Demographic Analysis of Fertility Using Data Mining Tools

Matjaž Gams and Jana Krivec
Department of intelligent systems
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: matjaz.gams@ijs.si, jana.krivec@ijs.si

Keywords: demography, fertility, data mining

Received: March 12, 2008

We used data mining techniques to discover which attributes have the highest impact on country fertility rates. The data was analyzed in various ways; altogether and joined in smaller, meaningful groups, such as sociological, economical, philosophical, biological, etc.. We separately analyzed different groups of countries, current state and fertility trends, and tested several class schemas. Most relevant decision trees are presented and interpreted showing some known and some new conclusions. The iterative use of data mining techniques again proved to be successful in finding complex relations, but still needing expert interpretation as any computer method.

Povzetek: Analizirani so poglavitni razlogi za premajhno oz. preveliko rodnost.

1 Introduction

Populations change through three major processes: fertility, mortality, and migration. A useful way to express the rate at which women have children is the Total Fertility Rate (TFR). TFR is the average number of children that would be born per woman if all women lived to the end of their childbearing years and bore children according to a given set of age-specific fertility rates [21, 3]. If the average woman has approximately 2 children in her lifetime, this is just enough to maintain the population [7].

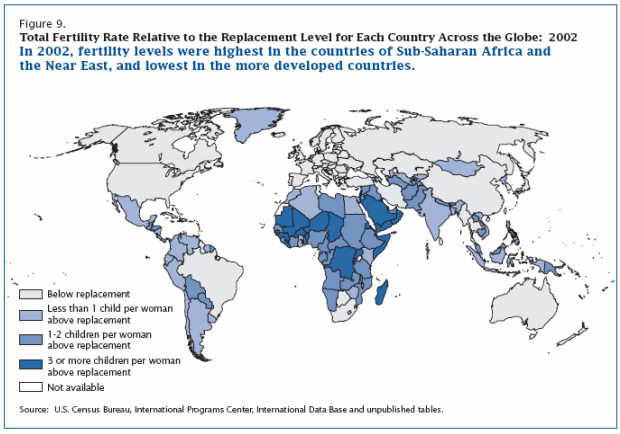

Figure 1: TFR in countries in 2002 [21].

As seen in Figure 1, some countries have high and some low TFR. In most European countries TFR in 2006 was below 1.5 children per women [19], which is far less than desired [15]. Namely, sustained low fertility rates can lead to a rapidly aging population and, in the long-run, may place a burden on the economy and the social security system because the pool of younger workers responsible for supporting the dependent elderly population is getting smaller. Tracking trends of fertility rates and factors that influence them helps to support effective social planning and the allocation of basic resources across generations [8].

In this article we present a demographic analysis of 147 worldwide countries described by 95 basic attributes that might affect fertility rate. Even though the idea is everything but new, our approach to this problem is. Namely, our research group has decades of experiences in developing and using data mining (DM) and machine learning (ML) systems such as Weka [25] and Orange [4], last being developed in our broader research group. We typically approach a problem domain in a specific manner, usually obtaining similar results than those of the best experts in the field. It was a particular challenge to test our methods on the demographic problem.

We use terms attributes, indicators and factors as synonyms.

2 Related work

So far scientific efforts in demography were devoted mainly to exploration and definition of the process of data collection and qualitative interpretation of the statistical results, consequently not putting emphasis on new data analyzing methods. Data is typically analyzed with event history regression methods, Markov transition models and Optimal matching method using common spread statistical packages like (SPSS, SAS, S-Plus, Stata, R, TDA, etc.) [14]. The hypothesis is that between these typical aggregate descriptions and causal analysis there is a deficit of research on complex relations. Several modern methods, including data mining, offer opportunities to fill this gap.

In the last decades, data mining tools for knowledge discovery from data (KDD) proved successful in various

fields. However, searching through the internet showed that these approaches have received little attention in demographic analyses. There are some publications, e.g. Blockeel et al. [2] showed how mining frequent item sets may be used to detect temporal changes in event sequences frequency from the Austrian FFS data. In Billari et al. [1], three of the authors experienced an induction tree approach for exploring differences in Austrian and Italian life event sequences. Oris et al. [12] initiated social mobility analysis with induction trees. Unlike the statistical modeling approach, the methods make no assumptions about an underlying process generating the data and proceeds mainly heuristically. The approach differs from ours because we study rather static data and do not yet apply sequential rule mining analysis on historical demographic data.

We have not noticed DM analyses on the level of countries, similar to ours.

3 Data mining for demography

Successful data mining is based on various investigations of the data using different methods, parameters, and data to find most meaningful relations.

3.1 Basic data description

Data for machine learning and data mining are most commonly presented in attribute-class form, i.e. in a "learning matrix", where rows represent examples and columns attributes [22]. In our case, an example corresponds to one country, and a class of the country, presented in the last column, denotes fertility rate. The first attribute is the name of the country. Alltogether there are 95 basic attributes and 147 countries. Attributes and their values were partially obtained from the demographic sources such as UN [20], Eurostat [5], and the Slovenian statistical database [16]. Several of the attributes were obtained from the internet, based on the assumption that they might show some interesting demographic relation. We were trying to get as many attributes as possible, nondiscriminatory whether positive or negative in terms of fertility rate.

Attributes in demographic literature are grouped into biological and social [9] since human fertility is a socially formed biological process [18]. Newer literature introduces more and more complex structures, based on detailed grouping of social factors. Malai [10] divided factors that impact fertility rate in six groups: (1) biological, (2) economical, (3) social, (4) cultural, (5) anthropological and (6) psychological. Our 95 basic attributes correspond to these six categories, e.g.: state politics towards maternity leave, homosexuality, religion, suicide, abortion, military etc. Some of our measurements were performed on specific groups like (2) economical, consisting of 12 attributes like unemployment rate, GDP ($) per habitant, GDP growth (%). Biological factors (1) include 6 attributes (number of habitants, life expectancy rate, number of men per 1000 women to mention a few of them). On top of these six we added a special category "education and R&D (research and development)" with 38 attributes. There are 11 binary attributes, 2 discrete and the rest numerical.

