.04	19.95	17.68	22.32	0.85
8			STD	30.90	30.92	27.43	35.56	1.60	34.34	33.97	30.06	42.52	2.37
			RI	12.37	12.35	9.91	15.76	0.98	18.15	18.20	15.10	20.82	1.14
			RIC	10.67	10.52	8.63	14.73	1.12	17.28	17.29	14.68	21.12	1.10
			MElog	9.45	9.39	7.91	11.35	0.74	16.42	16.37	14.16	18.60	0.92
9	Small		STD	12.86	12.64	10.21	17.48	1.45	31.81	31.15	29.00	37.92	1.85
			RI	11.12	10.73	8.87	16.57	1.62	31.36	30.93	28.12	36.82	1.86
			RIC	11.04	10.62	8.50	16.19	1.65	31.35	30.83	28.24	36.12	1.85
			MElog	8.79	8.73	7.77	10.44	0.50	29.01	28.99	26.98	30.78	0.71
10			STD	25.42	25.18	21.48	28.76	1.58	39.02	38.90	34.26	46.26	2.59
			RI	13.11	13.05	10.67	15.86	1.16	25.98	25.89	22.42	29.12	1.4
			RIC	12.54	12.48	10.23	15.16	1.09	25.84	25.72	22.74	29.82	1.38
			MElog	10.41	10.34	8.89	12.84	0.71	24.24	24.39	21.24	26.94	1.19

* STD : Standard tree ; RI : Random intercept tree ; RIC : Random intercept and covariate tree ; MElog : Mixed effect logistic

ARTICLE III

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

3.1 Abstract

This paper presents an extension of the well known random forest method to the case of clustered data. The proposed "mixed effects random forest" method is implemented using a standard random forest algorithm within the framework of the expectation-maximization (EM) algorithm. The simulation results show that the proposed mixed effects random forest method provides substantial improvements over standard random forest when the random effects are non negligible.

Keywords : Clustered data, mixed effects, regression tree, random forest.

3.2 Introduction

Tree based methods are well known and well 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 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. Bagging and the more general random forest algorithms (Breiman, 1996, 2001) are well known and very powerful ensemble methods for trees.

Using the mixed effects approach, Hajjem, Bellavance and Larocque (2008, 2010) extended the well known CART algorithm (Breiman, Friedman, Olshen and Stone, 1984) to the case of clustered data. They proposed the mixed effects regression tree (MERT) algorithm for a continuous outcome and the generalized mixed effects regression tree (GMERT) algorithm for discrete outcomes in clustered data settings. Simulation results showed that these methods provide 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 consists in the use of a standard regression tree algorithm within the framework of the expectation-maximization (EM) algorithm. MERT is basically an iterative call to the standard regression tree algorithm. At each iteration, the standard regression tree (SRT) is applied to the original response from which the current estimate of the random effect component is removed.

Following the same idea, one possibility for generalizing the standard random forest to clustered data consists in replacing the SRT within each iteration of the MERT algorithm with a standard forest of regression trees. The goal of the present paper is to introduce this new random forest method, named "mixed effects random forest" (MERF), and to investigate its performance with a simulation study. For that matter, the predictive performance of MERF will be compared to that of five alternative models, including the standard random forest, by varying some key features related to the strength of both the total and the random effects and to the dependence between the predictors. The main finding is that MERF seems to be more appropriate than a standard random forest for clustered data, particularly when the random effects are non-negligible.

The remainder of this article is organized as follows : Section 3.3 describes the proposed MERF approach ; Section 3.4 presents a simulation study to evaluate the performance of MERF ; Section 3.5 gives some concluding remarks.

3.3 Mixed Effects Random Forest Approach

We define the mixed effects random forest (MERF) of regression trees as follows :

$$\begin{gathered} y_i = f(X_i) + Z_i b_i + \epsilon_i, \\ b_i \sim N_q(0, D), \epsilon_i \sim N_{n_i}(0, R_i), \\ i = 1, ..., n, \end{gathered} \tag{3.1}$$

where $y_i = [y_{i1}, ..., y_{in_i}]^T$ is the $n_i \times 1$ vector of responses for the n_i observations in cluster i, $X_i = [x_{i1}, ..., x_{in_i}]^T$ is the $n_i \times p$ matrix of fixed-effects covariates, $Z_i = [z_{i1}, ..., z_{in_i}]^T$ is the $n_i \times q$ matrix of random-effects covariates, $\epsilon_i = [\epsilon_{i1}, ..., \epsilon_{in_i}]^T$ is the $n_i \times 1$ vector of errors, $b_i = (b_{i1}, ..., b_{iq})^T$ is the $q \times 1$ unknown vector of random effects for cluster i, and the unknown function $f(X_i)$ is estimated using a standard forest of regression trees. 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-clusters observations are independent. Hence, the covariance matrix of the vector of observations y_i in cluster i is $V_i = Cov(y_i) = Z_i D Z_i^T + R_i$, and $V = Cov(y) = diag(V_1, \ldots, V_n)$, where $y = [y_1^T, ..., y_n^T]^T$. We will also assume that the correlation is induced solely via the between-clusters variation, that is, R_i is diagonal ($R_i = \sigma^2 I_{n_i}, i = 1, ..., n$).

Basically, the MERF algorithm is the MERT algorithm (Hajjem et al., 2008) where the single regression tree structure used to estimate the fixed part of the model is replaced by an ensemble of unpruned regression trees (i.e. a forest). The out-of-bag estimates of the standard forest are used to predict the response fixed part.

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

ii) Let $\hat{f}(X_i)_{(r)}$ be an estimate of $f(X_i)$ obtained from the out-of-bag predictions of a standard random forest algorithm with $y^*_{i(r)}$ as the training set responses, X_i, $i = 1, \ldots, n$, as the corresponding training set of covariates, and taking as inputs a selected number of bootstrap training samples drawn with replacement from the training set $(y^*_{i(r)}, X_i)$, $i = 1, ..., n$.

iii) $\hat{b}_{i(r)} = \hat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} \left(y_i - \hat{f}(X_i)_{(r)} \right), i = 1, ..., n,$

where $\hat{V}_{i(r-1)} = Z_i \hat{D}_{(r-1)} Z_i^T + \hat{\sigma}^2_{(r-1)} I_{n_i}, i = 1, ..., n.$

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

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

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

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^*, i.e., 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_i^*, X_i) to build a forest of trees. To minimize over fitting, 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). Third, it updates $\hat{b}_i$. At step 2, it updates the variance components $\hat{\sigma}^2$ and $\hat{D}$ based on the residuals after the estimated fixed component $\hat{f}(X_i)$ is removed from the raw data y_i It 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 :

$$\begin{aligned} GLL(f, b_i|y) = \sum_{i=1}^{n} \{ & [y_i - f(X_i) - Z_i b_i]^T R_i^{-1} [y_i - f(X_i) - Z_i b_i] \\ & + b_i^T D^{-1} b_i + \log|D| + \log|R_i| \}. \end{aligned} \tag{3.2}$$

To predict the response for a new observation that belongs to a cluster among those used to fit the MERF model, we use both its corresponding population-averaged random forest prediction and the predicted random part corresponding to its cluster. 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 random forest prediction.

