The Tree Based Linear Regression Model for Hierarchical Categorical Variables

Carrizosa, Emilio ; Mortensen, Laust Hvas; Romero Morales, Dolores ; Sillero-Denamiel, M. Remedios

Document Version
Final published version

Published in:
Expert Systems with Applications

DOI:
10.1016/j.eswa.2022.117423

Publication date:
2022

Expert Systems With Applications 203 (2022) 117423

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

The tree based linear regression model for hierarchical categorical variables

Emilio Carrizosa [a,b], Laust Hvas Mortensen [c,d], Dolores Romero Morales [e], M. Remedios Sillero-Denamiel [f,b,*]

[a] Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, Seville, Spain
[b] IMUS, Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain
[c] Statistics Denmark, Copenhagen, Denmark
[d] Department of Public Health, University of Copenhagen, Copenhagen, Denmark
[e] Department of Economics, Copenhagen Business School, Frederiksberg, Denmark
[f] School of Computer Science & Statistics, Trinity College Dublin (TCD), Dublin, Ireland

ARTICLE INFO

Keywords:
Hierarchical categorical variables
Linear regression models
Accuracy vs. model complexity
Mixed integer convex quadratic problem with linear constraints

ABSTRACT

Many real-life applications consider nominal categorical predictor variables that have a hierarchical structure, e.g. economic activity data in Official Statistics. In this paper, we focus on linear regression models built in the presence of this type of nominal categorical predictor variables, and study the consolidation of their categories to have a better tradeoff between interpretability and fit of the model to the data. We propose the so-called Tree based Linear Regression (TLR) model that optimizes both the accuracy of the reduced linear regression model and its complexity, measured as a cost function of the level of granularity of the representation of the hierarchical categorical variables. We show that finding non-dominated outcomes for this problem boils down to solving Mixed Integer Convex Quadratic Problems with Linear Constraints, and small to medium size instances can be tackled using off-the-shelf solvers. We illustrate our approach in two real-world datasets, as well as a synthetic one, where our methodology finds a much less complex model with a very mild worsening of the accuracy.

1. Introduction

Categorical variables are increasingly present in a number of real-world applications. For example, in the healthcare field, data may contain high-cardinality categorical variables describing diagnoses and prescriptions (Jensen, Jensen, & Brunak, 2012). They may also appear in social and economic studies (Johannemann, Hadad, Athey, & Wager, 2020; Pauger & Wagner, 2019) or in Natural Language Processing (Mikolov, Chen, Corrado, & Dean, 2013), to name a few. Interpreting and visualizing information extracted from complex data is at the core of Data Science (Bertsimas, O'Hair, Relyea, & Silberholz, 2016; Carrizosa, Guerrero, & Romero Morales, 2018; Fang, Liu Sheng, & Goes, 2013; Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2017; Martens, Baesens, Gestel, & Vanthienen, 2007; Ustun & Rudin, 2016), and this is also the case for categorical variables where the information may be disaggregated across many categories. Mathematical Optimization is an important tool to build, in an efficient manner, data analysis models that can achieve a high accuracy (Bottou, Curtis, & Nocedal, 2018; Carrizosa & Romero Morales, 2013; Fountoulakis & Gondzio, 2016; Fu, Golden, Lele, Raghavan, & Wasil, 2003; Goodfellow, Bengio, & Courville, 2016), while being able to incorporate desirable properties, such as being parsimonious (Benítez-Peña, Blanquero, Carrizosa, & Ramírez-Cobo, 2019; Bertsimas & King, 2016; Bertsimas, Pauphilet, & Parys, 2020; Bertsimas & Van Parys, 2020; Blanquero, Carrizosa, Jiménez-Cordero, & Martín-Barragán, 2019; Carrizosa, Guerrero, & Romero Morales, 2020; Lin, Zhong, Hu, Rudin, & Seltzer, 2020), or tackling multiple objectives, such as the *bias–variance tradeoff* (Hastie, Tibshirani, & Friedman, 2009).

In the linear regression setting, to enhance the interpretability of the model and reduce the risk of overfitting in the presence of high-cardinality categorical variables, some works have fused categories, i.e., they are forced to share the same estimated coefficient, see Carrizosa, Galvis Restrepo, and Romero Morales (2021) and Stokell, Shah, and Tibshirani (2021) and references therein. In this paper, we are interested in the fusion of categories for a special case, those variables that have a hierarchical structure in their categories.

This kind of variable appears in different fields of research, such as nested spatial data in Spatial Statistics (Gotway & Young, 2002), as

* Corresponding author at: School of Computer Science & Statistics, Trinity College Dublin (TCD), Dublin, Ireland.
E-mail addresses: ecarrizosa@us.es (E. Carrizosa), LHM@dst.dk (L.H. Mortensen), drm.eco@cbs.dk (D. Romero Morales), sillerom@tcd.ie (M.R. Sillero-Denamiel).

https://doi.org/10.1016/j.eswa.2022.117423
Received 11 June 2021; Received in revised form 24 January 2022; Accepted 25 April 2022
Available online 4 May 2022

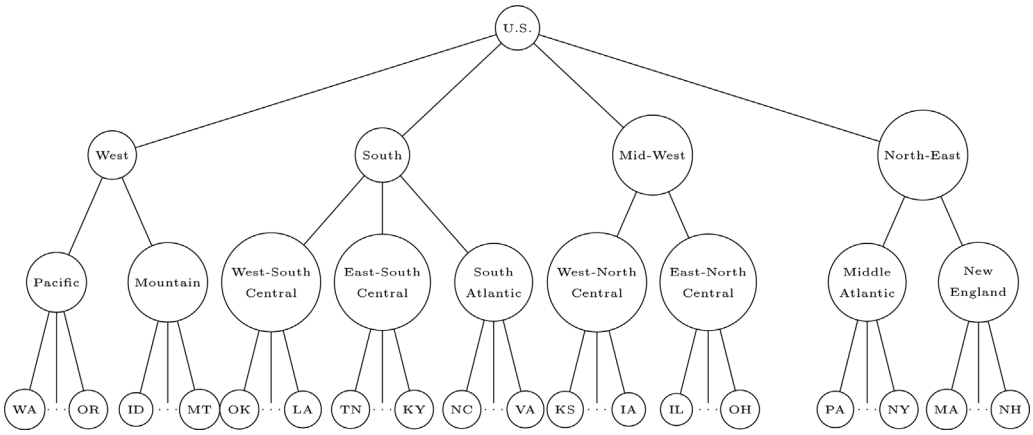


Fig. 1. Tree representation of the variable *geography* in the `cancer-reg` dataset.

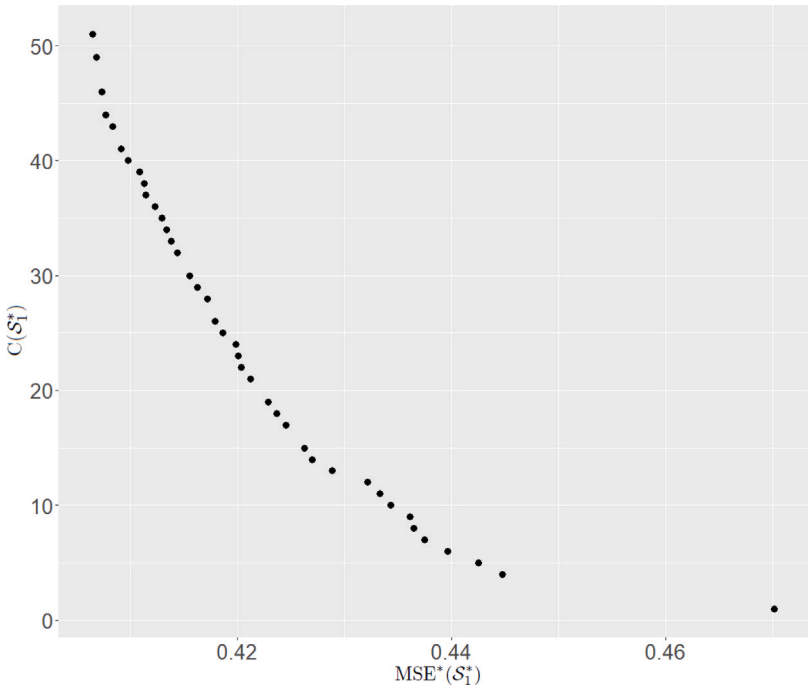
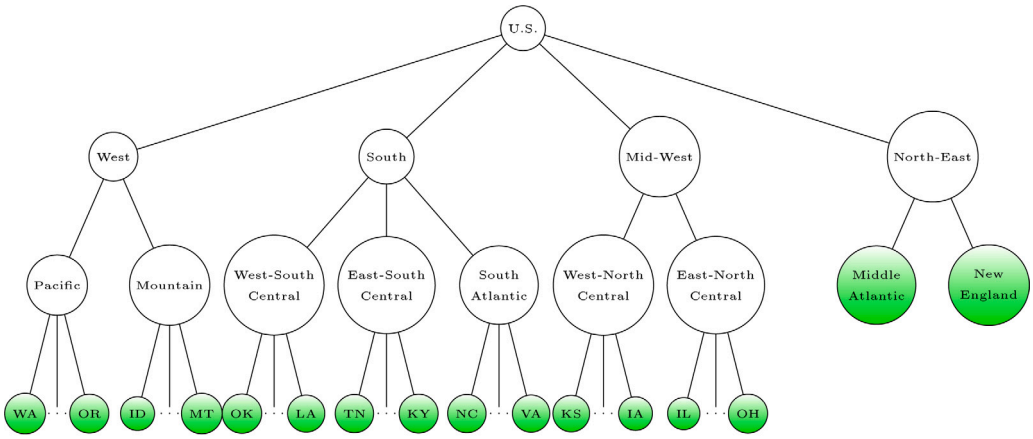