For the basic class we have chosen Total Fertility Rate (TFR), discretized into two values: high (>2) an low (<2). The branching point 2 was chosen because it represents the replacement level of the population. In reality, replacement level is a bit higher, around 2.1, but this number depends on several other parameters such as mortality rate and immigrations, and furthermore only two countries have fertility rate between 2 and 2.1.

3.2 Data modifications

3.2.1 Attribute modifications

By attribute modifications we denote eliminating some columns in the learning matrix, and adding new columns, i.e. attributes. Subgroups of columns were chosen based on the demographic categories, and by DM methods. There were 5 new attributes added during the process of DM, thus bringing the total number of attributes to 100. Around half of the experiments were performed on 100 attributes.

3.2.2 Class modifications

Besides the basic class discretization into two values, we tried three values of TFR as well: low (< 2), middle (2-3) and high (>3).

In another attempt we classified countries according to decrease or increase of TFR. We first calculated average UN predicted TFR for years 2005-2010 and subtracted average TFR for years 2000-2005. The obtained value was discretized into two classes: $\Delta TFR>0$, $\Delta TFR\leq 0$; or three classes: decreasing ($\Delta TFR<0.5$), stable ($-0.5<\Delta TFR<0.5$), and increasing ($\Delta TFR>0.5$).

3.2.3 Modifications of learning examples

Learning examples consisted of 147 countries, each represented by a row in the learning matrix. Modifications were performed as eliminating or choosing specific rows to form a new learning matrix. A typical example would be a subgroup "developed countries", consisting only of countries with high gross domestic product (GDP) or Failed States Index (FSI). GDP is defined as the total market value of all final goods and services produced within a given country or region in a given period of time (usually a calendar year) [24]. FSI on the other hand consists of several attributes, describing the strength of central government, provision of public services, level of corruption and criminality; percentage of refugees and involuntary movement of populations, and an amount of economic decline. Since 2005, the index has been published annually by the United States think-tank, the Fund for Peace and the magazine Foreign Policy [23]. GDP review extracted two groups of countries: well developed countries with GDP above 1000$ per habitant (39 countries), and developing countries with GDP less than 1000$ per

habitant (108 countries). Examination of FSI revealed three groups of countries: developed with FSI lower then 39.45 (29 countries), moderately developed with FSI 39.45-61.4 (21 countries), and developing with FSI>61.4 (97 countries).

We also prepared data for analysis based on the geographical region of the country. In correspondence with UN regional classification [21], we grouped our cases in 6 regions: Asia, Africa, Latin America and the Caribbean, Oceania, Europe and Northern America, and 20 sub regions.

3.3 Methods

Machine learning and lately data mining are among the most successful artificial intelligent application areas. Whenever there are lots of learning examples, these systems learn properties of the domain and make predictions about future cases. These systems not only compete with statistical methods in terms of accuracy, they also introduce several new approaches such as cooperation between systems and humans. The constructed knowledge is often in the form of readable, understandable trees, rules and other representations thus enabling further study and fine tuning. Two examples of successful scientific and engineering DM tools are Weka [27] and Orange [4]. Both systems provide tens of DM systems, several data preprocessing and visualization tools. From the ML and DM techniques available in Weka and Orange we have chosen J48, the implementation of C4.5 [25], a method used for induction of classification trees. This method is most commonly used when the emphasis is on transparency of the constructed knowledge. In our case this was indeed so, since the task was to extract most meaningful relation from hundreds of constructed trees.

Most meaningful relations are those most significant to humans with best classification accuracy at the same time. To estimate the accuracy of the trees, we used 10-fold cross-validation, built in the system. The estimated accuracy of a classification tree corresponds to a probability that a new example will be correctly classified.

A short description of decision trees is presented in this paragraph for readers not familiar with classification trees. Classification trees are built in a top-down manner. The first task is to choose the most informative attribute which will be placed at the root of the classification tree. The next step is to add branches according to the values of the attribute. For a discrete attribute, there are as many branches as there are different values. In case of a numeric attribute, there are only two branches, one that represents values less or equal than the border value as proposed by the system, and the other branch with greater values. The set of examples is divided into subsets corresponding to the branches. Now the process can be repeated recursively for each branch, using only those instances at each particular branch. If at any time all instances at a node have the same classification, further branching is stopped and the classification into that class is proclaimed. The splitting process is usually stopped as soon as sufficient statistical significance is obtained, classifying into the majority class. Classification is performed by starting at the top of the tree and choosing appropriate attribute values to proceed with the chosen branch. At the leaf, the numbers represent all examples and those with different class.

Experiments were performed with various method parameters, mainly changing levels of pruning. However, it turned out that default parameters were most successful.

Unlike in our typical DM session, we did not modify sources of the DM programs. J48 turned out successful enough.

4 Experiments

Tens of trees were created in a systematic way, as presented in Figure 2. First experiments were performed with TFR and ΔTFR, then with all and only developed countries. Finally, several selections of the attributes were tested: all, economical, direct, social, economical, and educational. These tests resulted in 24 basic trees. In addition, various further experiments were performed.

Due to lack of space, only most interesting trees are presented in this paper, those with most meaningful relations to humans and with best classification accuracy at the same time.