3.4 Simulation

We investigate the performance of the proposed mixed effects random forest of regression trees through a simulation study. We compare the predictive mean squared error (PMSE) of the MERF to that of five alternative models, namely, 1) the standard random

forest (SRF) of regression trees, 2) the mixed effects regression tree (MERT), 3) the standard regression tree (SRT), 4) the linear mixed effect (LME) model, and 5) the linear model (LM).

We implemented the proposed MERF algorithm in R (R Development Core Team, 2007) using the package *randomForest* (Liaw and Wiener, 2002). The function *randomForest* implements Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code) for classification and regression. Except for the parameter *ntree* which corresponds to the number of trees to grow within the forest, all the other default settings of the function *randomForest* are used. To save overall computing time for the simulation, we set the value of the parameter *ntree* to 300 instead of the default value of 500. 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 1/3 of the original sample on average. The SRT and MERT models are also fitted with the default settings of the function *rpart* (Therneau and Atkinson, 1997).

For MERF convergence, we suggest to force for a minimum number of iterations to avoid early stopping then keep iterating until the absolute change in GLL is less than a given small value (e.g. 1E-06). For the simulation, we however fixed the total number of iterations to 300, regardless the behavior of GLL. Preliminary simulation runs showed that GLL stabilizes between 50 and 250 iterations for the settings considered (see Subsection §3.1). The final MERF model is the one at the last iteration. For MERT models, we force a minimum of 50 iterations, then keep iterating while the absolute change in GLL is not less than 1E-06 or we reach a maximum of 300 iterations. Once the stopping criterion is met, we run an additional 50 iterations. The mixed 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 tree models (Hajjem et al. 2008).

3.4.1 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 test sample. Consequently, the trees are built with 500 observations nested within 100 unbalanced clusters having 1, 3, 5, 7, or 9 observations. The remaining 4500 observations form the test set.

The data generating process is as follows. Nine random variables are first generated from a multivariate normal distribution $(X_1, ..., X_9) \sim N_9(0, \Sigma)$ 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 linear model, using only the first three random variables :

$$\begin{gathered} y_{ij} = m \times g(x_{ij}) + b_i + \varepsilon_{ij}, \\ g(x_{ij}) = 2x_{1ij} + x_{2ij}^2 + 4(x_{3ij} > 0) + 2\log|x_{1ij}|x_{3ij}, \\ b_i \sim N(0, \sigma_b^2), \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \\ i = 1, ..., 100, j = 1, ..., n_i, \end{gathered} \tag{3.3}$$

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

The proportion of total effects variability (PTEV) of the model in (3.3) is given by

$$PTEV = \frac{\sigma_{Fixed}^2 + \sigma_b^2}{\sigma_{Fixed}^2 + \sigma_b^2 + \sigma_\varepsilon^2} \times 100, \tag{3.4}$$

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

$$PREV = \frac{\sigma_b^2}{\sigma_{Fixed}^2 + \sigma_b^2} \times 100. \tag{3.5}$$

We consider 12 different data generating processes (DGP), summarized in Table 3.I. 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).

Insert Table 3.I about here

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 (3.4) and (3.5) to obtain the values of m and σ_b^2 for each DGP in Table 3.I.

The simulation results are obtained by means of 100 runs.

3.4.2 Simulation Results

Table 3.II presents for each data generating process (DGP) the summary statistics of the predictive mean squared error (PMSE) on the test set of the six fitted models. The PMSE is computed as :

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

Insert Table 3.II about here

Table 3.III presents for each data generating process (DGP) the summary statistics of the relative difference (RD) in PMSE between each alternative model and the MERF model :

$$RD = \frac{PMSE_{Alternative} - PMSE_{MERF}}{PMSE_{Alternative}} \times 100.$$

Insert Table 3.III about here

Figures 3.1 to 3.5 show for each DGP the distribution of the RD.

Insert Figure 3.1 about here

Insert Figure 3.2 about here

Insert Figure 3.3 about here

Insert Figure 3.4 about here

Insert Figure 3.5 about here

The primary interest of this simulation study is the comparison of MERF and SRF (Figure 3.1). The main finding is that for a given value of PTEV and ρ, the benefit of MERF over SRF increases as PREV increases. This can be seen by looking at the progression of RD between DGP 1, 2 and 3 (PTEV = .9 and $\rho = 0$), between DGP 4, 5 and 6 (PTEV = .6 and $\rho = 0$), between DGP 7, 8 and 9 (PTEV = .9 and $\rho = .4$) and finally by looking at the progression of RD between DGP 10, 11 and 12 (PTEV = .6 and $\rho = .4$). This result was intuitively expected but this simulation study helps revealing how crucially the performance of SRF depends on the PREV. The SRF is just not able to compensate for its omission of taking the random effects into account and its performance worsen as the relative importance of the random effects increases.

If we look into more details at the results, we can see that, except for few runs in settings with relatively large noise and small random effects (i.e. PTEV = 60% and PREV = 10% in DGPs 4 and 10), there is always some improvement (i.e. minimum RD > 0) of MERF over each alternative model considered (Table 3.III, Figures 3.1 to 3.5). In all cases, MERF did on average better (i.e. average RD > 0) than all the alternatives.

In all cases where the random effects are relatively small (i.e. PREV = 10% in DGPs 1, 4, 7, and 10), MERF average improvement over the alternative models vary between 16.02% and 46.05%, except the ones over SRF which are much lower but still non negligible with an average RD varying between 2.5%, and 10.65%. A failure to account for the correlation among the observations may result in much less predictive performance, even in relatively large noise and small random effects settings.

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 effect (i.e. LM, SRT, and SRF) appear in settings with large random effects (i.e. PREV = 50% in DGPs 3 and 9), the average RD 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.

In comparison to MERF improvement over LME, MERF improvement over MERT seems to be relatively less affected by the PREV (see Figures 3.2 and 3.4). One additional and interesting point to notice is the relatively huge variability of the improvement over MERT in comparison to that over LME ; the standard deviations of the improvement over MERT are more than twice those of the improvement over LME (Table 3.II, Figures 3.2 and 3.4).