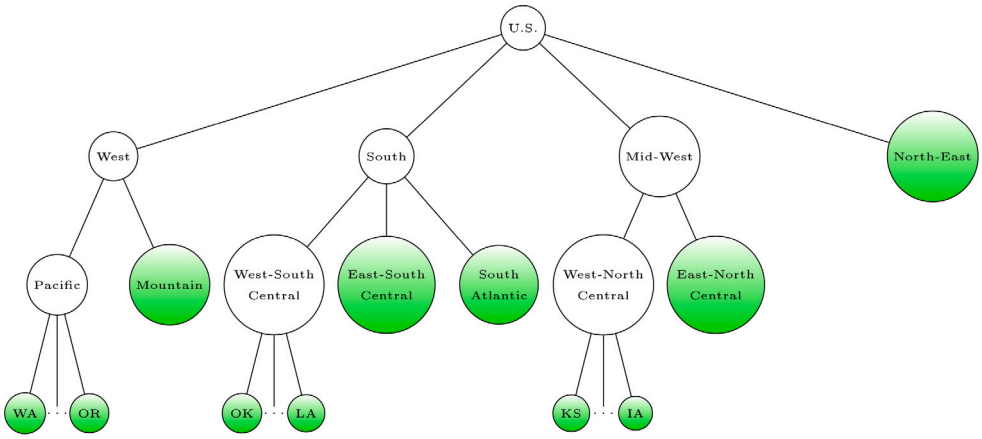
Fig. 2. Pareto frontier for MSE versus the number of coefficients to be estimated in the reduced model for the hierarchical categorical variable *geography* in the `cancer-reg` dataset.

for example the European Union with the NUTS classification (nomenclature of territorial units for statistics), where the small regions for specific diagnoses are consolidated at basic regions for the application of regional policies and these, in turn, are consolidated at major socio-economic regions. They also appear in behavioural data in Retail Business Analytics (Griva, Bardaki, Pramatari, & Papakiriakopoulos, 2018), since each retailer chain maintains a product hierarchy, which is necessary to conduct business processes such as store replenishment. Economic activity data in Official Statistics (European Commission, 2008; Katz-Gerro & López Sintas, 2019) is another example of hierarchical categorical variable, where the interdependency of activities forms a hierarchy. Thus, in this paper, we study the mathematical optimization problem that trades off, in linear regression models, accuracy and model complexity, while exploiting the structure of the nominal hierarchical categorical variables.

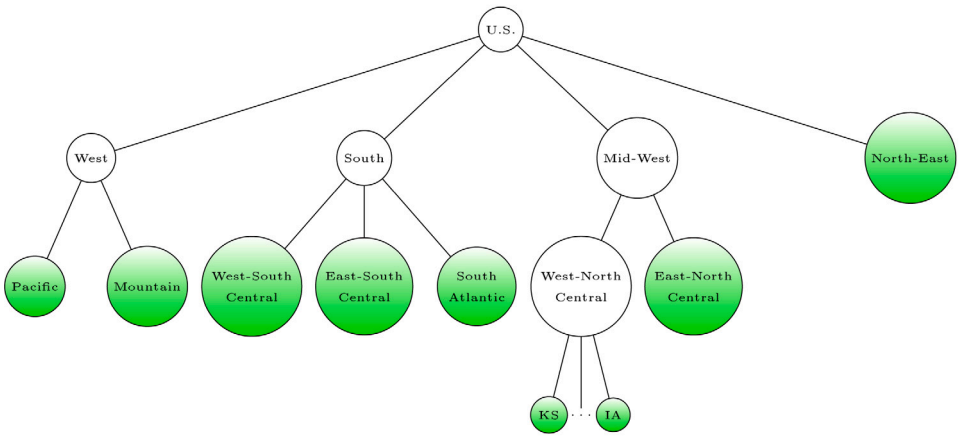
Let $\mathcal{J}'$ be the set of continuous and dummy predictor variables, whereas $\mathcal{J}$ the set of hierarchical categorical predictor variables. Throughout this paper, we will use the popular one-hot encoding for categorical predictor variables. Then, consider the random vector $(\mathbf{X}', \mathbf{X}, Y)$, where $\mathbf{X}'$ denotes the vector of the predictor variables in $\mathcal{J}'$, $\mathbf{X}$ denotes the vector of categorical predictor variables in $\mathcal{J}$, and Y denotes the response variable. In the real-world dataset `cancer-reg` (Rippner, 2017) used in the numerical section, with individuals from the United States of America (U.S.), *geography* is a categorical variable with a hierarchical structure. According to the *U.S. Department of Commerce Economics and Statistics Administration* and the *U.S. Census Bureau*, *geography* can be coded using the states (51 in total), which is the highest level of granularity for which information is available in the dataset. This means that 51 coefficients need to be estimated for this variable, where individuals in the same state share the same coefficient in the linear regression model. The variable *geography* can alternatively be coded using the subregions, such as *East-South Central*, *Middle Atlantic* and *New England*, where each state belongs to exactly one of the 9 subregions. Consolidating individuals at the subregions, sharing the same coefficient, yields a lower level of granularity for *geography*, where, instead of 51, only 9 coefficients need to be estimated and



(a) $\mathcal{S}_1^*$ when $\text{MSE}^*(\mathcal{S}_1^*) = 0.408$ and $c = 44$



(b) $\mathcal{S}_1^*$ when $\text{MSE}^*(\mathcal{S}_1^*) = 0.421$ and $c = 21$



(c) $\mathcal{S}_1^*$ when $\text{MSE}^*(\mathcal{S}_1^*) = 0.427$ and $c = 14$

Fig. 3. Less granular representations for the *geography* variable in the `cancer-reg` dataset.

interpreted. The individuals can be further consolidated into 4 regions, namely *West*, *South*, *Mid-West* and *North-East*, where only 4 coefficients would be associated to *geography* in the reduced linear regression model. Using these regions, one has the least granular representation of *geography*. This paper is devoted to trading off accuracy of the linear regression model and its complexity, measured as a cost function of the level of granularity used to represent each of the hierarchical categorical variables.

The categories of hierarchical categorical variable $j \in \mathcal{J}$ can be arranged as a directed tree $\mathcal{T}_j$, i.e., a directed graph with a root node, $r(\mathcal{T}_j)$, and a unique path from each node to $r(\mathcal{T}_j)$. In addition, let $\mathcal{V}(\mathcal{T}_j)$ denote the set of nodes in the tree and $\mathcal{L}(\mathcal{T}_j) \subset \mathcal{V}(\mathcal{T}_j)$ the set of leaf nodes. See Fig. 1 for the tree associated with the categories of *geography*, where the leaf nodes correspond to the states, going upstream we find the subregions and then the regions, which, in turn, are directly connected with the root node. Let $(\mathbf{x}_i', \mathbf{x}_i, y_i)$ be the vector associated with individual i, with $\mathbf{x}_i' = (x_{ij'}')$ and $\mathbf{x}_i = (x_{ijv})$, where x_{ijv} is

Table 1
Notations.

Data and parameters	
n	The sample size
$\mathcal{J}'$	The set of continuous and dummy predictor variables
$\mathcal{J}$	The set of hierarchical categorical predictor variables
$\mathcal{T}_j$	Directed tree related to the hierarchical categorical predictor variable $j \in \mathcal{J}$
$r(\mathcal{T}_j)$	The root of the directed tree $\mathcal{T}_j$
$\mathcal{V}(\mathcal{T}_j)$	The set of nodes in the tree $\mathcal{T}_j$, which represents the categories of the hierarchical categorical predictor variable j
$\mathcal{L}(\mathcal{T}_j)$	The set of leaf nodes in the tree $\mathcal{T}_j$, $\mathcal{L}(\mathcal{T}_j) \subset \mathcal{V}(\mathcal{T}_j)$
$\mathcal{P}_{jl}$	The set of categories associated with the unique path in $\mathcal{T}_j$ from its root node $r(\mathcal{T}_j)$ to the leaf node $l \in \mathcal{L}(\mathcal{T}_j)$
c_{jv}	The cost associated to node $v \in \mathcal{V}(\mathcal{T}_j)$
$x'_{ij'}$	The value of predictor variable $j' \in \mathcal{J}'$ for individual i
x_{ijv}	It takes value 1 if individual i belongs to category $v \in \mathcal{V}(\mathcal{T}_j)$ of variable $j \in \mathcal{J}$; 0 otherwise
y_i	Response variable for individual i
$(\mathbf{x}'_i, \mathbf{x}_i, y_i)$	Vector associated with individual i, where $\mathbf{x}'_i = (x'_{ij'})$ and $\mathbf{x}_i = (x_{ijv})$
c	Threshold on the complexity of the model
Decision variables	
$\beta'_0 \in \mathbb{R}$	The independent term in the model
$\beta'_{j'} \in \mathbb{R}$	The coefficient of variable $j' \in \mathcal{J}'$
$\beta_{jv} \in \mathbb{R}$	The coefficient of category $v \in \mathcal{V}(\mathcal{T}_j)$, $j \in \mathcal{J}$
S_j	Subtree of $\mathcal{T}_j$, with the same root: $r(S_j) = r(\mathcal{T}_j)$
$z_{jv} \in \{0, 1\}$	It takes value 1 if node associated with category v of the hierarchical categorical variable j is selected as a leaf node of S_j; 0 otherwise
$\mathbf{z} = (z_{jv})$	The vector of the binary decision variables

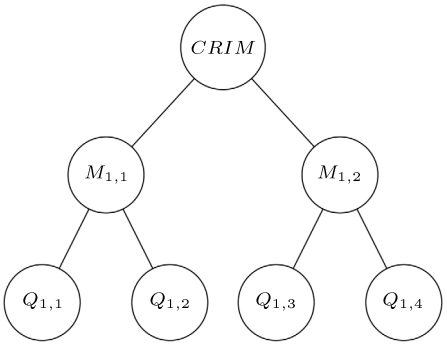


Fig. 4. Tree associated with the variable *CRIM* in the `housing` dataset after being discretized.