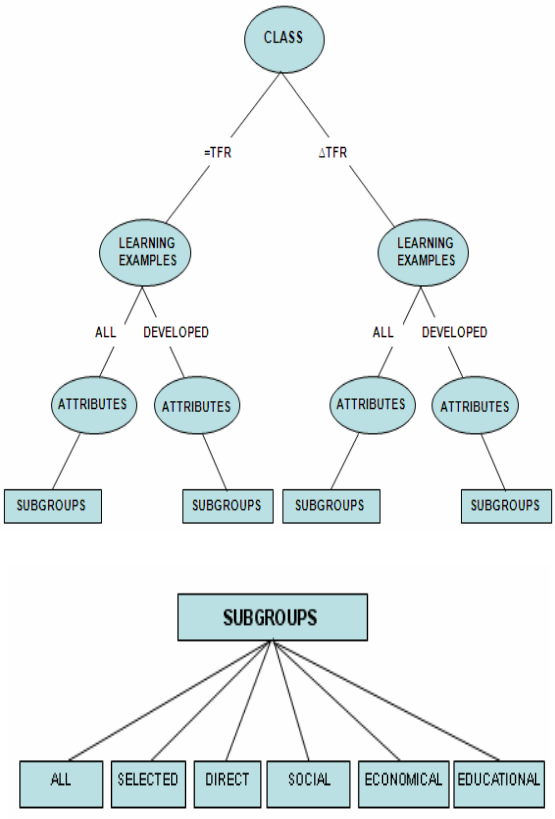

Figure 2: Structure of experiment.

4.1 TFR class

Firstly, the analysis was based on TFR as a class, with 2, 3 or more values. Only experiments with 2 or 3 values were interesting enough to be presented in this paper.

4.1.1 All countries

4.1.1.1 All attributes

In the first fertility rate analysis all 147 countries and all 95 available attributes were taken into consideration. The obtained tree is presented in Figure 3, showing that the most important indicator for high TFR is the number of stillborn children per 1000 births. More than 11.5 stillborn children per 1000 births is a strong supporting factor in favor of high TFR of the country and vice versa. The results are consistent with practically all literature in the demographic field and experts' opinions, who claim that death of newborns is in tight connection with social and economical status of mothers who need to have several children to compensate for those dead. According to experts, higher educated mothers usually have less children and lower newborn mortality, low percentage of stillborns is supposed to be related to the costs of child life-support , different life condition of the urbanized and industrialized society, changes of the attitude towards women, decaying of old patriarchal community etc. as the main reasons for fertility decline. As the tree in Figure 3 shows, these relations are indeed statistically most relevant. However, the tree shows additional relations in a structured way with appropriately weighted leaves, i.e. nodes at the bottom of the tree. For example, the top right leaf "high (104/16)" includes 88 countries with high TFR and 16 with low TFR. The bottom left leaf, on the other hand, encapsulates only 2 countries with high TFR, rendering this information as statistically less important. Therefore, in the tree there is just another statistically strongly confirmed relation: when number of dead born children is less than 11.5 and majority religion is Christianity and there are fewer men than women then TFR is low (35/1). This relation shows another crucial matter regarding interpretations of the tree. Why should Christian majority be negative for fertility rate while Christians give high emphasis on families, strong marriages and devotion to children? Indeed, further analysis show, as pointed out by demographic experts long time ago that population in these countries have high divorce rates etc. meaning that people do not follow church directions, but live according to their own desires. The bottom right part of the tree, starting with low percentage of women in the population is statistically rather meaningless, however, density and number of inhabitants gives some indication that these are among relevant attributes. Therefore, reading and interpreting trees demands some understanding of statistics, trees and demographic literature.

At each Figure title, there is cross-validation accuracy estimate. For Figure 3 it is over 80%, which is a reasonably good result. Default accuracy obtained by classifying only into the majority class is 89,4%.

In another attempt we divided the class values into three groups: low, moderate and high TFR rate. Immediately it should be noted that the accuracy of such a tree seemingly decreases. Namely, the tree in Figure 3

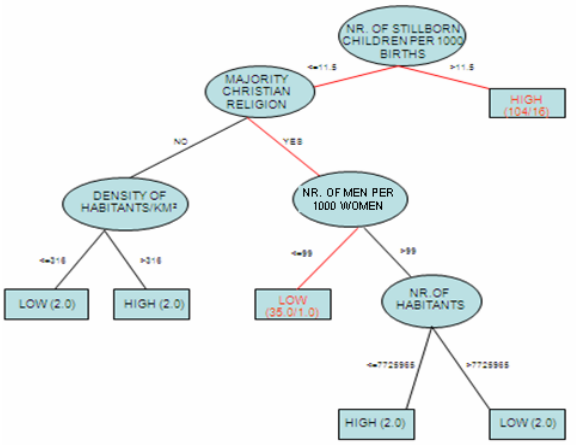

Figure 3: Two-valued TFR classification tree, all attributes (accuracy is 80,3%).

classifies into two classes, therefore, accuracy of blind guessing is 50%. For three classes, blind guessing results in 33% accuracy, and the default accuracy is 36,7% for the majority class. Having these statistics in mind, the obtained classification tree in Figure 4 achieves even better classification accuracy with 74,8%.

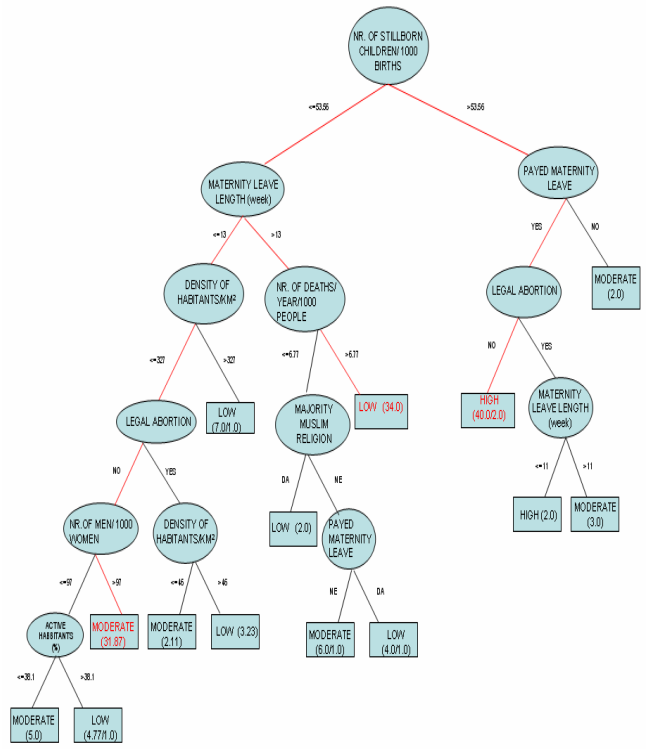

Figure 4: TFR classification tree with three-valued class, considering all attributes (accuracy is 74,8%).

