University of BRISTOL

DEPARTMENT OF COMPUTER SCIENCE

Data Mining and Analysis of Global Happiness: A Machine Learning Approach

Louise Millard

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering

September 2011 | CSMSC-11

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Louise Millard, September 2011

Executive Summary

The aim of the project was to explain and predict global happiness. This was a cross sectional study including 123 countries, each having a mean happiness value. This ground truth was established from survey data, using the answer to a life satisfaction question. The features used to predict happiness were collected from online sources, and were chosen using background knowledge gained from a review of previous work.

Initial data analysis was performed to discover patterns in the data using PCA, visualisations and correlations. PCA found an interesting convex relationship between life satisfaction and gender equality. The data was prepared firstly by imputing missing values with the k-nearest neighbour method. Some variables were transformed using log where their relationship with life satisfaction was found to be exponential.

Prior to performing feature selection life satisfaction prediction was assessed, comparing an initial feature set with economic variables using a t-test of the correlations of cross validation folds. The feature set was found to be as predictive as the economic variables. Feature selection was performed using lasso, least squares and decision trees. The significance of results was determined by finding test statistic thresholds using permutation testing and bootstrapping. A key feature set was identified as:

Life expectancy   Income distribution   Proportion of women in parliament
Freedom   Primary education enrolment   Secondary education enrolment
Mortality rate

These features were used with several learners to construct models to predict happiness. Model trees performed the best with a mean correlation of 0.86 across the cross validation folds. This was not significantly more predictive than lasso ($p = 0.52$), indicating the relationship of the key features with life satisfaction is highly linear. The key feature set was significantly more predictive than using the larger original feature set ($p = 2.12 \times 10^{-16}$), highlighting the benefits of feature selection. The performance of our key feature set was compared against economic variables and gave a significantly better performance for both lasso and decision trees.

Bayes networks were used to assess the relationships of the variables using the following measures of performance; percentage correct, ROC curve area and degrees of freedom. No improvement in performance was found when connecting GDP per capita directly with life satisfaction. This supports the notion that GDP allows other variables to occur which in turn impact on life satisfaction, rather than GDP having a direct and significant impact.

Finally, Chernoff faces proved an effective visualisation method for our multivariate dataset. This is an intuitive representation where happier faces correspond to higher life satisfaction, and hence patterns and anomalies can be easily identified. An interactive visualisation of results can be found at `http://intelligentsystems.bristol.ac.uk/research/Millard/`.

In summary, key highlights of this project are:

- PCA uncovered an interesting relationship between gender equality and life satisfaction
- Decision trees proved an effective method in both feature selection and life satisfaction prediction
- Our key features performed significantly better than economic variables
- Graphical models helped investigate variable relationship structure
- An effective data visualisation method was used to demonstrate results

The work previously completed consists of a review of previous work and methods, which contribute to sections 2 and 3 respectively. Also, initial survey data analysis was performed to determine a ground truth of life satisfaction (contributing to section 4).

Acknowledgements

I wish to thank, first and foremost, my supervisor Professor Nello Cristianini, for his guidance and advice throughout this project. This has been an invaluable and enjoyable experience and I am grateful for his continued enthusiasm and interest in this project.

I am grateful to Dr Tijl De Bie, for a thoroughly interesting and useful Pattern Analysis and Statistical Learning module. Thank you for your patience and time answering my questions. Also, many thanks to my Personal Tutor, Dr Tim Kovacs, for his time and advice.

Finally, thanks to my family and friends. I would like to thank my mother Alex, for her support whilst I returned to my studies. Thanks to my close friend Joanne for her continued encouragement. To Chris, thank you for your thoughts, advice and support.

Contents

1 Introduction

1.1 Overview

Happiness economics is an active area of research. The aim of this work is to use a machine learning approach, to explain and predict global happiness. At present countries mainly use GDP to determine progress[1] but this has come under increasing criticism as measuring "everything, in short, except that which makes life worthwhile". [Robert Kennedy, 1968]

We investigate a broad range of variables to discover more appropriate measures of progress, that impact on happiness. A focus is given to variables over which we have control and can change in order to improve quality of life, and those that governmental policy can directly affect such as education and health. Environmental variables such as weather are also considered as these are affected by issues such as climate change on which government policy has an impact. This leads us to state our meaning of happiness for the purposes of this project. We define happiness as synonymous with life satisfaction and corresponding to the longer term notion of the quality of one's life.

Figure 1.1 shows a happiness map generated in 2006 using various survey sources, and large global variations of happiness are very apparent. Whilst a strong correlation between GDP and happiness has been shown previously this is certainly not the whole picture, and other variables are likely to play a key role; "Economic factors are not goals in themselves but are means for acquiring other types of social goals" [25].

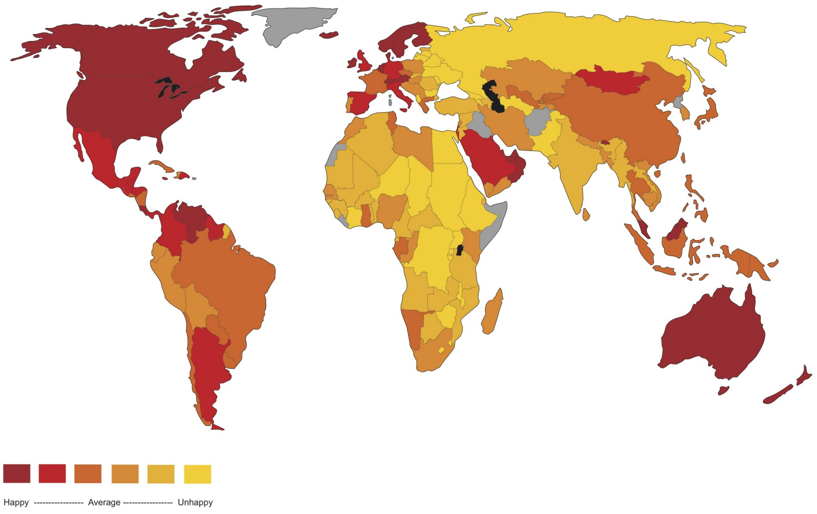


Figure 1: Happiness map [52]

1.2 Methodology

Our contribution from a new perspective using machine learning techniques aims to provide a new insight into this area. Previous research has used statistical methods predominantly linear regression. Machine learning techniques will allow the application of powerful methods to the dataset to find new patterns, such as decision trees and support vector machines, each having different benefits.

[1] One exception is Bhutan, a small country in South Asia, which uses an alternative measure called Gross National Happiness (GNH) to assess progress. This was established in 1972 by the fourth Dragon King, in order to better reflect the important aspects of his peoples lives, centred on Buddhism and spirituality. [38] More recently even the UK is considering alternative measures with David Cameron announcing in 2010 plans to measure happiness; "I believe a new measure wont give us the full story of our nation's well-being ... but it could give us a general picture of whether life is improving"[11].

1.3 Report Structure

A review of previous work is given in section 2, followed by an overview of key statistical and machine learning methods in section 3. Sections 4 - 11 are investigatory. A ground truth of happiness is established in section 4 and the reliability of this measure is also investigated. Data analysis is performed to uncover patterns, described in section 5.2. The predictive capabilities of our initial feature set compared to economic variables is assessed in section 7. Feature selection is performed in two stages (sections 8 and 9). Several methods are investigated to find the best happiness prediction in section 10, and structural relationships are assessed in section 11. Finally, our results are visualised using an effective representation (section 12).

2 Research Review

2.1 Measuring Happiness

Previously it was thought that you could not measure happiness but only behaviour.[31] However, there have been independent strands of research which provide strong evidence to the contrary. Biologically, it has been found that specific areas of the brain are active when a person experiences happiness. An interesting piece of research ([30]) looked at brain activity with respect to the emotions of happiness, sadness and disgust. They found correlations between these emotions with activity in multiple regions of the brain, some of which were shared for the different emotions while others were specific to the emotion and stimulus. The potential to assign a value of happiness depending on brain activity highlights that happiness is a more tangible and measurable property than once thought.

A pertinent issue is the contribution of the genome to ones happiness as this could significantly affect results (the "nature versus nurture" debate). We are interested in the "nurture" causes of happiness from living in different countries. Research published in 1996 ([33]) looked at the degree to which happiness may be encoded in our genes. They did this using data of a happiness survey question but where the recipients were twins, both monozygotic and dizygotic. Monozygotic twins have an identical genome, whereas dizygotic twins do not. 127 sets of twins were surveyed twice, with a ten year gap. To investigate the degree to which genomes are responsible for happiness the results were correlated within twins: $(person_1, survey_1) \rightarrow (person_2, survey_2)$ and $(person_2, survey_1) \rightarrow (person_1, survey_2)$. The results were significantly different for the twins types, where monozygotic and dizygotic twins gave correlation of 0.4 and 0.07 respectively. The twins with identical genomes had a much stronger correlations between happiness levels. They conclude that happiness in adults is determined in equal measure by both genes and environmental factors.

A more recent paper by Frey [23] aimed to find specific genes which code for happiness, using longitudinal survey data and gene association. Gene association is the process of investigating the correlation between a genes' expression and a particular trait (or phenotype). The gene 5HTT was selected as the candidate gene as it is known to be involved in brain development and so may have implications on happiness. They describe happiness as having a baseline which is specific to each individual and encoded in their genes, and happiness fluctuates about this value in response to their environment and experiences. This paper highlights the complexity of happiness, where there is likely to be a complex combination of factors that contribute, including the cumulative effect of multiple genes in addition to environmental factors.

A happiness coding gene however could potentially affect the results. The central consideration for this is whether the global genetic distribution is clustered such that there would be an uneven spread across the globe. This is however not thought to be the case, where there has been sufficient human migration to give an even distribution. There may still be effects on results as for instance, the existence of a baseline happiness may cause the variance of a countries happiness to remain fairly constant but the mean may change, as peoples happiness alters with respect to their own baseline values.

A final point of consideration is how well subjective happiness represents the happiness of a person. Previous research compared the consistency of happiness measures given from different sources. A 2009 study by Sandvik et al. ([45]) looked at the correlation between self-reported happiness and the values given by family and friends. They found these three assessments of a persons' happiness to be highly consistent.

This discussion has highlighted the complexity of this trait, but also shown happiness to be less abstract and intangible than was previously thought. Establishing a ground truth for happiness is an important aspect of this project, which must accurately represent our notion of happiness. The following section investigates deriving a ground truth from the survey data.

2.2 Establishing a Ground Truth

A general consensus is that answers to survey questions provide representative values of happiness or life satisfaction. The three main global surveys used in this type of research are the World Value Survey (WVS), Gallup World Poll (GWP) and World Database of Happiness (WDH). A typical question takes the form "how satisfied are you with your life as a whole these days?" with an answer scale typically 0 - 10 [16]. The GWP is a large scale survey carried out on a daily basis. However, the data is not freely available and we only have access to some limited data through the results of another study[2]. The WDH is a compilation of happiness values from different surveys, predominantly the WVS.

The ground truth we use can be either from a single source or a combined value of multiple sources. The choice we make will be dependent on results of data analysis. Previous work has tended to keep different happiness labels separate (choosing one or working with several simultaneously) rather than fusing multiple sources. Using a single source however has short comings due to the validity of each source. Firstly, the WDH is a secondary source, a compilation of results from multiple survey sources. We prefer to use a primary source such that we have a better knowledge of the survey data. Secondly, the different surveys vary with respect to the countries studied and various details regarding the survey format. This means that one cannot simply be preferred over the other. Therefore, we will begin with primary data and establish a ground truth from these values (which may mean several alternative labels). Section 4 investigates the available survey data and establishing a ground truth. Section 2.4.1 looks at conflicting results of previous work, and possible causes from differences in surveys.

The Happy Planet Index ([3]) developed by The New Economics Foundation (NEF) is a global measure indicating sustainable happiness, the happiness component of which came from the GWP life satisfaction question. This work is of particular interest as data from multiple sources was used to extend the dataset, using regression to infer a value for the GWP for additional countries. 112 countries came from Gallup, 16 from WVS and 14 from GWP's ladder of life question. This method relies on countries with values for multiple surveys which means a relationship between the survey values can be inferred. 68 countries had results for both GWP and WVS. Regression was performed using GWP and WVS as the dependent and independent variables respectively, such that GWP values are inferred from WVS values. [3]

Stepwise regression was used to improve the correlation between the survey datasets using additional variables. This is a recursive regression where a variable is added at each step . If this variable provides a significant improvement to the inference of the GWP value then it is kept, otherwise it is ignored.

The use of stepwise regression improved correlation, where they found that "four variables are able to predict 91% of the variance" which included the Human Development Index and an education index amongst others. It is interesting to assess the variables that improved the correlation between the surveys, as they may indicate the differences between them. For instance [16] noted that WVS tended to bias their surveying towards more intelligent people in certain countries to make the sample "more comparable". This may be consistent with the use of an education variable here, showing that the variance between the surveys may be accounted for to some degree by education levels. WVS has in effect controlled for education by reducing the difference across countries.

Stepwise regression was also used to correlate the GWP life satisfaction data with the GWP ladder of life data. This found different variables improved the correlation between the datasets including life expectancy and geographical variables. [3] The ladder of life question asks the respondent their life satisfaction relative to their attainable levels; "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?". This

[2] Source is Happy Planet Index v2.0

highlights the affect of seemingly small differences in question format can have on the responses given. [3]

The use of additional variables improved the correlation between surveys, and gave a larger set of countries with GWP values. This method is appropriate where the overlap between the surveys is large enough such that correlation can be accurately inferred. However, we suggest some concerns of the validity of this. Firstly, the assumption of a linear correlation between the two surveys. Secondly, the spread of data may affect the results, as it seems that the WVS countries are much sparser in the lower range of the GWP data.

The use of additional variables is not appropriate for our project because our happiness label is used to find a predictive feature set. Using additional variables to derive the label would clearly bias the subsequent correlations. Hence we would only be able to regress using solely survey data.

2.3 General Research

A 2004 paper by Gundelach et al. ([25]) is of particular interest for two reasons. Firstly, it includes the use of graphical models to analyse the causal relationships. Secondly, a distinction is made between life satisfaction and happiness, where they showed that these concepts are actually two distinct labels, with different independent variables. Gundelach et al. are critical of previous studies, where mostly life satisfaction and happiness are treated synonymously instead of accounting for their differences in meaning. [25] shows that GDP and these labels have a “rather high but far from perfect correlation” [25], indicating there is certainly more to the determinants of happiness than GDP alone.

The methods used seem distinct from other research in this area. Firstly, they discuss the importance of looking at distribution of values not just the means. This is because the survey answers are options over a range of values, usually between 0 and 10. Analysis has typically focused on the mean values of each country.

A key difference between this research and our study is the choice of entity, where individual respondents are used rather than countries. Variables always correspond to each individual entity, and hence in this study come from the survey data together with each happiness value. It may be useful to work with data corresponding to individual respondents rather than values for countries as a whole, but this has the disadvantage that survey answers have additional uncertainty as answers people give may not always be truthful or representative. The use of variables from global datasets such as the World Bank removes this source of inaccuracy. Further to this, our aim is to predict happiness without the need for surveys and hence using individual respondents as entities would be inappropriate.

This work included using “chain graphical models”, to produce a Markov graph (figure 2). The graph contains edges where an edge between two variables means they are directly related to each other. A Markov graph is one where each node is conditionally independent on all other nodes give the nodes directly connected to it. To construct the graph it is initially constructed by hand, where variables have no edge connecting them if it is certain there is no dependence relationship between them. This therefore requires domain knowledge to reduce the number of edges in this initial graph. The dependence relationships in this graph are then tested using log linear analysis to test dependence, and if they are found to be independent the edge is removed. This results in a graph showing the dependent relationships only.

This work found some interesting results and in particular, the resulting model (figure 2) shows differences between the causality of life satisfaction and happiness respectively. Both happiness and life satisfaction correlate with country (with correlation coefficients of 0.40 and 0.22), although this is stronger for happiness. Life satisfaction is found to correlate with “life control” whereas happiness correlated with “stable relationship” (these variables come from the survey data also). Of particular note, income is not included in the model as it was not found

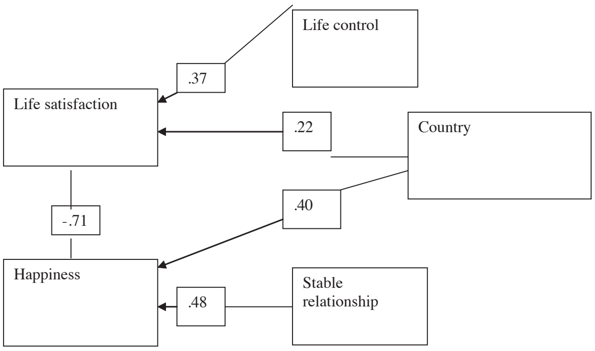


Figure 2: Causal network [25]

to correlate strongly (with a correlation above 0.16) with either life satisfaction or happiness. However, Gundelach et al. explain this with the choice of sample where only European countries were used and hence the variance of wealth may be fairly small. However, we question this as since this study is on an individual basis it seems there would still be a fair amount of variance of wealth within a population. This may therefore indicate that wealth is not a key factor in happiness or perhaps is evidence supporting the Easterlin Paradox (see section 2.4).

The differences in correlations for life satisfaction and happiness were attributed to the differences in meaning of the two terms where "happiness is more emotional and life satisfaction is more cognitive" [25], which intuitively makes sense with regards to the particular correlates of each.

We are not concerned with the semantic differences of these variables as we think both contribute to our notion of happiness for this project, and consider both types of label.

This work ([25]) shows strong evidence for the fact that happiness is not solely determined by wealth. The use of just EU countries was an effective way of controlling for wealth such that other correlations could be determined more easily. It should be noted however, that construction of such graphical models do not determine the direction of causality, and in this sense the directed graph produced in [25] (figure 2) is misleading.

Bayesian networks (see section 3.3.8) are an appropriate machine learning approach as an alternative to this method of graph construction, which instead use conditional probabilities.

2.4 Economic Research

The common assumption that wealth causes happiness has led a large amount of research to focus on this particular area. This research varies in many aspects such as: coverage (global, EU or a single country), type of data used (longitudinal or single instance), topic focused on (both general analysis and investigations of specific theories or previous findings). For a survey of such papers the reader is directed to [13]. It is generally agreed that life satisfaction is strongly and positively correlated with income shown in much previous work such as [16].

A much debated theory is that of the Easterlin Paradox. Richard Easterlin, a Professor of Economics has contributed much to this area, most notably a paper written in 1974 entitled "Does Economic Growth Improve the Human Lot? Some Empirical Evidence" [17] Easterlin used two survey questions, and investigated them separately. The analysis performed included both within country (America) and cross country data. A key finding was the relationship between income and happiness, which Easterlin found to be strongly positive for data within countries. However, the cross country analysis found a threshold of income above which there was no correlation. These findings were named the Easterlin Paradox.

The Easterlin Paradox has been a central focus of debate since then, and the findings are not decisive. Easterlin notes; "China's growth rate implies a doubling of real per capita income in less than 10 y ... one might think many of the people in these countries would be so happy, theyd be dancing in the streets" [18]. An early view by Robbins in 1938 suggests an opposing view, where after income has reached a certain level it allows happiness values to increase more, because then additional factors can contribute to life satisfaction levels as improvements to quality of life become affordable. The work by Robbins has been supported by more recent research such as [16], a 2008 study using the Gallup World Poll (GWP). This work showed that life satisfaction was affected more in the rich countries, shown by a noticeable change in the regression curve.