The effect of the correlation between the predictors on the relative improvement of MERF is basically absent in large noise settings, and small in small noise settings with different trends depending on the alternative models. The average improvements over SRF, MERT, and SRT seem to be slightly higher when the predictors are correlated than when they are independent. In contrast, the average improvement over LME seems to be slightly higher when the predictors are independent than when they are correlated. There is no clear effect in the case of the alternative LM.

3.5 Concluding Remarks

One key feature of the random forest 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 intra-cluster 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 showed that this approach seems reasonable, at least in the scenarios used. A possibility for future work would be to investigate the robustness of the proposed approach

when the intra-cluster 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 (Field and Welsh, 2007). 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). One possibility for future work would be to investigate these strategies and compare them to the approach proposed in this paper.

Finally, the proposed method is appropriate for a continuous outcome. Other types of outcomes could be handled by using GMERT (Hajjem, Bellavance and Larocque, 2010) instead of MERT. Specifically, one could replace the single weighted regression tree used to estimate the pseudo-response fixed part in the doubly iterative GMERT algorithm with a forest of weighted standard regression trees. Investigating this idea is left for future work.

The objective of this paper was to propose a way to build a forest of trees with clustered data and to explore its performance. The results of the simulation study are promising and the 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.

3.6 References

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees.* Wadsworth International Group. Belmont, California.

Breiman, L. (1996). Bagging predictors. *Machine Learning,* 24, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning,* 45, 5-32.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,* 39, 1-38.

Field, C. A. and Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society, Series B,* 69, 369-390.

Hajjem, A., Bellavance, F., and Larocque, D. (2008). Mixed Effects Regression Trees for Clustered Data. Submitted. *Les Cahiers du GERAD,* G-2008-57.

Hajjem, A., Bellavance, F., and Larocque, D. (2010). Generalized Mixed Effects Regression Trees. *Les Cahiers du GERAD,* G-2010-39.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News,* 2, 18-22.

McLachlan G. J. and Krishman T. (1997). *The EM algorithm and extensions.* Wiley. New York.

R Development Team (2007). *R : A Language and environment for statistical computing.* R Foundation for Statistical Computing : www.R-project.org.

Therneau, T. M. and Atkinson, E. J. (1997). *An introduction to recursive partitioning using the rpart routines.* Technical Report 61, Department of Health Science Research, Mayo Clinic, Rochester.

Table 3.I Data generating processes (DGP) for the simulation study.

DGP	ρ	PTEV*	PREV**	σ^2_{Fixed}	m	σ^2_b	ICC***
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

*Proportion of Total Effects Variability $= \frac{\sigma^2_{Fixed}+\sigma^2_b}{\sigma^2_{Fixed}+\sigma^2_b+\sigma^2_\varepsilon} \times 100$

**Proportion of Random Effects Variability $= \frac{\sigma^2_b}{\sigma^2_{Fixed}+\sigma^2_b} \times 100$

***Intra Cluster Correlation $= \frac{\sigma^2_b}{\sigma^2_b+\sigma^2_\varepsilon} \times 100$

Table 3.II Results of the predictive mean squared error (PMSE) of MERF, SRF, MERT, SRT, LME, and LM models based on 100 simulation runs.

DGP	MERF					SRF					MERT					SRT					LME					LM				
	Avg.	*Med*	*Min*	*Max*	*Std*	*Avg.*	*Med*	*Min*	*Max*	*Std*	*Avg.*	*Med*	*Min*	*Max*	*Std*	*Avg.*	*Med*	*Min*	*Max*	*Std*	*Avg.*	*Med*	*Min*	*Max*	*Std*	*Avg.*	*Med*	*Min*	*Max*	*Std*
1	4.06	4.03	3.40	4.80	0.28	4.55	4.55	3.60	5.53	0.31	5.42	5.37	4.26	9.28	0.71	6.00	5.95	4.57	8.91	0.73	7.12	7.11	6.28	7.89	0.31	7.52	7.54	6.39	8.41	0.34
2	3.67	3.65	3.10	4.57	0.26	5.86	5.85	4.72	7.46	0.51	4.84	4.69	3.63	8.13	0.67	7.24	7.10	5.52	9.71	0.78	6.08	6.08	5.62	6.64	0.23	7.98	7.95	6.75	9.35	0.49
3	3.01	2.97	2.53	3.69	0.20	7.28	7.26	4.73	9.27	0.77	3.86	3.83	3.01	5.58	0.51	8.55	8.62	6.29	11.09	0.94	4.73	4.73	4.29	5.19	0.19	8.56	8.55	6.05	10.53	0.69
4	1.63	1.62	1.45	1.83	0.07	1.67	1.67	1.51	1.89	0.07	1.95	1.94	1.63	2.32	0.14	2.02	2.01	1.69	2.41	0.14	2.07	2.06	1.89	2.29	0.07	2.10	2.11	1.94	2.33	0.07
5	1.60	1.60	1.44	1.75	0.06	1.91	1.90	1.69	2.17	0.10	1.89	1.89	1.63	2.27	0.14	2.25	2.25	1.87	2.56	0.15	1.92	1.92	1.77	2.10	0.05	2.20	2.19	2.01	2.45	0.09
6	1.53	1.53	1.38	1.66	0.06	2.14	2.12	1.86	2.45	0.13	1.76	1.77	1.57	2.08	0.09	2.41	2.38	2.04	2.81	0.16	1.74	1.74	1.60	1.90	0.05	2.31	2.30	2.06	2.58	0.12
7	3.22	3.20	2.78	4.06	0.22	3.75	3.71	3.29	4.56	0.27	4.45	4.41	3.65	5.78	0.42	5.02	4.92	4.20	6.18	0.48	5.22	5.19	4.76	5.72	0.21	5.70	5.67	5.21	6.42	0.26
8	2.93	2.93	2.55	3.43	0.17	5.31	5.27	4.18	6.62	0.51	3.96	3.93	3.12	5.48	0.45	6.60	6.53	5.10	8.40	0.73	4.50	4.49	4.10	5.07	0.19	6.67	6.65	5.47	8.05	0.52
9	2.51	2.51	2.19	3.00	0.15	6.87	6.92	5.07	10.03	0.81	3.36	3.33	2.73	5.13	0.37	8.11	8.10	6.08	11.47	0.94	3.65	3.65	3.33	4.02	0.14	7.68	7.75	5.98	10.63	0.79
10	1.47	1.47	1.35	1.59	0.05	1.53	1.52	1.37	1.67	0.05	1.78	1.78	1.51	2.04	0.10	1.85	1.84	1.56	2.10	0.12	1.75	1.75	1.62	1.89	0.05	1.80	1.80	1.64	1.92	0.05
11	1.48	1.48	1.36	1.61	0.05	1.79	1.78	1.56	2.10	0.09	1.75	1.74	1.50	2.02	0.10	2.08	2.07	1.82	2.39	0.11	1.68	1.68	1.56	1.81	0.05	1.97	1.96	1.77	2.26	0.09
12	1.43	1.43	1.34	1.56	0.04	2.02	2.01	1.70	2.44	0.15	1.66	1.64	1.47	1.97	0.09	2.25	2.24	1.90	2.69	0.16	1.56	1.56	1.47	1.67	0.04	2.12	2.09	1.81	2.51	0.15

Table 3.III Relative difference (RD*) in PMSE between MERF and each one of the alternative models : SRF, MERT, SRT, LME, and LM.

