Random effects regression trees for the analysis of INVALSI data

Giulia Vannucci, Anna Gottard, Leonardo Grilli, Carla Rampichini

1. Introduction

Multilevel data structures, where data are typically clustered in nested levels, are common in many fields. An emblematic example consists of students, that are grouped in classes and schools (individual cross-sectional data) or children growth evaluated at several time points (repeated measures). Multilevel data require specific models referred to as *multilevel*, *random effects* or *mixed* (Snijders and Bosker, 2012).

Model specification is a challenging task in mixed models. Typically, a linear model is assumed, although non-linearities and interaction effects are undeniably of interest. A worthwhile approach exploits regression trees and the CART algorithm (Breiman et al., 1984) to capture non-linearities and high-order interaction effects. In particular, regression trees are a statistical learning algorithm that shapes the regression function as piece-wise constant over a recursively found partition of the covariate space. The graphical display of the recursive partition provides an easy interpretation of this predictive algorithm. The procedure, however, assumes statistical units to be independent, which is not the case of clustered data.

Regression trees have been extended to clustered data by Hajjem et al. (2011), who proposed to model fixed effects with a decision tree while accounting for random effects via a linear mixed model in a separate, subsequent, step. In particular, they first apply the CART algorithm as if data were not clustered to estimate the fixed effects. It is shown that random effect regression trees are less sensitive to parametric assumptions and provide improved predictive power compared to linear models with random effects and regression trees without random effects. The literature has thereon grown with variants and extensions. Among others, see Sela (2012); Hajjem et al. (2014); Miller et al. (2017).

In this work, we propose a further variation of the mixed effects regression tree, where the fixed and the random part parameters are estimated jointly, using a backfitting algorithm. To ease the interpretation, our proposal incorporates a linear component additively to the regression trees. Consequently, the general trend of dependence is captured by the linear component, while the tree part captures interactions and non-linearities.

The proposed algorithm is then applied to data collected by the national institute for the evaluation of the educational system and training (INVALSI: Istituto Nazionale per la VALutazione del Sistema educativo di Istruzione e di formazione) in Italy. The study aims to compare schools' educational effectiveness impartially by measuring students' progress over their careers. We focus on test scores in Mathematics, given some characteristics of the school and the pupil. The proposed model is able to take into account the student clustering in schools and to capture interesting interactions between student-level covariates and school-level covariates.

The rest of the paper is organised as follows. Section 2 illustrates the model proposed, together with the backfitting algorithm. Section 3 describes the application of the proposal to INVALSI data. A brief section of final remarks concludes the paper.

Giulia Vannucci, University of Florence, Italy, giulia.vannucci@unifi.it, 0000-0003-3569-6274
Anna Gottard, University of Florence, Italy, anna.gottard@unifi.it, 0000-0002-8246-4962
Leonardo Grilli, University of Florence, Italy, leonardo.grilli@unifi.it, 0000-0002-3886-7705
Carla Rampichini, University of Florence, Italy, carla.rampichini@unifi.it, 0000-0002-8519-083X

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giulia Vannucci, Anna Gottard, Leonardo Grilli, Carla Rampichini, *Random effects regression trees for the analysis of INVALSI data*, pp. 29-34, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-304-8.07, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci, *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the opening conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-304-8 (PDF), DOI 10.36253/978-88-5518-304-8

2. A tree embedded linear mixed model

We propose a random effect model, called *Tree Embedded Linear Mixed* (TELM) *model*, able to treat both non-linear and interaction effects and cluster mean dependencies. Motivated by the application of interest, we consider in particular a two-level random effect model. Hence, we will denote as *level 1 units* the statistical units (e.g. students) and *level 2 units* the groups (e.g. schools).

The model is a piecewise-linear regression function, consisting of the sum of a tree component and a mixed effect linear component. The proposal is the mixed effect version of the semi-linear regression trees (Vannucci, 2019). It can be ideally divided into three parts: a fixed effect linear part, a fixed effect non-linear part based on a tree and a random effect part. The resulting model can be formulated as

$$Y_{ij} = \beta_0 + \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\gamma} + T(\mathbf{X}_{ij}, \mathbf{Z}_j) + U_j + \epsilon_{ij} \tag{1}$$

where Y_{ij} is the response variable for level 1 unit i belonging to level 2 unit j, β_0 is the (fixed-effect) regression intercept, $\mathbf{X}_{ij}$ is the vector of the level 1 covariates, $\boldsymbol{\beta}$ the associated fixed effect coefficients, $\mathbf{Z}_j$ is the vector of the level 2 covariates, $\boldsymbol{\gamma}$ the associated fixed effect coefficients. Here, $T(\mathbf{X}_{ij}, \mathbf{Z}_j)$ is the tree based component depending on some or all the level 1 and the level 2 explanatory variables. Finally, $U_j \sim N(0, \sigma_u^2)$ is the random intercept for level 2 unit j and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ are the regression errors.