Easterlin's most recent paper on this topic was that of "The happiness-income paradox revisited" in 2010 [18]. This is an extension of previous work that focused on America. This research looks at happiness over a number of years and for a wide range of countries (54 in total), using time series data from four survey sources. Regression was performed both with the datasets as a whole, and also with subsets where the countries were split into groups. The three groups were; developing, developed and those in a state of change from communism to capitalism. OLS regression was used and this showed no significant correlation between rate of growth and life satisfaction in all regressions performed. Easterlin shows that while happiness does fluctuate with economic conditions, there is no correlation in the long term for the diverse range of countries studied. Previous research by Stevenson and Wolfers in 2008 finding that life satisfaction and growth are positively correlated but Easterlin comments that they use only short term datasets. He shows that while this is the case there is no evidence of a longer term correlation. [18]

This paper highlights the affect that a few possibly unrepresentative data points can have on correlations. As an example, the results by Stevenson and Wolfers in 2008 were repeated removing a small number of results. A correlation that had been found for a dataset of 17 countries, was re-tested with 2 data points removed and this change meant no significant correlation was then found. Easterlin also did this for another test with 32 countries, removing transition countries and finding this again removed the significance. However, here 11 countries were removed which is a large proportion and hence would more likely affect the correlations found.

This paper has provided a valuable insight into the dual nature of correlation between growth and happiness (short term and long term), rather than the rather simplistic view that growth causes happiness as found previously.

> Solely GDP growth and time series data was used in this study. Our research is interested in the happiness differences between countries and hence it is more appropriate to use point of time data. We also look at several alternative indicators such as GDP and relative income, as well as GDP growth rate.

A possible issue with this research is the use of financial satisfaction as a label for Latin American countries. However, this was due to a lack of reliable life data and therefore this was probably the most suitable alternative. A positive correlation would not prove a correlation between life satisfaction and growth, but no correlation with financial satisfaction would render a positive correlation with life satisfaction highly unlikely as financial satisfaction is the component of life satisfaction most related to GDP growth. In fact, financial satisfaction showed no relationship with GDP growth which is of much surprise, especially where Latin American countries are growing at 1 - 3% per year. [18]

2.4.1 Survey data and conflicting results

The debate regarding the Easterlin paradox may be due to differences in survey data. It is important to understand possible affects that different survey procedures may have to be aware of possible bias in results. Recent work in 2008 by Deaton ([16]) and Bjrnskov ([7]) respectively has looked at the possible reasons that results from WVS support the Easterlin Paradox but results from the GWP conflict.

There are some fundamental differences of the surveys noted by Bjrnskov. Firstly, WDH is an accumulation of multiple sources (although primarily WVS). Secondly, the time point of the surveys is different, having taken place at around 2000 for WVS compared to 2006 for GWP. [7] This is a large time gap and there could potentially have been noticeable changes in life satisfaction during this time.

Deaton notes how the GWP shows a much gentler slope for the correlation between happiness and GDP, whereas the WVS gives a much steeper rise for low income countries. Deaton accounts this to several attributes of the surveys. Firstly, WVS includes a smaller set of low income countries and this part of the correlation may be skewed by the fact that these countries are mainly post Soviet Union. These countries may have much lower LS because of this, and without other poor countries to show a balanced view this has caused the sharp rise at the start of the graph.

An additional possible cause is the sampling performed by the WVS, which in some cases is not representative of the population as a whole. The population sample in some countries was taken from those of higher intelligence.[3] [16] This is quite surprising and means the data could be skewed with higher happiness values for these countries. Deaton claims that this in combination with the Soviet countries mentioned previously has created two types of country in the low income range, where they are either unusually satisfied or unusually unsatisfied. [16]

> The results in [16] are interesting as they do not indicate that one survey is more reliable than the other, but instead highlight the limits that smaller datasets place on the results and correlations that can be found. With this in mind a label with a large sample is preferred, and hence using multiple sources is also considered.

It is noted in [3] that GWP values are more conservative than WVS values such that "It tends to find higher life satisfactions for rich countries and lower life satisfactions for poor countries than, for example the World Values Survey does" [3]. Two possible explanations are given for this. Firstly, possible differences of the coverage within a country, where for instance one survey may reach more rural areas. Secondly, the question order is different between GWP and WVS, where the life satisfaction question is after economic and governmental question in GWP but at the beginning for WVS. This seems likely to cause the difference where people in poorer countries are likely to give a lower answer after thinking about the poor state of their countries economy, with an opposite effect to this for richer countries. Bjrnskov also suggests this, additionally noting that the answer scale of the question can affect results. [7]

The wording of a question may also affect the results. Bjrnskov discusses this comparing the ladder of life and the more standard life satisfaction question. The key difference is that the ladder of life bounds of values are referred to as the respondents bounds, rather than more general bounds, using wording such as best possible life for you. Bjrnskov says "the anchoring technique employed by the GWP is likely to produce smaller scores on average, as responses are probably anchored in comparison with an ideal situation instead of the weaker anchoring in a cognitive state of 'complete satisfaction' used by the WDH"[7].

Possible issues also exist with regards to normalising question meanings across the globe such that each respondent is answering the same question. Specifically they note how the word

[3] This has been checked for 2008 (sampling info at: `http://www.wvsevsdb.com/wvs/WVSDocumentation.jsp?Idioma=I`) and no bias can be found.

‘happy’ needs to be translated accurately across different languages and “the English word ‘happy’ is notoriously difficult to translate”. [7]

Whilst this work has focused on the correlation between GDP and happiness, it has highlighted many general differences of surveys which could affect correlations of many other variables.

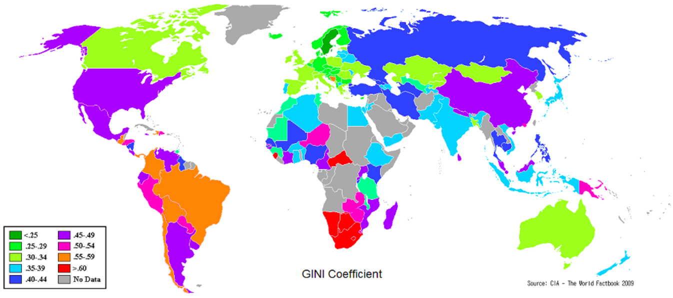
> Small differences in survey format can severely affect results. The WVS and GWP surveys are both not ideal for different reasons; the selection of respondents for WVS, and the position of the life satisfaction question in the GWP.

2.4.2 Relative wealth

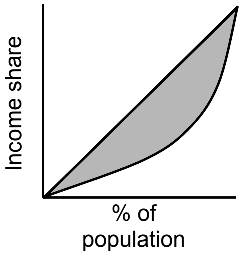
A Lorenz curve (figure 3(b)) is a graphical representation of the distribution of wealth. The curve indicates the proportion of the wealth at each percentage of the population. The more convex the curve the greater the inequality, where a larger proportion of the population has a smaller proportion of the wealth. A uniform distribution (straight line) represents complete equality. The Lorenz curve is the idea behind a common deprivation measure known as the Gini index, which is defined as:

$$Gini(A, B) = \frac{A}{A + B}, \tag{1}$$

where B is the area under the curve and A is the shaded area above the curve. The values range between zero and one from uniform to highly unequal respectively.



(a) Gini Index, Global Map [9]



(b) Lorenz Curve [24]

Figure 3(a) shows global Gini index values from the CIA World Factbook[4] 2009. This shows a large amount of variation between countries and even a visual comparison between global values of Gini index and happiness (figures 3(b) and 1.1 respectively) shows some similar patterns. For instance, countries in Europe generally has higher happiness levels and a lower Gini index than those in Africa.

> Distribution of wealth may be a valuable indicator of happiness.

2.4.3 A study of relative income

Relative variables such as the Gini index may be valuable indicators of happiness. The variables we will consider will represent relativity on a country level. Work by Mayraz et al. in 2009 ([34]) investigated the affects of relative income but considered different groups as the comparison namely friends, neighbourhood and colleagues. The data used were survey questions asking to give values for their relative income in relation to the different groups mentioned above, and also the importance they gave this comparison. In fact, these questions were designed by Mayraz specifically for this research.

[4] World Factbook is a global dataset of information on each country and can be found at https://www.cia.gov/library/publications/the-world-factbook/

Linear regression was used, with slightly different regression equations to answer different questions. The regression methods are reviewed in detail as they show the flexibility of linear regression as a tool for data mining (see section 3.3.4).

The basic regression equation is:

$$H_i = \alpha + \beta Y_{R_i} + \beta Y_i + \sum_k \lambda_k X_i^k + \varepsilon_i \tag{2}$$

This includes dependent variable H for life satisfaction, constant α and error ε. The regression variables include relative income Y_R, absolute income Y_i and a set of control variables X_i. This is a standard regression function using control variables to account for variation in H due to other factors such as age and education.

Firstly, the correlation between relative income and life satisfaction was investigated using a regression equation as above but using log absolute income. The regression was repeated for each group and also with and without the absolute income variable. The regression results show quite surprising results, where a significant correlation is found between relative income and life satisfaction for men but not for women. The coefficient values are smaller when log absolute income is included in the regression, which is expected as this shows that absolute income makes some contribution to a persons' life satisfaction valuation.

A second investigation in this research looked at the relationship between a persons' happiness, the importance they subjectively place on relative income, and the actual importance of relative income with respect to happiness. This is important to show that the subjective value is an adequate representation of the actual importance (as perhaps people perceive it to be important but in reality it does not affect their happiness). To do this they use the regression formula of equation 3, where the key difference is the use of an interaction term, which is the product of relative income and subjective importance (see section 3.1.1 for details of interaction terms).

$$H_i = \alpha + \beta_j Y_{R_i}^j + \beta_j' I_{R_i}^j + \beta_j'' Y_{R_i}^j I_{R_i}^j + \lambda log Y_i + \sum_k \delta_k X_i^k + \varepsilon_i \tag{3}$$

Put simply, this equation is asking whether the correlation between actual income and happiness is dependent on the subjective importance of relative income. A result where the coefficient of the interaction term is high, and the coefficient of the other terms is reduced, would show that the interaction between relative income and perceived importance accounted for the variation of well being better than the individual variables. This would infer that a persons' subjective importance does bear relation to the actual importance of relative income with respect to happiness. Note how this contrasts to the standard regression equation with only independent variables where each contribute individually through summation to the dependent variables value. The results of this regression gave very low coefficient values for the interaction term. This shows that the relationship between relative income and happiness is not governed by a persons perceived importance of relative income.

One potential weakness of the methods used was the use of surveys to retrieve values for all variables including that of relative income. A more robust method may have been to survey each persons income, and record the individuals within the groups so that relative income can be determined precisely. This however, would be a fairly arduous task. Mayraz tackles this issue by analysing the causality between happiness and subjective relative income. The concern was whether a higher relative income caused people to be happier or whether happier people were perhaps more optimistic in their estimations of relative income. To test this, relative income was used as the dependent variable and an interaction term of happiness and importance was included. This was to test if the relationship between relative income and importance is dependent on happiness, but the results showed that this was not the case. However this does not show whether subjective relative income is representative of actual relative income on a more fundamental level.

A final investigation by Mayraz looked at whether the difference in happiness from the mean is of equal magnitude either side for higher and lower relative income values. For instance, given mean relative income m_i with happiness m_h, and person X with relative income $mI + p$ and happiness $mH + q$. If a person Y has relative income $m_i - p$ is his happiness $m_h - q$? The hypothesis tested is that those with low relative income lose more happiness that a person gains who has high relative income, meaning that with each unit increase in relative income, the change in happiness decreases. Testing this with linear regression required transforming this hypothesised correlation into a linear one, by way of a quadratic transformation (illustrated in figure 3). This is the third term in the regression formula (equation 4 below).

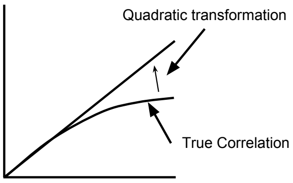


Figure 3: Graphical explanation of quadratic regression

$$H_i = \alpha + \beta_j Y^j_{R_i} + \beta'_j (Y^j_{R_i})^2 + \lambda log Y_i + \sum_k \delta_k X^k_i + \varepsilon_i \tag{4}$$

Mayraz notes that the coefficient for this quadratic variable will be negative if there is an asymmetric relationship. This is because the relationship hypothesised is concave and these result in negative coefficient values, in contrast to convex correlations which have positive coefficient values. To see why take a simple regression equation $y = ax^2 + b$ and differentiate with respect to x to give $\frac{\partial y}{\partial x} = 2ax$. It can be seen that negative values of a give a decreasing rate of change as x increases and hence a concave curve. The results for each group however give values very close to zero, and in fact only 3 of these are negative values. This indicates that the rate of change of happiness with relative income does not decrease as relative income values increase.

This research has demonstrated the versatility of linear regression to answer a range of questions regarding variable relationships, using both interaction terms and logarithmic and quadratic transformations. Of particular note is the use of interaction variables to analyse causality between variables, showing an additional approach to using longitudinal data.

Other research looks at non economic variables, such as a 2005 paper by Oswald et al. "Does happiness adapt? A longitudinal study of disability with implications for economists and judges" [39]. Oswald et al. (2005) analysed time series data to investigate the affect a disability has on well-being. The results showed a clear trend, where life satisfaction dropped significantly on gaining a disability, but remarkably bounced back to just below the original value showing the strength of our ability to adapt in adversity.

2.5 Research of Other Indicators

2.5.1 Health

Health may have a marked effect on happiness, a notion supported by previous research. We suggest a distinction between health variables depending on whether they are recoverable, due to the psychological differences between the two situations. For instance the paper above discussing disabilities showed happiness values returning to previous levels. A long standing but non permanent health issue however, may have a higher affect on happiness, due to the belief that they could be healthier than they are. A person becoming disabled does not have this hope and therefore adapts to this permanent change in their life.

Causality relationships with health may be particularly unclear, where conditions such as hypertension could be effects rather than causes of happiness. This does not affect a variables viability as an indicator and both types will be considered for this project.

Blanchflower & Oswald have investigated the relationship between hypertension (high blood pressure) and happiness [8], with an aim to incorporate hypertension as a variable in a well-being index. This research used life satisfaction and happiness responses from the Eurobarometer survey, and the hypertension values were also derived from the survey data. Correlations were assessed using both Pearsons and Spearmans rank tests, and OLS and logit regression[5] were used. The R^2 was used to validate the regression model (see section 3.3.1).

Individuals were used as entities with country dummy variables as controls (a binary variable for each country where the country of the individual is 'on' and the others are 'off'). The dependent variables of the regression tests were hypertension and life satisfaction. The results showed correlation between hypertension and happiness. For instance, when looking at the very satisfied responses the countries with the lowest blood pressure gave significantly higher happiness values (48.5% compared with 22.5%).

While the results are encouraging, some aspects of this research may not be ideal. Firstly the data sample is not quite small, consisting of just 16 countries. Also, the hypertension data came from the question "Would you say that you have had problems of high blood pressure?", which makes this a subjective measure. We question whether hypertension can really be estimated accurately through self assessment, as for instance it does not distinguish between high blood pressure and hypochondria. However, this was necessary for this research because the entities were individuals and hence the data needed to consist of happiness and hypertension values for each person. It may be worth a study into the accuracy of this self assessment through asking this question to individuals where actual blood pressure readings have been taken and can thus be compared.

Our project, with country entities uses more concrete health indicators derived from country records rather than self assessment such as life expectancy.

2.5.2 Light

Light may be an interesting correlate because of possible links with depression. Seasonal Affective Disorder (SAD) is a type of depression where the individuals are affected only in the winter months. SAD is thought to be caused by the lack of light, which affects the release of certain chemicals in the brain causing a change in mood [40]. Depression levels of SAD sufferers can vary widely throughout the year and this suggests light may have be a contributor to happiness levels, although no previous data mining work could be found regarding this. Previous research has however, found correlation between prevalence of SAD and latitude, such as [44].

Light may correlate with happiness and hence is considered as a feature.

2.5.3 Climate

Previous work has found evidence for a relationship between climate variables and happiness. The implications of a correlation between happiness and climate are vast. The threat of climate change causing large shifts in weather patterns means that a relationship between climate and happiness would likely cause changes in happiness levels. Climate variables add much complexity to this problem as they are likely to be closely related to other variables of interest. A key

[5] Regression where the data is fit to a logistic curve

example of this is health, where weather causes an increase of some illnesses. For instance, flu is known to be higher in the colder months[6].

Rehdanz & Maddison performed a comprehensive investigation into the relationship between climate and happiness using a wide range of variables [42]. The choice of variables includes some interesting options. Firstly the climate variables used were annual mean values of temperature and rainfall and indicators of extremes such as the precipitation of the wettest month. Absolute latitude was used to represent amount of daylight. Other variables were also included to control for other differences. Of particular interest is the construction of a variable to control for the countries that were previously communist, necessary as their initial analysis showed that these countries were the ones with the lowest temperature. A variable is needed to control for this or else the results would be biased and incorrect. Other variables included are taken from previous research such as religion and life expectancy.

Of particular note was the pre-processing of the variables to produce appropriate indicators. This is needed to ensure the values are representative of the country as a whole, because climate variables vary across countries and this is independent of the distribution of populations. The values assigned should be values representative of the populated areas. [42] accounted for this by taking the weighted average of several cities of a country, weighted by the population.

> Care should be taken to ensure the variables are representative of a country, such as by using a weighted average as in the example above.

Several regression tests were performed using different groups of climate variables, together with the control variables. This segregation was needed to remove collinearity, which would affect results. For example, two variables are the average mean temperature and the number of months where the temperature exceeds $20°C$, and these variables clearly have a close relationship. A regression equation with both these variables would attempt to attribute the same variance of the label to both of these variables.

This research performed separate regression using the different variable groups as explained above, and therefore it is interesting to compare these results. The results showed that the max and min temperature variables correlate negatively and positively respectively, which is expected as this infers people prefer medium temperatures to extremes. However, the regression using count variables of the number of cold and hot months, showed a negative correlation between number of cold months and happiness and a positive correlation between the number of hot months and happiness, indicating a preference for warmer climates. This is supported by the final regression which used mean values, and found a strong positive correlation between annual mean temperature and happiness. The t-statistic was however lower than that of mean rainfall in this regression showing that this has a stronger correlation than mean temperature.

> The variable groups represented similar concepts but resulted in different regression models, highlighting the sensitivity of variables selection and the importance of trying different alternatives that represent similar concepts.

The regression with maximum / minimum values proved to be the best representation, with the highest R^2 value of 0.7918 (possible values range between 0 and 1 where 1 means the variables completely predict the dependent variable), although this is only marginally larger. In fact, all three models showed significant correlations with happiness, with a highest f-statistic of 0.0081. Also, all models generated using the different groups of climate variables passed the RESET test showing these models were able to represent correlation with happiness.

[6] Search trends show annual cycles indicating correlations between health and climate. Search for flu - `http://www.google.com/trends\?q=flu&ctab=0&geo=gb&geor=all&date=all&sort=0`

These results show interesting results, primarily that climate indicators are good correlates with happiness.

2.6 Causal Inferences

Causality of happiness is an interesting subject as there are many possible correlates where the direction of causality is unclear. There has been only a limited amount of work looking at this.

Investigating causality is beyond the scope of this project.

Previous work has found married people have higher happiness levels and research by Stutzer et al. in 2006 looked at the relationship between marriage and happiness [47]. Establishing causality required the use of longitudinal data[7]; a dataset recording values related to the same entity over time. The relationship between changes in variable values over time can indicate causality. For instance, does happiness increase after marriage to infer that marriage makes people happy? Or are happier people married because one is more likely to find love if they are happy?