DGP	SRF					MERT					SRT					LME					LM				
	Avg.	*Med*	*Min*	*Max*	*Std*	*Avg.*	*Med*	*Min*	*Max*	*Std*	*Avg.*	*Med*	*Min*	*Max*	*Std*	*Avg.*	*Med*	*Min*	*Max*	*Std*	*Avg.*	*Med*	*Min*	*Max*	*Std*
1	10.65	10.75	1.49	19.55	3.35	24.34	23.87	4.20	51.41	7.92	31.70	30.45	17.87	49.37	7.03	42.98	43.35	34.09	48.72	2.85	46.05	46.59	37.91	51.47	2.91
2	36.95	36.75	20.62	50.82	6.08	23.29	22.98	5.66	53.71	7.94	48.82	48.91	25.66	60.07	5.85	53.85	53.84	42.02	61.07	3.22	53.85	53.84	42.02	61.07	4.09
3	58.28	58.96	37.81	69.04	4.64	21.23	21.48	2.65	39.62	8.10	64.48	64.83	53.10	74.12	3.97	36.38	36.31	25.20	42.84	3.12	64.66	64.86	51.35	72.73	3.37
4	2.50	2.46	-2.60	6.84	1.65	16.02	15.55	5.99	26.67	4.54	18.93	18.63	9.26	28.46	4.53	21.09	21.06	15.44	27.09	2.08	22.54	22.26	16.80	29.25	2.16
5	16.13	15.93	7.54	25.32	3.65	14.99	15.84	0.47	29.83	5.41	28.49	28.34	12.83	38.11	4.43	16.75	16.92	9.71	22.17	2.44	27.25	27.06	19.62	34.33	3.23
6	28.35	28.68	19.51	37.46	4.12	13.03	13.59	2.93	23.05	4.11	36.36	36.06	26.30	45.04	4.07	12.44	12.44	5.11	17.46	2.39	33.72	33.71	24.57	42.18	3.60
7	14.12	13.70	5.92	24.67	4.19	27.27	27.24	12.16	46.35	6.88	35.45	35.31	21.15	51.25	6.36	38.36	38.39	27.95	45.97	2.99	43.54	43.49	34.92	52.03	3.20
8	44.39	44.89	27.50	56.75	5.42	25.31	25.08	7.66	44.08	7.02	55.14	55.61	42.36	64.84	4.86	34.91	34.75	29.40	41.75	2.55	55.83	55.97	44.63	65.48	3.76
9	63.01	63.42	52.58	74.53	4.34	24.52	24.37	6.54	41.55	7.20	68.68	69.09	58.60	76.83	3.67	31.11	31.08	23.43	37.35	3.09	66.99	67.42	56.56	75.85	3.54
10	3.47	3.40	-0.40	8.95	1.77	17.18	16.84	3.76	27.37	4.25	19.99	19.63	6.81	29.49	4.58	15.73	15.74	10.39	20.67	2.32	18.00	18.12	11.33	23.40	2.38
11	17.05	17.08	8.87	27.76	3.61	15.22	15.40	6.87	25.67	3.77	28.50	28.20	18.04	36.34	3.56	11.97	11.88	7.39	16.38	1.83	24.64	24.65	16.93	32.86	3.16
12	28.81	28.75	15.85	42.34	5.06	13.50	13.30	4.92	25.50	3.83	36.19	36.03	25.61	47.87	4.39	8.42	8.69	3.54	14.09	2.11	32.07	32.46	20.89	44.09	4.68

*Relative Difference in PMSE $= \frac{PMSE_{Alternative} - PMSE_{MERF}}{PMSE_{Alternative}} \times 100$