The model is additive in its components where the tree-component acts as a region-specific categorical variable. This can be seen in the following alternative specification

$$Y_{ij} = \beta_0 + \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\gamma} + \sum_{m=1}^{M} \mu_m \mathbb{I}\{(\mathbf{X}_{ij}, \mathbf{Z}_j) \in R_m\} + U_j + \epsilon_{ij}, \tag{2}$$

where $R_1, \ldots, R_M$ is the partition of the predictor space corresponding to the tree-component. When the unknown regression function can be assumed to be quasi-linear (Wermuth and Cox, 1998), the number of leaf nodes M can be kept small to avoid overfitting.

To account for the contextual effects of level 1 predictors, we add the cluster mean $\overline{W}_j = (1/n_j)\sum_{i=1}^{n_j} W_{ij}$ to the set of level 2 predictors $\mathbf{Z}_j$ (Snijders and Bosker, 2012).

An iterative, backfitting-like procedure obtains model fitting. First, the tree is initialised at the mean of the response variable and the partial residuals Y^* are computed by subtracting to Y the tree prediction. Secondly, a linear random intercept effect model is fitted on Y^* and explanatory variables at the individual and group level. The corresponding partial residuals Y^{**} are obtained by subtracting to Y model predictions. These partial residuals Y^{**} are employed in the next step to fit a new tree, using the CART algorithm (Breiman et al., 1984) with a short depth. We iterate alternating the two fitting steps until convergence is reached. At the end of the procedure, model (2) is fitted by a linear random effect model using the partition associated with the tree selected at convergence. The leaf node parameters μ_m are estimated jointly with the other model parameters β_0, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, σ_u^2, σ_ε^2.

The main difference of our procedure with respect to previous proposals (Hajjem et al., 2011; Sela, 2012), is the inclusion of the linear component $\mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\gamma}$ in the random effect model (2). In the presence of quasi-linear relationships, this inclusion allows us to avoid overfitting and helps interpretation. Moreover, since the μ_m are jointly estimated in the final step, standard hypothesis tests and confidence intervals can be used for model selection and evaluation, together with the mean squared error computed on a test data set for prediction accuracy evaluation.

3. Application: Invalsi tests in Italian schools

We apply the TELM model outlined in the previous section to data on students' achievement collected by INVALSI. The Institute yearly carries out standardised tests to assess students' achievement in mathematics and reading and evaluate the overall quality of the educational offering of schools and vocational training institutes. See Arpino et al. (2019) for a discussion on this set of data.

As an illustration, we are here focusing on data on students who participated in the Maths tests at 5^{th} and 8^{th} grades. Specifically, the dataset is obtained by linking data on students who attended the 5^{th} grade in 2013-2014 with data on students who attended the 8^{th} grade in 2016-2017. The number of students who participated on both occasions of the Maths test is $409\,528$. They are grouped into $5\,773$ schools. We aim to predict the Maths test score, while understanding which of the included variables may be associate to the final score. Table 1 lists the considered explanatory variables. As shown in the table, we include both student level and school-level covariates, denoted in (1) as $\mathbf{X}_{ij}$ and $\mathbf{Z}_j$ respectively. Among the school level variables, we consider, in addition, the average of 5^{th} grade Maths test and the average of the Socio-economic status index for each school. We are denoting these variables CM_MATH5 and CM_SES.

Table 1: Student and school level variables (INVALSI data years 2014 and 2017).

