Schmollers Jahrbuch 127 (2007), 127 – 138
Duncker & Humblot, Berlin

# Determinants of Fertility – An Application of Machine Learning Techniques

By Christin Schäfer and Christian Schmitt*

## Abstract

The paper at hand applies machine learning techniques to investigate first birth transitions. The methods do not rely an distribution assumptions and require only few preconditions for application. The results are compatible with contemporary demographic research, highlighting – among other factors – the Status of relationship, income and the distribution of labour in the family. Machine learning techniques may thus be used as explorative method in the social sciences as well as tool for an in-depth analysis in future research as they are especially suited to process large data sets.

*JEL Classification: C49, J13*

## 1. Introduction

This contribution offers an exemplification of machine learning techniques in order to reflect their value for application in the social sciences. The methods applied rely on only few statistical presuppositions while at the same time offering the ability to simultaneously process extensive sets of data. With these characteristics, machine learning techniques might proof to be a well suited tool for explorative data analysis in the social sciences. Especially research fields where theoretical foundations are controversial and areas in which research is still in the early stages might benefit from this approach.

To test such an applicability of machine learning techniques, we deploy the methodological repertoire to investigate a topic, which already benefits from fullfilling elaboration but still displays some remaining blind spots. A successful application of machine learning techniques on such a topic should produce results which offer a high level of congruency to research findings made so far in that respective field.

* We are grateful to the Federal Ministry of the Family, Senior Citizens, Women and Youth, and very particularly to Minister Renate Schmidt, for financial support for this project. Furthermore we would like to thank an anonymous reviewer for useful hints and suggestions. All remaining errors are our own.

The topic covered, pertains to the area of fertility dynamics. The research question, examined for that purpose is: What is preventing couples from fullfilling their desire to have children? To arrive at an answer, it is important to first identify the factors, describing a couple's situation when deciding *for* or *against* having a child (for an overview of up to date research findings see Lesthaeghe / Moors, 2000).

Modeling this question is, like the subject itself, complex. Classic statistical methods such as logistic regression can estimate a model, explaining the difference between mothers-to-be and women who remain childless, in practice however, these methods can only take a very limited range of factors into account. Limitations exist in the form of theoretical assumptions, in the number of variables which can be handled statistical assumptions regarding the empirical model (like distribution assumptions, e.g.). So before modeling, it is necessary to make a selection of the variables on a theoretical basis. But this selection requires focusing on certain groups of causes. So, right from the start, the possible empirical results and insights that will be achieved are restricted. In order to evade this model immanent determinism, an ex ante selection of variables has to be avoided. I.e. an available data base has to be deployed with as little limitation as possible.

„Machine learning" methods make it possible to handle very large data sets with numerous variables. Machine learning includes a whole toolbox of methods (for a more concise description see Mitchell, 1997; Hastie et al., 2003; Duda et al., 2001). We give an idea in the next section. Typically, machine learning procedures make extensive use of computational power for extracting patterns of interest and for unraveling complex interrelations in the available data. Machine learning approaches have been applied successfully in gene finding, automatic digit recognition and aspects of automatic pattern recognition, as well as in biological, chemical and medical research. We will apply this toolbox to an exciting research question in the social sciences: that of the socio-economic determinants of fertility. If the analysis is capable of highlighting already well known links between social determinants and fertility decisions, machine learning techniques could in fact pose a promising method for explorative analysis in the social sciences. Ideally the method should also unravel previously unknown patterns in fertility decisions.

## 2. Data

Our analysis is based on micro-data from the German Socio-Economic Panel (SOEP). The SOEP currently includes 20,000 adult respondants, who have given birth to nearly 4,000 children over the years.

For all the participants, we not only have data on their personal histories and socio-demographic situations over a number of years but also on subjec-

tive attitudes toward various aspects of their lives. This enables us to compile a data set for the analysis that contains observations, both at the time of the decision for parenthood as well as in previous years.