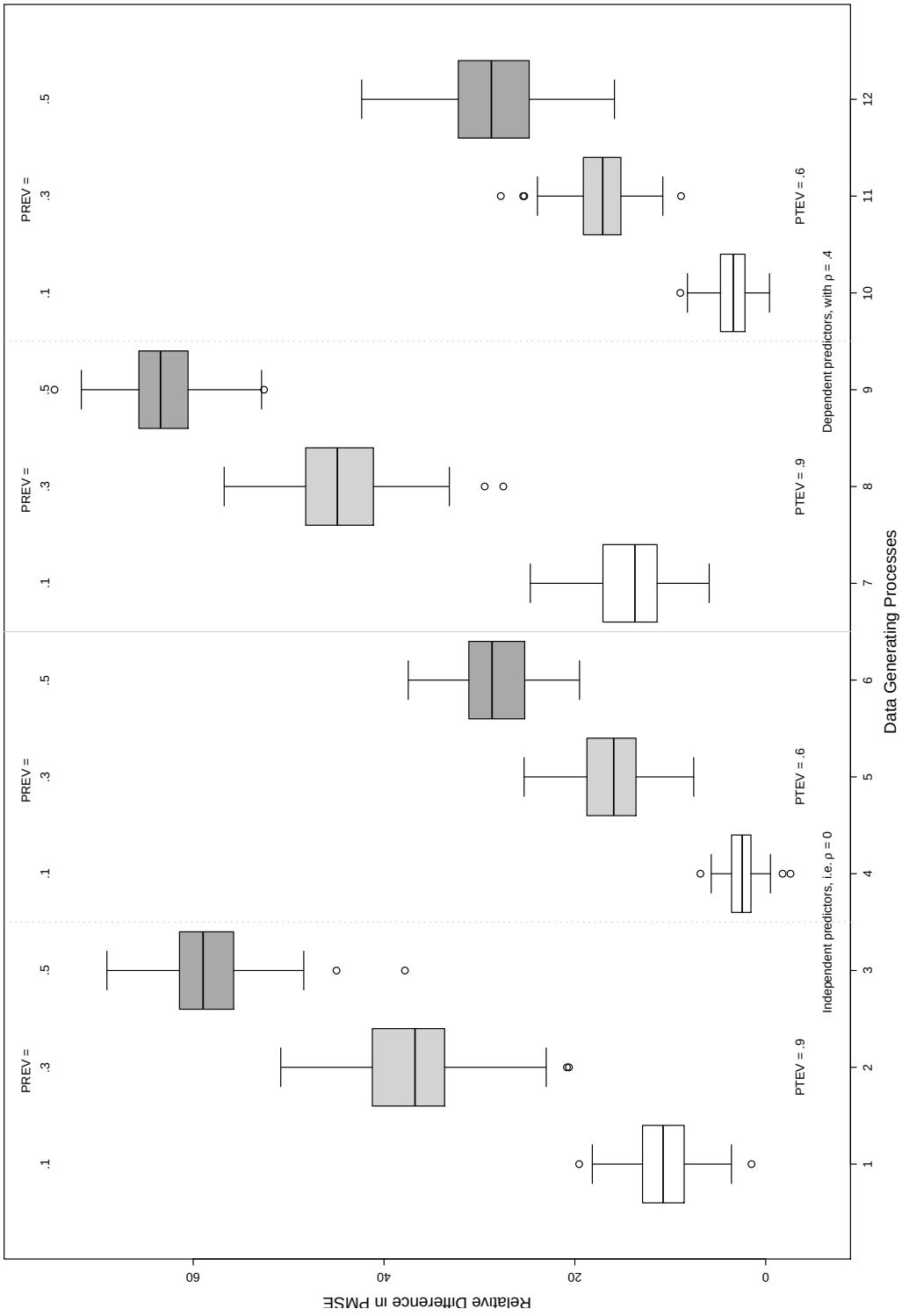

Figure 3.1 Distribution over the 100 simulation runs of the relative difference in PMSE between MERF and SRF

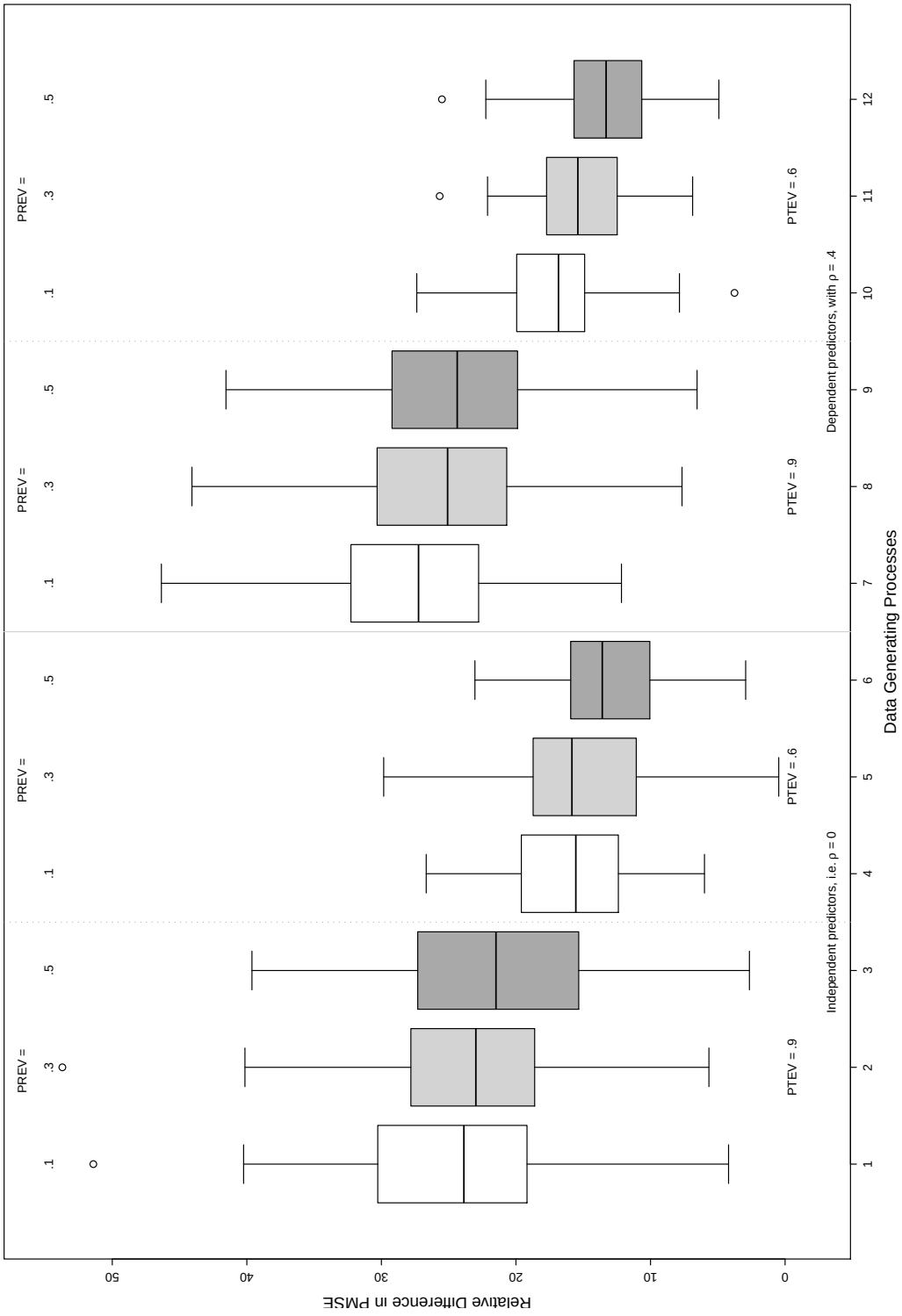

Figure 3.2 Distribution over the 100 simulation runs of the relative difference in PMSE between MERF and MERT