The dataset was split into three groups; remaining single, married and marry later in life. The life satisfaction scores were adjusted for different factors such as age and gender, but the methods for this are not specified. Even so, the results are fascinating, and simply through graphing the data and visual examination interesting results can be detected. For instance, at age 20 the people who go on later to marry have much higher life satisfaction than those that stay single. Also for these people, there was a steady increase in happiness in the years prior to marriage, which returns to the previous value in approximately the same length of time afterwards. Additionally they find this trend is also found for people who marry and subsequently divorce but at lower levels of happiness. This shows that people who are less happy are less likely to marry and more likely to divorce, indicating happiness causes marriage.

Longitudinal data analysis is a valuable way of determining causality and as can be seen from this paper the causality can often be very apparent. However, this has also highlighted the fact that happiness is often not changed suddenly, but gradually over a period of time. For instance, increased happiness caused by marriage is not suddenly altered but gradually increased over a number of years prior to marriage, perhaps due to the hope and thoughts of marrying in the future or being in a long term stable relationship. This means that analysis needs to be over a wide time span to fully investigate causal affects.

This indicates that variables may have better correlation if several years are used. One option is to construct a weighted average where more distant years have less contribution.

[7] Survey source: German Socio-Economic Panel Study (GSOEP)

3 Methods Overview

3.1 Data Preparation

The success of classifiers and prediction methods depend heavily on the quality and choice of the data used. [5] Domain knowledge is important to be able to choose features that are likely to bring good results, and previous work reviewed in section 2 provides important background information of this subject area. Data preparation and feature selection are two critical parts of a data mining project.

3.1.1 Feature construction

The data collated can be used directly or can be manipulated to provide alternative features which may be more effective in models. This stage is an opportunity to use domain knowledge to construct appropriate features for analysis. This includes simple techniques such as combining values into a single feature, or more complex methods such as principal component analysis (PCA).

Feature variables can be constructed in the following ways:

- **Time correlations** Time series data can be used to calculate the relative change of a variable, which can be used as a feature.
- **Transformations** Data transformations such as log, square-root, square and PCA (see below).
- **Interactions** Features can be constructed using mathematical functions on several variables. Many of these are readily available such as GDP per capita.

Transformations Transformations are common in linear regression, as the linear nature would otherwise be restrictive when non-linear relationships exist. Performing transformations allows non linear correlations to be found when using linear techniques.

Transformations with logs is particularly common, and additionally this gives an additional property beyond just changing the relationship between the variables. Standard variables correlate with the dependent variable in the standard way of the formulation $a = bx + c$, such that when the value of x changes by 1 the value of a changes by b. However the values of the coefficients produced depend on the units of the variables. Introducing logs changes this so that the relationship between the log of the independent variable and the dependent variable is now in percentages such that it represents the relative rather than the average change. [20]

Interaction terms Interaction terms are useful when the relationship between several variables needs investigating, such as where the correlation between an independent and the dependent variable may additionally depend on a third variable (see section 2.4.3 for an example using this). This is a simple regression function with two variables X and Y that are also interaction terms:

$$f(X) = \beta_0 + X_j\beta_j + X_k\beta_k + \sum X_jY_j\beta_j, \tag{5}$$

3.1.2 Principle Component Analysis (PCA)

PCA is a powerful method to find hidden patterns in datasets, by performing an orthogonal[8] transformation of the dataset into a set of new variables called principal components. Given a set of data points we can view these as points in a multidimensional space, where the axes are

[8] Orthogonal refers to the fact that the axes of the principal components are perpendicular to each other, just as the x and y axis in a typical 2 dimensional space

just arbitrarily defined. The axes can be moved such that they correspond to the directions of highest variance, and this can reveal hidden patterns in the data.

To perform PCA it is important to first normalise the data, so that all variables have mean zero and variance 1. If variables with different variance are used this would affect the PCA results. The problem involves solving an eigenvector/eigenvalue problem and the derivation is given below.

PCA derivation Given a data matrix $\mathbf{X}$, where each row is the transpose of a data point:

$$\mathbf{x}_i = \begin{bmatrix} x_i(1) \\ x_i(2) \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_1(1) & x_1(2) \\ x_2(1) & x_2(2) \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \end{bmatrix},$$

PCA finds vectors $\mathbf{w}$ representing new axes, along which the variance of the data points is maximised. The position of $\mathbf{x}_i$ along this axis is found by projecting onto $\mathbf{w}$: $\mathbf{x}^T\mathbf{w}$. For example, if $x = \begin{bmatrix} 2 & 1 \end{bmatrix}$ when projected onto vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is:

$$\begin{bmatrix} 2 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 3$$

The variance of a set of values is given by $Var(X) = E(X - \mu)^2$. Under the assumption that the data is centred ($\mu = 0$) then $Var(X) = E(X)^2$. Therefore, the variance of $\mathbf{X}$ along $\mathbf{w}$ is $E(\mathbf{x}^T\mathbf{w})^2$. Therefore the aim is to find:

$$max_{\mathbf{w}} \frac{1}{n} \sum ((\mathbf{x}_i^T)\mathbf{w})^2, \qquad [15] \qquad (6)$$

where n is the number of data points. This equates to:

$$max_{\mathbf{w}} \frac{1}{n} \mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} \qquad [15] \qquad (7)$$

Increasing the values of $\mathbf{w}$ will always increase the results of this however, so $\mathbf{w}$ is bounded by $\mathbf{w}^T\mathbf{w} \leq 1$. This value will always equal 1 in solutions because for any $\mathbf{w}$ where $\mathbf{w}^T\mathbf{w} < 1$ a higher value of equation 6 will always be found by increasing $\mathbf{w}$ such that $\mathbf{w}^T\mathbf{w} = 1$.

This is incorporated into the problem equation using a Lagrange multiplier. These are commonly used to find the solutions of problems seeking to minimise or maximise a function subject to a particular constraint. In this case we are maximising equation 6 subject to the constraint $\mathbf{w}^T\mathbf{w} \leq 1$. Equations using Lagrange multipliers take the following form:

$$f(\mathbf{W}, \mathbf{w}) - \lambda(g(\mathbf{X}, \mathbf{w}) - k), \qquad [6] \qquad (8)$$

where f is the original function, g is the constraint and k is a constant such that $g(\mathbf{X}, \mathbf{w}) - k = 0$. Hence in this case $k = 1$ because $\mathbf{w}^T\mathbf{w} - k = 0$ and $\mathbf{w}^T\mathbf{w} = 1$.

Therefore the PCA formulation becomes:

$$max_{\mathbf{w}} \frac{1}{n} \mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1) \qquad (9)$$

This is maximal where the gradient w.r.t. $\mathbf{w}$ is 0. The following equations state some basic rules for matrix differentiation:

$$\frac{\partial \mathbf{w}^T A \mathbf{w}}{\partial \mathbf{w}} = 2A\mathbf{w} \quad \frac{\partial b^T \mathbf{w}}{\partial \mathbf{w}} = \mathbf{w}$$

Differentiating and equating to zero results in equation 10, which is an eigenvalue problem where λ is an eigenvalue for matrix X. There number of d $\lambda/\mathbf{w}$ pairs equals the original dimensionality of the data space.

$$\frac{1}{n}(\mathbf{X}^T\mathbf{X})\mathbf{w} = \lambda\mathbf{w} \qquad (10)$$

The $\mathbf{w}$ with the highest variance corresponds to the largest eigenvalue because:

$$\frac{1}{n}\frac{(\mathbf{X}^T\mathbf{X})\mathbf{w}}{\mathbf{w}} = \lambda \tag{11}$$

$$\frac{1}{n}\frac{\mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w}}{\mathbf{w}^T\mathbf{w}} = \lambda \tag{12}$$

and $\mathbf{w}^T\mathbf{w} = 1$, hence:

$$\frac{1}{n}\mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} = \lambda, \tag{13}$$

and the left side equates to the variance (we are back to the original equation 6). [15] The eigenvector $\mathbf{w}$ is a new axis and it's eigenvalue λ represents the degree of variance along it.

$\mathbf{W}$ is a matrix of vectors of the directions of greatest variance, ordered by eigenvalue. Those $\mathbf{w}$ with low λ typically have little variance and can often be ignored.

$$\mathbf{W} = \begin{bmatrix} w_1(1) & w_2(1) \\ w_1(2) & w_2(2) \end{bmatrix} \tag{14}$$

The original data points can be transformed into the new space by projecting onto the vectors in $\mathbf{W}$:

$$\mathbf{X}_{new} = \mathbf{X}\mathbf{W} \tag{15}$$

The relationship between eigenvalue and variance means that each principal component p_1 to p_d, ordered by eigenvalue contains the highest amount of information (variance) not accounted for by p_1 to p_{d-1}. This is important for data analysis as the key principal components are often highly informative. Additionally, the dimensionality can often be reduced by using the PCs instead of the original data and ignoring PCs containing a low proportion of the variance of the original data.

3.1.3 Missing values

The data used in this project will contain some missing values, as it is collated from different sources. Machine learning methods often require that all the features have values, although many have inbuilt methods for dealing with this. We may prefer to choose to remove missing values prior to using ML methods and there are three main ways of doing this; reduction of the dataset, indicator variables, and imputation. Reducing the dataset involves removing entities where the data is incomplete. This is not feasible for this project as this will remove whole countries from the analysis (although the coverage will be considered when choosing potential features).

In some cases there is an underlying reason why values are missing from the dataset and this may in itself contribute valuable information. As an example, time series data of a country may have missing values during periods of conflict or natural disaster, and this may itself be a correlate with happiness. Data analysis using background knowledge is important to determine the significance of a null value and whether it may have relevance. If so the null values can be replaced with an indicator variable to represent the underlying cause. [53] However, for this project it will usually be the case that the data is missing because a country was simply not included, as each data source will have slightly different coverage. Indicator variables may be appropriate in a minority of cases but inference of missing values (imputation) is likely to be of most use.

3.1.4 Imputation

Imputation is the technique of using the data available to infer a value that is missing from the dataset. Features are used to complete the data of another feature, which can then be used in the data mining methods. Two well known options are mean (or mode) or nearest neighbour. Regression can also be used to infer values, an example of this is given in section 2.2 where linear regression was used to infer values of GWP from other survey sources. Often ML methods have built in methods for imputation. For example, decision trees have in built methods to infer a missing values, such as the surrogacy method of the M5 algorithm (section 3.3.6).

Mean or mode This is a naive approach of replacing a missing value with the mean or mode values of the entire dataset. The value assigned is the same for all instances and no attempt is made to infer a value more representative using other variables of the instance.

Weighted k-nearest neighbour The nearest neighbour algorithm is a simple inference method for numeric variables where the missing value is assigned the value of the entity closest to it according to a specified distance measure. The k-nearest neighbour (KNN) is an extension where it is assigned the average value of the k nearest elements, and weighting gives closer nodes a larger contribution to the value. [53]

Previous work has investigated the performance of weighted k-nearest neighbour imputation. Work by [29] included testing on several datasets, by removing values to create data with artificial missing values. The missing values were then imputed and could be compared against the true values by calculating the average error of the predicted values. The tests were performed both with the Euclidean and Manhattan distance measures. The Euclidean distance is given by:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}, \tag{16}$$

and the Manhattan distance by:

$$d(X,Y) = \sum_{i=1}^{n}(X_i - Y_i), \tag{17}$$

The Manhattan distance is the sum of the individual distances between the variables, whereas the Euclidean is the distance moving directly from point X to point Y in the variable space. The key difference is that the Euclidean distance prefers *fewer* larger deviations to *many* smaller deviations. As an example consider two feature sets X and Y with two features. Consider the following two situations for the distance d_i between X and Y for the two features; $\{d1_1 = 1$, $d1_2 = 3\}$ and $\{d2_1 = 2$, $d2_2 = 2\}$. The distance between both these feature sets for the Manhattan distance is 4. However, the Euclidean distance gives different values; $e_1 = 3.16$ and $e_2 = 2.8$. In this situation the first feature set would be deemed closer according to the Euclidean distance, but this may not be ideal as one of the variables has a larger distance and is hence less similar.

Imputation was also performed with the mean method to provide a comparison. KNN with Euclidean distance was found to predict the values better than both KNN with Manhattan and the mean imputation method. Both NN methods showed better results than the mean method. This was expected, but in fact NN was better in 71% of cases and we might expect this to have been higher. The Manhattan distance proved better than the Euclidean, although these results were quite similar. Manhattan may be preferable because of the bias in the Euclidean measurement for longer distances.

An important consideration when using KNN is the variable units, as KNN is sensitive to the scales used. For instance, a distance variable may be given in miles or metres and this affects the KNN results. Therefore normalisation needs to be performed so that the scales are consistent and comparable. Also, the features may have different importance. Therefore it is necessary to determine which features are most relevant. Incorporating features that bear no correlation to the missing value can skew results. Variables may have different relative significance, and it may be preferable to weight the distances according to this. [53]

A further difficulty is the choice of K, which needs to be optimised manually. A K that is too small means that only a few instances are used which may not give a good approximation of the missing value as it over fits the data. A K that is too large will mean instances that are quite far (and hence perhaps quite different) contribute to the instances value. Finding the optimal value of K requires testing with different values. [4]

KNN will be useful to impute missing values in our dataset. The data available to us contains variables representing different aspects of human life such as health, education and wealth with multiple variables of each. We ideally would use a single variable representing a single concept to prevent dependence between features. The missing values of a feature could be imputed from the set of features representing the same concept, which is appropriate as the features are likely to have a good correspondence with these variables. For example, using several educational variables impute missing values of another education variable may be effective.

3.2 Statistical & Machine Learning Methods

We use both statistical and machine learning methods throughout this work. ML techniques can be divided into two main types; black box or white box. The methods used to classify or regress can be understood and analysed with white box techniques, which is useful to gain understanding of the relationships between the input and target variables. We intend to use several statistical and ML methods, and describe those of particular interest.

3.3 Statistical Tests

The main statistical methods to determine results significance are correlation with R^2, t-tests and permutation testing.

3.3.1 Pearson Product Moment Correlation Coefficient (R)

R is a measure of correlation of two datasets with values ranging from -1 to 1 where -1 and 1 represent perfect negative and positive correlations respectively. More specifically, this measure represents the variance of the label that is explained by the model relative to the unexplained variance. A value of 0 indicates that there is no correlation between the variables. R is the covariance of the variables relative to their individual variance, given by the formula:

$$r_{x,y} = \frac{C_{x,y}}{\sigma_x^2 \sigma_y^2}, \tag{18}$$

where C and σ^2 are the covariance and variance respectively.

Quantitative measure for results analysis A consistent measure is needed to compare results and the correlation coefficient (R) will be used. This is appropriate because it is a measure of the goodness of fit of a model. The correlation value is not affected by the number of features of the model, which is important as we are comparing tests involving a variable number of features.

P-value A p-value indicates the significance of a result, representing the probability the result would occur by chance. A low value means the result is unlikely to occur randomly and hence is statistically significant. Threshold values of 5% and 1% are commonly used, below which a result can be stated as significant. Statistical tests can be relative to test parameters such as the size of the dataset, and a p-value provides a comparable value that takes into account such aspects of the data.

R^2 and Adjusted R^2 R^2 is also known as the coefficient of determination, and is an extension of the R value. It represents the proportion of variation in the label that can be accounted for by the regression model. However R is relative to the number of variables used in the model and therefore is not comparable when this differs. Adjusted R^2 takes into consideration the number of variables used in the regression.

3.3.2 T-Test

A t-test calculates the likelihood that two datasets are generated from the same probability distribution. We use this measure to compare results such as the performance of two learners. A t-test performed on the results of 10 fold cross validation for instance determines the likelihood that these two results sets come from the same distribution. A result indicating they are from different distributions indicates that one performs significantly better.

3.3.3 Significance testing with permutation testing

Permutation testing is used to compute a p-value of results when the probability distribution is not known. For instance, suppose we find a model with a correlation value and we wish to know the likelihood that a model with this correlation would occur by chance for data from the same distribution. The null hypothesis states that the correlation is likely to occur on random data. Therefore the null hypothesis is rejected when the result is unlikely on random data and hence is significant.

Permutation tests amount to sampling from the distribution of a data set by randomly permuting the data. The samples are taken many times and the test performed and this gives a probability distribution of results for the random data. The proportion of times the results are 'better' than that originally found is the p-value:

$$p - value = \frac{\#p \in P : T(p) \geq v}{\#p \in P}, \tag{19}$$

where v is the original test value, P is the set of permutations and T is the test performed. Equivalently, a threshold value t_σ such that at most 5% of the permutation tests give a value 'better' than this value can be found:

$$\frac{\#p \in P : T(p) \geq t_\sigma}{\#p \in P} \leq 0.05, \tag{20}$$

This value can then be compared against all results to determine if it is significant. This is a useful alternative where the significance of several results needs assessing, because a single threshold value can be compared against many results rather than calculating a specific p-value for each.

3.3.4 Linear Regression

Linear regression is a method to model the output value as a linear contribution of the inputs. With respect to this project this can be used to determine a function $f : X \rightarrow Y$, where X is a vector of feature values and Y is a happiness ground truth. The linear nature of the representation produced can sometimes be too simplistic, as the target may not be linearly correlated with the feature set. However, linear regression is well known to produce accurate results and can also outperform more complex non linear methods.

A typical regression formula is:

$$f(X) = \beta_0 + \sum_{i=1}^{n} X_j \beta_j, \qquad [27] \tag{21}$$

with bias β_0 and weight β_j of feature j in X. The feature space is a multidimensional space and the aim is to find a line in this space that best fits the data points. β_0 would then represent the intersect on the axis of $f(X)$.

There are several methods to estimate the coefficients β_j for the model. These include; least squares, principal component regression, ridge regression and the lasso.

Least squares The least square is a simple and common measure of error of a regression model given by;

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - f(x_i))^2. \qquad [27] \tag{22}$$

with n elements in the training set, target value y_i and regression output $f(x_i)$. Linear regression using this measure aims to minimise the sum of the squared error (or residue) on the training data. The square of the error is used instead of just the absolute error to give a preference for smaller deviations from the target values.

3.3.5 The Least Absolute Shrink and Selection Operator (Lasso)

A significant weakness of OLS regression is that it cannot regress to the correct function where the feature variables are not independent. There are three key methods which improve OLS regression; subset selection, ridge regression and lasso. [49] Lasso, was originally proposed in 1996 by Tibshirani as an alternative to ridge regression and subset selection. [49]

Subset selection and ridge regression have some drawbacks and Tibshirani's aim, described in [49], was to combine the best of both, and he describes the following shortfalls with these techniques. Subset selection includes methods such as stepwise regression, and these methods involve fairly large discrete changes to the model, as variables are added or removed. Also, this method is quite sensitive to changes to the input data. Ridge regression is an improvement on subset selection because it uses continuous changes to the model by altering the size of the coefficients in the regression function. However, it is unable to reduce any coefficient to zero and therefore does not remove any completely from the feature space. [49]

Ridge regression and lasso are both a specific type of regularisation method known as shrinkage methods. Regularisation methods constrain the search space by using penalties, and specifically for shrinkage methods this reduces the size of the coefficient in the regression function. The search space is the set of all possible regression functions for a feature set, with all possible coefficient values. As an example, the error calculation could incorporate the sum of squares of all coefficients, such that when the coefficients have lower values the error is also lower (this example is the penalty for ridge regression as detailed below). [6] Regularisation prevents over fitting by imposing pressure towards simpler functions with small coefficients.