The subject of this study is the transition from childlessness to first parenthood,[1] that is, the point in time when an individual decides to start a family. Although childbearing decisions reveal different gender-specific patterns, we focus on women aged 25 to 29. Given that the advantage of the machine learning processes lies in their ability to process large data sets (i.e. many variables per case), we see the greatest potential in the evaluation of a three-year spell: This, in our view, should yield the most interesting results since it enables us to model important life patterns, existing immediately prior to the childbirth decision. Our longitudinal data set, which takes into account the situation at the time of the decision as well as spells from the previous three years, is based on data from 1990 to 2002. The data set includes observations of 6,108 female panel members in 732 dimensions, and 315 births.

## 3. The Process of Analysis

This section gives an account of the method used as part of the machine learning process.

### 3.1 Feature Selection and Variable Selection via Classification

Using the 6,108 observations in 732 dimensions (variable characteristics) at our disposal, we focus on dimensions in which differences between future mothers and women who will remain childless can be recognized.[2] In the first step of the analysis, we choose those dimensions that enable the two groups to be contrasted and differentiated most clearly.

The task of differentiating between two groups using the machine learning process is a classification problem. Our analysis starts with the problem of how to differentiate between mothers and childless women. The main emphasis here is not to achieve the best possible classification result, that is, to obtain a perfect distinction between the two groups. Rather, we want to obtain the classification function and thus acquire information on the importance of a

[1] While the decision to have a second child or more children greatly depends on the time of earlier births (see Kreyenfeld / Huinink, 2003) the analysis of the first birth offers particular insight into exogenous factors.

[2] For the sake of simplicity we refer in this paper to „mothers" and çhildless women". By mothers we mean the women of whom we know, on the basis of projections of panel data, that they will have a child within 10 months. Childless women in this differentiation are all of those who will still be childless at the end of the current period examined ($t_0 + 10$ months).

variable for classification. The aim is to use only those variables in the ensuing steps of the analysis that are of high importance for classification – i.e. variables that embody the differences between the two sub-populations. All other factors of minor importance are automatically ignored and left unconsidered in the following steps of analysis.

The classification procedure uses a linear programming machine (LPM, see Bennett / Mangasarian, 1992) that creates a linear classification function with maximal margin. Each factor in our data set forms one dimension in the data space. Therefore the linear classification function is a weighted linear combination of the factors, since what is learned are the weights for each variable. LPM provides „sparse" results. That is, optimization is performed with as few variables as possible. This provides the desired contrast between dimensions that are relevant respectively irrelevant for differentiating between mothers and childless women.

To compensate for unbalancedness – the fact that the group of childless women is much larger than the group of mothers – the LPM training is repeated one hundred times on sub-samples of the data. Each sub-sample is constructed as follows: all the observations of mothers are included in each sample while the same number of observations is chosen at random from the group of childless women in order to balance the population size. Each training of a subsample results in a sparse assignment of weights to each variable and a hundred repetitions of this process finally provide a distribution of weights.

The final selection of variables is accomplished, using a statistical test. For each variable, a test is applied, to determine, whether the average of the weight distribution assigned deviates significantly from zero. As no assumption can be made on the distribution, we have chosen the non-parametric sign test for the level of significance alpha = 5%.

Of the 732 dimensions, 25 are essential for classification. They are the basis for the ensuing examinations, and they cover the range of the results from which our conclusions will be drawn.

In Table 1 the second column shows the sum of the weights assigned to the dimension by the LPM training. The third column shows the year of the observation and last column names the variables. Does the nature of these variables in the year of the birth or one or two previous years play a role?

Table 1 shows that no variable on educational attainment proved influential.[3] Occupational qualifications are important, this is revealed indirectly by job status and amount of time since leaving full-time education or vocational education. Of importance is also whether the respondent is in a steady rela-

[3] Most likely educational attainment unfolds its effect rather in determining related variables like access to the labour market, job status and income which all proved to be influential in our analysis.

tionship or single. Variables 1 and 9 (partner's income and age) also code this status indirectly and are thus in part correlated to the partnerships status. The analysis will have to take into account whether effects measured can actually be traced to the partner's income (or age) or whether this simply constitutes an indirect measure of the *presence* of a partner.