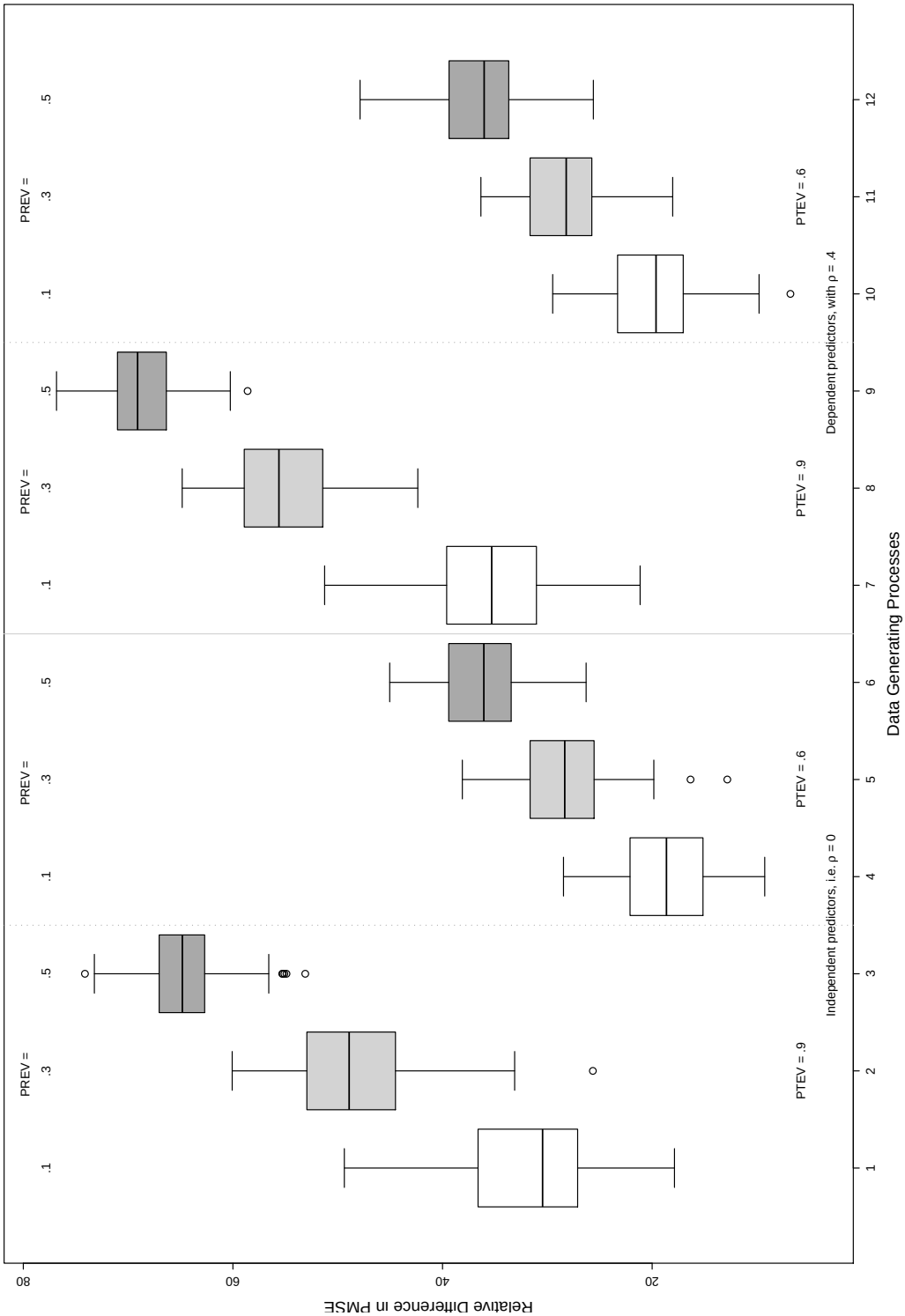

Figure 3.3 Distribution over the 100 simulation runs of the relative difference in PMSE between MERF and SRT

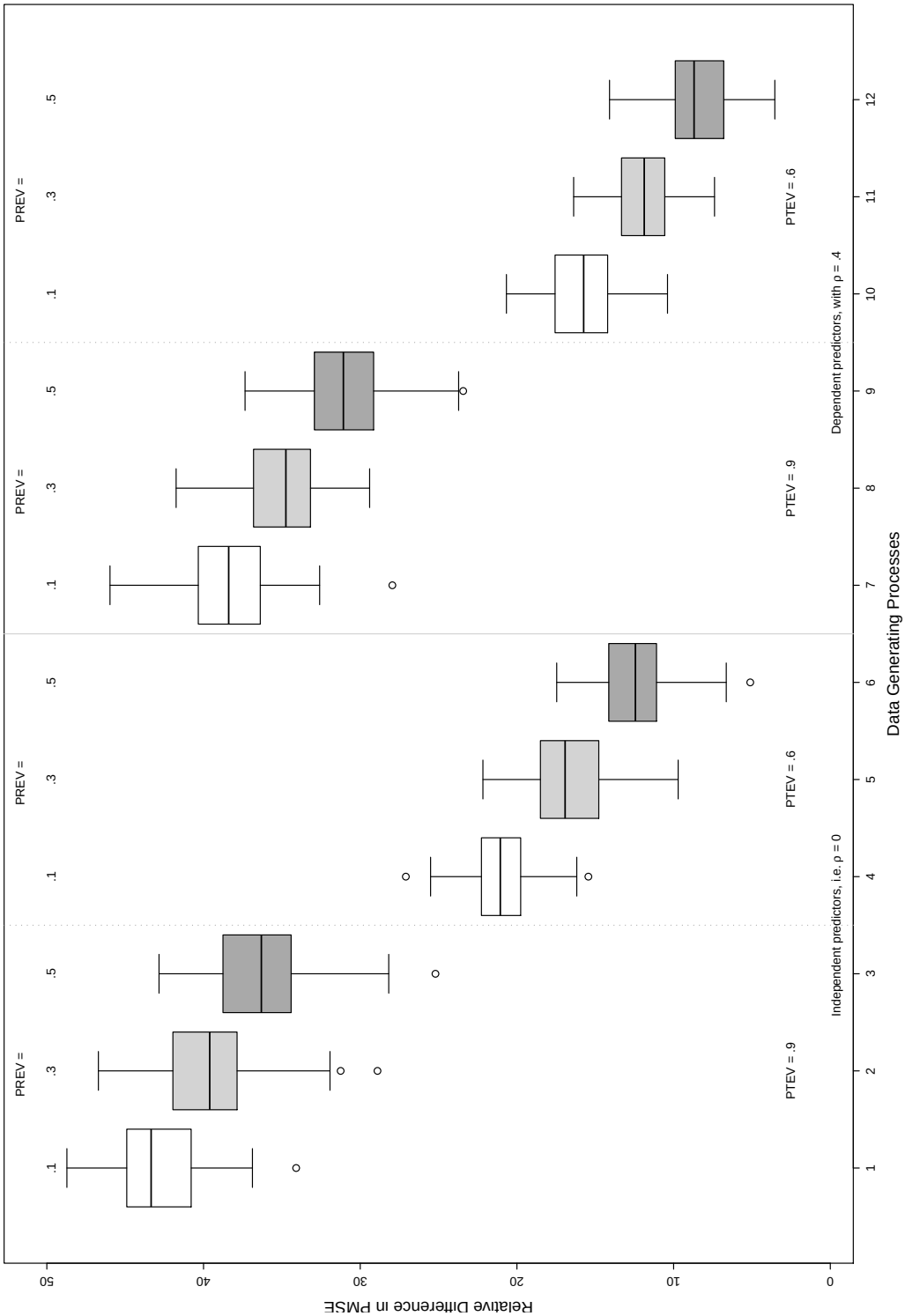

Figure 3.4 Distribution over the 100 simulation runs of the relative difference in PMSE between MERF and LME

