Population Influences Using Genetic Programming

Patrick Tierney

School of Computer Science & Informatics
University College Dublin

Abstract

A census is used by most countries to record its population. The information obtained from censuses carried out by the Central Statistics Office of Ireland will be used to view the factors of society that influence the population of the Republic of Ireland, It will be seen if whether work based factors (e.g. unemployment rate) or natural factors (e.g. number of births) affect the population more using the Genetic Programming system ECJ.

1. Introduction

The aim is to discover which of the details that are contained in a census, influence the population the most, work based factors or natural based factors. First all the record details from all the previous censuses must be acquired from the Central Statistics Office. This information will then be sorted into different tables and categories. By using a Genetic Programming (GP) system called Evolutionary Computation Java (ECJ), these details will be processed and will find which factors that influence the population the most<1,2>.

Much work has been done in this area, especially when considering the populations of Third World countries as opposed to First World countries like the Republic of Ireland. Many charities and governments use this information to decide on a course of action to take to increase, decrease or steady the population of certain countries. For example, China limits the number of births to 1 child per mother, because it has nearly reached its population capacity.

1.1 Central Statistics Office

When a census is taken, many details of each persons life is recorded, from their age to their current employment. Not only is this information used to record the amount of people in the country at the time and their lifestyle but it is also used to predict the future population and future aspects of the country. The Central Statistics Office of Ireland conducts the census collection in the Republic of Ireland. The Central Statistics Office is an independent Office established in 1949 <3>. It operates under the aegis of the Department of the Taoiseach (Government) to guarantee its statistical independence and the confidentiality of the data it collects. The government uses this information to predict many of the problems they will face

in the future. For example, if the government predicts that the number of deaths will fall in the next few years, then the government can prepare for this by building bigger hospitals, building more elderly homes and training more doctors. Censuses are carried out every 5 to 10 years.

1.2 Evolutionary Computation Java

ECJ is a tree based research Evolutionary Computation (EC) system which was developed at George Mason University's Evolutionary Computation Laboratory. This EC system is written in Java as it was designed to be highly flexible, with nearly all classes (and all of their settings) dynamically determined at runtime by a user-provided parameter file. All structures in the system are arranged to be easily modifiable. Even so, the system was designed with an eye toward efficiency <4>. It is one of the most popular programs used in EC as it is also used extensively in GP.

2. Experiments

2.1 Data Recording

All details were first taken from the CSO over a ninety five year period from 1910 to 2005. The reason that 1910 was chosen as the start year is that it is the earliest year that all the data used in the experiment were all recorded. For each year the parameters used in the experiments were population, number of births, number of deaths, average age, number of working people, average unemployment and average weekly income.

As censuses are only recorded every 5 to 10 years some parameters such as the population, had to be modified for most of the years. So to counteract this problem, the population from the previous census was used. For example, for the year 1964 the population was set as 2,818,341 because it is the population that was recorded in the previous census taken in 1961. Another modification that had to be made to the data was setting the average unemployment rate. The reason for this was that for most years the unemployment rate was recorded quarterly (every 3 months). So for the yearly rate, the average of the recorded data was taken.

All the data then sorted according to year and type. In all 3 different parameter files were used; the first file (*File 1*) used all the parameters, the second file (*File 2*), used the number of births, number of deaths, and the average age, while the third file (*File 3*), used number of working people, average unemployment and the average weekly income. A population of 1000 was used during the tests and each test was run for 20 generations. A mutation rate of 0.5% was also used.

2.2 Testing Procedure

Each parameter file was passed into the ECJ, along with a fitness function and function set. The ECJ then returned a fitness value based on the parameters given. The fitness function combined each of the parameters with the function set and a value was returned. This value was selected as it was the best value returned by the fitness function. This value was then multiplied by a constant and then compared to the population of that year. A percentage error was then returned as the value of that year. An error of 0% indicates that the population was calculated correctly.

2.3 Results

All though there were very few results, they occurred as expected when comparing to reports in this area of work. Overall the results proved that the first parameter file out performs the other two files when they are compared. This means that all of the parameters have a huge effect on the population.

Table 1. Results from 1981-1985 using File 1

Year	**Error**
1981	1.8034%
1982	2.0565%
1983	2.4770%
1984	3.1716%
1985	4.6466%

The year in which the census was taken can be clearly above seen in Table 1. It is also demonstrated how as the data is less accurate we get a worse error returned. The error for 1981 was 1.8034%. Where as the error for 1985 is 4.6466%, the year before the next census is taken. This is a big difference between the given years.

Table 2. Results from 1981-1985 using File 2

Year	**Error**
1981	2.0958%
1982	2.8012%
1983	3.5386%
1984	4.4553%
1985	6.0566%

Table 3. Results from 1981-1985 using File 3

Year	**Error**
1981	2.7656%
1982	3.6333%
1983	4.1788%
1984	5.7824%
1985	6.5728%

The results proved that natural based factors have a greater effect on the population than work based factors, as can be seen in the two Tables above. Its been noticed when comparing results from File 2 and File 3 that while natural based factors have more of an effect on the population than work based factors, there have been certain years where the work based factors have outperformed the natural based factors. This was especially true in the years from 1993 onwards. This fact compares to the reports on population influences for 2006, looked at by Glenn and Gordon, all though they consider many more factors (5).

Another factor that could be included when analysing these results is the amount of emigration that occurred in Ireland during these years. Immigration and emigration factors affected the results from File 3, as can be seen in Table 3, mostly due to the fact that it was based on work related factors. Immigration was high in most years up to the 1990's, where emigration became more prevalent due mainly to Irelands Celtic tiger.

Table 4. Results from 1996-2000 using File 1

Year	**Error**
1996	1.6041%
1997	1.8142%
1998	2.0553%
1999	2.5794%
2000	3.1258%

Table 5. Results from 1974-1978 using File 1

Year	**Error**
1974	5.9273%
1975	7.6922%
1976	8.2823%
1977	9.4289%
1978	10.5603%

The best solution returned by the ECJ was the year 1996, which returned an error of 1.6041%. When analysing this result, we can see that 1996 was a year in which a census was taken, so all data for the parameters is accurate. Also we notice from above the rise in emigration in Ireland during that period. So it is true to think that 1996 was the best result. The worst solution returned however, was the year 1978, which returned an error of 10.5603%. The main factor of this was the massive change in population between 1971 and the next census taken in 1979, which was a rise of just under 400,000, even with all the emigration during this period.

3. Conclusions

It can be seen that the modified data did affect the results greatly. It was shown that the results are far better for the years that the censuses were taken in, as opposed to the following years. It was surprising to find that another aspect to the results is how immigration and emigration affect the population. Future work could be carried out in this area especially.
This paper does not take into account the effect factors such the climate, education, accessibilities to facilities (e.g. food, water, medicine), life expectancy, etc. has on a population. Future work could include these factors as well.

References

1. Koza J.R.: Genetic Programming, *Encyclopedia for Computer Science and Technology* (1997)
2. Banzhaf, W., Nordin, P., Keller, R.E., Francone, F.D: *Genetic Programming: An Introduction* (1998), Morgan Kaufmann
3. CSO Website: http://www.cso.ie/
4. ECJ Website: http://cs.gmu.edu/~eclab/projects/ecj/
5. Glenn J.C., Gordon T.J.: *State of the Future*. (2006) AC/UNU