*Table 1*

**The Variables Relevant for Classification shown by Importance**

| No of Variable | Sum (alpha) | ...years ago | Variable |
|---|---|---|---|
| 1 | 81.18 | 0 | Partner's income betw. 0 and 250 euros |
| 2 | –43.83 | 2 | I do the housework |
| 3 | 35.12 | 0 | Married |
| 4 | 23.93 | 0 | Work betw. 35 and 40 hours / week |
| 5 | 23.55 | 1 | Full-time job |
| 6 | 22.51 | 1 | Car available in household |
| 7 | –22.09 | 2 | Housewife |
| 8 | –18.28 | 2 | Married |
| 9 | 17.29 | 0 | Partner aged betw. 25 and 30 |
| 10 | 16.62 | 1 | Housework shared with partner |
| 11 | 15.75 | 2 | No fears for job security |
| 12 | 13.38 | 0 | Qualified betw. 60 and 84 months ago |
| 13 | –13.17 | 2 | Qualified betw. 84 and 120 months ago |
| 14 | 12.63 | *0* | Housework shared with partner |
| 15 | 12.04 | 0 | Very satisfied with own state of health |
| 16 | 11.11 | 0 | Qualified 12 to 24 months ago |
| 17 | 10.22 | 0 | Spent 5 to 10 nights in hospital last year |
| 18 | –9.38 | 0 | Single |
| 19 | 9.09 | 1 | Smoker |
| 20 | 7.47 | 1 | Satisfied with available |
| 21 | 6.48 | 2 | Housework shared with partner |
| 22 | 6.21 | 0 | Very satisfied with job |
| 23 | 4.68 | 0 | Very satisfied with standard of living |
| 24 | 4.62 | 1 | Very satisfied with housing |
| 25 | 3.54 | 0 | Job very important |

*Source: SOEP* 1990 – 2002; own calculations.

There are also problems with Variable 17 in the table: the number of nights spent in the hospital. Owing to the backdating of the decision for parenthood it is possible, in certain constellations, for the event of birth to be included here indirectly. It is significant that four of the dimensions shown to be important in a decision for parenthood contain patterns of behavior in the distribution of the housework. In addition, six variables refer to employment.

The analysis of the data described so far is a first step. Rather than being based on a theoretical socio-economic model, it was designed to produce an automatic selection of variables through classification by means of an LPM. In the next step, we will try to identify prototypical life paths, and then examine which of these typify the transition to motherhood and which tend to make it more difficult.

### 3.2 Clustering: Identifying Prototypes

We start from a data set with 6,108 observations in 25 dimensions. We make no distinction at all between future mothers and childless women but examine all the data. Our task is then to identify clusters of mothers in the data.

A cluster is evident when the data points contained are very similar. Observations assigned to different clusters are dissimilar. The cluster assignment and the number of clusters are undetermined. Thus we face a so-called unsupervised learning problem.

In studies of social data, the data structure is generally as little known as the answer to the question, which cluster process is best suited to the analysis. Before displaying an overview of the technique of cluster evolution – and before describing how to solve the three questions of proper clustering process, proper number of clusters, and proper cluster assignment without further assumptions – we need to say a few words about the distinctions that need to be made between prototypes and clusters.

The general methodological procedure of the clustering process is to summarize a given number of data-points into several exclusive sub-groups. The aim of the clustering process is to achieve a maximum of homogeneity within groups and of heterogeneity across groups. Clustering assigns each observation to a cluster as shown in the simple examples presented above. However, as seen in the left plot of the following diagram, the search for prototypes does not require that all of the points form a single cluster. Rather we are interested in the position and nature of the prototypes, as indicated schematically in the right plot. It is therefore not the aim of the analysis to assign each observation to a cluster and thus link it to a prototypical pattern, but instead to identify those areas in the data with a high density. These areas define our prototypes.

This procedure appears to be the most appropriate means of approaching the crucial issue of the present study. While many women share similar life his-

tories, certainly not every life history is prototypical. Furthermore, attempting a mathematical formulation of the learning problem described above would only be able to provide an approximate solution to this complex question.