The experiment once again revealed the most important attribute: "number of stillborn children". However, the branching point leading to high TFR is in this case much higher: 53.56 children per 1000 births. In this tree, there are three major groups all from 30 to 40 countries: high, moderate and low. The major attribute distinguishing between moderate and low TFR countries is the length of the maternity leave. At this point one should be aware that such attributes are semantically potentially misleading - countries with low TFR probably introduced lengthier maternity leave as a consequence and not as cause. The tree therefore shows most important relations without knowing the nature of them.

After obtaining the first tree, in a series of tests seemingly most important attributes are being eliminated in order to test if other attributes can replace them and still obtain similar accuracy. Instead of "number of stillborn children" several attributes can be used: human development index (HDI), life expectancy rate, literacy rate, etc. all denoting the same concept. It is generally accepted that in these, developing countries, TFR is high.

For the maternity leave, the elimination of the attribute results in lower accuracy 68%. Although this attribute is obviously important, we are not able to establish the type of relation. Whatever the case, countries with short maternity leave have moderate TFR, and those with long maternity leave low TFR.

Although the rest of the relations are not so significant, they represent a bigger share than in the previous tree and they seem to have two common denominators: developmental status and value system.

Altogether, analysis so far indicate that the developed countries have low TFR, e.g. most of the European and north American countries, developing countries have high TFR, and moderately developed countries like Botswana, Bolivia, Honduras, Jamaica, etc. have moderate TFR.

4.1.1.2 Selected attributes

We further filtered attributes according to the algorithms in DM tools. Again, as seen in Figure 5, the most distinctive attribute regarding TFR rate appears to be the number of stillborn children per 1000 births. When this number is lower or equal to 11.55, the TFR is low (under 2), with the exception of the countries that do not ensure appropriate delivery treatment and invest most of its educational foundation in a primary sector.

On the other hand, TFR is low despite high number of stillborn children in the case when the human development index (consists of life expectancy rate, literacy rate, educational rate and standard of living) is high, abortion is allowed and unemployment rate is low (under 13.9 %), or if abortion is not allowed, but the country invests most of its educational foundation in a primary sector and has long maternity leave (more than 11 weeks). The discovered relations indicate a meta attribute - developmental status of the country.

4.1.1.3 Direct attributes

The demographic experts classify fertility attributes, i.e. factors, on direct and indirect [10]. Direct factors have direct influence on fertile persons. In this context we built a decision tree including 4 attributes: legality of demanded abortion, number of abortions per 1000 people, percent of married women (between 15 and 49 years old) that use contraception and percent of elders infected with HIV virus or AIDS. The obtained 82,31% accurate tree is presented in Figure 6. Legal abortion associated with low percent of HIV infected elderly relates to low TFR while illegal abortion and lower percent (less then 70) of women using contraception leads to high TFR. These attributes again seems to correlate to the meta attribute - developmental state of the country and to the value system. The other derivation could be that the value system plays an important role. The accuracy is very high indicating that these attributes are meaningful.

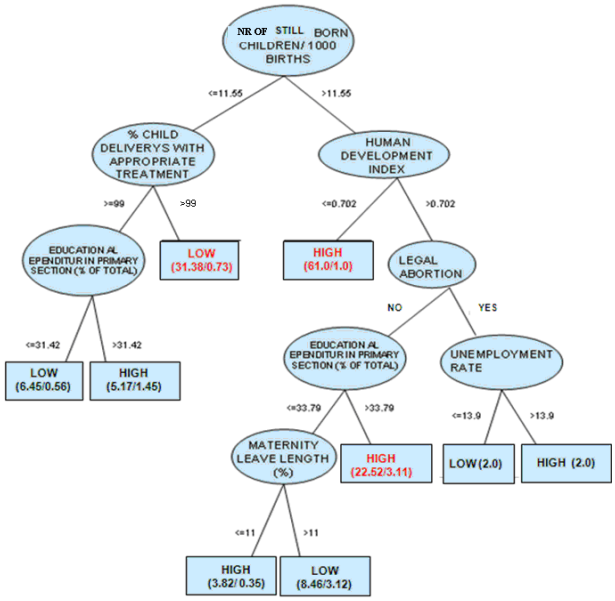

Figure 5: TFR classification tree with two-valued class, considering only automatically selected attributes (accuracy is 81,6%).

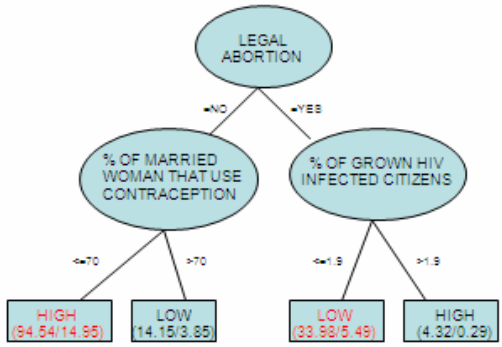

Figure 6: TFR classification tree with two-valued class considering only direct attributes (82,3%).

4.1.1.4 Social attributes

Since many experts in the field agree [10] that only direct factors can not explain the fertility rate determination, we further examined influence of the indirect TFR factors. We analyzed 11 attributes that express the society attitude towards general life questions: legality of homosexuality, legality of homosexual marriages, possibility of adoptions to homosexuals, number of suicides per 10000 persons (men only, women only, altogether), legality of abortion, number of abortions per 1000 people, number of divorces per 1000 persons, percent of women in the parliament.