Student level variables (level one)	
MATH8	(*Response*) Test score at the 8^{th} grade (0-100)
MATH5	Test score at the 5^{th} grade (0-100)
SES	Socio-economic status
FEMALE	$1 =$ Yes, $0 =$ No
ENROLLED	School enrolment ($1 =$ Regualrly enrolled, $2 =$ Enrolled one year in advance, $3 =$ Enrolled one year later)
IMM	Citizenship ($0 =$Italian, $1 = 1^{\text{st}}$ generation immigrant, $2 = 2^{\text{nd}}$ generation immigrant)
School level variables (level two)	
AREA	Geographical area ($0 =$ NE, 5 categories)
TOWN	Provincial capital
CLSIZE	Average num of students per class
SCSIZE	Number of classes in the school
PUBLIC	Type of school ($0 =$ Private, $1 =$ Public)

The proposed model takes into account both linear and non-linear effects and can detect the presence of both within level and cross-level interaction effects. In particular, the tree component $T(\mathbf{X}_{ij}, \mathbf{Z}_j)$ in (1) is modelling non-linearities and interactions at once via a piece-wise linear function. Estimates for model parameters are reported in Table 2, while the tree component is also illustrated in Figure 1. The two terminal nodes without label in the plot have been automatically set in the reference category.

Individual and school level covariates not selected by the algorithm in the tree component have the usual interpretation. For example, controlling for the model covariates, females have, on average, around 1.5 points less than males in the score of math at the $8th$ grade.

Besides the usual interpretation of the coefficients of the linear components, it seems here interesting to focus on the covariates selected in the tree component of the model, namely the math score at grade 5 (MATH5) and the geographical area of the school (AREA). In particular, the tree component algorithm splits the values of MATH5 into three intervals: below 33 (2% of the observations), between 33 and 72 (55%) and above 72 (43%). Moreover, the algorithm

Table 2: TELM model fitted on INVALSI data: parameter estimates, standard errors and t-test.

	Estimate	Std. Error	t value
Student level			
(Intercept)	31.0733	0.9518	32.6462
MATH5	0.6263	0.0027	232.3463
SES	2.5246	0.0270	93.5909
FEMALE	-1.5021	0.0467	-32.1616
ENROLLED_2	1.8029	0.2053	8.7802
ENROLLED_3	-3.5558	0.2027	-17.5432
IMM_1	-1.1107	0.1779	-6.2434
IMM_2	-1.4869	0.1127	-13.1934
School level			
CM_MATH5	-0.2765	0.0131	-21.1820
CM_SES	1.2876	0.2260	5.6971
AREA_2 (NW)	0.4675	0.2574	1.8160
AREA_3 (Centre)	-1.9239	0.2562	-7.5080
AREA_4 (South)	8.1865	0.3559	22.9993
AREA_5 (Islands)	8.5293	0.3612	23.6133
CLSIZE	0.1629	0.0088	18.6063
SCSIZE	0.0613	0.0394	1.5562
PUBLIC	-2.1495	0.3773	-5.6971
TOWN	0.0275	0.1981	0.1388
Tree nodes			
N1: $33 \leq$ MATH5< 73 & AREA$= 4, 5$	-7.1138	0.2465	-28.8633
N2: MATH5 ≥ 73 & AREA$= 4, 5$	-11.9902	0.2806	-42.7304
N3: MATH5 ≥ 73 & AREA$= 1, 2, 3$	4.4819	0.0903	49.6540
N4: MATH5 < 33 & AREA$= 1, 2, 3$	1.9711	0.2691	7.3250
Residual variances			
School level (Intercept)		35.75	
Student level		218.35	
Number of students: 409528	Number of schools: 5773		

splits the schools into two groups depending on AREA: schools placed in North or Center Italy, and schools placed in South Italy and Islands. Thus, the algorithm suggests the presence of an interaction effect between these two variables, with the effect of AREA depending on the interval of MATH5 and vice versa. For example, for a pupil living in a region of NW of Italy, the expected difference with respect to a pupil with same characteristics living in the NE of Italy (baseline) is 2.4386 if MATH5< 33, it decreases to 0.4675 if $33 <$MATH5< 73, and it rises up to 4.9494 if MATH5 ≥ 73.

Note that the ordinary mixed effect regression model, whose parameter estimates are reported in Table 3, is nested with the TELM model. The Likelihood Ratio test comparing these two models obtains a test statistic equal to 10168, with 4 degrees of freedom, in favour of the TELM model. The variation between the estimates in the two models is due to the inclusion of the tree component, that relaxes the assumption of linearity and includes interaction effects. An interesting variation concerns the AREA coefficients estimates. Ignoring the AREA and MATH5 interaction, and the MATH5 non-linearity, completely reverse the main effect of AREA for South and Islands.