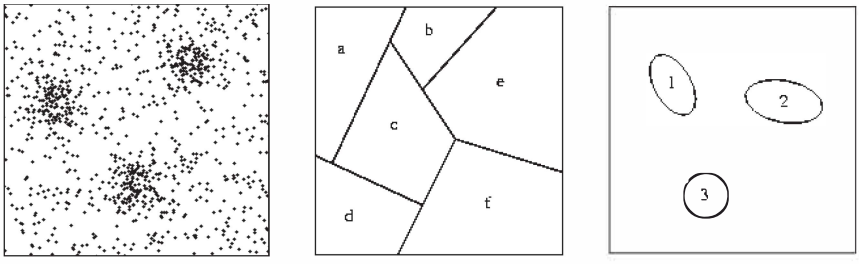


Figure 1: Cluster prototypes selection

Figure 2 shows how the cluster evolution method works; it is inspired by the work of Blatt et al. (1996). Instead of specifying the unknown quantities (number of clusters, cluster assignment, cluster process and cluster observations) ex ante, any possible cluster process is used, and the solutions are calculated successively with a growing number of clusters, while the development of the number of observations assigned to the largest, second-largest, third-largest, etc. cluster is recorded. In most practical applications, individual observations each form a cluster on their own initially. Only after all the deviant single observations have their own cluster, the structure of interests is revealed in the form of a cloud of observations. In the above schematic diagram, this is done with the Sten cluster solution which first recognizes and separates the two Gaussian clouds that are arranged close together in the center of the observed area.

As a solution, a number of clusters that provides a stable result is chosen (in the above plot, the area between $k = 5$ and $k = 11$). It should be kept in mind that it is not our aim to assign as many data points as possible to a prototypical cluster. Rather, we want to find *prototypical* clusters whose specific characteristics appear in particularly clear contrast to others.

Hence, if a solution is chosen – for example the solution shown with $k = 11$ clusters – all the points in the two major clusters are classified as prototypical observations. All the other observations are regarded as remaining and unassigned clusters and left out of any further consideration.

Let us return to our study. The cluster evolution of childless women aged 25 to 29 in the three-year spell data is less clear than the schematic representation above. The figure below only shows those paths in the cluster evolution that pass the 5 % limit (horizontal line), that is, containing more than 306 observations in at least one solution (out of 6,108 observations in total). We intro-

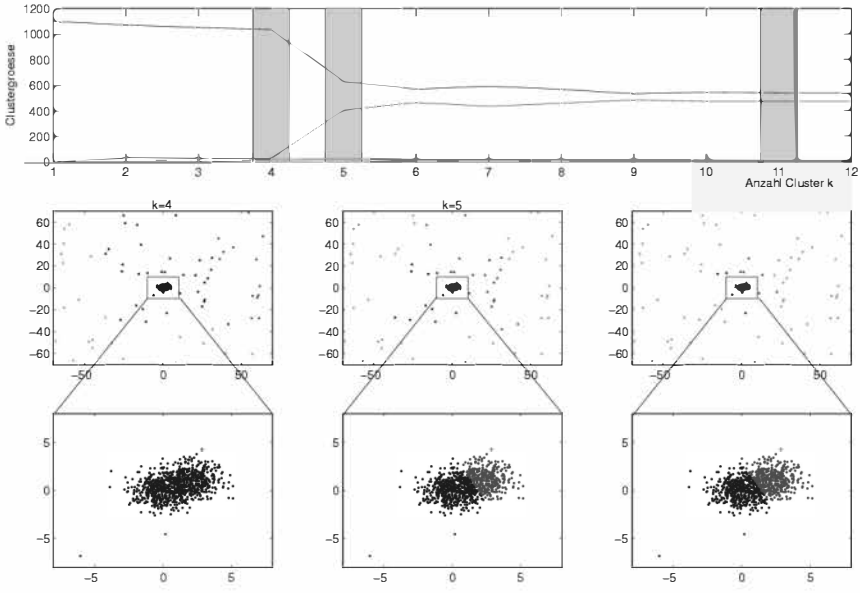


Figure 2: Progression with cluster evolution

duced the 5 % limit in order to guarantee a minimum size for possible prototypical clusters. Choosing a solution with $k = 24$ and using the k-means clustering algorithm, five prototypical clusters are revealed. The quality of this solution was confirmed using a cluster stability analysis (for more details on this process see Roth et al., 2002).