equal to 1 if individual i belongs to category $v \in \mathcal{V}(\mathcal{T}_j)$ of variable $j \in \mathcal{J}$. If we were to use the most granular representation of the hierarchical categorical variables, we would need to use the categories associated with the leaf nodes $l \in \mathcal{L}(\mathcal{T}_j)$, i.e.,

$$\hat{y}_i = \beta'_0 + \sum_{j' \in \mathcal{J}'} \beta'_{j'} x'_{ij'} + \sum_{j \in \mathcal{J}} \sum_{l \in \mathcal{L}(\mathcal{T}_j)} \beta_{jl} x_{ijl}, \tag{1}$$

where β'_0 is the independent term, $\beta'_{j'}$ is the coefficient of variable $j' \in \mathcal{J}'$, whereas β_{jl} is the coefficient of category $l \in \mathcal{L}(\mathcal{T}_j)$ of hierarchical categorical variable $j \in \mathcal{J}$. In the ordinary least squares (OLS) paradigm, the coefficients are obtained by minimizing the mean squared error (MSE). The corresponding OLS model reads as follows

$$\begin{aligned} \mathrm{MSE}^*((\mathcal{T}_j)_{j \in \mathcal{J}}) = \min_{\beta'_0, (\beta'_{j'})_{j' \in \mathcal{J}'}, (\beta_{jl})_{l \in \mathcal{L}(\mathcal{T}_j), j \in \mathcal{J}}} & \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta'_0 - \sum_{j' \in \mathcal{J}'} \beta'_{j'} x'_{ij'} \\ & - \sum_{j \in \mathcal{J}} \sum_{l \in \mathcal{L}(\mathcal{T}_j)} \beta_{jl} x_{ijl})^2, \end{aligned} \tag{2}$$

where n is the sample size. In the `cancer-reg` dataset, with the most granular representation of *geography*, we have an in-sample MSE of 0.4065. It should be highlighted that this MSE is obtained without exploiting the hierarchical structure of *geography*. The question arises as to whether that level of granularity is necessary, or whether we can merge categories at the bottom of the tree into a broader category upstream in the tree. With this, we can eliminate the state information for all the individuals of same subregion, respectively from the same region, and report the subregion, respectively the region. We have done this for the states in the subregions *Middle Atlantic* and *New England*, yielding the subtree in Fig. 3(a) of the tree in Fig. 1. All individuals in the descendants leaf nodes of *Middle Atlantic* are consolidated in its parent node *Middle Atlantic* and, therefore, they share the same coefficient in the linear regression model, and the same for *New England* node. With this representation, the in-sample MSE increases from 0.4065 to 0.408. This mild worsening in accuracy corresponds to an improvement in the complexity of the linear regression model, with a reduction from 51 to 44 in the number of coefficients to be estimated and interpreted for the *geography* variable.

Reducing the granularity of the representation of hierarchical categorical variables has several advantages. First, and as illustrated above, it is a step towards enhancing the interpretability of the linear regression model, where fewer coefficients need to be estimated and interpreted (Carrizosa et al., 2021; Carrizosa, Nogales-Gómez, & Romero Morales, 2017). Second, if the samples of individuals associated with categories are homogeneous enough, a very granular representation would yield an overparameterized model. Instead, we could merge these categories into a broader one upstream the tree, thus having more observations to estimate fewer coefficients. The homogeneity together with the increase in sample size ensure lower errors in the estimation of the coefficients of the broader categories (LeBlanc & Tibshirani, 1998). Third, and again if the samples of individuals associated with categories are homogeneous enough, a very granular representation will yield higher data gathering costs (Carrizosa, Martín-Barragán, & Romero Morales, 2008; Turney, 1995), if, for instance, the surveying costs are asymmetric. Indeed, we would need to ensure a large enough sample for each category in the representation, even though the cost of surveying may be high for some of these categories. By merging homogeneous categories into a broader one upstream the tree, we can sample from a larger subpopulation lowering these data gathering costs. Fourth, our methodology can identify where j is an irrelevant predictor (Bertsimas et al., 2020; Blanquero, Carrizosa, Molero-Río, & Romero Morales, 2020; Carrizosa, Olivares-Nadal and

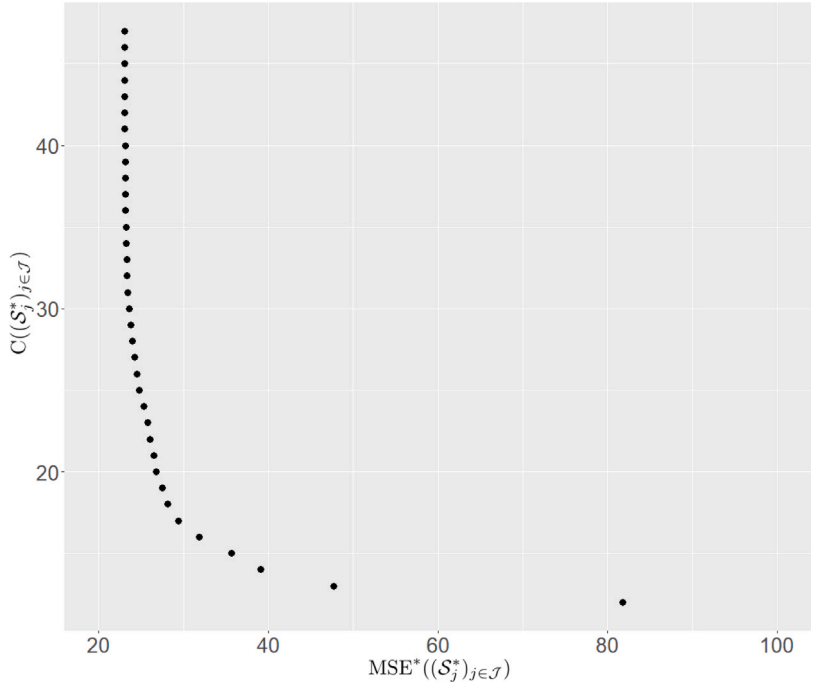


Fig. 5. Pareto frontier for MSE versus the number of coefficients to be estimated in the reduced model for the hierarchical categorical variables in the `housing` dataset.

Ramírez-Cobo, 2016) by consolidating individuals at the root node $r(\mathcal{T}_j)$. Finally, the consolidation of information is important when having data privacy considerations, Li and Sarkar (2009) and Lu, Zhu, Liu, Liu, and Shao (2014), since it is well-known that more detailed information is linked to confidentiality concerns (Baena, Castro, & Frangioni, 2020).

The remainder of this paper is structured as follows. In Section 2, we introduce the Tree based Linear Regression (TLR) model, a *constrained* problem, in which we minimize the accuracy of the reduced linear regression model, measured by its MSE, subject to a complexity constraint, where a threshold is imposed on the cost of granularity of the representation of the nominal hierarchical categorical variables. This problem is then formulated as a Mixed Integer Convex Quadratic Problem with Linear Constraints. Section 3 illustrates our approach in two real-world datasets as well as in a synthetic one, where the entire set of non-dominated outcomes to the problem is obtained solving the constrained problem for the different values of the threshold. To end, some conclusions and lines for future research are provided in Section 4.

2. The tree based linear regression model

In this section, we first model the two objectives under consideration when building the reduced linear regression model. We then provide a Mixed Integer Convex Quadratic formulation with Linear Constraints for the constrained problem, hereafter the so-called Tree based Linear Regression (TLR) model. We end the section with a discussion on the values of the threshold parameter to find all possible non-dominated outcomes to our problem. Before that, we introduce below some notation, see Table 1 for a summary of it.

Consolidating the information of hierarchical categorical variables is equivalent to finding, for each $j \in \mathcal{J}$, a subtree $\mathcal{S}_j$ of $\mathcal{T}_j$, with the same root as $\mathcal{T}_j$, $r(\mathcal{S}_j) = r(\mathcal{T}_j)$. The accuracy of the reduced linear regression model, with individuals consolidated at the leaf nodes $\mathcal{L}(\mathcal{S}_j)$, will be measured by its MSE, while its complexity will be measured by

$$\mathrm{C}((S_j)_{j\in\mathcal{J}}) = \sum_{j\in\mathcal{J}} \sum_{l\in\mathcal{L}(S_j)} c_{jl}, \tag{3}$$

where $c_{jv} \geq 0$ represents the cost associated to node $v \in \mathcal{V}(\mathcal{T}_j)$.

With this, our problem reads as follows:

$$\min_{(S_j)_{j\in\mathcal{J}}} (\mathrm{MSE}^*((S_j)_{j\in\mathcal{J}}), \mathrm{C}((S_j)_{j\in\mathcal{J}})), \tag{4}$$

where $\mathrm{MSE}^*((S_j)_{j\in\mathcal{J}})$ is defined as in (2) with $\mathcal{L}(S_j)$ replacing $\mathcal{L}(\mathcal{T}_j)$. Note that Problem (4) performs akin to the pruning of a regression tree (Sherali, Hobeika, & Jeenanunta, 2009; Su, Wang, & Fan, 2004). In our case, we have one tree per hierarchical categorical predictor in the dataset, and the pruning of all these trees needs to be performed simultaneously to properly trade off the accuracy and the complexity of the reduced linear regression model.

Non-dominated outcomes to Problem (4) are obtained by solving the Tree based Linear Regression (TLR) model:

$$\begin{aligned} \min_{(S_j)_{j\in\mathcal{J}}} \quad & \mathrm{MSE}^*((S_j)_{j\in\mathcal{J}}) \\ \text{s.t.} \quad & \mathrm{C}((S_j)_{j\in\mathcal{J}}) \leq c, \end{aligned} \tag{TLR}$$

where c is a threshold on the complexity of the model.