The aim of lasso is to perform constrained regression, such that the minimum number of features are used that are required to explain the data. This is done by finding features that highly correlate and removing all but one of these from the feature space. It does this by reducing the coefficient size in the regression function, such that some can be reduced to zero and hence are removed from the feature space. Therefore lasso combines the gradual changes of shrinkage methods with the ability to reduce the feature set, the valuable aspects of ridge regression and subset selection respectively.

Equation 23 shows the lasso optimisation problem, where n and m are the number of data points and features respectively.

$$\begin{aligned} & min_{\alpha,\mathbf{w}} \sum_{i=1}^{n} (y_i - \alpha - \sum_{j}^{m} w_j x_{ij}) \qquad [49] \\ & s.t \sum_{j} |w_j| \leq t \end{aligned} \tag{23}$$

w_j is the coefficient of variable j in the model, and regularisation constrains the sum of these coefficients. α is the bias in model, such that under the assumption of standardised data such that $\sum_i \frac{x_{ij}}{N} = 0$ the data is centred and the α can be omitted. [49] This results in the following equation:

$$min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \mathbf{w}) + \lambda \sum_{j}^{m} |w_j| \qquad [49] \tag{24}$$

It is interesting to note just how similar this is to ridge regression, where the regularising component of the equation is instead $\lambda \|\mathbf{w}\|^2$. [49] This difference gives lasso the key property of often reducing coefficient values to zero providing much value for feature selection.

How lasso works Lasso is not a greedy approach but uses least angle regression (LARS) to find the model. LARS is similar to forward stepwise regression (FSR). FSR works by firstly starting with an empty model, finding the variable most correlated with the label and adding this to the model. The residual is taken, which is the difference between the label value and the value predicted by the model. Then recursively the variable is found that is most correlated with the residual and this is added to the model. Thus gradually the model accounts for more of the variation in the model by incorporating more features.[48]

LARS is different because it does not add a variable to the model with the real coefficient value. Instead the coefficient is the lowest value such that it is not the variable most correlated with the label. This is done by increasing the coefficient gradually and at each step taking the residual. The correlations of this residual with each variable is taken until another variable has the same correlation with the residual as the original variable. LARS maintains a variable set, and the process is repeated with this gradually growing set of variables. The coefficients of the variables in the set are increased together (equiangular such that it is proportionate to their current coefficient values), until another variable has the same correlation with the residual, which is then added to this set.[19] This is continued gradually adding more variables until all predictor have been incorporated. This can be demonstrated with a simple example, for the relationship $y = a + 2b$ and the example data of table 4(a). The starting model is: $y = \beta_0 + \beta_a a + \beta_b b$, where all β are zero.

y	3	5	5	6	8	7	9
a	1	1	3	2	2	3	3
b	1	2	1	2	3	2	3

(a) Example data

r	corr a	corr b
0.5	0.7026	0.8525
0.7	0.7409	0.8222
0.8	0.7618	0.8038
0.9	0.7837	0.7828

(b) Step 1

β_{ab}	corr
0.5	0.9909
1.5	0.3235
1.6	0.1216
1.7	-0.0708

(c) Step 2

Figure 4: Example lasso walk through

The steps taken to create a model are:

- Find correlations of each variable with the label (y): $a = 0.6358$ and $b = 0.8954$.
- b has the highest correlation. The coefficient of b is increased until the correlation with the residual is higher for variable a (table 4(b)): $\beta_b = 0.9$.
- Find β_{ab}: $\beta_{ab}(a + 0.9b)$, increasing until correlation is reduced to zero (which is the correlation of the bias variable because it is constant) (table 4(c)): $\beta_{ab} = 1.7$.
- All predictors are now in the model therefore we can find the bias. Take the mean of the residual: -0.5600.
- Therefore the model is: $y = 1.7a + (1.7 * 0.9b) - 0.56 \equiv y = 1.7a + 1.53b - 0.56$.
- The sum of squared error (ss) is: 2.6056. Comparing this with a model where β_{ab} is not increased far enough and the error is higher: $beta_{ab} = 1.0$ then $ss = 4.8400$.

Lasso vs OLS Lasso is more stable than OLS regression because where several variables are similar they are both included equally in the model, rather than one representing both and the other being excluded.[49] The lasso uses LARS to generate the model but regularises to bound the total size of the coefficients.[48] Regularising reduces overfitting, as an unregularised model can have any coefficient values and hence may fit the training data too closely. Lasso uses 10 fold cross validation to find the best λ value to regularise such that the error on the test set is minimised. A λ value that is too large does not regularise sufficiently and the model overfits the training data. A λ value that is too restrictive such that the model cannot represent the stable patterns (that are present in both the training and test data).

3.3.6 Decision Trees

Decision trees are an effective ML method, with a simple and intuitive construction algorithm. They are a white box technique where the generated tree can be analysed to gain understanding.

Tree construction takes a vector of instances as input, consisting of a set of input variables and a target variable. The tree is constructed by starting at the root and choosing a variable at each node to split the dataset, a process known as recursive partitioning. At each node down the

the tree this is repeated, with progressively smaller subsets of the original dataset. This process ends when a stopping condition is met, such as a minimum number of instances at a node, and this node is then a leaf of the tree.

Decision trees use a greedy approach where the best split at each node is used without looking at the tree as a whole. This is necessary as the search space of all possible trees is too large to search exhaustively. This property means the tree may be sub-optimal, but also that it is a simple and fast method that can be applied to big datasets.

Decision trees automatically perform feature selection where an impurity measures determines the variable chosen at each node, and hence variables that do not help predict the target value are automatically ignored. This selection of variables with the highest information gain also means there is a bias for smaller trees. Decision trees therefore seek to model relationships in the simplest way. Also, there is no assumption of a specific type of correlation between the variables and the target value, unlike other algorithms such as linear regression. [53]

The tree structure can give valuable insight into the strength of the correlation between each variable and the target value. In addition, the tree can be converted to a set of rules, known as classification rules. This is done by starting at the root and constructing a rule for the path from the root to each leaf.

There are several algorithms for constructing decision trees, differing in how they determine which attribute to use at each node to split the dataset. ID3 and C4.5 are common examples but these are not suitable for our needs as they are classifiers and cannot be applied to regression tasks.

Regression trees There are several alternative construction algorithms for regression trees where the main difference is the measure used to split the data at the internal nodes. A classification and regression tree (CART) uses the Gini index as an impurity measure, which is similar to entropy[9] but for continuous values. The Gini gain indicates the increase in inequality in the child nodes compared to the parent node, when splitting on a particular feature. Alternatively variance can be used such as the implementation of REPTree[10] by Weka.

Model trees Regression trees can be further extended to contain linear regression functions at each leaf node, and these are known as model trees. [53] Model trees are built in the same way as regression trees except that additionally linear regression is performed at each leaf using the instances assigned to it. The m5 algorithm is a model tree generation algorithm. A further alternative is logistic model trees, using logistic regression at each leaf rather than linear regression. This may be preferable where a logistic correspondence between features and labels is more representative than a linear relationship.

Model trees have several benefits over regression trees including smaller trees and greater accuracy. This is because a single leaf represents a range of values and thus a model can fit these better than a single value. In addition, they can also be used for extrapolation, as the linear function generalises to values that are outside those of the original dataset.

Missing values Decision tree algorithms handle missing values automatically, using different methods to do this (or imputation can be performed prior to using ML, see section 3.1.3). CART and M5 algorithms use a technique called surrogate splitting. This involves a second variable at a node, that is used to split the instances where an instance does not have a value for the main variable. The split used for the second variable is chosen such that it is most representative of the split of the first variable. [22]

[9] Entropy is used by ID3 and C4.5 and represents how well a particular attributes divides the dataset with respect to the target variable.

[10] class weka.classifiers.trees.REPTree

Overfitting Decision trees are prone to overfitting, where a tree fits the training data well but does not generalise well to the unseen case and hence performs poorly on test data. Pruning is a common technique which can be carried out either during or after tree construction, named pre-pruning and post-pruning respectively. [53] Prepruning constrains the tree size while it is being constructed, using a stopping condition to prevent further child nodes. Commonly, the number of instances at each node is used, which ensures the rules in the tree are generalised to cover several instances. Prepruning model trees means the regression model at each leaf is formed using more instances.

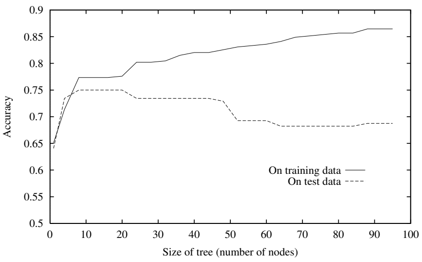


Figure 5: Overfitting [37]

Post-pruning is carried out after the tree is built. This involves using a separate test set and checking each node from the leaves to the root by comparing the error of each node with that of it's parent. If the parent has a lower error the child nodes can be pruned. This can be seen in figure 5, where a large tree has a low accuracy on the test set and this improves by reducing its size until the test and training error values are similar (around size 20 in this example). Post-pruning can be more effective than pre-pruning as it uses an independent dataset to test the rules of the tree.

Decision tree critique Decision trees can be used for feature selection as explained above. Here we look in more detail at the algorithms behind tree construction to highlight any limitations of using this learner for this purpose. We look specifically at Weka's ([26]) M5P tree[11] which implements the standard M5 model tree algorithm. [41] One concern for this algorithm is the construction of the linear models at the leaves as we would like to infer information from these models, but if a linear model includes dependent variables the model is unstable and cannot be used in this way.

The M5P algorithm does make use of the information gained from tree construction to create the leaf models. These models are generated after post-pruning using only the features part of the sub-tree just pruned, and this reduces the likelihood of dependent variables within leaf models. Each leaf will use only the features most appropriate for that portion of the dataset in its model.[51] However, dependent and irrelevant attributes can still occur, due to differences in subtrees. For instance, given a subtree such as that in figure 6, where pruning is performed such that node A is a leaf, and subtrees B and C are pruned. Attributes used at nodes B and C may be highly correlating dependent variables where small differences in the instances down these branches meant one was preferred (marginally) over the other. These could be two variables representing similar concepts, health expenditure and immunisation rate for instance. Additionally, as the depth increases in tree construction the subsets at each node become smaller and thus it becomes more likely that irrelevant / non optimal features are chosen.

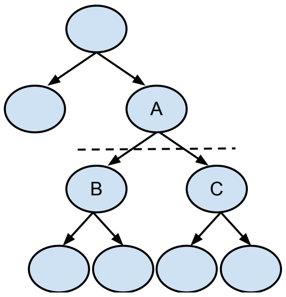


Figure 6: Tree pruning limitations

The pruned nodes consist of rules that perform badly on unseen data and so are a poor generalisation. The features above this node in the tree are more stable splitting criteria because, they haven't been pruned and so these rules are good generalisations to unseen data (largely because they were constructed using a larger proportion of the dataset because they are higher in the tree). Therefore, it is unclear why the variables of the pruned nodes are used to construct the models rather than those of the nodes above.

M5P does attempt to drop terms from the leaf models with the aim of minimising the expected error on unseen data (to prevent overfitting of the training data). This is done using equation 25,

[11] class: weka.classifiers.trees.M5P

where n is the number of training examples and v is the number of features in the regression model. [51]

$$Error_{expected} = Error_{train} \times \frac{n+v}{n-v} \tag{25}$$

The VC dimension[12] of a linear model increases with the number of variables, which increases the likelihood of overfitting. Incorporating this value into the error estimation therefore gives a preference for models with low VC dimensions. The method used for removing variables in the model in order to minimise $Error_{expected}$ is a greedy algorithm, where the features are tested sequentially to see if removing them reduces $Error_{expected}$.

It is clear that this method, although not intended for this use, may be quite effective at removing the dependent variables. For instance, returning to the example of figure 6, this algorithm would try removing A and B in turn. If the variance of the label accounted for by A is also accounted for by B then removing A would have little affect on the error, but $Error_{expected}$ will be lower because the feature set is smaller, and thus A is removed from the model.

This is however quite a crude approach, with the use of a greedy algorithm. Assessing the variables sequentially can cause a variable to be removed when it may be preferable to remove an alternative one. For instance, suppose that removing variable B gives a lower $Error_{expected}$ to removing variable A. Since A is considered first this means that A is removed rather than B, and this is not the optimal solution. This problem is heightened by the division of the data set such that each leaf has only a small subset, which adds instability because a small dataset is more likely to affected by variations or anomalies.

Therefore, decision trees can be used (after pruning) for feature selection using the internal nodes. However, the features used in the leaf models cannot be used. This is because the variables included in the leaf model are from pruned nodes, and the method of reducing the expected error is a crude greedy approach which may produce models containing a suboptimal feature set.

> During feature selection regression trees may be preferred to model trees, to construct larger and more informative trees.

Summary Decision trees have may beneficial properties. Firstly the automatic feature selection is useful where there are dependent variables. Also, the tree structure that is generated can be highly informative. Decision trees can represent non linear patterns, where different segments of the trees can contain different rules. The variations of decision trees such as regression and model trees, and the ability to use both numeric and nominal features, provides much flexibility. These features in combination with the ability to analyse the tree structure makes this a useful learner for this project.

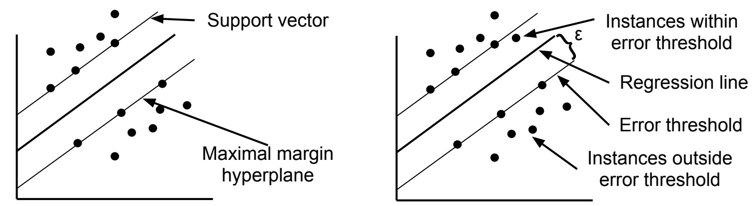
However, we have also highlighted some areas of concern. Firstly, further down the tree smaller subsets of the data are used and hence it becomes more likely to select irrelevant attributes. Additionally, dependent attributes can still be selected where they are preferable to split the data at different points in the tree. Therefore, when using model trees dependant attributes can still be used to construct the leaf models.

3.3.7 Support Vector Machines

Support Vector Machines (SVMs) were created by Vladimir Vapnik in 1992 and are a powerful machine learning technique that can be used for both regression and classification problems. This method uses theory and techniques from machine learning and statistical theory that were already well known. [14]

Support vector regression (SVR) is a special type of SVM. The standard SVM is used for classification and will be introduced briefly (but SVR involves some alternate methods). The SVM

[12] VC dimension specifies the degree of complexity of a hypothesis that can be generated by a learner



Support vector
Maximal margin hyperplane
Instances within error threshold
ε
Regression line
Error threshold
Instances outside error threshold

(a) Support Vector Machine for Classification

(b) Support Vector Regression

Figure 7: Graphical representation of support vector instance space

hyperplane is positioned to give the largest margin between itself and the nearest classification points (and hence is known as a maximum margin hyperplane). The points sitting on the margin constitute the support vectors, and since a support vector is the 'front' of an output value (see figure 7(a)), there must be at least one for each class value. For example, a boolean output will have two support vectors, with the maximal margin hyperplane sitting directly between them.

SVMs seek to find the parameters of the hyperplane given by a weight vector **w**, which amounts to a constrained quadratic optimisation problem: [53]

$$min\frac{1}{n}\sum_{i=1}^{n}\xi + \gamma\|\mathbf{w}\|^2, \qquad [15] \tag{26}$$

where ξ is a slack variable, used to remove the minimisation over two variables:

$$\xi = max(0, 1 - \mathbf{y}_i\mathbf{x}_i^T\mathbf{w}) \qquad [15] \tag{27}$$

$\mathbf{y}_i$ is a points classification with a value of either +1 or -1. $\mathbf{y}_i\mathbf{x}_i^T\mathbf{w}$ is negative when a point is classified incorrectly, which also means that $\xi > 1$. No cost is associated to a point where it sits on the correct side of the hyperplane and beyond the margin, otherwise they are linear in the distance from the hyperplane.

The most useful property of SVMs for this project is the ability to transform the data, such that any linear patterns found in the new data space will correspond to a non linear pattern in the original space, as illustrated in figure 8. This is possible because the problem can be reformulated in terms of inner products of the data items, which is known as the 'kernel trick'[13]. Kernel versions provide several useful features such as that the resulting dimensions are not dependent on the size of the feature space and therefore can be used on data sets with a large number of features. A kernel function is specified to transform the data and hence kernel versions can be used to find complex patterns. This also gives great flexibility with the ability to choose the most appropriate kernel for a particular problem.

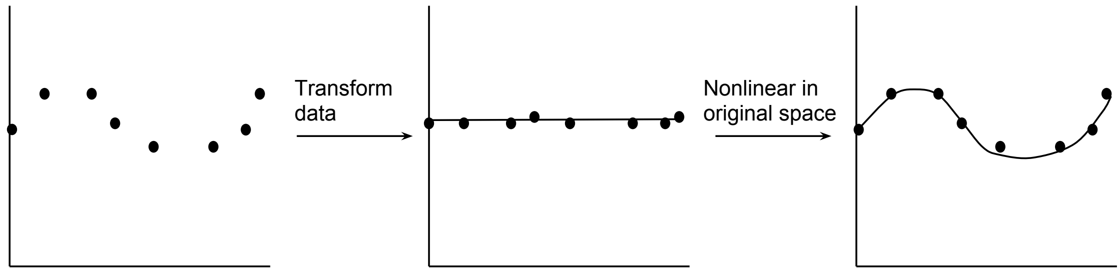


Figure 8: Kernel transformation

[13] Many other learners also have a kernel version such as Fisher Discriminate Analysis (FDA), PCA and k-means clustering

Two common kernel functions are the Radial Basis Function (RBF) and polynomial (equations 28 and 29 respectively).

$$kernel_{RBF}(x_i, x_j) = exp - \left(\frac{\|x_i - x_j\|^2}{2\sigma^2} \right) \tag{28}$$

$$kernel_{POLY}(x_i, x_j) = \left(x_i^T \cdot x_j + 1 \right)^d \tag{29}$$

Support vector regression differs slightly from the classification case, as a maximally margin hyperplane is not used. Alternatively, an ϵ-insensitive loss function (equation 30) is used to minimise the error on the data. An ϵ value is specified which is a threshold distance from the hyperplane and instances are only said to result in an error if the error value falls outside of this threshold. Errors within this threshold are small enough and are thus synonymous with a correct classification and impose no cost. [10] This provides a way of defining an error for regression problems that allow for some noise. The value chosen as ϵ determines the acceptable error size and therefore the degree to which the function fits the data. The cost of a point ξ, is given by:

$$\xi = max(0, |\mathbf{Xw} - y| - \epsilon) \tag{30}$$

The cost therefore depends on the distance from the hyperplane, and these two properties always hold; $\xi \geq 0$ and $\xi \geq |\mathbf{Xw} - y| - \epsilon$. A point nearer than ϵ to the hyperplane incurs no cost, and for those past this point the cost is linear in the distance (and these points constitute the support vector).

In addition to the loss, a parameter called the cost should also be optimised. This cost parameter specifies the magnitude of the cost relative to the distance from the hyperplane, where increasing the cost means the fit will be closer and overfitting may occur. If the cost is too low the fit is not representative of the data as the penalty for being far from the hyperplane is too low. The RBF kernel additionally has a gamma value (σ^2 in equation 28), which is a regularisation parameter that specifies how closely the fit to the data. The polynomial kernel requires an exponent parameter (d in equation 29).