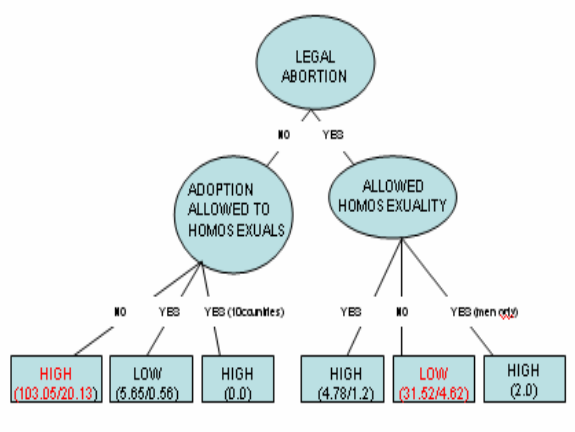

Figure 7: TFR classification tree with two-valued class, considering only social attributes (81,6%).

From Figure 7 one can conclude that TFR is high in more conservative countries that don't allow abortion and adoptions to homosexuals. TFR is high also in the countries that allow abortions but prohibit homosexuality. By this view more liberal countries have low TFR. The accuracy is as high as of the tree constructed on all the attributes, selected by the system. Again it seems that the value system plays an important role.

4.1.1.5 Economical attributes

Experts generally find low TFR strongly related to the economical factors, society modernization and liberalization [19]. We wanted to established the nature of economic relations by extracting 13 economical attributes that refer to the field of unemployment, GDP, public health and social protection expenditure, number of working ours per week and inflation rate. The constructed tree is presented in Figure 8.

The tree indicates that high GDP, low unemployment rate and high inflation GDP deflator relate to low fertility rate, while low GDP per capita usually relates to a high TFR.

As David Heer said [13], economical progress should positively influence fertility rate. Overall statistics significantly disconfirm the hypothesis at least in the modern world where food is not scarce. Our analysis indicates that direct economical attributes are not very relevant for fertility on their own, at least not as other groups of attributes. For example in figure 8 in some cases high GDP per capita leads to high and in others to low TFR. Becker (1981) [17] presents a plausible explanation of such GDP-TFR relations. He claims that TFR depends on the disposable expenses and expected usefulness of the children. To uphold the thesis he gives an example of the rural family that used to have more children in order to assure help for maintaining the family. Human resources were urgent for working on the fields, in the woods, etc. Nowadays, agriculture has become more and more automated, thus reducing the need for human forces. Consequently, the cost benefit of the children dropped drastically and families began to shrink. Besides, factors like higher educational level, lower child mortality rate, and the desire for career making among young people, pushes TFR even lower. This linkage between income and fertility is typical for developed countries, where despite constant income growth, TFR is continually decreasing. Whereas in developing countries, low income does not influence fertility rate.

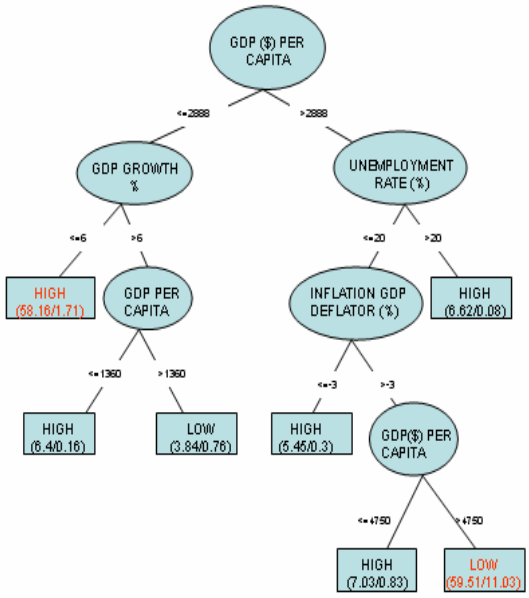

Figure 8: TFR classification tree with two-valued class, considering only economical attributes (78,2%).

In any case, the tree from Figure 8 is only 78,23 % accurate, which is low in comparison with trees based on other attributes. This indicates that direct economical factors are not the main cause for the distinction among countries with low and countries with high fertility rate.

4.1.1.6 Educational attributes

Analyzing the relation between educational factors and TFR resulted in the tree presented in Figure 9. High percentage of enrolment in primary educational level is in general related with high fertility rate, whereas low TFR is more related to enrolment in secondary or tertiary educational factors. As observed by experts before, high education, especially of women, decreases TFR.

4.1.1.7 Developed countries

While developing countries have problems with too high TFR, developed countries, especially in Europe, have problems with low TFR. Mark Steyn, a conservative polemicist, argues that Europe is quickly becoming a barren, ageing, enfeebled place [6]. In the decades after the second world war, rich countries everywhere experienced similar trends. The bonds of traditional family life began to slacken, more women got jobs, people sought enjoyment and satisfaction more and more through individual pursuits rather than in families. This social transformation, which is occurring also in America and East Asia, led to a demographic bonus (a bulge of people working) and to what might be called "the postponement of everything". People left school later, left home later, married later, had children later, they also died later [6]. Even though these interpretations are not uniformly accepted, they seem to be statistically quite well grounded.