To formulate Problem (TLR) as a Mixed Integer Convex Quadratic Problem with Linear Constraints, we note that finding a subtree $\mathcal{S}_j$ of $\mathcal{T}_j$, with $r(\mathcal{S}_j) = r(\mathcal{T}_j)$, is equivalent to finding its leaf nodes. Therefore, we introduce binary decision variables $\mathbf{z} = (z_{jv})$, such that $z_{jv} = 1$ if the node associated with category v of the hierarchical categorical variable j is selected as leaf node of $\mathcal{S}_j$, and $z_{jv} = 0$ otherwise. If node v is selected, all individuals in its descendant leaf nodes are consolidated at v, and these individuals will share the same coefficient in the reduced linear regression model.

We need additional constraints to ensure that $\mathbf{z}$ is well defined. For this, we make use of the structural properties of the unique path $\mathcal{P}_{jl}$ in $\mathcal{T}_j$ from its root to leaf node $l \in \mathcal{L}(\mathcal{T}_j)$, $j \in \mathcal{J}$. It is easy to see that $\mathbf{z}$ is well defined if and only if there exists exactly one v such $z_{jv} = 1$ for each path $\mathcal{P}_{jl}$. With this, $\sum_{v\in\mathcal{V}(\mathcal{T}_j)} z_{jv}x_{ijv}$ represents the observed value for hierarchical predictor variable j in individual i, $\sum_{v\in\mathcal{V}(\mathcal{T}_j)} z_{jv}x_{ijv}\beta_{jv}$ is the contribution of j towards the predicted response for individual i, and $\sum_{v\in\mathcal{V}(\mathcal{T}_j)} c_{jv}z_{jv}$ is the contribution of j towards the cost in (3).

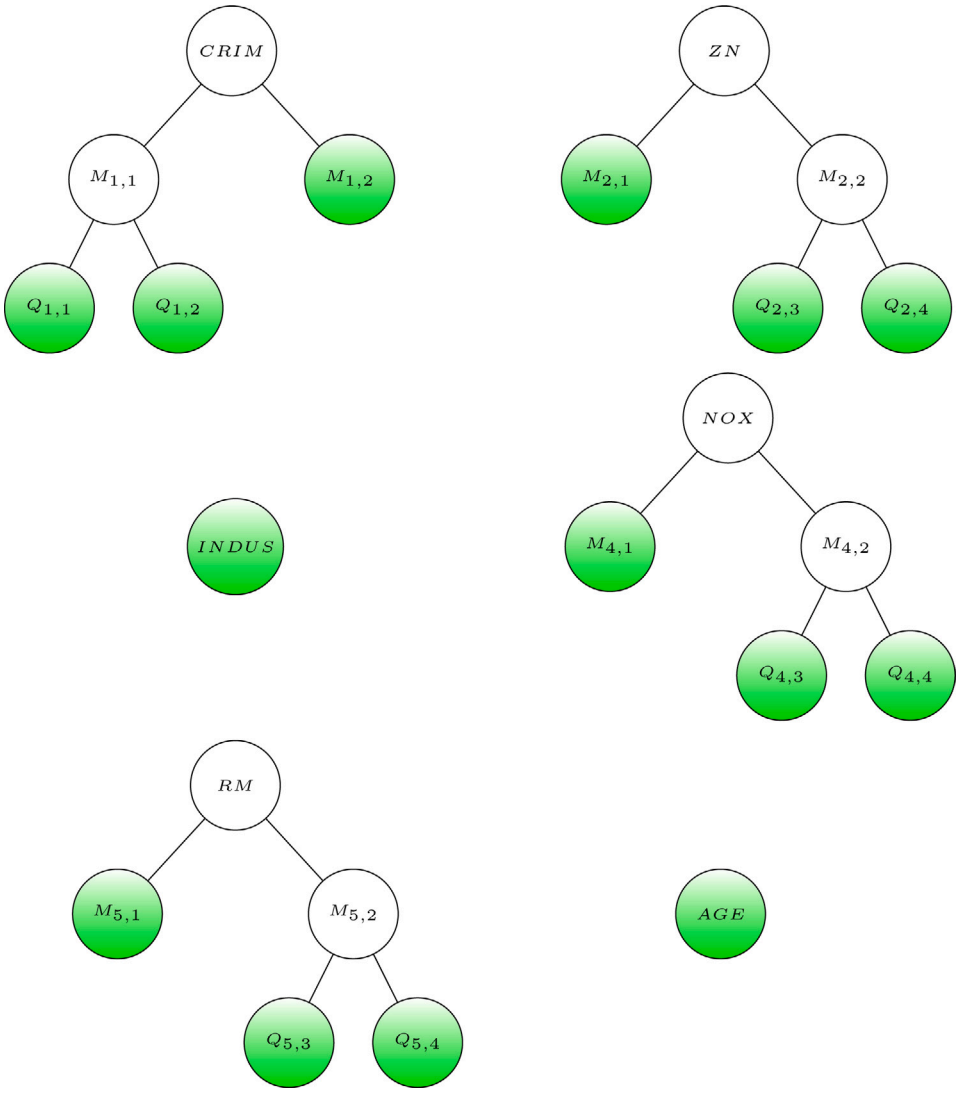


Fig. 6. Less granular representations for the first six hierarchical categorical variables in the `housing` dataset for the solution in Fig. 5 with $\text{MSE}^*((\mathcal{S}_j^*)_{j\in\mathcal{J}}) = 23.371$ and $c = 32$. Note that this is the solution that achieves the minimum AIC.

Therefore, Problem (TLR) can be formulated as follows:

$$\min_{\mathbf{z},\beta_0',(\beta_{j'}')_{j'\in\mathcal{J}'},(\beta_{jv})_{v\in\mathcal{V}(\mathcal{T}_j),j\in\mathcal{J}}} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0' - \sum_{j'\in\mathcal{J}'} x_{ij'}'\beta_{j'}' - \sum_{j\in\mathcal{J}}\sum_{v\in\mathcal{V}(\mathcal{T}_j)} z_{jv}x_{ijv}\beta_{jv})^2 \tag{5}$$

$$\text{s.t.} \quad \sum_{v\in\mathcal{P}_{jl}} z_{jv} = 1, \ \ l \in \mathcal{L}(\mathcal{T}_j), \ \ j \in \mathcal{J}, \tag{6}$$

$$\sum_{j\in\mathcal{J}}\sum_{v\in\mathcal{V}(\mathcal{T}_j)} c_{jv}z_{jv} \le c, \tag{7}$$

$$z_{jv} \in \{0,1\}, \ \ \forall v \in \mathcal{V}(\mathcal{T}_j), \ \ j \in \mathcal{J}, \tag{8}$$

$$\beta_0', \beta_{j'}', \beta_{jv} \in \mathbb{R}, \forall j' \in \mathcal{J}', \forall v \in \mathcal{V}(\mathcal{T}_j), \ j \in \mathcal{J}. \tag{9}$$

The objective function (5) is the MSE of linear models. The linear constraints (6) model that only one node is selected per path, becoming thus a leaf node of the subtree sought. Constraint (7) imposes the threshold c on the complexity of the reduced linear regression model. Constraints (8) and (9) impose the range of the decision variables.

Since the objective function (5) has semi-continuous variables, $z_{jv}\beta_{jv}$, a smooth formulation can be obtained using big M constraints:

$$\min_{\mathbf{z},\beta_0',(\beta_{j'}')_{j'\in\mathcal{J}'},(\beta_{jv})_{v\in\mathcal{V}(\mathcal{T}_j),j\in\mathcal{J}}} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0' - \sum_{j'\in\mathcal{J}'} x_{ij'}'\beta_{j'}' - \sum_{j\in\mathcal{J}}\sum_{v\in\mathcal{V}(\mathcal{T}_j)} x_{ijv}\tilde{\beta}_{jv})^2$$

$$\begin{aligned} \text{s.t.} \quad & (6)\text{–}(8), \\ & -Mz_{jv} \le \tilde{\beta}_{jv} \le Mz_{jv}, \ \ \forall v \in \mathcal{V}(\mathcal{T}_j), \ j \in \mathcal{J}, \\ & \beta_0', \beta_{j'}', \tilde{\beta}_{jv} \in \mathbb{R}, \forall j' \in \mathcal{J}', \forall v \in \mathcal{V}(\mathcal{T}_j), \ j \in \mathcal{J}. \end{aligned} \tag{10}$$

This is the formulation that will be used in the numerical section. Note that we can sharpen the value of M by imposing an upper bound on the coefficients of the categories of hierarchical variables. This can be seen as a regularization, thus preventing overfitting and allowing for sparser models (Carrizosa, Nogales-Gómez and Romero Morales, 2016). Other types of regularization can be easily incorporated into our model,