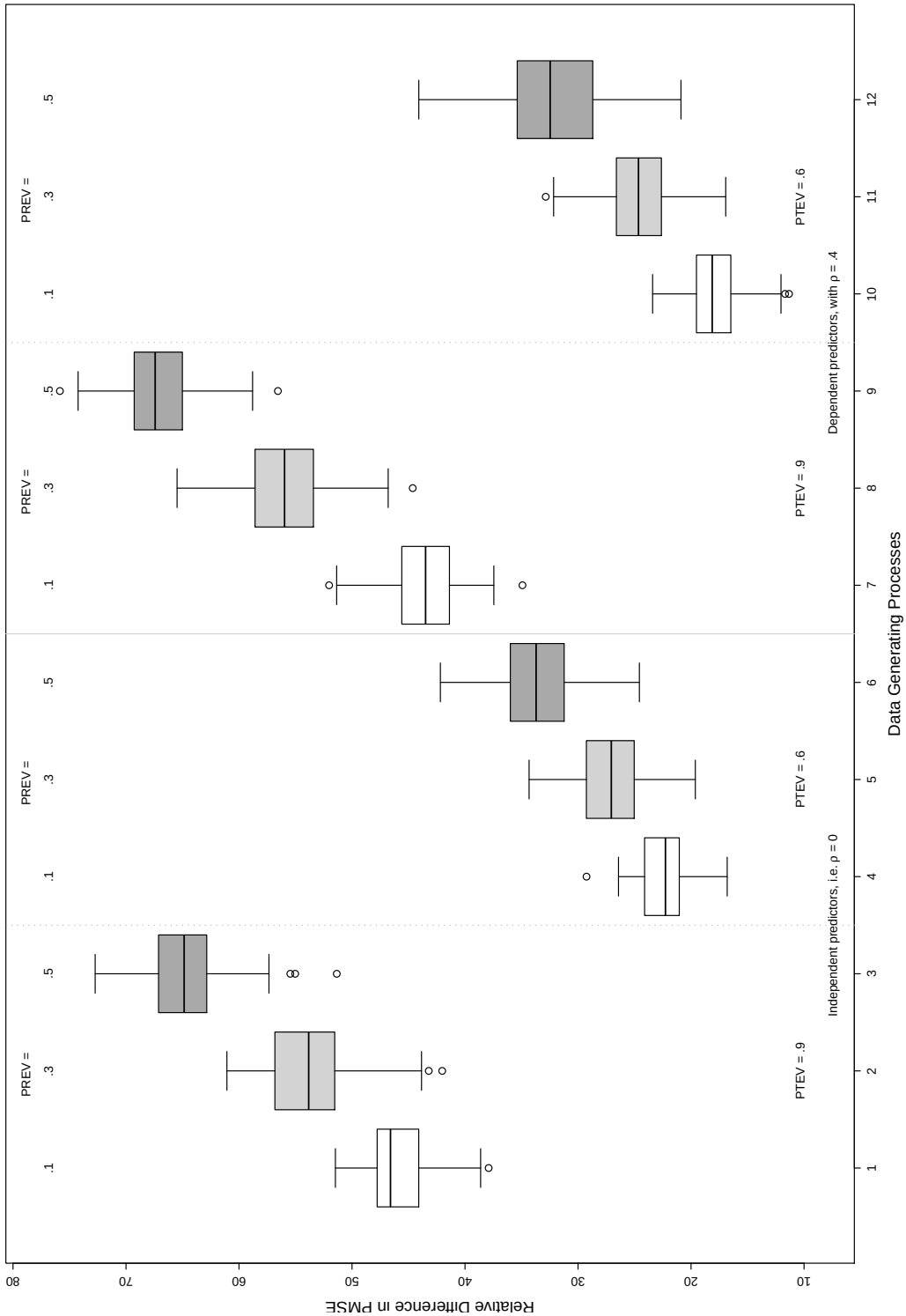

Figure 3.5 Distribution over the 100 simulation runs of the relative difference in PMSE between MERF and LM

CONCLUSION GÉNÉRALE

Dans cette thèse par articles, nous avons proposé une approche simple pour rendre plus appropriées les méthodes d'arbres et de forêts aléatoires standards lorqu'on veut les appliquer aux données hiérarchiques. Il s'agit des arbres et des forêts aléatoires à effets mixtes.

Dans le premier article, nous avons proposé la méthode d'arbre nommée "mixed effects regression tree" (MERT). Elle étend la méthode d'arbre de régression standard aux données hiérarchiques avec une réponse continue. Sur la base d'une étude de simulation, nous avons pu démontrer que ne pas tenir compte de la dépendance des données nuit à la capacité de l'algorithme standard d'identifier le vrai lien entre la réponse et les covariables. En tenant compte de cette dépendance, MERT réussit mieux ce défi. En modélisant la partie fixe de la variable réponse par une structure d'arbre, MERT a assouplit l'hypothèse de la linéarité de la partie fixe dans le modèle de régression linéaire à effets mixtes. En procédant toujours par une structure d'arbre, des travaux futurs pourraient tenter d'assouplir aussi l'hypothèse de la linéarité de la partie aléatoire, et/ou celle de l'additivité de ces deux parties.

Dans le deuxième article, nous avons proposé une méthode nommé "generalized mixed effects regression tree" (GMERT). Elle étend la méthode MERT à d'autres types de réponses (réponses binaires, données de comptage, réponses catégorielles ordonnées, réponses multi-catégorielles nominales). Tout comme le modèle linéaire généralisé (McCullagh and Nelder, 1989), le modèle GMERT transforme la réponse espérée en utilisant une fonction de lien appropriée selon le type de la variable réponse, et l'apparie à une fonction d'arbre des effets fixes en plus d'une fonction linéaire des effets aléatoires. Le modèle GMERT est par conséquent un modèle conditionnel qui génère des estimations conditionnelles et non pas des estimations marginales.

Dans le troisième article, nous avons proposé une méthode de forêt aléatoire à effets mixtes. Nous avons nommée cette méthode “mixed effects random forest” (MERF). Nous l’avons implémenté en utilisant une forêt d’arbres de régression standards à l’intérieur de l’algorithme EM. Plus précisément, à chaque itération de l’algorithme EM, les prédictions “out-of-bag” d’une forêt aléatoire standard sont utilisées pour estimer la partie fixe de la variable réponse mesurée sur une échelle continue. Il serait certainement utile d’étendre MERF à d’autres types de réponses. Une idée simple serait de remplacer l’arbre de régression standard pondéré, utilisé pour estimer la partie fixe de la pseudo-réponse dans l’algorithme doublement itératif GMERT, par une forêt d’arbres de régression standards pondérés. Il serait aussi intéressant de comparer GMERT à une approche alternative qui consisterait dans la construction d’une forêt aléatoire d’arbres MERT en utilisant des stratégies de rééchantillonage appropriées pour des données hiérarchiques, comme par exemple un rééchantillonnage au niveau groupe, ou un rééchantillonnage en deux étapes, c.à.d, un rééchantillonnage au niveau groupe suivi d’un rééchantillonnage au niveau observation à l’intérieur des groupes déjà échantillonnés.