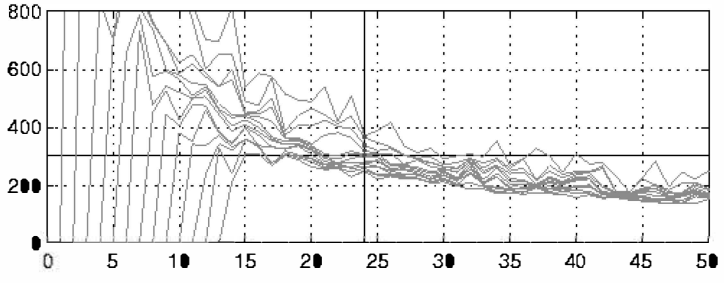

Figure 3: *n* of clusters selection

This step basically concludes the procedure. However, it is still not clear how the prototypical clusters differ in contents. In concrete terms this means: How do typical life histories and structural conditions appear in the context of the transition to parenthood within these clusters?

### 3.3 Extraction of characteristic patterns

In order to reveal the characteristics of the individual prototypes and select relevant data, we use a number of explanation and visualization techniques as well as statistical tests. For the latter, a hierarchical approach is used:

First, a binomial test determines whether the features of a variable differ significantly among the prototypes. If not, the variable is excluded from further tests. In the second step, we analyze whether the features of the remaining variables of a prototype differ significantly from the rest of the observations, or if the differ significantly from another cluster. For example, it is possible for all the childless women in one cluster to be married, to be in full-time jobs, or to be single. Our analysis reveals that in our study, 24 out of 25 dimensions are relevant for the differentiation between prototypes.

Using principal component analysis (PCA) and correspondence analysis (CA, see Nakayama, 2001), an embedding of data is undertaken to allow a visualization of the similarities between observations (PCA) or between observations and variables (CA). The center of a cluster represents the prototype in question. A decision tree procedure is used to refine the selection (see Blanchard et al., 2006). The next section gives examples of the most important results obtained.

## 4. Empirical Results

The five prototypes extracted in the preceding analysis are distinguished by the following characteristics. We have chosen those variables for which clear differences are most evident. The selection is determined by the test statistics of the binomial test mentioned in the preceding section.

*Table 2*

| | PT 1 | PT 2 | PT 3 | PT 4 | PT 5 | Rest | | Variable |
|---|---|---|---|---|---|---|---|---|
| | 372 | 366 | 340 | 321 | 306 | 4403 | | # Obs |
| 1 | +-+ | – | – – | – – | – – | o | 0 | Married |
| 2 | +-+ | – – | – – | – – | – – | o | 2 | Married |
| 3 | ++ | - | - - | + + | - - | o | 0 | Joint household |
| 4 | +-+ | – – | – – | + + | - - | o | 1 | Joint household |
| 5 | ++ | – | – – | + | – – | o | 2 | Joint household |
| 6 | + | o | - - | + + | o | o | 1 | Full-time job |
| 7 | – – | – – | o | – – | o | – | 0 | Single |

PT = Prototype; + + Very frequent feature, + frequent, – seldom, – – very seldom observed.
*Source: SOEP* 1990 – 2002; own calculations.

The first prototype contains mainly childless women who have been married for a long time; most of them work full-time and share the housework with their partner, so the role division in for these couples is an egalitarian one. Prototype 4 is very similar to Prototype 1; the difference between them is that in Prototype 4 there are no married childless women, but again egalitarian living arrangements. Prototypes 3 and 5 are quite different: they contain a very high share of childless women living alone, so by definition there can be no egalitarian distribution of labour within their household. The difference between these two clusters is in the extent of employment. Prototype 2 is very similar to the two single clusters – these are unmarried childless women, most of whom have a partner but who widely perform the housework in a traditional role setting i.e., they specialize in housework (women) and gainful employment (men).

*Table 3*

| PT | # Obs | # Event | Prob(event) | Test statistics | Significance |
|---|---|---|---|---|---|
| 1 | 372 | 33 | 0.0887 | 3.2388 | + + |
| 2 | 366 | 20 | 0.0546 | 0.2658 | o |
| 3 | 340 | 5 | 0.0147 | –3.0737 | + + |
| 4 | 321 | 27 | 0.0841 | 2.6361 | + + |
| 5 | 306 | 6 | 0.0196 | –2.5282 | + + |
| Rest | 4403 | 224 | 0.0509 | –0.2092 | o |
| Population | 6108 | 315 | 0.0516 | – | – |

*Source: SOEP* 1990 – 2002; own calculations.