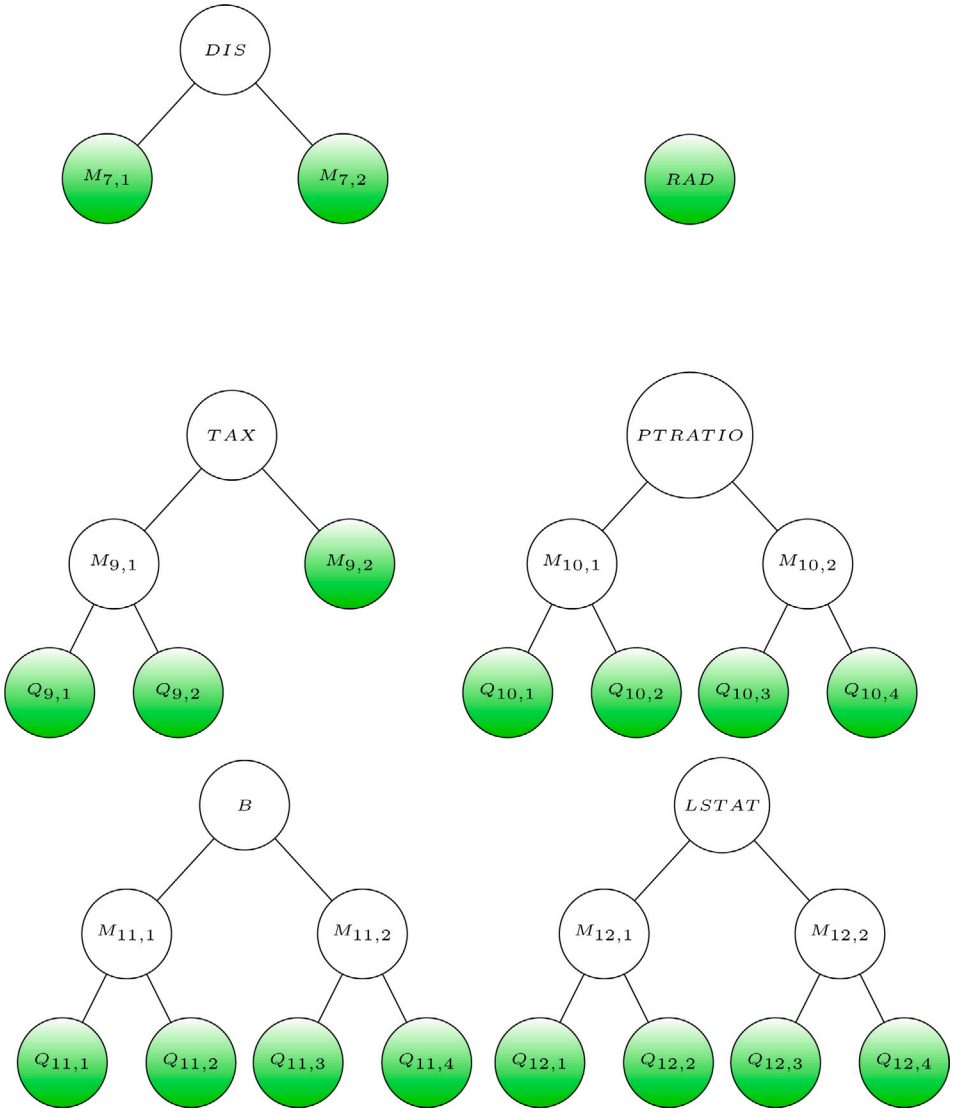


Fig. 7. Less granular representations for the last six hierarchical categorical variables in the `housing` dataset for the solution in Fig. 5 with $\text{MSE}^*((S_j^*)_{j\in\mathcal{J}}) = 23.371$ and $c = 32$. Note that this is the solution that achieves the minimum AIC.

such as those in Simon, Friedman, Hastie, and Tibshirani (2011) and Yuan and Lin (2006).

We now discuss the choice of values for threshold c. It is easy to show that if c_{jv} are integer numbers, it is enough to consider integer values for c too. Moreover, it is easy to define lower ($c^{\min} := |\mathcal{J}|$) and upper ($c^{\max} := \text{C}((\mathcal{T}_j)_{j\in\mathcal{J}})$) bounds on c. By varying the threshold value c among this finite set of values, we obtain the entire set of non-dominated outcomes to Problem (4).

Non-dominated outcomes to Problem (4) can also be obtained by solving the alternative constrained problem:

$$\begin{aligned} \min_{(S_j)_{j\in\mathcal{J}}} \quad & \text{C}((S_j)_{j\in\mathcal{J}}) \\ \text{s.t.} \quad & \text{MSE}^*((S_j)_{j\in\mathcal{J}}) \leq f, \end{aligned} \tag{11}$$

where f is the threshold value on the MSE of the reduced linear regression model. The advantage of constraining $\text{MSE}^*((S_j)_{j\in\mathcal{J}})$ is to have full control on the accuracy of the model and to allow the user to define meaningful values of f, Blanquero, Carrizosa, Ramírez-Cobo, and Sillero-Denamiel (2021). Therefore, this option is recommended when the constrained problem is solved only for a few values of f.

A lower bound on f is

$$f^{\min} := \text{MSE}^*((\mathcal{T}_j)_{j\in\mathcal{J}}), \tag{12}$$

which is the MSE that we achieve for the highest level of granularity on all the hierarchical categorical variables. An upper bound on f is found by removing all the variables $j \in \mathcal{J}$. This corresponds to

$$f^{\max} := \min_{\beta'_0,(\beta'_{j'})_{j'\in\mathcal{J}'}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta'_0 - \sum_{j'\in\mathcal{J}'}\beta'_{j'}x'_{ij'})^2, \tag{13}$$

where we consider the subtree with only the root node, i.e., $S_j = \{r(\mathcal{T}_j)\}$ $\forall j \in \mathcal{J}$. In this case, by varying the threshold value f in a grid of $[f^{\min}, f^{\max}]$, we obtain a collection of non-dominated outcomes to Problem (4).

3. Numerical experiments

In this section, we illustrate our approach using two real-world datasets and a synthetic one. Our aim is to depict the tradeoff between the accuracy of the reduced model and its complexity, measured by the number of coefficients to be estimated for the hierarchical categorical

Table 2
Coefficients associated with four different representations for *geography* variable in the `cancer-reg` dataset.

				Tree Fig. 1	Optimal tree Fig. 3(a)	Optimal tree Fig. 3(b)	Optimal tree Fig. 3(c)
U.S.	West	Pacific	WA	0.058	0.065	0.086	0.109
			AK	0.720	0.716	0.651	
			CA	−0.149	−0.138	−0.114	
			HI	−0.978	−0.992	−0.990	
			OR	−0.015	−0.010	0.002	
		Mountain	ID	−0.302	−0.313	−0.199	−0.161
			WY	0.143	0.143		
			NV	0.481	0.499		
			UT	−0.459	−0.466		
			CO	−0.300	−0.303		
			AZ	−0.473	−0.459		
			NM	−0.273	−0.263		
			MT	−0.131	−0.144		
	South	West-South Central	OK	0.692	0.683	0.705	0.567
			AR	0.686	0.678	0.722	
			TX	0.416	0.408	0.432	
			LA	0.355	0.345	0.378	
		East-South Central	TN	0.529	0.525	0.575	0.631
			MS	0.545	0.539		
			AL	0.330	0.330		
			KY	0.681	0.683		
		South Atlantic	NC	0.099	0.098	0.306	0.368
			DE	0.043	0.056		
			FL	0.284	0.282		
			GA	0.129	0.121		
			MD	0.325	0.337		
			SC	0.373	0.369		
			WV	0.364	0.359		
			DC	0.350	0.393		
			VA	0.535	0.538		
	Mid-West	West-North Central	KS	0.638	0.659	0.538	0.580
			MN	0.302	0.327	0.201	0.230
			MO	0.581	0.575	0.574	0.611
			NE	0.069	0.073	0.069	0.100
			ND	0.123	0.132	0.097	0.103
			SD	0.019	0.019	0.016	0.014
			IA	−0.077	−0.072	−0.088	−0.057
		East-North Central	IL	0.201	0.211	0.291	0.336
			IN	0.527	0.526		
			MI	0.239	0.243		
			WI	0.145	0.148		
			OH	0.386	0.389		
	North-East	Middle Atlantic	PA	−0.099	−0.102	−0.076	−0.025
			NJ	0.074			
			NY	−0.183			
		New England	ME	0.327	0.121		
			VT	0.241			
			MA	−0.053			
			RI	0.102			
			CT	−0.311			
			NH	0.183			

variables, which corresponds to $c_{jv} = 1$ in (3). To solve Problem (10) for all possible values of $c \in \{c^{\min}, \ldots, c^{\max}\}$, we use the solver Gurobi (Gurobi Optimization, 2018) for mixed-integer quadratically-constrained problems. In particular, the Gurobi R interface is used in this work to obtain all numerical results, where M is set to 1000. The experiments have been run on Intel(R) Core(TM) i7-7500U CPU at 2.70 GHz 2.90 GHz with 8.0 GB of RAM.

3.1. Cancer trials dataset: a real-world dataset

Consider again the real-world dataset `cancer-reg` introduced in Section 1. This dataset aims to look for relationships between the socioeconomic status in U.S. and the mean per capita cancer mortality (response variable). It has a sample of size $n = 3047$ with 32 predictor variables: one hierarchical predictor variable ($|\mathcal{J}| = 1$) and 31 non-hierarchical predictor variables ($|\mathcal{J}'| = 31$), where continuous predictors have been standardized and, as commented in Section 1, the one-hot encoding has been used for the categorical variable. This database was collected from the American Community Survey (`census.gov`), `clinicaltrials.gov` and `cancer.gov` sources. As mentioned in Section 1, the only hierarchical categorical variable is *geography*, see Fig. 1, and contains information on the state linked to the individuals.

We solve Problem (10) for the 51 values of c in the set $\{1, \ldots, 51\}$. Fig. 2 reports the Pareto frontier for the MSE and the number of coefficients to be estimated in the reduced model for the hierarchical categorical variable. In particular, the point with maximum MSE (in the bottom right corner) is the case when the variable *geography* is not considered by the model, that is, when the individuals have been consolidated at the root node and then $C(S_1^*) = 1$. The point with minimum MSE (in the top left corner) shows the result when the tree structure of the hierarchical categorical variables is ignored and thus the one-hot encoding of the most granular representation of the variable is considered. As can be observed from the top left corner of

Table 3
The predictor and the response variables in the `housing` dataset.