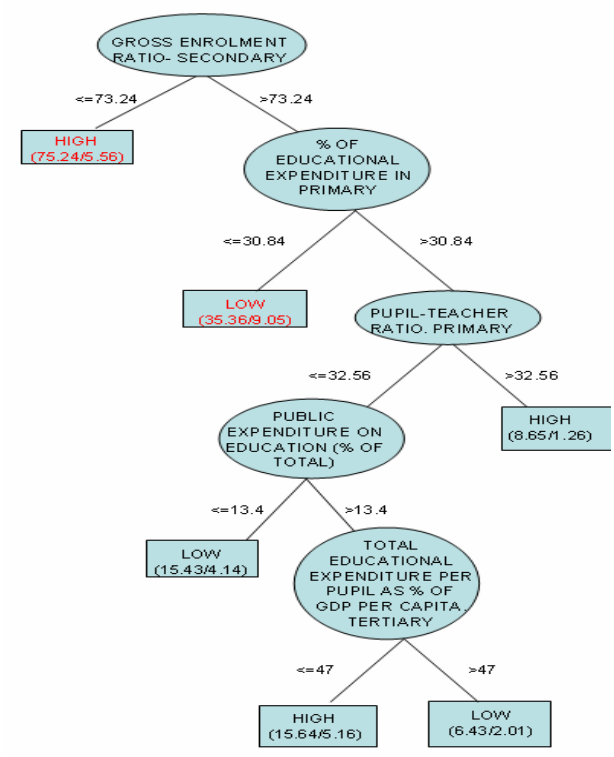

Figure 9: TFR classification tree with two-valued class, considering only educational attributes (78,2%).

Having that in mind, the relevant question is: Why do some rich countries still have high TFR?

In the following experiments we denoted 39 countries with high GDP as rich.

4.1.1.8 Selected attributes

The tree in Figure 10 indicates that exceptions to the low fertility rate have poor education and social system. Further analyses showed that these countries rely on natural resources such as oil.

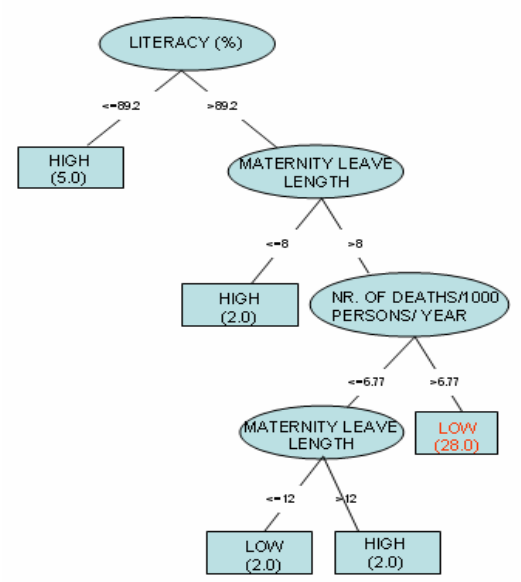

Figure 10: TFR classification tree with two-valued class and automatically selected attributes (78,2%).

4.1.1.9 Social attributes

Analyses of the obtained tree presented in Figure 11 revealed that countries with oil are rich and have Muslim religion. But the relation can be interpreted originally as follows: when Islam is the prevailing religion of the country, then TFR is most likely to be high, while otherwise, TFR decline is the more likely option. Results are consistent with the previously observed relations that TFR is higher in more conservative countries, which Islam countries certainly are.

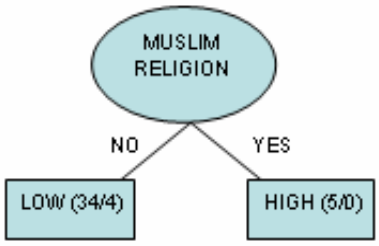

Figure 11. TFR classification tree with two-valued class considering only social attributes (89,7%).

4.2 ΔTFR class

The newest studies of Worldwatch Institute conclude that there is so much variability in fertility rates that we can not know with any confidence how many people the future holds [11]. Indeed, it seems reasonable that ΔTFR analyses are a bit less relevant as those with TFR, since they measure the amount of change and not the obtained situation. Even though, our next attempt was to established factors that might influence TFR growth and decline. In the next section a few of the most interesting and accurate trees are presented.

4.2.1 All countries

4.2.1.1 All attributes

Again, literacy seems to be an important indicator of TFR trends (see Figures 12 and 13). Countries with low percent of literate habitants generally have increase in TFR. Countries with high percent of literate citizens (above 97.9%) and low unemployment rate (below 9.6%) on the other hand have decreasing TFR trend.

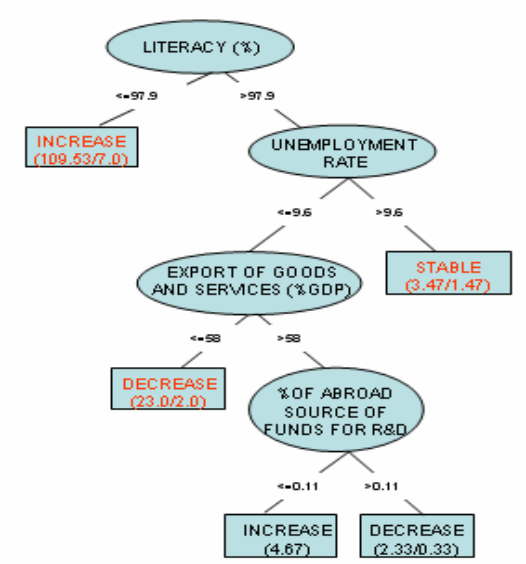

Figure 12: ΔTFR classification tree with three-valued class (81,1%).

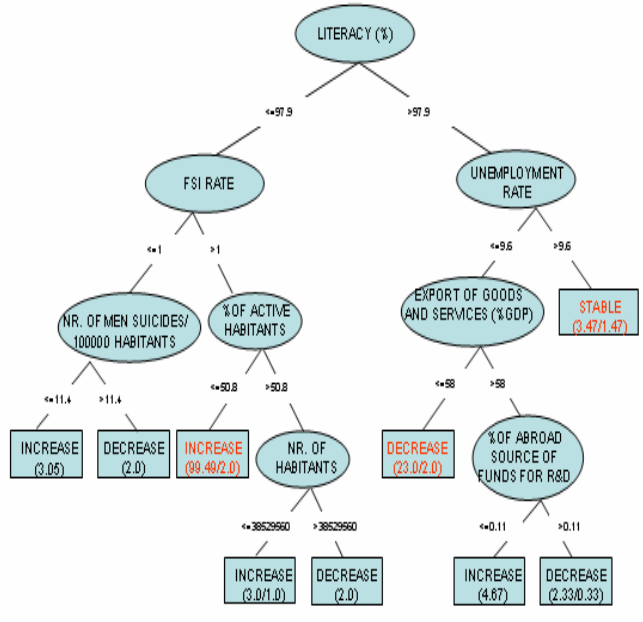

Figure 13: Unprunned ΔTFR classification tree with three-valued class indicator (83,2%).

