Experiences from a socio-economic application of induction trees

Fabio B. Losa[1], Pau Origoni[1], and Gilbert Ritschard[2]

[1] Statistical Office of Ticino Canton, CH-6500 Bellinzona, Switzerland
fabio.losa@ti.ch
[2] Dept of Econometrics, University of Geneva, CH-1211 Geneva 4, Switzerland
gilbert.ritschard@themes.unige.ch

Abstract. This paper presents a full scaled application of induction trees for non-classificatory purposes. The grown trees are used for highlighting regional differences in the women's labor participation, by using data from the Swiss Population Census. Hence, the focus is on their descriptive rather than predictive power. Trees grown by language regions exhibit fundamental cultural differences supporting the hypothesis of cultural models in female participation. The explanatory power of the induced trees is measured with deviance based fit measures.

1 Introduction

Induced decision trees have become popular supervised classification tools since [1]. Though their primary purpose is to predict and to classify, trees can be used for many other relevant purposes: as exploratory methods for partitioning and identifying local structures in datasets, as well as alternatives to statistical descriptive methods like linear or logistic regression, discriminant analysis, and other mathematical modeling approaches [5].

This contribution demonstrates such a *non-classificatory* use of classification trees by presenting a full scaled application on female labor market data from the Swiss 2000 Population Census (SPC). The use of trees for our analysis was dictated by our primary interest in discovering the interaction effects of predictors of the women's labor participation. The practical experiment presented is original in at least two respects: 1) the use of trees for microeconomic analysis, which does not appear to be a common domain of application; 2) the use of induction trees for a complete population census dataset.

Note that since our goal is not to extract classification rules, but to understand — from a cross-cultural perspective — the forces that drive women's participation behavior, we do not rely on the usual misclassification rates for validating the trees. Rather, we consider some deviance based fit criteria [7, 8] similar to those used with logistic regression.

Before presenting our experiment, let us shortly recall the principle of classification trees. They are grown by seeking, through successive splits of the learning data set, some optimal partition of the predictor space for predicting the outcome

class. Each split is done according to the values of one predictor. The process is greedy. At the first step, it tries all predictors to find the "best" split. Then, the process is repeated at each new node until some stopping rule is reached. For our application, we used CART [1] that builds only binary trees by choosing at each step the split that maximizes the gain in purity measured by the Gini index. CART uses relatively loose stopping rules, but proceeds to a pruning round after the preliminary growing phase. We chose CART for our analysis, despite the gain in purity seems less appropriate for a non-classificatory purpose than, for example, the strength of association criterion used by CHAID [2]. Indeed, the great readability of the binary CART trees was decisive when compared with the n-ary CHAID trees that had, even at the first level, a much too high number of nodes to allow for any useful interpretation.

2 The applied study

We begin by setting the applied research framework, then we sketch our global analysis procedure and, finally, we present selected findings.

Female labor market participation reveals significant differences across countries. In Europe, scholars often identify at least two general models: a Mediterranean one (Italy, Greece, Portugal, etc.) versus a model typical for Central and Northern Europe [6]. The first is represented by an inverse L-shaped curve of the *activity* or *participation rate* by age, where after a short period of high rate (at entry in the labor market) the proportion of women working or seeking work begins to steadily decline up to retirement. The same graph depicts a M-shaped curve in Central and Northern European countries, characterized by high participation at entry, followed by a temporary decline during the period of motherhood and childbearing, and a subsequent comeback to work, up to a certain age where the process of definite exit starts.