Variable	Name	Description	Type	Discretized
Predictor	CRIM	Crime rate by town	Continuous	Yes
	ZN	Proportion of residential land zoned for lots greater than 25.000 square feet	Continuous	Yes
	INDUS	Proportion of nonretail business acres per town	Continuous	Yes
	NOX	Nitrogen oxide concentrations	Continuous	Yes
	RM	Average number of rooms	Continuous	Yes
	AGE	Proportion of owner units built prior to 1940	Continuous	Yes
	DIS	Weighted distances to five employment centres	Continuous	Yes
	RAD	Index of accessibility to radial highways	Continuous	Yes
	TAX	Full value property tax rate ($/$10.000)	Continuous	Yes
	PTRATIO	Pupil–teacher ratio by town school district	Continuous	Yes
	B	Black proportion of population	Continuous	Yes
	LSTAT	Proportion of population that is lower status	Continuous	Yes
	CHAS	1 if tract bounds river; 0 otherwise	Binary	No
Response	MEDV	Median value of owner-occupied homes (in $1000's)	Continuous	No

Fig. 2, a mild worsening of the MSE implies an improvement in the fusion of categories for *geography*. For example, Fig. 3(a) is related to the point that returns an increase in the MSE of 0.34% and a decrease of 15.91% in the number of coefficients to be estimated, with respect to the results under not exploiting the tree structure. This behaviour is also observed in Fig. 5 for the `housing` dataset. Clearly, our methodology can find a much less complex model with a very mild worsening of the accuracy, but it is ultimately the decision of the user as to which reduced model to choose.

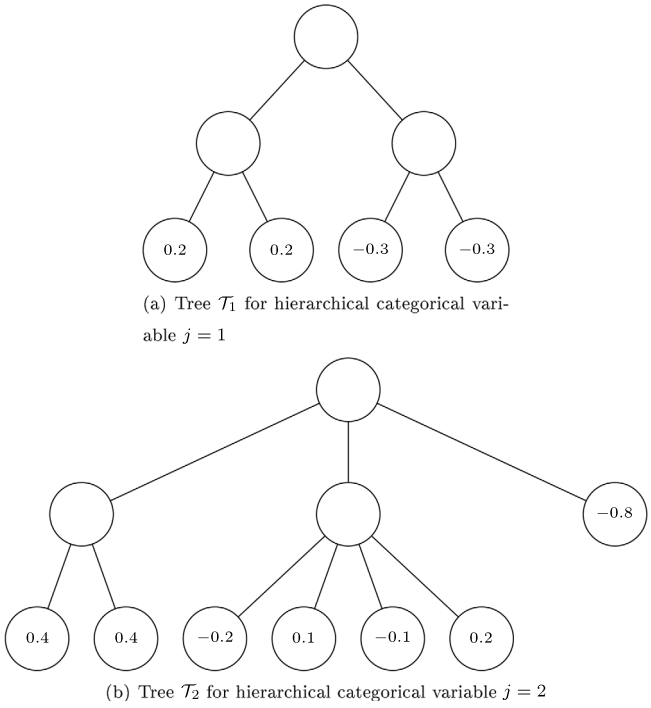
Fig. 3 plots the selected subtree S_1^* associated with *geography* for three of the solutions in Fig. 2. In particular, Fig. 3(a) is the representation associated with the model that achieves the minimum Akaike information criterion (AIC) metric (Akaike, 1998), whereas Fig. 3(c) the one with the minimum Bayesian information criterion (BIC) (Schwarz, 1978), which are two measures for model selection that compute the tradeoff between the in-sample fit and the number of parameters involved.

Table 2 presents the coefficients of *geography* for four of the solutions in Fig. 2, namely the most complex model when all the leaf nodes in Fig. 1 are considered, as well as the three reduced models with less granular representation of *geography* in Fig. 3. We can see that when categories are merged into one upstream the tree, the single coefficient that needs to be estimated for that broader category is within the range of the coefficients obtained with the most granular representation.

3.2. Boston housing dataset

The well-known `housing` dataset (Harrison & Rubinfeld, 1978) contains information concerning the price of the houses in the area of Boston, which was collected from the U.S. Census Service. See Table 3 for a description of its predictor variables, as well as the response. It has a sample of size $n = 506$ with 13 predictor variables: 12 continuous, which have been discretized yielding 12 hierarchical predictor variables ($|\mathcal{J}| = 12$), and 1 binary one ($|\mathcal{J}'| = 1$). Fig. 4 illustrates the discretization of *CRIM*, the first continuous variable. Similar ones have been implemented for the other 11 continuous variables. First, we split the observations of *CRIM* into two groups: those whose values are below (node $M_{1,1}$) and above (node $M_{1,2}$) the median. Second, the quartiles are used to subdivide $M_{1,1}$ (nodes $Q_{1,1}$ and $Q_{1,2}$) and $M_{1,2}$ (nodes $Q_{1,3}$ and $Q_{1,4}$) into two nodes. This way we examine the thresholds of the continuous predictor variables required to predict the response variable.

When solving Problem (10) for the 37 values of c in the set $\{12, \dots, 48\}$, we obtain the Pareto frontier in Fig. 5. The MSE of the model with the highest granularity for all hierarchical variables is 23.064. When we start reducing the granularity the MSE remains approximately the same. Actually, when c is reduced from 48 to 30, the accuracy is barely damaged but the complexity of the linear regression model is dramatically improved.



(a) Tree $\mathcal{T}_1$ for hierarchical categorical variable $j = 1$

(b) Tree $\mathcal{T}_2$ for hierarchical categorical variable $j = 2$

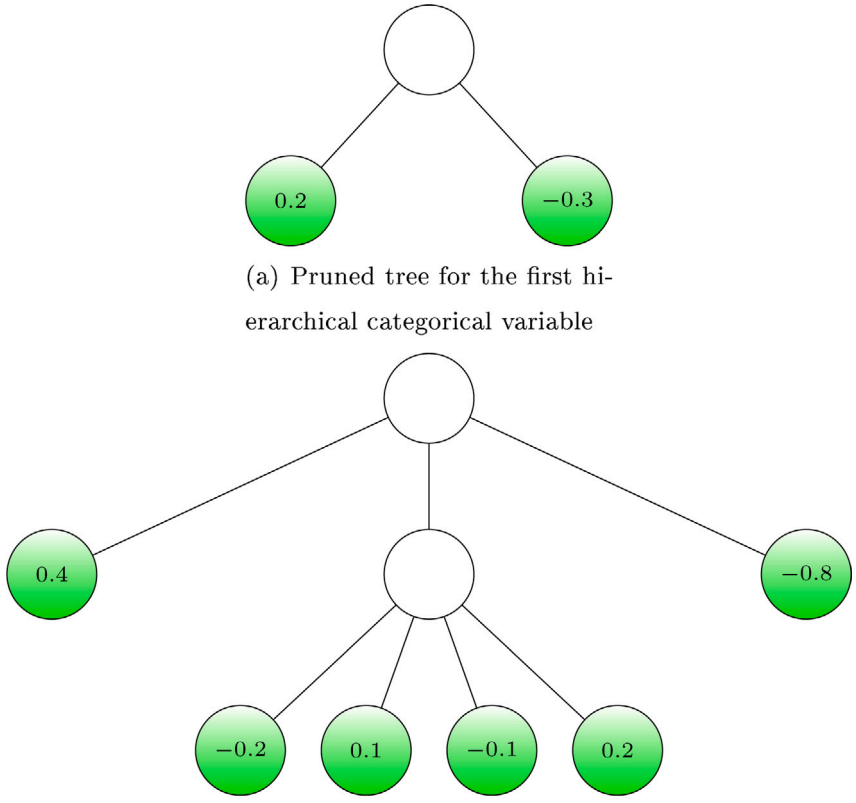
Fig. 8. Trees associated with the two hierarchical categorical variables in the synthetic dataset together with β_{jl}^{S}, $l \in \mathcal{L}(\mathcal{T}_j)$.

Figs. 6–7 show the subtrees S_j^* for all $j \in \mathcal{J}$ for the solution in Fig. 5 that achieves the minimum AIC. In this solution, we can observe how variables *INDUS*, *AGE* and *RAD* are eliminated from the linear regression model, as their root node is the only one selected with a coefficient equal to zero. By contrast, we require the highest level of granularity for *PTRATIO*, *B* and *LSTAT*. For *DIS*, the linear regression model only needs to know whether the predictor variable is below the median. For the remaining predictor variables, leaf as well as non-leaf nodes are selected.

3.3. The synthetic data

In this section we illustrate our approach on synthetic data. The data generating model is

$$y_i = \sum_{j \in \mathcal{J}} \sum_{l \in \mathcal{L}(\mathcal{T}_j)} \beta_{jl}^{S} x_{ijl} + \beta_1' x_{i1}' + \varepsilon_i, \; i = 1, \dots, n, \tag{14}$$



(a) Pruned tree for the first hierarchical categorical variable

(b) Pruned tree for the second hierarchical categorical variable

Fig. 9. Pruned tree and less granular representation of the two hierarchical categorical variables in Fig. 8 from the synthetic dataset.

where $|\mathcal{J}| = 2$ and $|\mathcal{J}'| = 1$. The values of the coefficients of the hierarchical categorical variables β^{S}_{jl}, $l \in \mathcal{L}(\mathcal{T}_j)$, are given in Fig. 8, whereas the coefficient β'_1 associated with the only continuous variable is equal to 1. Note that the first two leaf nodes of $\mathcal{T}_1$ have the same coefficient, and the same holds for the other two leaf nodes. Therefore, the tree can be pruned to avoid unnecessary splits, yielding the subtree in Fig. 9. The same holds for $\mathcal{T}_2$. We construct the continuous variable X'_1 such that it depends on the first hierarchical categorical variable (i.e., the variable in Fig. 9(a)). Indeed, $x'_{i1} \sim N(0,1)$ and $x'_{i1} \sim N(2,1)$, respectively for each leaf node of the pruned tree. Finally, the error is taken $\varepsilon_i \sim N(0,\sigma^2)$ for different values of σ^2 given below. We have $n = 3000$ individuals, evenly distributed across the different combinations of categories $l_1 \in \mathcal{L}(\mathcal{T}_1)$ and $l_2 \in \mathcal{L}(\mathcal{T}_2)$. The purpose of this section is twofold. First, we illustrate how our approach is able to recover the pruned tree underlying our synthetic data. Second, we carry out an *out-of-sample* study.