Similar conclusions can be drawn from the tree on Figure 14, when attributes were automatically selected. This tree has surprising high accuracy.

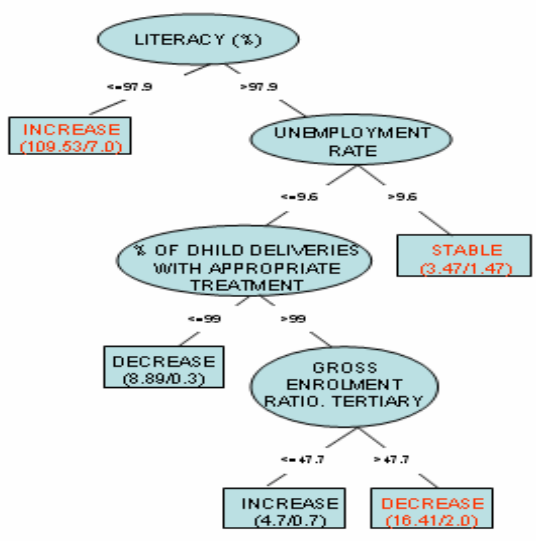

Figure 14: ΔTFR classification tree with three-valued class, automatically selected attributes (85,3%).

4.2.1.2 Social attributes

Considering only social attributes, the same tree as in the case of TFR class appeared (see Figure 7), again exposing the importance of conservative politics of the country for the TFR growth trend. Countries that don't allow abortion and adoption to homosexuals have TFR growth trend, whereas countries that allow abortion and homosexuality have TFR decline trend. Accuracy in this case is 78.32%, much lower than in the tree presented in Figure 14.

4.2.2 Developed countries

In this case our criteria for dividing countries by their developmental status was FSI. A country was classified as well developed if FSI index was less than 39.45, resulting in 27 countries. Analyses were performed on the attributes separately merged in smaller groups.

4.2.2.1 All attributes

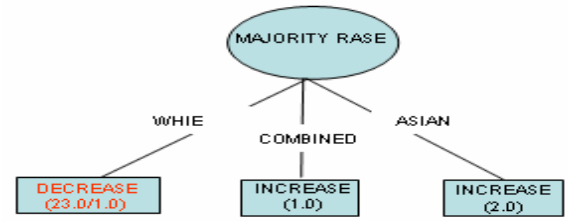

Figure 15: ΔTFR classification tree with two-valued class (accuracy is 84.6 %).

Race appeared to be an important factor of TFR trend (see Figure 15). In nations with prevalent Asian and combined race, TFR is likely to increase, while in countries with a majority of white race, TFR is declining. The nature of this genetic relation is not cleat at this point.

4.2.2.2 Economical attributes

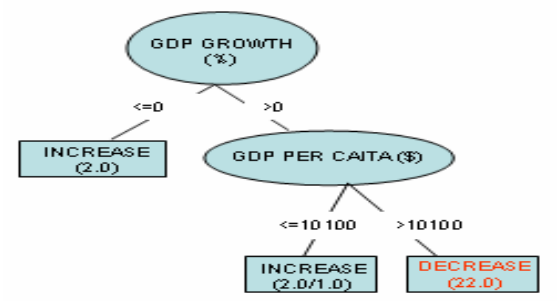

Figure 16: ΔTFR classification tree with two-valued class, considering only economical attributes (84.6 %).

We can see that highly economical developed countries with more than 10100 GDP per capita ($) have TFR decline. This thesis is for example not in agreement with the Worldwatch Institute study noting that fertility rate is rising in the United States [11]. However, this study is violating the age-old dictum that rich countries do not make lots of babies as well [11]. The tree based on economical attributes is this time quite accurate. Therefore, ΔTFR analyses gave more statistical relevance on economical attributes than analyses with TFR.

4.2.2.3 Social attributes

When selecting only social attributes, the accuracy of ΔTFR classification trees dropped drastically (on 76.9%) what means that these factors are not good indicators for TFR trends.

5 Conclusion and discussion

In fertility analyses, the data mining tools again proved their major asset: the constructed knowledge is in a transparent form, enabling human comprehension of relevant relations in complex forms. In this way, an interactive and interaction process is enabled between computers and humans, exploiting best properties of the two most advanced information machines. Computers fast examine vast search spaces with their advanced speed and accuracy while humans make conclusions and guide search with the advanced cognitive skills.

To readdress the problem, let us restate that the space of all potential hypothesis for 100 binary attributes and a single binary class is $2^{2^{100}}$. This number is far larger than the number of all atoms in our universe, which is according to Wikipedia around 10^{80}, i.e. 2^{266}. Therefore, there is no way humans can analyze any meaningful share of all the hypotheses. But we can examine results of one search, make conclusions and redo the search changing specific details of the search. In this way humans can "mine" for relevant hypothesis.