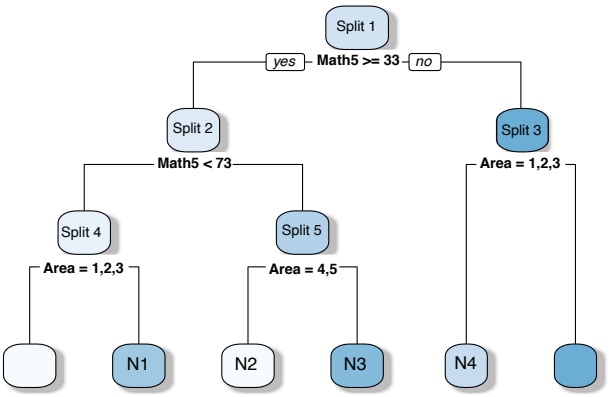

Figure 1: Graphical representation of the tree component of TELM model in Table 2 (nodes with a label correspond to a parameter in the model; the proportions of level 1 observations at each node are: left white node 0.35, N1 0.20, N2 0.17, N3 0.26, N4 0.01, right blue node 0.01)

Table 3: Random intercept model fitted on INVALSI data: parameter estimates, standard errors and t-test.

	Estimate	Std. Error	t value
Student level			
(Intercept)	36.2031	0.9438	38.3569
MATH5	0.6328	0.0015	412.2164
SES	2.5509	0.0273	93.3900
FEMALE	-1.6821	0.0473	-35.5926
ENROLLED_2	1.5694	0.2079	7.5486
ENROLLED_3	-3.5092	0.2052	-17.1025
IMM_1	-1.5243	0.1801	-8.4648
IMM_2	-1.8652	0.1140	-16.3553
School level			
CM_MATH5	-0.3240	0.0131	-24.8104
CM_SES	1.2809	0.2262	5.6620
AREA_2 (NW)	0.4945	0.2575	1.9203
AREA_3 (centre)	-1.9364	0.2564	-7.5529
AREA_4 (south)	-2.7964	0.2611	-10.7116
AREA_5 (islands)	-2.3695	0.2694	-8.7939
CLSIZE	0.1486	0.0089	16.7787
SCSIZE	0.0572	0.0394	1.4505
PUBLIC	-2.1165	0.3779	-5.6012
TOWN	0.1113	0.1982	0.5613
Residual variances			
School level (Intercept)		35.58	
Student level		223.91	
Number of students: 409528		Number of schools: 5773	

4. Conclusions

Tree Embedded Linear Mixed (TELM) models extend random effect models by including both a linear component and tree component in the regression function. The proposal increases the flexibility and the predictive ability of ordinary random effects models by handling simultaneously linear and non-linear associations and interactions.

A TELM model has the following characteristics: (1) it can handle clusters with different numbers of observations (unbalanced clusters); (2) it allows the inclusion of level 1 and level 2 covariates in the splitting process; (3) it allows observation-level covariates to have random effects. Besides, our proposal extends random effect regression trees in two directions: (i) incorporating a linear component in the final random effect model, and (ii) allowing to take into account contextual effects of level 1 covariates.

The application on INVALSI data is an illustrative example of TELM models that shows how the inclusion of a tree component helps highlight cross-level interactions.

References

Arpino, B., Bacci, S., Grilli, L., Guetto, R., Rampichini, C. (2019) Issues in prior achievement adjustment for value added analysis: an application to Invalsi tests in Italian schools. pp.17-20. In *ASA Conference 2019. Statistics for Health and Well-Being. Book of Short Papers* - ISBN:978-88-5495-135-8

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, (CA).

Hajjem, A., Bellavance, F., Larocque, D. (2011). Mixed Effects Regression Trees for Clustered Data. *Statistics and Probability Letters*, **81**, pp. 451–459.

Hajjem, A., Bellavance, F., Larocque, D. (2014). Mixed-effects Random Forest for Clustered Data. *Journal of Statistical Computation and Simulation*, **84**, pp. 1313–1328.

Miller, P.J., McArtor, D.B., Lubke, G.H (2017). METBOOST: Exploratory Regression Analysis with Hierarchically Clustered Data. arXiv:1702.03994v1 [stat.ML]

Sela, R.J., Simonoff, J.S. (2012). RE-EM Trees: A Data Mining Approach for Longitudinal and Clustered Data. *Machine Learning*, **86**, pp. 169–207.

Snijders, T.A.B., Bosker, R.J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling (2nd ed.)*. SAGE Publications Ltd.

Vannucci, G. (2019). Interpretable Semilinear Regression Trees. PhD Thesis, FLOrence REsearch repository.

Wermuth, N., Cox, D.R. (1998). On Association Models defined over Independence Graphs. *Bernoulli Society for Mathematical Statistics and Probability*, **4**, pp. 477–495.