The following picture emerges in regard to our central question of the conditions which form the framework for the transition to first motherhood:

Prototypes 3 and 5 contain childless women who are significantly less inclined toward parenthood than the childless women in the other prototypes.[4] This is not surprising insofar as most of these are childless women living alone. But it is surprising that the likelihood of motherhood is clearly higher for Prototypes 1 and 4, who are either married or unmarried. It becomes clear that the variables on the share of household work are major indicators of a positive or negative environment to perform the transition to parenthood. Prototypes 1 and 4 have in common joint housekeeping and employment – in contrast to Clusters 2, 3 and 5, which, like Cluster 4, contain unmarried childless women. However, unlike the childless women in Prototypes 1 and 4, these childless women do not share the housekeeping with a partner.

[4] As based on the binomial test.

It should be noted that the close relations of housekeeping to the transition to motherhood was also revealed to be a decisive factor for the childless women in the analysis of the age groups 20 to 24, and 30 to 34, which is not discussed in more detail at this juncture.

## 5. Conclusion

The application of machine learning techniques to a research question of fertility dynamics could prove that the employed methods have a high potential for explorative data analysis in empirical social research. According to our analysis, the factors, which determine fertility decisions, include the status of the relationship, income, the role of qualifications as well as labour market participation. Although these results are far from being unexpected or spectacular, they are in perfect compliance with contemporary demographic research. This proves that the method of machine learning is capable of highlighting *relevant* factors without extensive theorizing or a substantial elaboration of the research topic.

While refined theorizing and in depth consideration of previous findings remain indispensable elements of social research, machine learning techniques are a promising method to offer impulses and initial guidelines in areas, where development of theory and empirical research has been widely neglected. This consideration is also underlined by the fact that our application of machine learning techniques stressed the significant importance of the distribution of labour in the household. While this area of gender roles and their impact on fertility has long been an area of controversial discussion in demographics (see e.g. McDonald, 2000), the empirical investigation of this relation remains underexposed. In that sense another role of machine learning techniques might be found in pointing out topics which might prove to be fruitful for future research.

## References

*Bennett,* K. P. / *Mangasarian,* O. L. (1992): Robust linear programming discrimination of two linearly inseparable sets, Optimization Methods and Software 1, 23 – 34.

*Blanchard,* G. / *Schäfer,* C. / *Rozenholc,* Y. / *Müller,* K.-R. (2006): Optimal Dyadic Decision Trees, Machine Learning, published online: http: // eprints.pascal-network.org / archive / 00001324 / 01 / BlaSchRozMue06. pdf

*Blatt,* M. / *Wiseman,* S. / *Domany,* E. (1996): Super-parametric clustering of data, Physical Review Letters 76 (18).

*Duda,* R. O. / *Hart,* P. E. / Stork, D. G. (2001): Pattern Classification, New York.

*Hastie,* T. / *Tibshirani,* R. / *Friedman,* J. H. (2003): The Elements of Statistical Learning, Berlin.

*Kreyenfeld,* M. / *Huinink,* J. (2003): Der Übergang zum ersten und zweiten Kind – Ein Vergleich zwischen Familiensurvey und Mikrozensus, in: W. Bien, / J. H. Marbach, Partnerschaft und Familiengründung – Ergebnisse der dritten Welle des Familien-Survey, Opladen, 43 – 64.

*Lesthaeghe,* R. / *Moors,* G. (2000): Recent Trends in Fertility and Household Formation in the Industrialized World. Review of Population and Social Policy 9, 167.

*McDonald,* P. (2000): Gender Equity Theories of Fertility Transition, Population and Development Review 26, 427 – 439.

*Mitchell,* T. M. (1997): Machine Learning, McGraw Hill.

*Nakayama,* T. (2001): Tests for redundancy of some variables in correspondence analysis, Hiroshima Mathematical Journal 31, 1 – 34.

*Roth,* V. / *Braun,* M. / *Lange,* T. / *Buhmann,* J. (2002): A resampling approach to cluster validation, Computational Statistics – COMPSTAT'02, 123 – 128.