In this respect, Switzerland is an interesting case. Firstly, Switzerland is a country placed in a nutshell across the Alps, which naturally divides Southern Europe from Central and Northern Europe. Secondly, there are three main languages, spoken by people living in three geographically distinct regions: French in the western part on the border with France, German in the northern and eastern parts on the border with Germany and Austria, and Italian south of the Alps in a region leading to Italy. The existence of three regions, with highly distinctive historical, social and cultural backgrounds and characters, and the fact that the Italian-speaking part is divided from the other two by the Alps highlight the very specific particularity of this country for a cross-cultural analysis of the female participation in the labor market. Moreover, the fact that the comparative analysis is performed amongst regions of the same country guarantees, despite differences stemming from the Swiss federal system, a higher degree of comparability on a large series of institutional, political and other factors than one would get with cross-country studies.

The idea of the research project was to verify the existence of differing cultural models of female labor market participation, by analysing activity rates

and hours worked per week — in terms of proportions of full-timers and part-timers — across the three linguistic regions in Switzerland, by using the SPC 2000 data.

To shortly describe the data, we can say that the Federal Statistical Office made us available a clean census dataset covering the about 7 millions inhabitants of Switzerland. For our study, only the about 3.5 millions women were indeed of interest. In the preprocessing step we disregarded young ($< 20,23\%$) and elderly ($> 61,18\%$) women, as well as non Swiss women not born in Switzerland (1.6%), i.e. about 43% of the women. This left us with about 2 millions cases. Finally, we dropped about 350000 cases with missing values, and hence included 1667494 cases into the analysis.

The empirical research design. The research procedure used classification trees at two different stages, with differing but complementary purposes. A tree was first grown in what we refer to as the *preliminary step*. Its main goal was to find a sound partition of the analysed population into a limited number of homogeneous groups — homogeneous female labor supply behavior in terms of activity and choice between full-time and part-time employment — over which a tailored analysis could be performed. This first step was run on the whole Swiss female population of age 20 to 61, using their *labor market status*[3] as outcome variable, and general socio-demographic characteristics (civil status, mother/non mother, ...) as predictive attributes. From this, a robust partition in three groups was chosen: the *non-mothers*, the *married or widowed mothers*, and the *divorced or single mothers*. The first group is composed by 609,861 women (36.6%), the second one by 903,527 (54.2%) and the third one by 154,106 (9.2%).

The second application of classification trees took place in the analysis of cross-cultural female labor supply behavior for each selected group. Here again the outcome variable was the *labor market status* of the women. A much broader series of predictive variables was retained however: age, profession, educational level, number of kids, age of last-born kid, type of household, etc. Classification trees have been produced separately for each region and then compared in order to analyse cultural patterns in the participation behavior of the three main language regions in Switzerland.

It is worth mentioning here that the final trees retained are simplified versions of those that resulted from the stopping and pruning criteria. They were selected on the basis of comprehensibility and stability factors. We checked for instance that the splits retained stayed the same when removing randomly 5% of the cases from the learning data set.

Results. In order to identify cultural models of female labor supply, three trees (one per region) were generated for each group. These — in combination with

[3] Labor market status is a categorical variable with four values: full-time active (at least 90% of standard hours worked per week), long part-time active (50% to 90%), short part-time active (less than 50%) and non active, where active means working or seeking for a job.

the results of the traditional bivariate analyses — were compared and thoroughly analysed in terms of structure and results. We give hereafter a very brief overview of the main results for the third group, i.e. divorced or single mothers. For details interested readers may consult the research report [3, 4].

A first obvious outcome is that opting for inactivity is much more frequent in the Italian speaking region. Italian speaking mothers without high education who have a last-born child less than 4 are most often inactive. For the other regions, part time activity is typical for women with a child less than 14, who either have at most a medium education level or work in fields such as health, education or sciences. A single overall splitting threshold at 54 for the age of the mothers in the Italian speaking region tends to confirm that this region conforms to the reversed L-shape of the Mediterranean model. In comparison, for the French speaking region, the major difference regarding the age of the mothers concerns only those whose last born child is more than 14. Among them, those who have a low or medium education level stop working at 60.