Les extensions antérieures des méthodes d’arbres standards aux données corrélées (Segal, 1992 ; Zhang, 1998 ; Abdolell, Leblanc, Stephens, and Harrison, 2002 ; Lee, 2005) ne permettent pas que les covariables du niveau observation entrent comme candidates dans le processus d’embranchement, et par conséquent, 1) aucun effet aléatoire ou spécifique au groupe n’est modélisable, et 2) toutes les observations provenant d’un même sujet restent ensemble tout au long de ce processus et ne peuvent pas être séparées dans des noeuds différents. La méthode d’arbre à effets mixtes que nous avons proposé traite de façon appropriée les effets aléatoires potentiels des covariables du niveau observation. En plus, ces dernières sont candidates dans le processus d’embranchement de l’arbre à effets mixtes. Par conséquent, les observations intra-groupe pourraient être séparées dans des noeuds différents. Toutefois, l’arbre à effets mixtes suppose que la corrélation découle uniquement de la variation inter-groupes. Il serait donc utile de la généraliser afin de permettre la modélisation de structures alternatives de covariances intra-groupe. Une idée à investiguer serait de rempla-

cer l'algorithme EM utilisé jusqu'ici par l'algorithme EM hybride de Jennrich et Schluchter (1986).

Notre implémentation de l'arbre à effets mixtes fait en sorte que tous les avantages de l'arbre standard comparativement aux modèles de régression paramétriques s'étendent naturellement aux arbres à effets mixtes. Par exemple, tout comme les arbres standards, les arbres à effets mixtes sont des modèles qui peuvent être représentés graphiquement et qui sont facilement interprétables. Il faut néanmoins rester prudent et investiguer la robustesse de cette approche lorsque certaines de ses hypothèses ne sont pas vérifiées, soient la non linéarité de la partie aléatoire, la non additivité de la partie fixe et aléatoire, la non normalité des erreurs, et/ou la présence d'une corrélation induite à la fois par la variation inter- et intra-groupes.

BIBLIOGRAPHIE

Abdollel, M., LeBlanc, M., Stephens, D. and Harrison, R. V. (2002). Binary partitioning for continuous longitudinal data : Categorizing a prognostic variable. *Statistics in Medicine,* 21, 3395-3409.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees.* Wadsworth International Group. Belmont, California.

Breiman, L. (1996). Bagging predictors. *Machine Learning,* 24, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning,* 45, 5-32.

Breslow, N.and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association,* 88, 9-25.

Bryk, A. S., and Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin,* 101, 147-158.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Mixed Effects Models for Repeated Measurement Data.* Chapman and Hall.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,* 39, 1-38.

Field, C. A. and Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society, Series B,* 69, 369-390.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied longitidunal analysis.* New York : Wiley.

Ghattas, B., and Nerini, D. (2007). Classifying densities using functional regression trees : Applications in oceanology. *Computational Statistics & Data Analysis,* 51, 4984-4993.

Goldstein, H. (2003). *Multilevel statistical models (3rd Edition).* Arnold, London.

Goldstein, H. and Rasbash, J. R. (1996). Improved Approximations for Multilevel Models with Binary Responses. *Journal of the Royal Statistical Society,* 159, 505-513.

Harville, D. A. (1976). Extension of the Gauss-Markov theorem to include the

estimation of random effects. *Annals of Statistics,* 4, 384-395.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association,* 72, 320-38.

Jennrich, R. I., and Schluchter, M. D. (1986). Unbalanced Repeated-Measures with Structured Covariance Matrices. *Biometrics,* 42, 805-820.

Kuhnert, P. M., Do, K.-A., and McClure, R. (2000). Combining nonparametric models with logistic regression : An application to motor vehicle injury data. *Computational Statistics and Data Analysis,* 34, 371-386.

Laird, N. M. and Ware J. H. (1982). Random-effects models for longitudinal data. *Biometrics,* 38, 963-974.

Lee, S. K. (2005). On Generalized multivariate decision tree by using GEE. *Computational Statistics & Data Analysis,* 49, 1105-1119.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News,* 2, 18-22.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models (2nd Edition).* Chapman & Hall/CRC. London.

McLachlan G. J. and Krishman T. (1997). *The EM algorithm and extensions.* Wiley. New York.

R Development Team (2007). *R : A Language and environment for statistical computing.* R Foundation for Statistical Computing : www.R-project.org.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical linear models : Applications and data analysis method (2nd Edition).* Sage. Newbury Park, CA.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., and du Toit, M. (2004). *HLM 6 : Hierarchical Linear & Nonlinear Modeling.* Scientific Software International, Inc.

SAS Institute Inc. (2008). *SAS/STAT 9.2 User's Guide : The GLIMMIX Procedure (Book Excerpt).* Cary, NC : SAS Institute Inc.

Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association,* 87, 407-418.

Simonoff, J. S. and Sparrow, I. R. (2000). Predicting movie grosses : Winners and losers, blockbusters and sleepers. *Chance,* 13(3), 15-24.

Therneau, T. M. and Atkinson, E. J. (1997). *An introduction to recursive par-*

titioning using the rpart routines. Technical Report 61, Department of Health Science Research, Mayo Clinic, Rochester.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S (Fourth edition).* New York : Springer-Verlag.

Wu, H. and Zhang, J. T. (2006). *Nonparametric regression methods for longitidunal data analysis : Mixed-effects modeling approaches.* Wiley. New York.

Yu, Y. and Lambert, D. (1999). Fitting Trees to Functional Data : With an Application to Time-of-day Patterns. *Journal of Computational and Graphical Statistics,* 8, 749-762.

Zhang, H., (1997). Multivariate Adaptive Splines for Analysis of Longitudinal Data. *Journal of Computational and Graphical Statistics,* 6, 74 - 91.

Zhang, H., (1998). Classification trees for multiple binary responses. *Journal of the American Statistical Association,* 93, 180-193.

Zhang, D. and Davidian, M. (2004). Likelihood and conditional likelihood inference for generalized additive mixed models for clustered data. *Journal of Multivariate Analysis,* 91, 90-106.