In summary, SVMs are a powerful and flexible method for regression problems, in comparison to standard linear techniques such as least squares. The two objectives, minimising both the error and the complexity of the hypothesis (called optimisation and regularisation respectively) means SVMs are able to learn a hypothesis with an inbuilt method to prevent overfitting the data. They have greater modelling capabilities due to the transformation of the feature space and hence can find non linear patterns. The appropriate kernel function can be chosen according to the problem. SVMs cope well with noisy data as they regularise to prevent overfitting and incorporate a margin to allow for a degree of deviation from the hyperplane. They are less affected by outliers as the cost function is linear in distance from the hyperplane, compared to quadratic for least squares.

3.3.8 Bayesian Belief Networks

A Bayesian belief network (BBN) is a directed acyclic graph (DAG). BBNs are useful to visualise the interaction between values, based on the conditional probabilities.

Each node represents one variable and contains a probability table. This states the conditional probability of each value given each value of the nodes that point to it. Each variable is conditionally independent of the other variables, given the value of its parents (the variables that lead directly to it).[14] [36]

The conditional probability tables are computed using the frequencies of the attribute values given each of the possible combinations of parent value. This means a variable with 4 nominal values having a parent with 3 would have 12 different probabilities in it's table. Once the network is constructed, the probability of a set of variable values called the joint probability can be found.

[14] X is conditionally dependent of Y given Z if where a value of Z is given X is not dependent on Y.

We will be able to determine the probability of a certain range of happiness given specific values of features. The joint probability[15] is the product of the conditional probability values of the variable values in each node. Formally:

$$P(y_1, \ldots, y_n) = \prod_{i=1}^{n} P(y_i \mid Parents(Y_i)) \tag{32}$$

The probabilities can be calculated using the frequencies of the variable values:

$$P(x = x_1 | y = y_1) = \frac{|instances_{x1,y1}| + n}{|instances_{y1}| + n \cdot N} \tag{33}$$

This creates a maximum likelihood BBN with respect to the probabilities (and not the structure). The n and N variables incorporate a Laplace correction into the probabilities to ensure that no values are ever zero in the probabilities. Data used to construct a BBN is only a small sample used to represent more general concepts and so zero frequencies may exist in the sample where the true probability is small but non-zero. Incorporating a correction prevents zero probabilities which would give a large bias such that any joint probability calculated with it will be zero. The correction effectively assigns each variable value to one extra instance. Therefore in this case N is $|x| \cdot |y|$, the number of possible combinations of these two variables, and n is the weight of this correction. A larger n value will reduce the 'free' probability assigned.

Discretization A Bayesian belief network can only use discrete values. Our variables including the happiness variable are continuous, and these need to be discretised prior to constructing the network. This involves defining a set of 'bins' which are categories (a range of values) of the variables, to convert from continuous to nominal values. Binning can be either equal interval or equal frequency.[53] Equal interval splits the range into equal sized segments, whereas equal frequency ensures the same number of elements occur in each bin. The latter is often preferable as where many elements are clustered around a similar value these values can be segmented to a finer degree and hence keep more information about the distribution of values.[16]

This is a possible drawback of BBN's where discretising the values will inevitably lose some information in the data. As an example, consider the following situation:

$Instance\ X:\ i = 10$
$Instance\ Y:\ i = 11$
$Instance\ Z:\ i = 19$

Suppose two bins were for ranges 1-10 and 11-20. X would be assigned a different bin to Y and Z. Instances X and Y are very close but after discretising it performed it is Y and Z that appear most similar. The number of bins chosen is important to represent the data most appropriately in nominal form, to reduce the amount of information that is lost. If too many or too few are used then the nominal values do not show some data patterns. Additionally, the degrees of freedom of a network (the number of changeable parameters) is exponential in the number of bins. This should be minimised to reduce the complexity and prevent overfitting.

[15] The joint probability is derived by the chain rule. Generally $P(A \cap B) = P(A \mid B)P(B)$. Given 3 nodes with the following structure: $A \rightarrow B \rightarrow C$. C is conditionally independent of A given B such that $P(C \mid A \cap B) = P(C \mid B)$. Therefore:

$$P(A \cap B \cap C) = P(C \mid A \cap B)P(A \cap B) = P(C \mid B)P(B \mid A). \tag{31}$$

[16] Weka provides a Bayesian network classifier.[43] All inputs must be discrete, pre-processing is not part of the tool. [43] Discretising can be done with Weka's discretise tool, which provides both binning types described here.

Missing values Weka's implementation of a BNN handles missing values by assigning them the mean of the dataset. This is a naive approach and hence it may be preferable to impute the values before constructing the network.

The following are points to consider when creating a BBN:

- The number of nominal values or 'bins' to choose
- Equal interval or equal frequency binning
- Preprocessing of missing values (see section 3.1.3)

BBN will provide a useful tool for this project to visualise and assess the relationships between the variables and happiness label. However, we must be aware that this method does not infer the direction of causality, but just dependence between variables.

3.3.9 Testing the accuracy of classifiers

Overfitting is a common issue with ML algorithms, and occurs when the learned hypothesis too closely represents the training data such that it is unable to generalise to unseen examples. This results in a low error on the training set but a much higher error on the test set. To determine the accuracy on unseen data the dataset should be split into training and test sets. This can often be problematic when there is limited data available.

With regards to this project, ideally survey data and feature variables from multiple years would be used, to provide independent sets of data on which to train and test. However, this is unlikely as there is limited data available. Alternatively, methods exist to work with separate training and test sets, such as 10 fold cross validation. This makes best use of the data available by testing the algorithms multiple times on the same dataset. The dataset is divided into 10 parts, and the algorithm is run 10 times. Each time 9 parts are combined and used for training the learner and 1 is used for testing. In this way the training and test sets are distinct and the accuracy can be assessed with 10 different tests.

4 Survey Data Analysis

There are several sources of survey responses with a question on life satisfaction. This analysis investigates the correlation between the survey data with two aims. Firstly, to ensure survey data is a reliable and representative measure of happiness. This can be shown where the data from independent surveys have a high degree of correlation. Secondly, analysis of the data will provide information in order to decide the best choice of happiness label.

The surveys are correlated using the corrcoef function in Matlab, which also provides p-values but under the assumption that the data has a multivariate normal distribution. However, the histograms of figure 10 show that this is not always the case and therefore permutation testing (see section 3.3.3) may provide more reliable results.

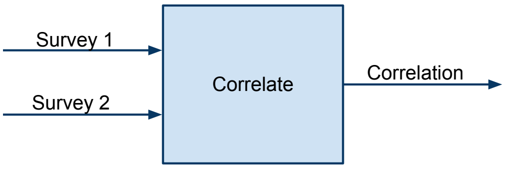


Figure 9: Survey analysis test design

Figure 9 shows the test design performed. Permutation testing is performed to find a threshold value above which the correlations are significant. The null hypothesis states that the correlation found between two surveys is likely to be found in the random case. The threshold is found for each pair of values because each survey has a unique probability distribution (the distributions shown in figure 10 are mostly quite different).

Eight surveys were analysed, which were carried out on or around 2008 (details in table 28,Appendix A), and this includes 4 global, 3 European and 1 Latin American survey. Prior to analysis the data was transformed to a consistent scale of 0 - 10 (for details see the 'Standardisation' column in table 28).

The descriptions in table 28 show several differences between the surveys. These can be summarised as:

- Respondent age
- Question
- Answer scale
- Countries surveyed
- Year
- Sample size
- Position of question in survey
- Sample methods

Table 1 details the sample size of each survey. The coverage varies largely and is an important consideration in choice of ground truth. Patterns found in larger samples are more reliable as small samples are less robust to anomalies and outliers and any patterns are more likely to occur by chance.

WVS	HPI GAL	OECD POS	OECD LOL	ESS	EUROB	EVS	LATINO
57	112	33	34	25	30	47	18

Table 1: Survey sample sizes

4.1 Results

The histograms in figure 10 show some interesting distributions. Two of the global surveys have approximately normal distributions. A normal distribution may be expected because we would expect many countries to have average LS values and fewer to have extreme values. It may be tempting to apply the central limit theorem (CLT) here which states that the mean of a large number of random variables will be normally distributed. Our country LS values can be thought of as the mean value of random variables (the LS of each recipient in the country). However, the CLT only holds under the assumption that the random variables are independent and identically distributed, and this is certainly not true in this instance.

The HPI data has a larger variance than the WVS data. This may be because of the position of the question in the survey, as discussed in section 2.4.1. The Gallup life satisfaction question is after governmental and economic questions and this may cause the rich and poor to give higher and lower life satisfaction values respectively.

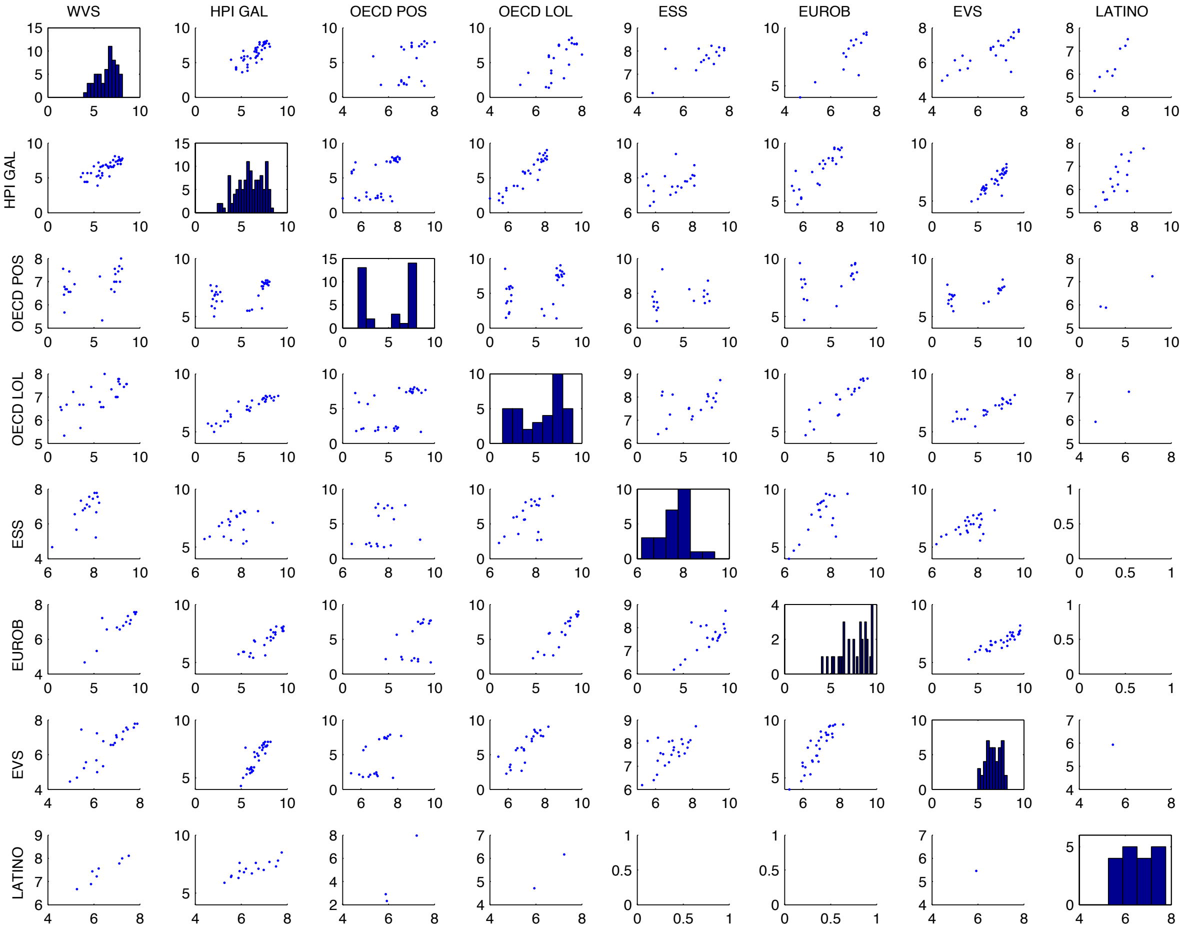


Figure 10: Survey data correlations & distributions

The two types of OECD data do not show a normal distribution, and in particular the Positive Experience Index (OECD_POS in column/row 3) shows two clusters centred around 2.5 and 7.0. It is unclear why there is such a clear split for this survey. This measure is very different from the others, as it is a compiled index created from OECD consisting of data from six questions related to what they call ‘positive experiences’. However, the answer scale was still 0 - 10 and a unimodal distribution was expected.

The distribution of Gallup *life satisfaction* (LS) and Gallup *ladder of life* (LOL) questions show the latter has a larger variance with lower values also. This is unexpected as we would expect the LOL questions to give higher values because the question is relative to ones situation; ‘best possible life for me’ rather than ‘best possible life’ for the LS question. The upper bound of ones’ attainable happiness decreases with decreasing life satisfaction, as the standard of life that is reachable also decreases.

The results show a high degree of correlation between the data, shown in figure 10. Table 2 shows the correlation values for each survey, and the p-values output from the corrcoef function. Each pair of surveys has a different number of countries in common and the p-value incorporates a correction to account for this. The significant p-values are highlighted indicating 20 survey cross-correlations are significant compared with just 4 that are not. 3 of these are for this OECD Positive Experience Index with the anomalous distribution discussed above. 4 results give no value where there is no overlap between the countries covered.

Permutation tests gave similar results, shown in table 3. A 1 value denotes a significant result, such that the surveys were significantly more correlated than the random case. The main differences are with regards to correlations involving the OECD surveys, and this is consistent

	WVS	HPI GAL	OECD POS	OECD LOL	ESS	EUROB	EVS	LATINO
WVS	-	0.79	0.406	0.695	0.650	0.858	0.792	0.941
HPI GAL	0.0000	-	0.468	0.940	0.344	0.870	0.865	0.82
OECD POS	0.054	0.0069	-	0.483	0.300	0.486	0.560	0.99
OECD LOL	0.0002	0.0000	0.0068	-	0.473	0.890	0.837	1
ESS	0.0064	0.138	0.242	0.0064	-	0.71	0.628	NaN
EUROB	0.0001	0.0000	0.0407	0.0000	0.0004	-	0.9	NaN
EVS	0.0000	0.0000	0.0102	0.0000	0.0010	0.0000	-	NaN
LATINO	0.0005	0.0001	0.082	NaN	NaN	NaN	NaN	-

Table 2: Survey correlations: Pearson coefficient values & p-values (significant results highlighted, 0.05 threshold)

	WVS	HPI	OECD POS	OECD LOL	ESS	EUROB	EVS	LATINO
WVS	-	0.258	0.515	0.503	0.595	0.500	0.385	0.770
HPI	1	-	0.337	0.340	0.396	0.349	0.291	0.461
OECD POS	0	1	-	0.687	0.861	0.777	0.537	0.974
OECD LOL	1	1	0	-	0.799	0.734	0.547	0.954
ESS	1	0	0	0	-	0.841	0.647	0.971
EUROB	1	1	0	1	0	-	0.607	0.991
EVS	1	1	1	1	0	1	-	0.786
LATINO	1	1	1	1	0	0	0	-

Table 3: Survey permutation test results: Threshold correlation values and significance (1: significant)

with the distributions which are not Gaussian. WVS, HPI(Gallup) and EVS all show Gaussian distributions and the results for these in tables 2 and 3 are highly similar. It is interesting to see how the correlation thresholds of the permutation testing results vary. Correlations involving surveys with smaller sample sizes need to be very high in order to be significant, because good correlations are more likely to occur purely by chance for smaller datasets. For instance, the thresholds of the two largest and two smallest surveys are 0.258 and 0.971 respectively.

The results found above and the survey coverage are used to establish the best choice of life satisfaction ground truth. The unusual distribution of the OECD POS survey leads us to discard this data source as a potential label. The LOL data is a small sample and much of the data missing is for countries having lower values for the LS question. In a country listing, ranked by Gallup LS value 26 countries (with LS values between 2.4 and 4.9) precede South Africa, which has the lowest LOL value of 2.05 (and having a LS value of 5.0). This adds further anomaly to our findings, as the LOL results have a lower minimum without the countries at the lower end of the scale. The close correlation of LS and LOL data indicates that if LOL did have values for these countries this would decrease the lower bound of LOL values further.

This indicates two reasons why we should discard LOL as a happiness label. Firstly, we do not find it appropriate to use a measure that anchors responses by attainable life satisfaction. Secondly, the OECD LOL dataset is much smaller and lacking values for countries at the lower end of the scale. Previous research has shown the sensitivity of analysis when using smaller samples. Furthermore, we will discard the three European and the Latin American dataset, as we prefer to use global data. This gives us two happiness labels to use; WVS and Gallup LS. These have very strong correlations to each other with $p - value < 0.0001$ (or equivalently a significance threshold of 0.258 with a high correlation of 0.79). However, there are fundamental differences in survey methods which mean one cannot simply be preferred to the other.

In conclusion, these results are very encouraging. The independent surveys have very strong and often significant correlations showing that survey data can provide a consistent measure of happiness. The HPI (Gallup) and WVS data have been selected as appropriate happiness labels, to be used in parallel throughout the study. These will be referred to as GAL and WVS throughout this report.

5 Data Collection & Description

Features are collected from freely available online sources[17], predominantly The World Bank. 20 features are used initially (including economic variables), detailed in table 4 (full details can be found in table 13.3 in appendix A). These features are based around key areas, based on knowledge gained from a review of previous work (section 2). These areas are: Climate, Economic, Lifestyle, Environmental, Equality, Freedom, Health.

Topic	Feature	Code	
ECONOMIC	GDP growth (annual %)	GDP-growth	EC1
	Employment to population ratio 15+, total (%)	employment	EC3
	GDP per capita (current US$)	GDP-per-capita	EC4
HEALTH	Life expectancy at birth, total (years)	life-expectancy	HE1
	Immunization, DPT (% of children ages 12-23 months)	child-immunisation	HE2
	Health expenditure per capita (current US$)	health-expenditure	HE3
	Mortality rate, under-5 (per 1,000)	mortality-rate	HE4
ENVIRON	CO2 emissions (metric tons per capita)	C02-emissions	EN1
	Mammal species, threatened	mammals-threatened	EN2
EQUALITY	Ratio of gender labor participation rate	gender-labour-ratio	EQ2
	Proportion of seats held by women in national parliaments (%)	proportion-women-parliament	EQ3
FREEDOM	Time required to start a business (days)	time-start-business	FR1
EDUCATION	Pupil-teacher ratio, primary	pupil-teacher-ratio	ED2
CLIMATE	Latitude weighted average by population (indicator for light)	light	CL2
LIFESTYLE	Population growth (annual %)	population-growth	LI2
	Urban population (% of total)	population-urban	LI3
	Population density (people per sq. km of land area)	population-density	LI4
CRIME	Intentional homicide, rate per 100,000 population	homicide	CR2
DISTRIB'N	Distribution of family income - Gini index	income-distribution	DI1
	Population over 65 (%) %	population-over-65	DI2

Table 4: Features collected and constructed

5.1 Feature Collection & Construction

Previous work (such as that described in section 2.5.3) has highlighted the importance of ensuring the variables used are representative of the intended notion. Locating sources that give appropriate representations of particular concepts, and there is difficulty locating good data sources with enough coverage.

Where required, features are constructed from the available data. For instance, to represent gender labour equality a ratio was created from the individual variables of male and female labour participation rates. Also, latitude is used to represent light. This variable was also adjusted using the population of cities in a country, to ensure the latitude was most representative of the location of the population in the country.