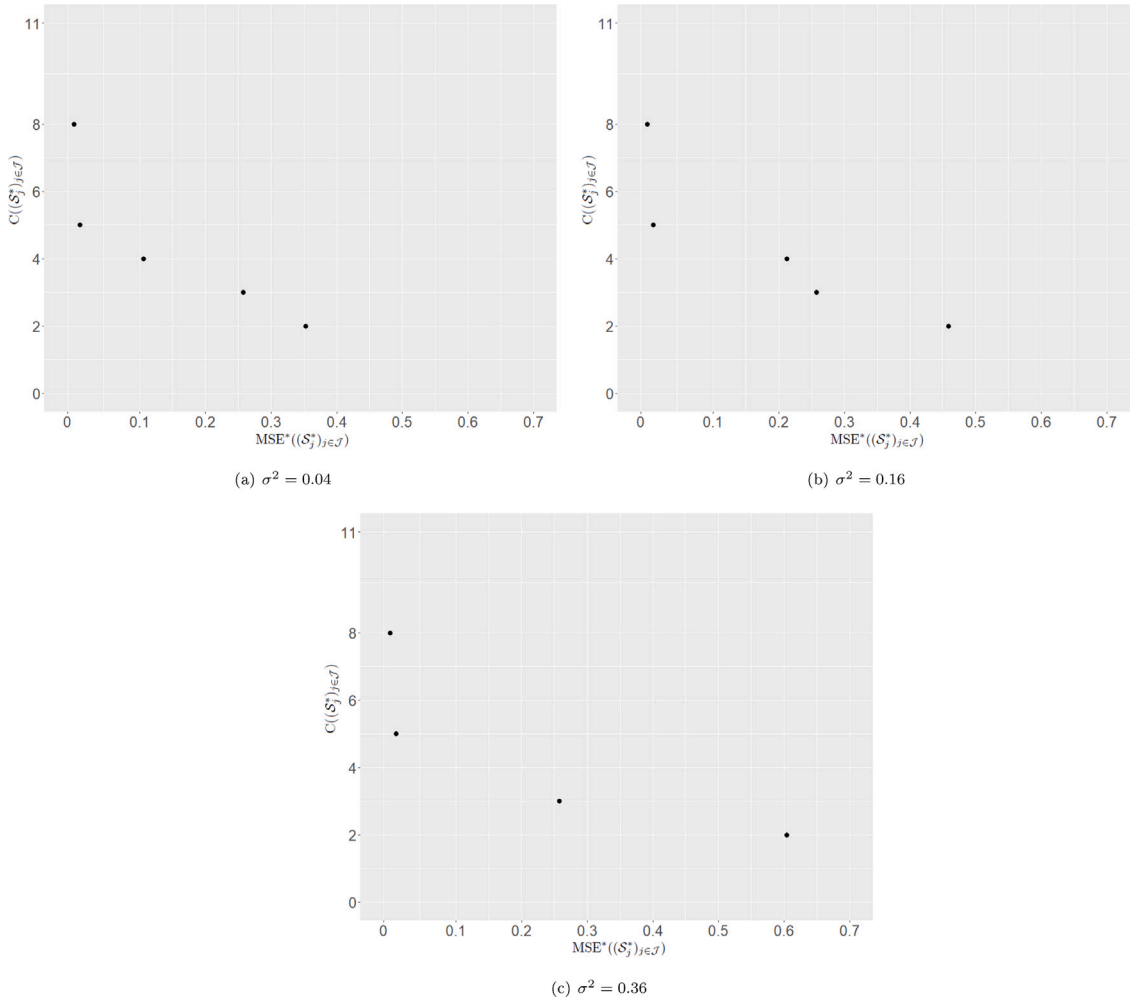
Let us consider $\sigma^2 = 0.04$ and solve Problem (10) for the 10 values of c in the set $\{2,\ldots,11\}$. Fig. 10(a) shows the Pareto frontier for the number of coefficients to be estimated in the reduced model versus the MSE. For small values of MSE, the chosen nodes are the 8 green leaf nodes in Fig. 9, which implies that our methodology is able to successfully detect the pruned tree underlying each hierarchical categorical variable in our data. Similar conclusions can be drawn when $\sigma^2 = 0.16$ (Fig. 10(b)) and $\sigma^2 = 0.36$ (Fig. 10(c)).

To end the numerical section, we provide an estimation for the MSE and the complexity of the reduced model using a 10-fold cross validation approach, showing that our procedure works properly with the available (*in-sample*) individuals, but also for future (*out-of-sample*) individuals. For each fold, the *in-sample* set is used to solve Problem (10) and get S^*_j, $j \in \mathcal{J}$. Once the subtrees are found, and thus the reduced linear regression model, we calculate its *in-sample* and *out-of-sample* MSE, which are plotted in Fig. 11(a)–(c) for the different values of σ^2. As can be observed, the *in-sample* MSE values (solid lines) are only slightly smaller than the *out-of-sample* values (dashed lines). Then, in view of results, we can conclude that our methodology generalizes well.

4. Conclusions and extensions

In this paper we have developed a novel methodology, within linear regression, to fuse categories of hierarchical categorical variables while respecting their tree structure. Through the TLR, a Mixed Integer Convex Quadratic Problem with Linear Constraints, we study the tradeoff between accuracy and model complexity. Our methodology has been tested on both real-world and synthetic datasets. The numerical section shows that much less granular representations for the hierarchical categorical variables can be found at the expense of slightly damaging the accuracy.

A number of extensions to this work are worth investigating. Firstly, when the number of categories is large, instead of solving Problem (10) considering all the categories at once, a sequential pruning can be used instead. The main idea is to consider subtrees in $\mathcal{T}_j$ and try to compress their categories solving Problem (10) sequentially. Another option to deal with large number of categories is to cluster them based on a dissimilarity, see Carrizosa et al. (2017) and Cerda, Varoquaux, and Kégl (2018) and references therein. Secondly, instead of use the objective function of the OLS method, it may be of interest to change it



(a) $\sigma^2 = 0.04$

(b) $\sigma^2 = 0.16$

(c) $\sigma^2 = 0.36$

Fig. 10. Pareto frontier for MSE versus the number of coefficients to be estimated in the reduced model for the synthetic dataset for different σ^2 values.

by that of the, e.g., elastic net, for the sake of dealing with strongly correlated predictors. Or even by that of the Lasso and its variants, to also get sparser solutions. In the same vein, our proposal can be run as a first step where the tree structure is exploited for hierarchical categorical variables, and then a Feature Selection procedure such as that introduced in Wang, Jiang, Huang, and Zhang (2013) could be performed for obtaining sparser solutions in terms of non-hierarchical categorical variables. Thirdly, this paper is based on the MSE as performance measure, but other strategies, such as those for robust estimation (see Jiang, Wang, Fu, & Wang, 2019; Wang et al., 2013), could be implemented. Finally, our methodology can be extended to generalized linear models (Tibshirani, 1996), where, instead of predicting the response variable as in (1), a non-linear relationship between the response variable and the predictors is through a linkage function. However, the last two extensions make the optimization problem highly nonlinear and its resolution is very challenging and outside the scope of this paper.