3 Explanatory power of our non-classificatory trees

Table 1 reports some of the quality figures we have computed for each of the three regional trees for divorced or single mothers: CHI for the Italian speaking, CHF for the French speaking and CHG for the German speaking region. The figures reported are q the number of leaves of the tree, c^* the number of different observed profiles in terms of the retained predictors, $D(m_0|m)$ the deviance between the induced tree and the root node, which is a likelihood ratio Chi-square measuring the improvement in explanatory power of the tree over the root node, d and sig, respectively the degrees of freedom and the significance probability of the Chi-square, Theil's uncertainty u, i.e. the proportion of reduction in Shannon's entropy over the root node, and $\sqrt{u}$ its square root, which can be interpreted as the part of the distance to perfect association covered by the tree. For technical details and justifications on the measures considered see [7].

The deviances $D(m_0|m)$ are all very large for their degrees of freedom. This tells us that the grown trees make much better than the root node and, hence, clearly provide statistically significant explanations. The Theil uncertainty coefficient u seems to exhibit a low proportion of gain in uncertainty. However, looking at its square root, we see that we have covered about 25% of the distance to perfect association. Furthermore, the values obtained should be compared with the maximal values that can be achieved with the attributes considered. These are about .5, i.e. only about twice the values obtained for the trees. Thus, with

Table 1. *Trees quality measures*

	q	c^*	n	$D(m_0\|m)$	d	sig.	u	$\sqrt{u}$
CHI	12	263	5770	822.2	33	.00	.056	.237
CHF	10	644	35239	4293.3	27	.00	.052	.227
CHG	11	684	99641	16258.6	30	.00	.064	.253

the grown trees that define a partition into q classes only instead of c^* for the finest partition, we are about half the way from the finest partition.

4 Conclusion

The experiment reported demonstrates the great potential of classification trees as an analytical tool for investigating socio-economic issues. Especially interesting is the visual tree outcome. For our study, this synthetic view of the relatively complex mechanisms that steer the way women decide about their participation in the labor market provided valuable insight into the studied issue. It allowed us to highlight cultural differences in the interaction effects of attributes like age of last-born child, number of children, profession and education level that would have been hard to uncover through regression analysis, for example.

It is worth mentioning that generating reasonably sized trees is essential when the purpose is to describe and understand underlying phenomenon. Indeed, complex trees with many levels and hundred of leaves, even with excellent classification performance in generalization, would be too confusing to be helpful. Furthermore, in a socio-economic framework, like that considered here, the tree should make sense from the social and economic standpoint. The tree outcomes should therefore be confronted with other bivariate analyses and modeling approaches. Our experience benefited a great deal from this interplay.

References

[1] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification And Regression Trees.* Chapman and Hall, New York (1984)
[2] Kass, G.V.: An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* **29** (1980) 119–127
[3] Losa, F.B., Origoni, P.: Partecipazione e non partecipazione femminile al mercato del lavoro. Modelli socioculturali a confronto. Il caso della Svizzera italiana nel contesto nazionale. Aspetti statistici, Ufficio cantonale di statistica, Bellinzona (2004)
[4] Losa, F.B., Origoni, P.: The socio-cultural dimension of women's labour force participation choices in Switzerland. *International Labour Review* **144** (2005) 473–494
[5] Murthy, S.K.: Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery* **2** (1998) 345–389
[6] Reyneri, E.: *Sociologia del mercato del lavoro.* Il Mulino, Bologna (1996)
[7] Ritschard, G.: Computing and using the deviance with classification trees. In Rizzi, A., Vichi, M., eds.: *COMPSTAT 2006 - Proceedings in Computational Statistics.* Springer, Berlin (2006) forthcoming
[8] Ritschard, G., Zighed, D.A.: Goodness-of-fit measures for induction trees. In Zhong, N., Ras, Z., Tsumo, S., Suzuki, E., eds.: *Foundations of Intelligent Systems, ISMIS03.* Volume LNAI 2871. Springer, Berlin (2003) 57–64