[17] This is to allow future integration of online visualisations with live data

	WVS	HPI	EC1	EC3	EC4	HE1	HE2	HE3	HE4	EN1	EN2	EQ2	EQ3	FR1	ED2	CL2	LI2	LI3	LI4	CR2	DI1	DI2
WVS Rescaled	1	0.79	-0.55	0.1	0.53	0.55	0.32	0.51	-0.47	0.36	0.11	0.11	0.22	0.03	-0.39	-0.1	-0.04	0.5	-0.06	0.01	0.16	0.25
HPI (gallup only)	0.79	1	-0.39	-0.34	0.63	0.83	0.38	0.63	-0.76	0.6	-0.03	-0.18	0.22	-0.2	-0.68	0.32	-0.35	0.69	0.09	-0.41	-0.16	0.55
EC1-GDPGROWTH-2008	-0.55	-0.39	1	0.21	-0.55	-0.42	-0.2	-0.55	0.32	-0.47	0.18	-0.01	-0.04	0.24	0.32	-0.33	0.25	-0.34	-0.09	0.18	0.25	-0.52
EC3-EMPLOY-RATIO-2008	0.1	-0.34	0.21	1	-0.15	-0.42	-0.22	-0.13	0.47	-0.26	0.25	0.56	0.23	0.23	0.64	-0.42	0.44	-0.46	0	0.32	0.23	-0.39
EC4-GDP-PER-CAPITA-2008	0.53	0.63	-0.55	-0.15	1	0.6	0.31	0.97	-0.47	0.68	-0.24	0.2	0.35	-0.31	-0.48	0.45	-0.19	0.55	0.14	-0.4	-0.44	0.67
HE1-LIFEEXP-2008	0.55	0.83	-0.42	-0.42	0.6	1	0.56	0.58	-0.92	0.54	-0.07	-0.18	0.1	-0.3	-0.86	0.48	-0.48	0.66	0.18	-0.59	-0.36	0.7
HE2-IMMUN-08	0.32	0.38	-0.2	-0.22	0.31	0.56	1	0.27	-0.63	0.34	-0.14	0.03	0.12	-0.23	-0.52	0.3	-0.39	0.38	0.06	-0.28	-0.19	0.43
HE3-HEALTH-EXP-08	0.51	0.63	-0.55	-0.13	0.97	0.58	0.27	1	-0.45	0.62	-0.23	0.22	0.37	-0.31	-0.45	0.44	-0.23	0.51	0.02	-0.39	-0.44	0.67
HE4-MORT-RATE-08	-0.47	-0.76	0.32	0.47	-0.47	-0.92	-0.63	-0.45	1	-0.52	0.04	0.16	-0.07	0.25	0.88	-0.44	0.6	-0.63	-0.1	0.5	0.3	-0.66
EN1-CO2EMIS-2007	0.36	0.6	-0.47	-0.26	0.68	0.54	0.34	0.62	-0.52	1	-0.17	0.06	0.15	-0.31	-0.56	0.38	-0.3	0.47	0.11	-0.35	-0.38	0.58
EN2-MAM-THREAT-08	0.11	-0.03	0.18	0.25	-0.24	-0.07	-0.14	-0.23	0.04	-0.17	1	-0.08	-0.12	0.32	0.13	-0.34	0.09	-0.14	-0.06	0.11	0.22	-0.26
EQ2-LABOR-RATIO-GEN-08	0.11	-0.18	-0.01	0.56	0.2	-0.18	0.03	0.22	0.16	0.06	-0.08	1	0.38	0.03	0.16	-0.02	-0.08	-0.15	-0.01	0.08	-0.18	0.2
EQ3-PARL-WOM-08	0.22	0.22	-0.04	0.23	0.35	0.1	0.12	0.37	-0.07	0.15	-0.12	0.38	1	-0.1	0.05	0.03	-0.04	0.07	0.05	0.01	-0.14	0.23
FR1-TIME-START-BUS-2008	0.03	-0.2	0.24	0.23	-0.31	-0.3	-0.23	-0.31	0.25	-0.31	0.32	0.03	-0.1	1	0.22	-0.38	0.08	-0.18	-0.11	0.28	0.34	-0.35
ED2-PUP-TEA-RAT-08	-0.39	-0.68	0.32	0.64	-0.48	-0.86	-0.52	-0.45	0.88	-0.56	0.13	0.16	0.05	0.22	1	-0.55	0.63	-0.64	-0.08	0.55	0.34	-0.71
CL2-LAT-WEIGHTED	-0.1	0.32	-0.33	-0.42	0.45	0.48	0.3	0.44	-0.44	0.38	-0.34	-0.02	0.03	-0.38	-0.55	1	-0.43	0.22	-0.02	-0.5	-0.71	0.58
LI2-POP-GROW-08	-0.04	-0.35	0.25	0.44	-0.19	-0.48	-0.39	-0.23	0.6	-0.3	0.09	-0.08	-0.04	0.08	0.63	-0.43	1	-0.31	0.19	0.3	0.37	-0.7
LI3-POP-URBAN-08	0.5	0.69	-0.34	-0.46	0.55	0.66	0.38	0.51	-0.63	0.47	-0.14	-0.15	0.07	-0.18	-0.64	0.22	-0.31	1	0.23	-0.35	-0.07	0.56
LI4-POP-DENS-08	-0.06	0.09	-0.09	0	0.14	0.18	0.06	0.02	-0.1	0.11	-0.06	-0.01	0.05	-0.11	-0.08	-0.02	0.19	0.23	1	-0.09	0	0.07
CR2-INTENT-HOMO-03-08	0.01	-0.41	0.18	0.32	-0.4	-0.59	-0.28	-0.39	0.5	-0.35	0.11	0.08	0.01	0.28	0.55	-0.5	0.3	-0.35	-0.09	1	0.64	-0.51
DI1-FAM-INCOME	0.16	-0.16	0.25	0.23	-0.44	-0.36	-0.19	-0.44	0.3	-0.38	0.22	-0.18	-0.14	0.34	0.34	-0.71	0.37	-0.07	0	0.64	1	-0.59
DI2-AGE-65-2008	0.25	0.55	-0.52	-0.39	0.67	0.7	0.43	0.67	-0.66	0.58	-0.26	0.2	0.23	-0.35	-0.71	0.58	-0.7	0.56	0.07	-0.51	-0.59	1

Table 5: R values of features / LS labels

	WVS	HPI	EC1	EC3	EC4	HE1	HE2	HE3	HE4	EN1	EN2	EQ2	EQ3	FR1	ED2	CL2	LI2	LI3	LI4	CR2	DI1	DI2
WVS	1	1.41E-10	1.14E-05	0.49	2.27E-05	1.63E-05	0.02	6.30E-05	0	0.01	0.42	0.43	0.11	0.81	0.01	0.48	0.76	8.49E-05	0.64	0.91	0.25	0.07
HPI	1.41E-10	1	2.73E-05	0	1.87E-13	3.68E-29	4.54E-05	1.89E-13	9.75E-22	2.21E-12	0.76	0.06	0.02	0.03	3.44E-12	0	0	1.00E-16	0.35	6.02E-06	0.1	2.79E-10
EC1	1.14E-05	2.73E-05	1	0.02	8.14E-11	1.91E-06	0.03	9.95E-11	0	8.35E-08	0.05	0.9	0.64	0.01	0	0	0.01	0	0.35	0.05	0.01	1.49E-09
EC3	0.49	0	0.02	1	0.11	1.34E-06	0.02	0.15	8.30E-08	0	0	2.23E-11	0.01	0.01	8.82E-12	1.35E-06	4.27E-07	1.11E-07	0.96	0	0.01	8.91E-06
EC4	2.27E-05	1.87E-13	8.14E-11	0.11	1	3.04E-13	0	6.40E-71	6.27E-08	1.21E-17	0.01	0.03	0	0	1.32E-06	3.07E-07	0.03	1.00E-10	0.14	5.75E-06	1.01E-06	7.68E-17
HE1	1.63E-05	3.68E-29	1.91E-06	1.34E-06	3.04E-13	1	2.07E-11	2.60E-12	7.20E-49	1.73E-10	0.46	0.05	0.28	0	1.48E-28	3.16E-08	2.87E-08	1.21E-16	0.05	1.23E-12	6.59E-05	3.29E-19
HE2	0.02	4.54E-05	0.03	0.02	0	2.07E-11	1	0	5.44E-15	0	0.13	0.71	0.18	0.01	1.01E-07	0	7.54E-06	2.08E-05	0.53	0	0.05	9.66E-07
HE3	6.30E-05	1.89E-13	9.95E-11	0.15	6.40E-71	2.60E-12	0	1	2.95E-07	3.00E-14	0.01	0.02	3.07E-05	0	7.37E-06	3.18E-07	0.01	1.56E-09	0.8	9.54E-06	1.03E-06	2.27E-17
HE4	0	9.75E-22	0	8.30E-08	6.27E-08	7.20E-49	5.44E-15	2.95E-07	1	9.06E-10	0.67	0.08	0.47	0.01	2.94E-30	4.30E-07	3.13E-13	1.16E-14	0.27	4.13E-09	0	2.21E-16
EN1	0.01	2.21E-12	8.35E-08	0	1.21E-17	1.73E-10	0	3.00E-14	9.06E-10	1	0.06	0.52	0.1	0	4.60E-09	1.72E-05	0	5.53E-08	0.24	0	2.60E-05	3.48E-12
EN2	0.42	0.76	0.05	0	0.01	0.46	0.13	0.01	0.67	0.06	1	0.37	0.2	0	0.22	0	0.32	0.12	0.5	0.23	0.02	0
EQ2	0.43	0.06	0.9	2.23E-11	0.03	0.05	0.71	0.02	0.08	0.52	0.37	1	1.78E-05	0.79	0.12	0.84	0.36	0.1	0.93	0.37	0.06	0.03
EQ3	0.11	0.02	0.64	0.01	0	0.28	0.18	3.07E-05	0.47	0.1	0.2	1.78E-05	1	0.3	0.63	0.7	0.68	0.48	0.58	0.93	0.13	0.01
FR1	0.81	0.03	0.01	0.01	0	0	0.01	0	0.01	0	0	0.79	0.3	1	0.03	2.03E-05	0.37	0.05	0.24	0	0	9.92E-05
ED2	0.01	3.44E-12	0	8.82E-12	1.32E-06	1.48E-28	1.01E-07	7.37E-06	2.94E-30	4.60E-09	0.22	0.12	0.63	0.03	1	8.79E-09	1.83E-11	3.11E-12	0.43	1.01E-08	0	1.95E-15
CL2	0.48	0	0	1.35E-06	3.07E-07	3.16E-08	0	3.18E-07	4.30E-07	1.72E-05	0	0.84	0.7	2.03E-05	8.79E-09	1	5.98E-07	0.01	0.79	3.79E-09	9.76E-19	1.52E-12
LI2	0.76	0	0.01	4.27E-07	0.03	2.87E-08	7.54E-06	0.01	3.13E-13	0	0.32	0.36	0.68	0.37	1.83E-11	5.98E-07	1	0	0.04	0	3.42E-05	3.44E-19
LI3	8.49E-05	1.00E-16	0	1.11E-07	1.00E-10	1.21E-16	2.08E-05	1.56E-09	1.16E-14	5.53E-08	0.12	0.1	0.48	0.05	3.11E-12	0.01	0	1	0.01	7.67E-05	0.46	2.07E-11
LI4	0.64	0.35	0.35	0.96	0.14	0.05	0.53	0.8	0.27	0.24	0.5	0.93	0.58	0.24	0.43	0.79	0.04	0.01	1	0.3	0.96	0.42
CR2	0.91	6.02E-06	0.05	0	5.75E-06	1.23E-12	0	9.54E-06	4.13E-09	0	0.23	0.37	0.93	0	1.01E-08	3.79E-09	0	7.67E-05	0.3	1	1.26E-14	2.68E-09
DI1	0.25	0.1	0.01	0.01	1.01E-06	6.59E-05	0.05	1.03E-06	0	2.60E-05	0.02	0.06	0.13	0	0	9.76E-19	3.42E-05	0.46	0.96	1.26E-14	1	3.11E-12
DI2	0.07	2.79E-10	1.49E-09	8.91E-06	7.68E-17	3.29E-19	9.66E-07	2.27E-17	2.21E-16	3.48E-12	0	0.03	0.01	9.92E-05	1.95E-15	1.52E-12	3.44E-19	2.07E-11	0.42	2.68E-09	3.11E-12	1

Table 6: P VALUES of features / LS labels

5.2 Data Analysis

5.2.1 Data correlations

WVS	GAL
GAL	life-expectancy
GDP-growth	mortality-rate
life-expectancy	population-urban
GDP-per-capita	GDP-per-capita
health-expenditure	health-expenditure
population-urban	C02-emissions
pupil-teacher-ratio	pupil-teacher-ratio
C02-emissions	WVS
child-immunisation	population-over-65
	homicide
	GDP-growth
	child-immunisation
	proportion-women-parliament
	time-start-business

Table 7: Ranking significant results (by p-value)

The features and labels were correlated, and the Pearson's r and p values are shown in tables 5 and 6 respectively. A significant correlation can be seen between many of the features and labels. This is to be expected and is a challenge of this research; finding informative and stable models where the variables are highly dependent.

Table 7 shows a ranking of features that are significantly correlated with the two surveys. Highlighted are the variables significant to all. All the features found to be significant for WVS were also significant for GAL. The p-values are lower for GAL, which is likely to be because it is a larger study.

This initial analysis highlights potential differences in the two labels. The ranking are in different orderings, and it is particularly surprising that GAL correlates better with some of the features than it does with WVS. The potential causes of the differences in happiness labels are discussed in section 2.

GDP-growth is shown to be significantly and negatively correlated with both labels. This is consistent with previous work on the subject. This is very encouraging as evidence of the unsuitability of GDP as a measure of success of a country.

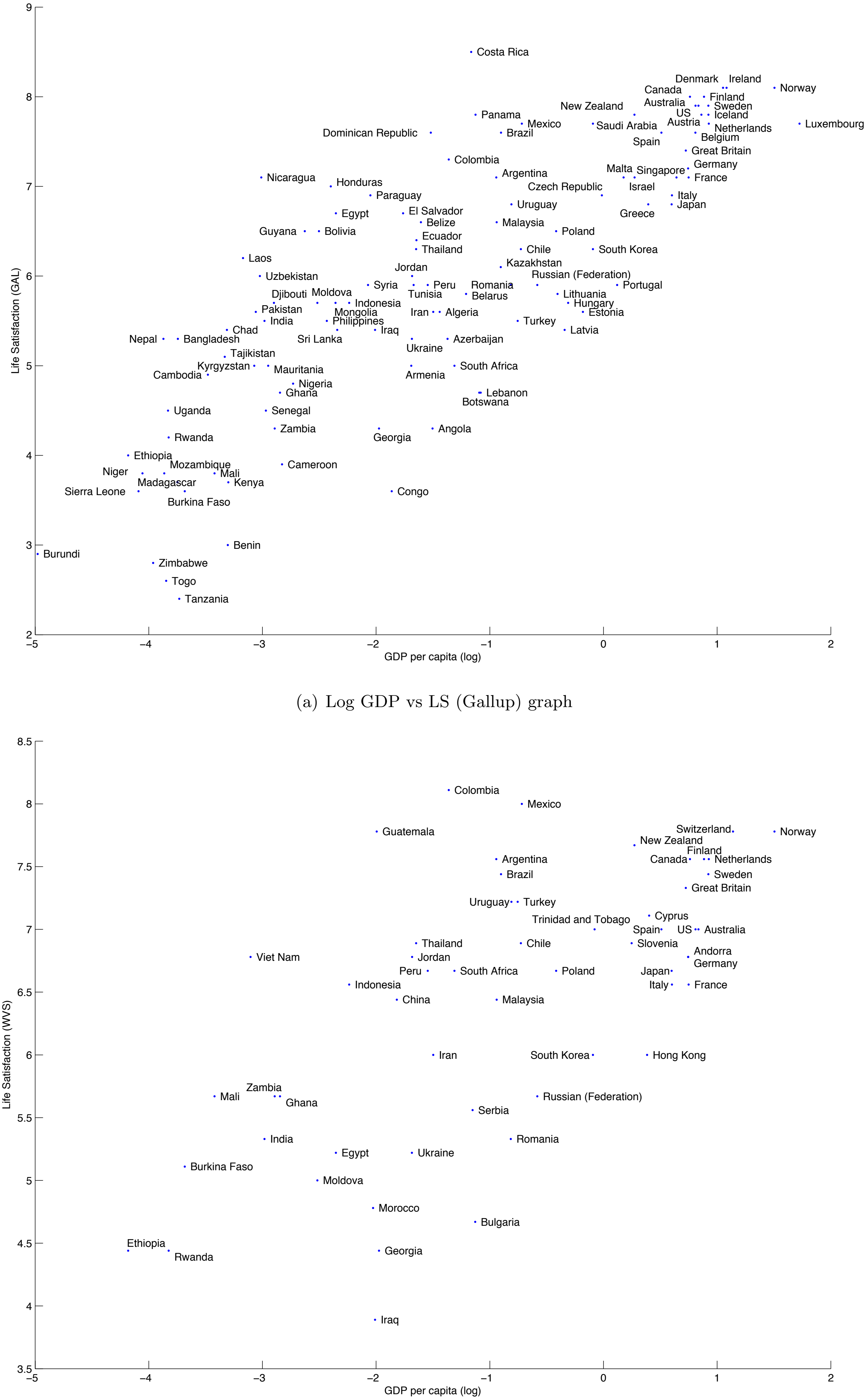
GDP-growth has a significant negative correlation with LS

GDP-per-capita is however strongly and positively correlated, which is expected, as a degree of wealth is necessary to enable other variables to occur. This is shown in figure 11(a) and figure 11(b)), and highlights an important difference in coverage between the two surveys, where WVS is particularly sparse at lower happiness / GDP values.

GDP-growth, life-expectancy, mortality-rate and population-urban are more significantly correlated than GDP-per-capita with at least one label

Four health indicators were used and all were found to be significant; 3 for WVS and 4 for GAL. Mortality rate (% under fives) was not significant for WVS, which perhaps corresponds to the fact that this variable relates to poverty and WVS lacks data for many of the developing countries (see figure 11(b) for graph of happiness vs GDP). Figure 26(a) shows that countries with a lower LS have a much greater variance of mortality-rate.

Life-expectancy was strongly correlated with both surveys, ranked 2nd and 1st for WVS and GAL respectively. This is particularly apparent for GAL having a correlation of 0.83 ($p = 3.68 \times 10^{-29}$), and this can be clearly seen in figure 26(b). Life expectancy has a higher correlation than health expenditure with LS, which is intuitive as it is more directly related to LS. A person is likely to be less concerned about their countries health expenditure than the quality of the health care it produces.