CRediT authorship contribution statement

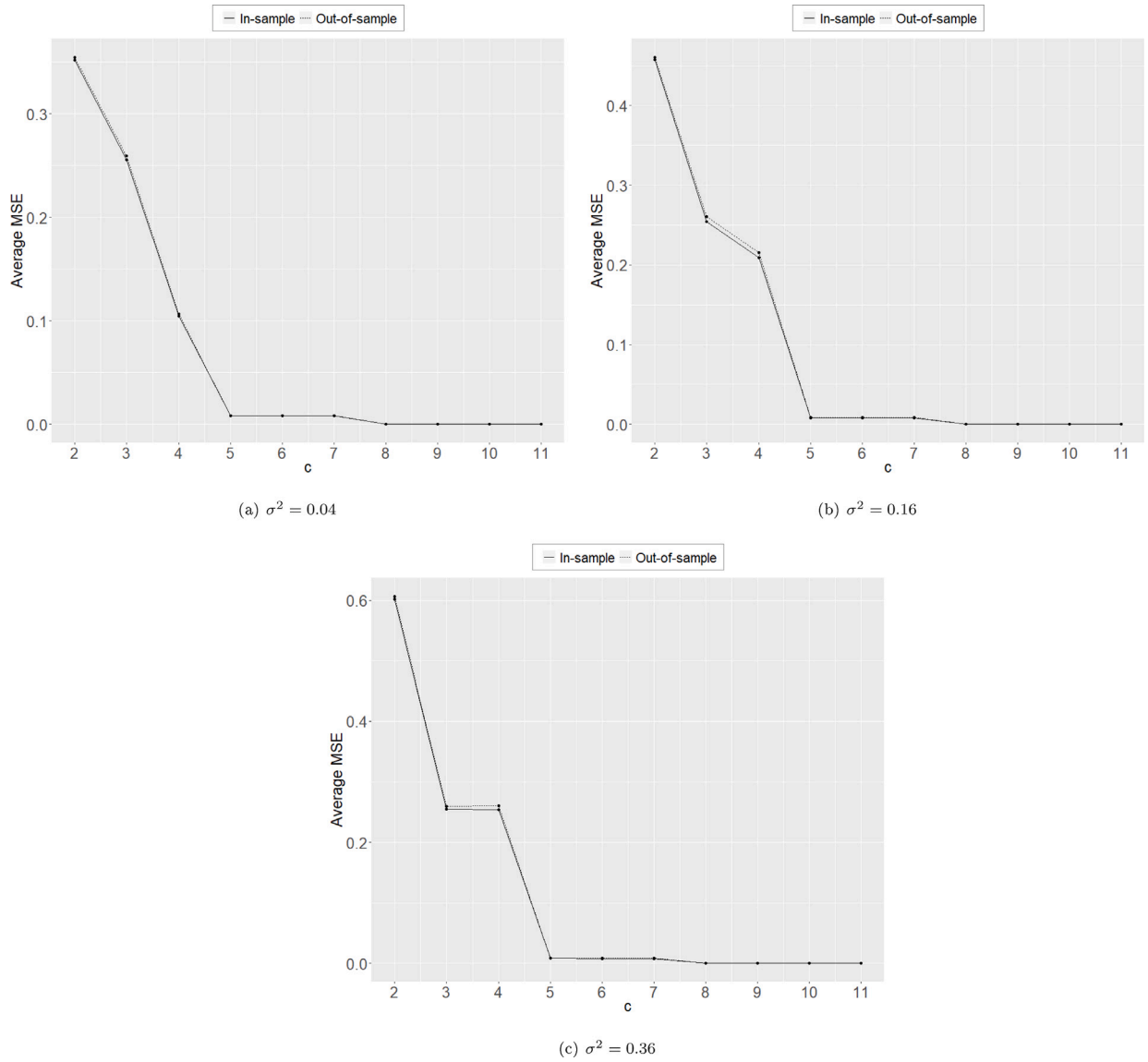
Emilio Carrizosa: Modelling and design of the experiments, Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Laust Hvas Mortensen:** Modelling and design of the experiments, Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Dolores Romero Morales:** Modelling and design of the experiments, Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **M. Remedios Sillero-Denamiel:** Modelling and design of the experiments, Implementing the code, Running the experiments, Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is partially supported by research grants and projects MTM2015-65915-R (Ministerio de Economía y Competitividad, Spain), PID2019-110886RB-I00 (Ministerio de Ciencia, Innovación y Universidades, Spain), FQM-329 and P18-FR-2369 (Junta de Andalucía, Spain), PR2019-029 (Universidad de Cádiz, Spain), Fundación BBVA, EC H2020 MSCA RISE NeEDS Project (Grant agreement ID: 822214) and The Insight Centre for Data Analytics (Ireland). This support is gratefully acknowledged.



(a) $\sigma^2 = 0.04$

(b) $\sigma^2 = 0.16$

(c) $\sigma^2 = 0.36$

Fig. 11. Average MSE (10-fold CV) versus the imposed threshold c when σ^2 changes.

References

Akaike, H. (1998). In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Information theory and an extension of the maximum likelihood principle* (pp. 199–213). New York, NY: Springer New York.

Baena, D., Castro, J., & Frangioni, A. (2020). Stabilized benders methods for large-scale combinatorial optimization, with application to data privacy. *Management Science*, *66*(7), 3051–3068.

Benítez-Peña, S., Blanquero, R., Carrizosa, E., & Ramírez-Cobo, P. (2019). Cost-sensitive feature selection for support vector machines. *Computers & Operations Research*, *106*, 169–178.

Bertsimas, D., & King, A. (2016). OR forum–an algorithmic approach to linear regression. *Operations Research*, *64*(1), 2–16.

Bertsimas, D., O'Hair, A., Relyea, S., & Silberholz, J. (2016). An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science*, *62*(5), 1511–1531.

Bertsimas, D., Pauphilet, J., & Parys, B. V. (2020). Sparse regression: Scalable algorithms and empirical performance. *Statistical Science*, *35*(4), 555–578.

Bertsimas, D., & Van Parys, B. (2020). Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *The Annals of Statistics*, *48*(1), 300–323.

Blanquero, R., Carrizosa, E., Jiménez-Cordero, A., & Martín-Barragán, B. (2019). Variable selection in classification for multivariate functional data. *Information Sciences*, *481*, 445–462.

Blanquero, R., Carrizosa, E., Molero-Río, C., & Romero Morales, D. (2020). Sparsity in optimal randomized classification trees. *European Journal of Operational Research*, *284*(1), 255–272.

Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., & Sillero-Denamiel, M. R. (2021). A cost-sensitive constrained lasso. *Advances in Data Analysis and Classification*, *15*, 121–158.

Bottou, L., Curtis, F., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, *60*(2), 223–311.

Carrizosa, E., Galvis Restrepo, M., & Romero Morales, D. (2021). On clustering categories of categorical predictors in generalized linear models. *Expert Systems with Applications*, *182*, Article 115245.

Carrizosa, E., Guerrero, V., & Romero Morales, D. (2018). Visualizing data as objects by DC (difference of convex) optimization. *Mathematical Programming, Series B*, *169*, 119–140.

Carrizosa, E., Guerrero, V., & Romero Morales, D. (2020). *On mathematical optimization for clustering categories in contingency tables*. Technical report. Universidad Carlos III, Madrid, Spain, https://www.researchgate.net/publication/341079651_On_mathematical_optimization_for_clustering_categories_in_contingency_tables.

Carrizosa, E., Martín-Barragán, B., & Romero Morales, D. (2008). Multi-group support vector machines with measurement costs: A biobjective approach. *Discrete Applied Mathematics*, *156*, 950–966.

Carrizosa, E., Nogales-Gómez, A., & Romero Morales, D. (2016). Strongly agree or strongly disagree?: Rating features in support vector machines. *Information Sciences*, *329*, 256–273.

Carrizosa, E., Nogales-Gómez, A., & Romero Morales, D. (2017). Clustering categories in support vector machines. *Omega*, *66*, 28–37.

Carrizosa, E., Olivares-Nadal, A. V., & Ramírez-Cobo, P. (2016). A sparsity-controlled vector autoregressive model. *Biostatistics*, *18*(2), 244–259.

Carrizosa, E., & Romero Morales, D. (2013). Supervised classification and mathematical optimization. *Computers & Operations Research*, *40*(1), 150–165.

Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, *107*(8), 1477–1494.

European Commission (2008). *NACE rev. 2 – statistical classification of economic activites in the European community*. Luxembourg: Office for Official Publications of the European Communities, https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF.

Fang, X., Liu Sheng, O. R., & Goes, P. (2013). When is the right time to refresh knowledge discovered from data? *Operations Research*, *61*(1), 32–44.

Fountoulakis, K., & Gondzio, J. (2016). A second-order method for strongly convex ℓ_1-regularization problems. *Mathematical Programming*, *156*(1), 189–219.

Fu, Z., Golden, B., Lele, S., Raghavan, S., & Wasil, E. (2003). Genetically engineered decision trees: Population diversity produces smarter trees. *Operations Research*, *51*(6), 894–907.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gotway, C. A., & Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, *97*(458), 632–648.

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications*, *100*, 1–16.

Gurobi Optimization, L. (2018). Gurobi optimizer reference manual.

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Jensen, P. B., Jensen, L., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*, 395–405.

Jiang, Y., Wang, Y.-G., Fu, L., & Wang, X. (2019). Robust estimation using modified Huber's functions with new tails. *Technometrics*, *61*(1), 111–122.

Johannemann, J., Hadad, V., Athey, S., & Wager, S. (2020). Sufficient representations for categorical variables. https://arxiv.org/abs/1908.09874.

Katz-Gerro, T., & López Sintas, J. (2019). Mapping circular economy activities in the European union: Patterns of implementation and their correlates in small and medium-sized enterprises. *Business Strategy and the Environment*, *28*(4), 485–496.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *Quarterly Journal of Economics*, *133*(1), 237–293.

LeBlanc, M., & Tibshirani, R. (1998). Monotone shrinkage of trees. *Journal of Computational and Graphical Statistics*, *7*(4), 417–433.

Li, X.-B., & Sarkar, S. (2009). Against classification attacks: A decision tree pruning approach to privacy protection in data mining. *Operations Research*, *57*(6), 1496–1509.

Lin, J., Zhong, C., Hu, D., Rudin, C., & Seltzer, M. (2020). Generalized and scalable optimal sparse decision trees. In *Proceedings of the 37th international conference on machine learning, Vol. 119* (pp. 6150–6160). PMLR.

Lu, R., Zhu, H., Liu, X., Liu, J. K., & Shao, J. (2014). Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, *28*(4), 46–50.

Martens, D., Baesens, B., Gestel, T. V., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, *183*(3), 1466–1476.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st international conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, workshop track proceedings*.

Pauger, D., & Wagner, H. (2019). Bayesian effect fusion for categorical predictors. *Bayesian Analysis*, *14*(2), 341–369.

Rippner, N. (2017). Cancer trials. Retrieved from http://data.world/exercises/linear-regression-exercise-1/workspace/file?filename=cancer_reg.csv.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Sherali, H., Hobeika, A., & Jeenanunta, C. (2009). An optimal constrained pruning strategy for decision trees. *INFORMS Journal on Computing*, *21*(1), 49–61.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*(5), 1–13.

Stokell, B. G., Shah, R. D., & Tibshirani, R. J. (2021). Modelling high-dimensional categorical data using nonconvex fusion penalties. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *83*(3), 579–611.

Su, X., Wang, M., & Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, *13*(3), 586–598.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.

Turney, P. (1995). Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, *2*, 369–409.

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, *102*(3), 349–391.

Wang, X., Jiang, Y., Huang, M., & Zhang, H. (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, *108*(502), 632–643.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *68*(1), 49–67.