Regarding the fertility relations, the DM tools enabled rediscovery of major properties. The authors are not experts in the fertility or demographic field, therefore verification of our conclusions by an expert and further analyses of interesting new patterns are a matter of further research. However, we report our impressions for further discussions:

- Firstly, we were surprised that there are so many distinctive hypothesis, i.e. patterns discriminating countries with low from those of high fertility rate. Rich countries are predominantly white, have good education, women live longer, literacy is high, the predominant race is white, people have no strong religion obligations, they are liberal etc. and vice versa for the developing countries.
- Secondly, according to the constructed trees, it is rather simple to influence the fertility rate – just improve literacy in women or just allow liberalization and decrease the influence of religion. And vice versa – to improve fertility, just e.g. improve moral values and decrease liberalization or decrease literacy or apply any of the remaining 10 or 15 attributes. Some are costly, some are unacceptable, e.g. decreasing literacy. According to some experts, practically all of these attributes are hard to implement in democratic countries. Still, the trees indicates that there are several mechanisms, some of them rather costless, that will change the process in European countries, leading first to economic problems and later to extinction of nations and cultures.
- We did not have time to study each particular attribute in detail, such as the length of maternity leave. While the trees so often show relevance of maternity leave for decreasing fertility, the trees do not show whether this is a cause or a consequence. At first, we thought that it is just a consequence of low fertility, just a mechanism of countries promoting higher fertility. In addition, mothers are generally in favor of longer leaves. But after so many trees, and having in mind that this is one of very costly mechanisms, it is becoming more than a suspicion that longer maternity leave is at least a controversial matter.

Our analyses dealt with countries and not with individuals. Obviously, several of the fertility and demographic matters are open for further investigation with DM techniques, next time with fertility experts.

Acknowledgement

In the analyses performed, contributions of the following computer-science students should be acknowledged: Barbara Tvrdi, Anže Jazbec, Marko Kovki, Domen Muren. We also thank Statistical office of the Republic of Slovenia for the answers and the IS 2007 program committee members for the constructive suggestions.

Reference

[1] F.C. Billari, J. Fürnkranz, and A. Prskawetz (2000). Timing, sequencing, and quantum of life course events: a machine learning approach. Working paper 010, Max-Plank-Institute for Demographic Research, Rostock.

[2] H. Blockel, J. Fürnkranz, A. Prskawetz, and F. Billari (2001). Detecting temporal change in event sequences: An application to demographic data. In L.D. Raedt and A. Siebes (eds.), *Principles of data Mining and Knowledge discovery: 5th European Conference*, PKDD 2001, Volume LNCS 2168, pp.29-41. Freiburg in Brisgau: Springer.

[3] M. Christenson, McDevitt, T., and Stanecki, K. (2004). *Global Population Profile: 2002. International Population Reports.* Health Studies Branch, International Programs Center, Washington Plaza II, Room 313A U.S. Census Bureau, Washington, DC 20233-8860.

[4] J. Demsar, and B. Zupan (2005). From Experimental Machine Learning to Interactive Data Mining, White Paper (www.ailab.si/orange), Faculty of Computer and Information science, University of Ljubljana.

[5] Eurostat http://epp.eurostat.ec.europa.eu/

[6] Economist http://www.economist.com/world/europe/displaystory.cfm?story_id=9334869

[7] M.Gams (2007). Osnovna demografska gibanja (Basic Demographic Dynamics). In J. Malačič, M. Gams (eds.), *Proceedings of the 10th International Multi-conference Information Society (volume B) Slovenian Demographic Challenges of the 21st Century*. Ljubljana: "Jožef Stefan" Institute, pp. 35-37.

[8] M. Gams, J. Krivec (2007). Analiza vplivov na rodnost (Analysis of Impacts on Fertility). In J. Malačič, M. Gams (Eds.), *Proceedings of the 10th International Multi-conference Information Society (volume B) Slovenian Demographic Challenges of the 21st Century*. Ljubljana: "Jožef Stefan" Institute, pp. 35-37.

[9] J. Malačič (2006). *Demografija: teorija, analiza, metode in modeli (Demography: theory, methods and models)*, (EF, Manual). 6. edition. Ljubljana: Faculty of Economy, pp. 339

[10] J. Malačič (2000). *Demografija– teorija, analiza, metode in modeli (Demography: theory, methods and models)*. 4. edition. Ljubljana.

[11] New York Times http://dotearth.blogs.nytimes.com/2008/03/14/earth-2050-population-unknowable/

[12] M. Oris, G. Ritschard, and A. Berchtold (2000). The use of Markow process and induction trees for the study intergenerational social mobility in nineteenth century Geneva. In *Social Science History Association Annual Meeting*, Baltimore.

[13] M. Potts, Society and Fertility. Estover: McDonald and Evans. 1979. pp. 384

[14] G. Ritschard, and M. Oris (2005). Dealing with Life Course Data in Demography: Statistical and Data mining Approaches, in R. Lévy, P. Ghisletta, J.-M. Le Goff, D. Spini et E. Widmer (eds.), *Towards an Interdisciplinary Perspective on the Life Course*, Advances in Life Course Research, Vol. 10. Amsterdam: Elsevier, pp. 283-314.

[15] Slovenija v številkah (Slovenia in Numbers). Ljubljana: Statistical office of the Republic of Slovenia, 2006. pp. 79

[16] Statistical office of the Republic of Slovenia; Slovenian Statistical Database: http://www.stat.si/

[17] N. Stropnik (1997). *Ekonomski vidik starševstva (Economical side of Parenthood).* Ljubljana, Scientific and publication center , pp.221.

[18] M. Šircelj (1991). *Determinante rodnosti v Sloveniji (Fertility Determinates in Slovenia).* Doctoral dissertation, Faculty of Art, Ljubljana.

[19] United Nations, Department of Economic and Social Affairs, Population Division (2007). World Population Prospects: The 2006 Review, Hilights, Working Paper No.ESA/P/WP.202.

[20] UN http://esa.un.org/unpp/

[21] U.S. Census Bureau, International Population Reports WP/02, *Global Population Profile: 2002*, U.S. Government Printing Office, Washington, DC, 2004.

[22] V. Vidulin, M. Gams (2006). "Vpliv investicij v izobraževanje in R&R na gospodarsko rast," *Elektroteh. vestn., vol. 73, nr. 5*, pp. 285-290.

[23] WIKI-1 http://en.wikipedia.org/wiki/Failed_States_Index

[24] WIKI-2 http://en.wikipedia.org/wiki/Gross_Domestic_Product

[25] I. H. Witten, E., Frank (2005). "Data Mining – Practical Machine Learning Tools and Techniques (sec. ed.)," Morgan Kaufmann.