(a) Log GDP vs LS (Gallup) graph

(b) Log GDP vs LS (WVS) graph

Figure 11: GDP correlations

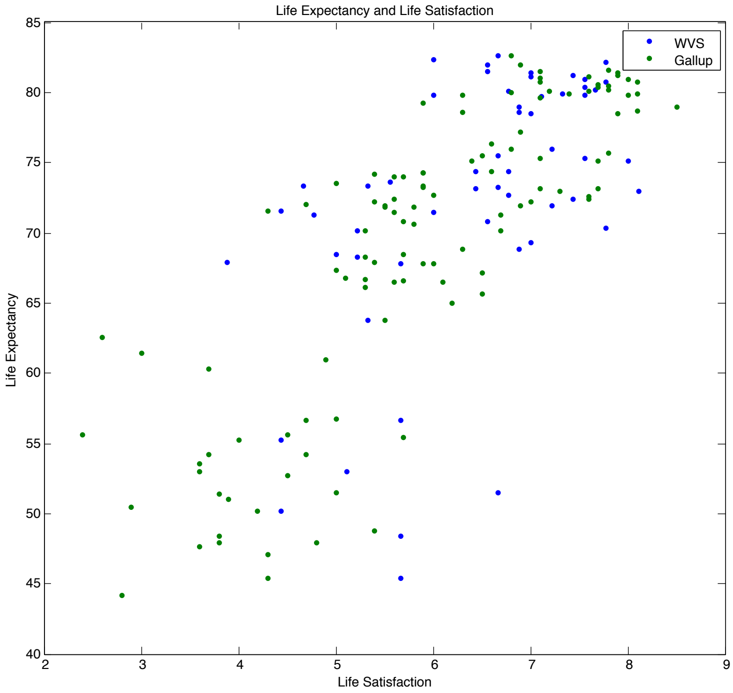


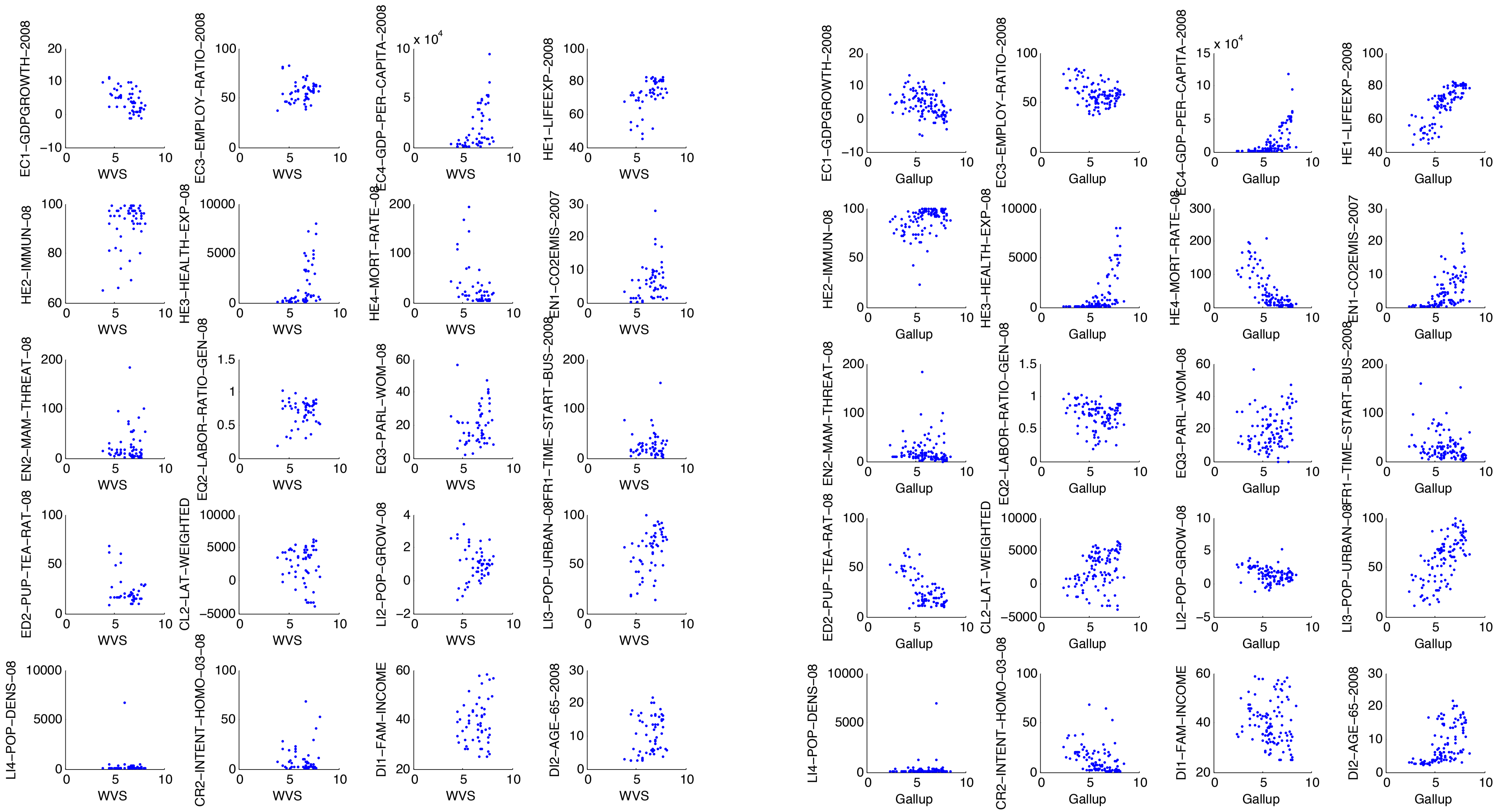
Figure 12: Life expectancy vs LS

Life expectancy was strongly correlated with both surveys

C02-emissions and mammals-threatened are strongly and positively correlated with LS. However this may be because they are strongly related to GDP per capita. C02-emissions is significantly correlated with both GDP-growth and employment, negatively and positively respectively. In fact, it is more strongly correlated with GDP growth ($p = 1.21 \times 10^{-17}$) than it is to WVS (p = 0.01) and Gallup ($p = 2.21 \times 10^{-12}$). The fact that it is more strongly correlated with GAL than WVS is also consistent with this as GAL has a stronger correlation with GDP-per-capita than WVS.

Variables may appear good indicators of LS due to confounding with GDP per capita

These correlations use a simple linear correlation but have given a useful insight into potential indicators. However, it is likely that nonlinear relationships exist, and investigations and methods in the following sections will look at both linear and nonlinear relationships.



(a) WVS against feature set

(b) GAL against feature set

Figure 13: Label / feature correlations

5.3 PCA

PCA is a useful method to uncover patterns in the data (see section 3.1.2 for details). PCA was performed on the feature set and figure 14 shows principal component (PC) 1 against PC2. These are the directions with the greatest variance in the data, where PC2 is the direction of greatest variance that is not accounted for by PC1. A threshold value of PC1 (approximately 0 because the data is centered) splits the data set such that either side shows a similar but opposed effect of a change in value of PC1.

The convex shape shows two sets of countries (at near the maximum value of PC2) where they have very different values for PC1 but the same values for PC2. There looks to be an interesting split of countries (the colors represent geographical proximity) between African with PC1 = -4 and European PC1 = +4 (approximately).

	PC1			PC2		
Feature	coef	R	p-value	coef	R	p-value
GDP-growth	-0.16	-4.35E-01	4.85E-07	-0.15	-2.29E-01	1.08E-02
employment	-0.19	-5.19E-01	7.45E-10	0.41	6.14E-01	4.16E-14
GDP-per-capita	0.27	7.60E-01	2.09E-24	0.32	4.86E-01	1.25E-08
life-expectancy	0.31	8.68E-01	1.19E-38	-0.11	-1.73E-01	5.62E-02
child-immunisation	0.19	5.17E-01	9.13E-10	-0.08	-1.23E-01	1.74E-01
health-expenditure	0.27	7.46E-01	4.57E-23	0.33	5.01E-01	3.54E-09
mortality-rate	-0.3	-8.33E-01	6.70E-33	0.18	2.73E-01	2.29E-03
c02-emissions	0.25	7.10E-01	4.02E-20	0.14	2.12E-01	1.86E-02
mammals-threatened	-0.1	-2.91E-01	1.09E-03	-0.17	-1.73E-01	5.61E-02
gender-labour-ratio	-0.02	-5.33E-02	5.58E-01	0.76	7.55E-01	5.93E-24
proportion-women-parliament	0.07	1.90E-01	3.55E-02	0.63	6.33E-01	3.76E-15
time-start-business	-0.15	-4.17E-01	1.56E-06	-0.05	-7.83E-02	3.89E-01
pupil-teacher-ratio	-0.31	-8.70E-01	6.62E-39	0.17	2.58E-01	3.91E-03
light	0.24	6.81E-01	4.67E-18	-0.03	-3.88E-02	6.70E-01
population-growth	-0.22	-6.00E-01	2.13E-13	0.14	2.13E-01	1.81E-02
population-urban	0.25	6.87E-01	1.68E-18	-0.08	-1.20E-01	1.88E-01
population-density	0.04	1.04E-01	2.50E-01	0.09	8.55E-02	3.47E-01
homicide	-0.23	-6.50E-01	3.90E-16	0.09	1.39E-01	1.25E-01
income-distribution	-0.23	-5.91E-01	6.08E-13	-0.03	-3.79E-02	6.77E-01
population-over-65	0.31	8.75E-01	5.12E-40	0.1	1.48E-01	1.02E-01

Table 8: Principal component results

PC1 and PC2 account for 99% of the variance (77 and 22 % respectively) and the coefficient values are shown in table 8. The individual features are correlated against the PCs to indicate the dominant features that constitutes each PC. PC1 correlates significantly with 18 of the 22 variables, the very low p-values such as for population-over-65, life-expectancy and pupil-teacher-ratio. PC2 however shows high coefficients for both gender-labour-ratio and proportion-women-parliament (both representing gender equality), and these are the only variables with a higher correlation for PC2 than PC1. This indicates that gender equality is not linearly (and positively) correlated with variables which can generally be thought of as representing living standards (GDP, health, education etc). Countries with average living standards such as Egypt tend to have poor gender equality, and increasing or decreasing GDP etc corresponds to an increase in gender equality.

The feature set includes latitude (light) and hence perhaps this is grouping the countries together. Removing the light feature and a highly similar graph is produced.

The correlation of GDP-per-capita with gender equality variables gender-labour-ratio and proportion-women-parliament is negative and positive either side of a GDP threshold value

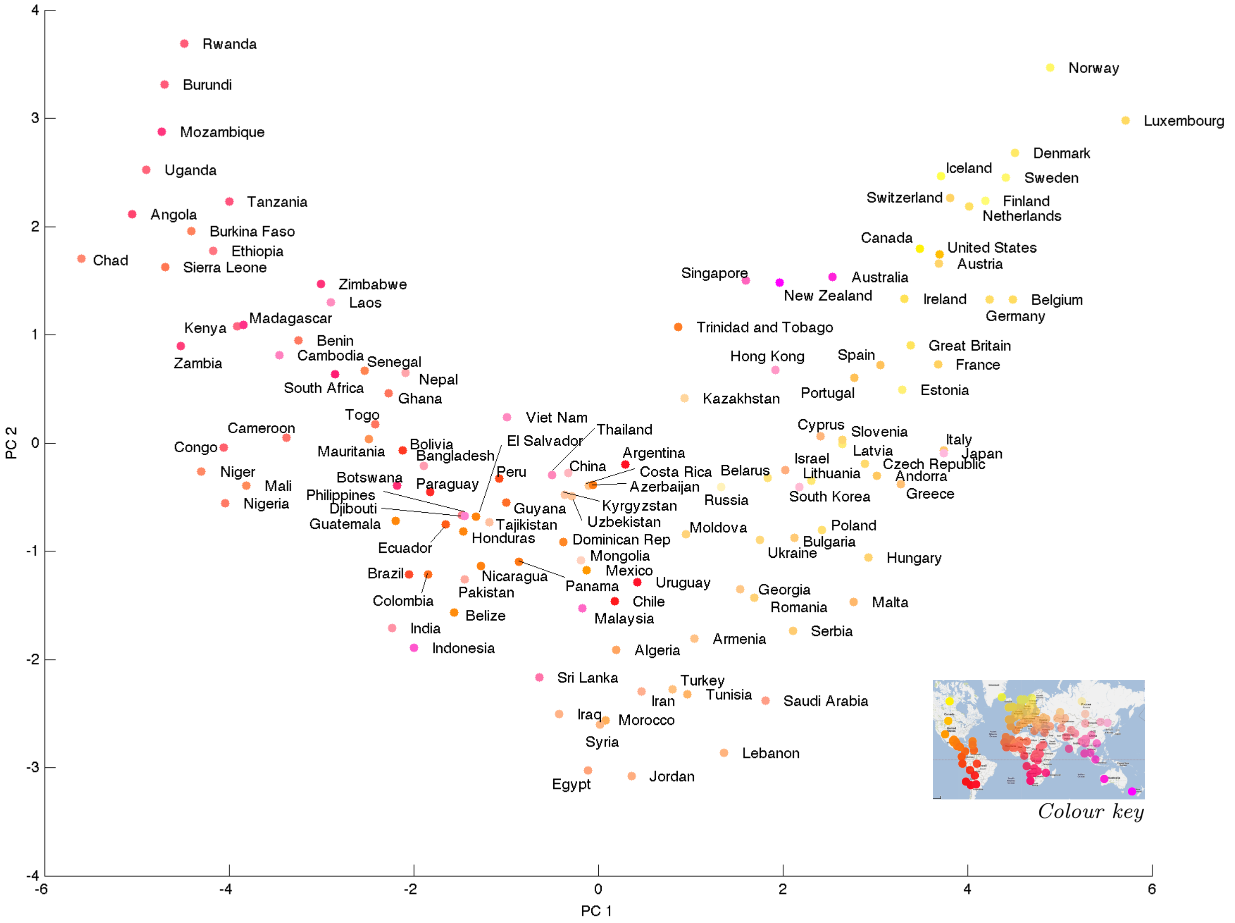


Figure 14: PC1 vs PC2

A note on representing equality The gender equality variables are naturally convex in the sense that inequality can exist where the minority is either males or females. However in reality the data is such that the women are almost always in the minority. Only the gender-labour-ratio of Rwanda and Burundi and proportion-women-parliament of Rwanda 'favour' women. We manipulate the data such that the variables truly represent equality. Gender-labour-ratio is changed to a value between 0 and 1, where 1 represents perfect equality:

$$Var_{new} = 1 - abs(var_{old} - 1) \tag{34}$$

The proportion-women-parliament was originally a % and this is changed to a value between 0 and 50, where 50 represents perfect equality (Since only Rwanda has a value above 50 this amounts to just changing this value from 56.3 to 43.7):

$$Var_{new} = 50 - abs(var_{old} - 50) \tag{35}$$

Data subset	R (gend-lab-ratio)	P (gend-lab-ratio)	R (prop-wom-parl)	P (prop-wom-parl)
$GDP <= 1.9973 \times 10^3$	-0.4059	0.0085	-0.2779	0.0785
$GDP > 1.9973 \times 10^3$	0.3823	0.0004	0.4666	9.9377×10^{-6}
$LS(GAL) <= 6.7$	-0.3594	0.0017	-0.0979	0.4065
$LS(GAL) > 6.7$	0.2885	0.0834	0.4727	0.0031

Table 9: GDP-per-capita vs gender equality (gender-labour-ratio and proportion-women-parliament)

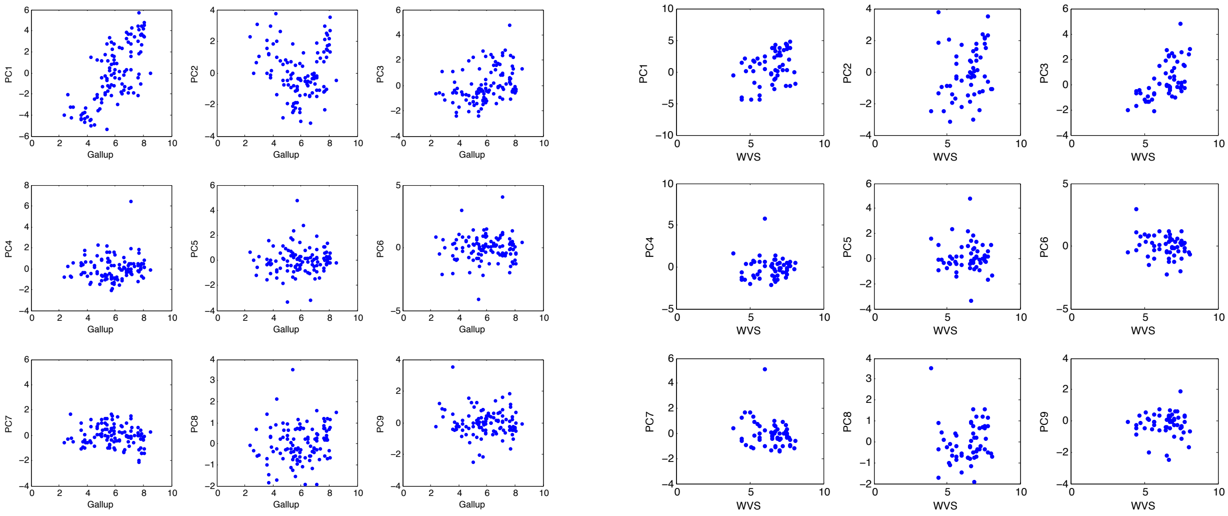
To further investigate equality the data was split into two subsets to correlate these with GDP-per-capita and LS separately. The split was done using Egypt's GDP value (1.9973×10^3)

and LS value (6.7) respectively as this country sits at the bottom of the convex relationship shown in figure 14. The results, shown in table 9, clearly show two different correlations for these subsets. 5 are significant at the 0.05 level and a further 2 at the 0.1 level with signs consistent with the PCA results.

The question is whether this is equality or inequality of a different form. For instance, is the female/male labour ratio 'better' for low GDP countries because they have no choice but to work in order to survive, in contrast to countries with better conditions where they are more comfortable and in effect have the choice of inequality. However, this reasoning is less appropriate for proportion-women-parliament and we cannot suggest a possible reason for a convex correlation of this feature. Investigating this is beyond the scope of this project, but this demonstrates the complexities of the semantics and relationships of these variables.

Figure 15 shows the correlations of labels against principal components. The GAL correlations are expected, having a linear and convex correlation for PC1 and PC2 respectively. LS is strongly and linearly correlated with many of the variables and as such the correlation with PC1 is linear. The WVS correlations are less clear because the sample is small. A convex relationship with PC2 is not clear because WVS does not have many countries with low LS values and so in effect may only be showing the upper portion of this relationship. The remaining PC's account for very little of the variance and so are not considered here.

The correlation of LS (GAL) with gender equality variables gender-labour-ratio and proportion-women-parliament is convex (negative and positive either side of a LS threshold value)



(a) Principal components vs GAL

(b) Principal components vs WVS

Figure 15: LS / principal component correlations

6 Data Preparation

6.1 Imputation

Many of the features have at least some missing values (see table 13.3). Particularly pupil-teacher ratio has 30 missing values and so it is important to keep the error as low as possible. The following section details work to attempt to impute the missing values with minimal error. This involves firstly testing the imputation to investigate how to minimise the error.

The KNN method is used because previous work has shown that this method performs better than mean values (see section 3.1.4). There are two aspects affecting the effectiveness of KNN, the value of K and also the subset of features used to impute each variable. KNN is affected by irrelevant attributes and therefore we will expect that a subset of relevant variables (specific for each feature) will give the lowest imputation error. A test script was used to test the imputation of each feature for each value of k (see pseudocode of algorithm 1). The experiment was first run on the whole feature set and then repeated using our knowledge of the variables to restrict the features used to impute each feature.

Table 10 shows the results, and table 11 shows a sample of results for imputing GDP-growth and income-distribution using various subsets. The RMSE is used to compare error results of a single feature, whereas the average error is preferred to compare results between features because they use a variable number of tests and the RMSE is relative to this. These results show that the error can be reduced by removing irrelevant attributes.

Algorithm 1 KNN test script

```
comment: impute for each K, feature and instance
for $k = 1 \to numK$ do
  for $i = 1 \to numFeatures$ do
    for $i = 1 \to numCountries$ do
      $value \leftarrow data(i, j)$
      comment: remove value then impute
      $data(i, j) \leftarrow NaN$
      $DO\ impute$
    end for
    $DO\ calculateError$
  end for
  $DO\ SET\ resultsMatrixRMSE(k, i),\ avgError\ in\ resultsMatrix(k, i)$
end for
return resultsMatrixRMSE
```

	EC1	EC3	EC4	HE1	HE2	HE3	HE4	EN1	EN2	EQ2	EQ3	FR1	ED2	CL2	LI2	LI3	LI4	CR2	DI1	DI2
# NaN	2	2	2	1	1	2	1	1	0	2	3	2	30	0	0	0	0	1	7	0
Best K	11	7	8	4	20	8	3	19	5	7	17	14	3	4	14	4	4	11	10	6
RMSE	0.88	0.70	0.54	0.50	0.90	0.55	0.45	0.82	0.94	0.80	0.94	0.91	0.49	0.74	0.74	0.70	0.96	0.82	0.64	0.40
AvgErr	0.67	0.69	0.67	0.71	0.67	0.67	0.72	0.67	0.69	0.69	0.67	0.67	0.72	0.71	0.67	0.71	0.71	0.67	0.67	0.70

Table 10: K value of each feature with lowest RMSE

Feature Imputed	Features used	RMSE	Avg err	K
GDP Growth	All	0.87	0.6575	11
	HE1, HE2 EN1, ED2, LI3, CR2, DI2	0.8524	0.6481	11
	HE1, HE3	0.9173	0.7342	11
	HE1, HE2 EN1 CR2, DI2	0.8581	0.6529	11
Income Distribution	EC4 HE1 EC4 EC4 CL2 LI2 CR2 DI2	0.6114	0.4857	8
	CL2 CR2	0.6570	0.5146	10
	All Signif (12 in total)	0.6584	0.5338	10

Table 11: GDP-growth & income-distribution imputation results

6.1.1 Imputation summary

We need the data to be as accurate as possible and KNN has obvious shortcomings to impute values of features. After imputing and visualising the results it is clear some errors may be quite large. We therefore opt to improve the data by completing some missing data values using alternative sources, to minimise the number of values that need imputing using KNN. The remaining features were imputed using the optimum K (feature specific) and a subset of features where this was found to give better inference during testing. This was done in a stepwise fashion starting with the features with the fewest variables missing such that latter imputations could use more features to infer the values. This is an area of concern, as independent data sources can have differences due to collection methods. However, the number of missing values is very small and so this will have a limited impact on the data.

Imputation with KNN did not give consistently low errors and therefore it was preferred to instead find alternative reliable data sources where possible

6.2 Transformations

The relationship of each feature with the labels was assessed visually (Figure 13), to find any clear nonlinear correlations, as these could be transformed to a linear relationship to use in our investigations. This is beneficial as we use several linear methods and additionally decision trees which have linear decision boundaries.

It can be seen that the population density of Hong Kong is an outlier by Hong Kong's extremely high population, and therefore these were re-plotted excluding Hong Kong to assess the correlation (shown in figure 16), but these show no obvious relationship however.

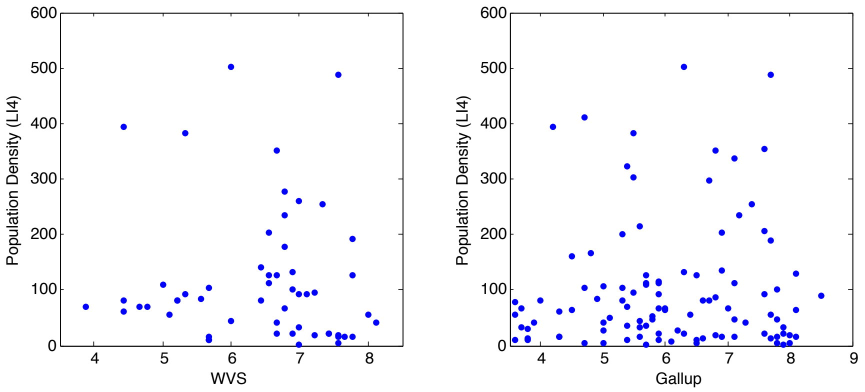
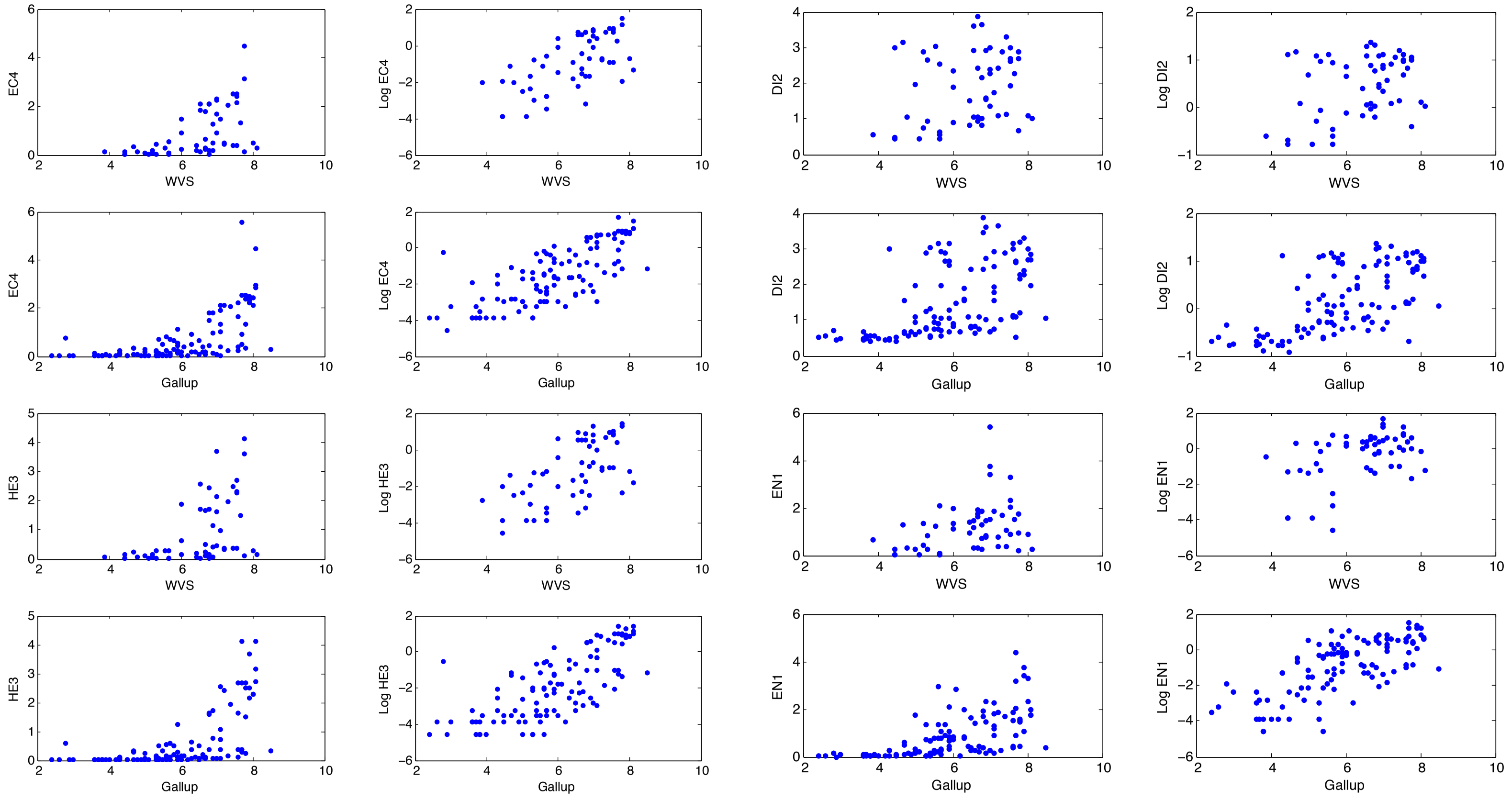


Figure 16: Population-density with Hong Kong removed

The following 4 variables were found to have non linear relationships: GDP-per-capita health-expenditure, population-over-65 and C02-emissions. The transformations are shown in figure 17(a) and figure 17(b). A new feature set was created using the logs of these 4 variables[18]. The transformations are used in conjunction with the standards features, as we will be using feature selection techniques which will enable the learner to choose itself which is more indicative.

[18] A value of zero has a log value of -Inf, and so this was altered to $log(1 \times 10^{-7})$ to solve the problem of dealing with this



(a) Log transformations 1

(b) Log transformations 2

Figure 17: Feature log transformations

7 Models to Predict Life Satisfaction

The aim of this section is to answer an initial question: Can LS be predicted using a feature set without GDP growth? More formally, the aim is to find a feature set such that the results of tests with and without GDP do not show a significantly worse result. This involves performing tests both with and without GDP and statistically comparing the results.

The main economic variable is GDP-per-capita. However, there are others that are likely to be representations of this variable also, and these are also removed. This is somewhat subjective depending on how directly they are felt to represent the intended concept to provide beneficial information in results, rather than a confounder for economic success. As an example, C02-emissions is closely related to industry and thus GDP, and hence finding that this variable helps to predict LS would likely mean that it confounds with GDP. Confounding is a big issue in this work, and for our next feature set (section 9) we will choose variables more carefully to attempt to avoid these problems. The variables removed are; GDP-growth, GDP-per-capita, population-growth, population-density, and C02-emissions. The choice of potentially confounding variables is selected cautiously, such that a larger set is preferred. This only makes our results more rigorous as a smaller feature set is less likely to be predictive than one also containing these features.

Three methods are used to investigate the predictive power of this feature set; decision trees, least squares, and lasso. 10 fold cross validation was used and a t-test (see section 3.3.2) performed on the correlations of the folds. The algorithm for DT's and least squares involve standard CV, but the lasso method is a little more involved.

Using lasso Lasso is a regularisation method and hence the regularisation parameter (λ) needs to be optimised. This should be done for each test, rather than the dataset as a whole as it is specific to the data given to lasso. The pseudocode is shown in algorithm 2, and figure 18 shows how the error varies with λ value, where the optimum λ value minimises the error on the test set. Sub-optimal λ values either constrain lasso too much such that it cannot model the patterns in the data, or not enough such that overfitting occurs.

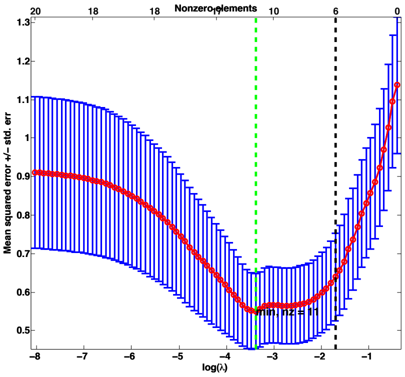


Figure 18: Lasso: optimising lambda

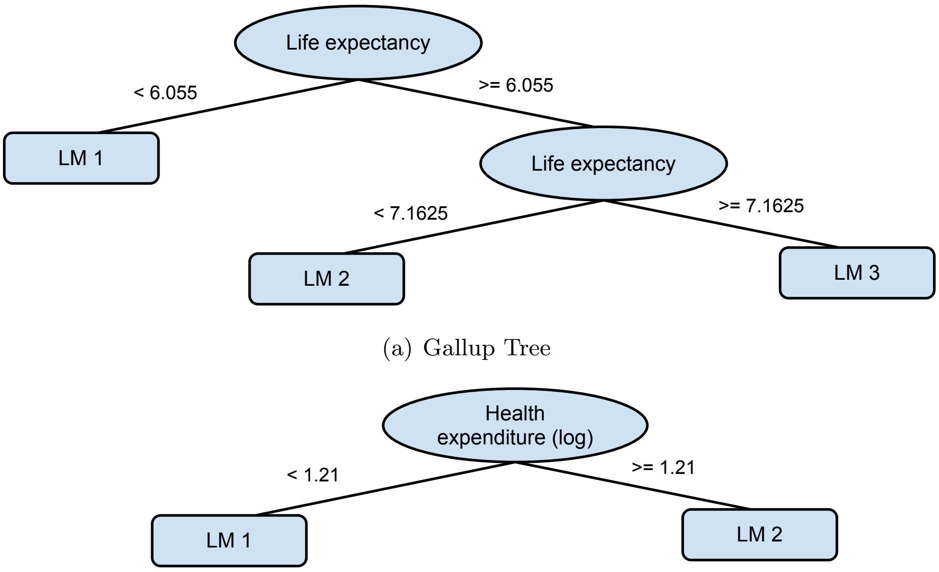
Algorithm 2 Lasso 10 fold cross validation

```
for $k = 1 \rightarrow numFolds$ do
    comment: train lasso on training set
    $lambda \leftarrow doFindLambda(trainData)$
    $model \leftarrow doLasso(lambda, trainData)$
    comment: compute correlation on test set
    $correlations(i) \leftarrow doCorrelation(testData)$
end for
return correlations
```

7.1 Results and Analysis

Decision Trees Model trees are used (see section 3.3.6) where each leaf has a regression model. This means that all trees generated are smaller as some of the data patterns are accounted for in the leaf models. The trees constructed including GDP (for both labels) consisted of a single node with a regression model. This infers that labels can best be explained by simply using a linear model rather than a decision tree structure.

The trees constructed when excluding economic variables are shown in figure 19. The trees have 3 and 2 leaf nodes respectively, splitting on health variables; health-expenditure and life-expectancy. This indicates the possibility of a more complex relationship between the features and labels. Subspaces of the attribute space are represented using separate models at the leaves. We will not analyse the coefficients of the linear models as they may contain sets of dependent variables and hence are likely to be unstable.



(a) Gallup Tree

(b) WVS Tree

Figure 19: Decision trees excluding GDP variables

Features selected are health variables; life-expectancy and log-health-expenditure

Transformations are effective, as log-health-expenditure was selected rather than the original health-expenditure variable

7.2 Statistical Methods

Table 12 shows the results of 10 fold cross validation using the three methods with and without GDP-per-capita. The models give different correlation values. A t-test is used to test the statistical significance of the difference between the results, for each method/label. The t-test compares the cross validation results to determine if these are likely to have come from the same distribution. The null hypothesis states that the difference between the results sets has a mean of zero, and hence a statistically significant t-test result indicates that the two result sets are significantly difference. Therefore, we aim to show that the results are comparable such that H_0 is not rejected (or the feature set excluding GDP-per-capita performs better and H_0 is rejected).

Learner	Survey	Corr coeff with GDP	Corr coeff without GDP*	test statistic	p-value
Dec Tree	GAL	0.88	0.89	0.55	0.59
Least Sq		0.8378	0.8285	1.9883	0.0780
Lasso		0.87	0.88	-0.8560	0.4142
Dec Tree	WVS	0.64	0.65	0.97	0.35
Least Sq		0.8323	0.7797	5.3932	0.00044
Lasso		0.66	0.78	-1.3204	0.2193

Table 12: Results of learners with and without GDP (* Excluding GDP excludes the variables: GDP-growth, GDP-per-capita, population-urban, population-density, C02-emissions)

7.2.1 Conclusions

Table 12 shows there is no significant difference ($p < 0.05$) between the results when removing economic type variables, for 5 of the 6 tests performed. This is very encouraging, as we are using just 15 variables to predict LS and are able to do this without a significant difference when GDP is removed. Its interesting to note that the correlation values of these tests are very similar, and in 4 of these tests actually increases with GDP-per-capita removed (although this difference is not significant). Additionally, it is likely that other variables that may be effective in a model to predict happiness and these will be included later.

Result: Feature sets excluding economic variables are able to predict LS as well as GDP-per-capita

8 Feature Selection (1)

We have shown that LS can be inferred when using a feature set excluding GDP. We now investigate which features in particular are key in these models. Two approaches are used for this; lasso and least squares regression. Permutation tests using bootstrapping are used to determine the significance of the results. This work amounts to a pattern discovery problem, where the patterns are feature subsets or individual features. Therefore the size of the pattern space is all possible patterns considered. There are 17 features (15 features plus 2 log versions) and hence the pattern spaces are 2^{17} for feature subsets and 17 for individual features. Significance of results is assessed using permutation testing to provide a threshold test statistic t_σ (see section 3.3.3) such that:

$$p(T \geq t_\sigma) \leq 0.05 \tag{36}$$

where T is the test statistic on permuted data.

8.1 Lasso Feature Selection

Lasso[19] finds a subset that gives a good correlation, but where there are several dependent variables lasso results are inherently unstable as small changes in the data set supplied to the learner can affect the variables chosen in the model. This is because when several variables have very similar relationships with the labels the learner chooses the best to include in the model, and repeating this on different data can cause the learner to select a different feature each time. Therefore performing lasso on different bootstrapped data subsets results in a set of models. Although the coefficient values are unlikely to be the same, often the same subset of features are used (the coefficients of the same feature set have not been reduced to zero). Therefore tests must be performed many times to improve the stability of the results, by finding frequent exact subsets.

Lasso is used here to find both dominant subsets of features and also individual features. To discover important feature sets we aim to identify key exact subsets where they are produced significantly often by lasso, which will indicate relationships between variables to predict LS.

The number of features is constrained to 6 in these tests[20], as we are interested in small predictive subsets of features. Although the correlations will be slightly reduced it is still very comparable to when not constraining (in fact some of the models produced when not constraining the model size are < 6)[21]. A t-test of constrained (< 6) against unconstrained gave p-values of 0.6914 and 0.9077 for GAL and WVS respectively, showing the results are comparable.

The lasso tests aim to answer three questions. These are all based on a test method involving bootstrapping 100 times, and returning an individual feature or feature subset and a test statistic using the results of the bootstraps. Three questions are posed:

1. Are there feature subsets that are generated frequently by lasso? We will call this frequency the vote.

   H_0: The most frequent subset generated is no more frequent than the most frequent subset found on random data

   Pattern: Feature subset

   Output: Most voted (exact) subset.

   Test statistic: Vote of most voted subset

[19] glmnet Matlab library was used

[20] default setting is $numFeatures + 1$

[21] The models are also constrained to > 0 features, which is necessary because in permutation tests where there is no correlation between (random)label and features lasso often prefers a model with no features (just a bias), but we are interested in the best model containing at least 1 feature

2. Are the most voted subsets significantly correlated with the label? When permuting the label is a correlation greater than this likely?

   H_0: Lasso models generated using the most frequent subset are no more correlated with LS than with random labels

   Pattern: Feature subset

   Output: Most voted (exact) subset.

   Test statistic: Correlation value

3. Is the highest feature frequency in the bootstrap models likely to occur by chance?

   H_0: The most frequent feature is no more frequent than the most frequent found with random data

   Pattern: Individual feature

   Output: Most frequent feature

   Test statistic: Feature frequency: number of bootstraps it occurs in (rather than the number of unique feature sets)

Methodology Figure 20 shows the test design and the three test statistics corresponding to the questions specified above. 100 bootstraps are performed and the 3 test statistics and variables are output. In actuality a sorted listing is returned to provide more information, sorted by test statistic so the significance threshold can be used to assess models that are lower in this ordering. For instance, if the second result is greater than the threshold then it can also be deemed significant.

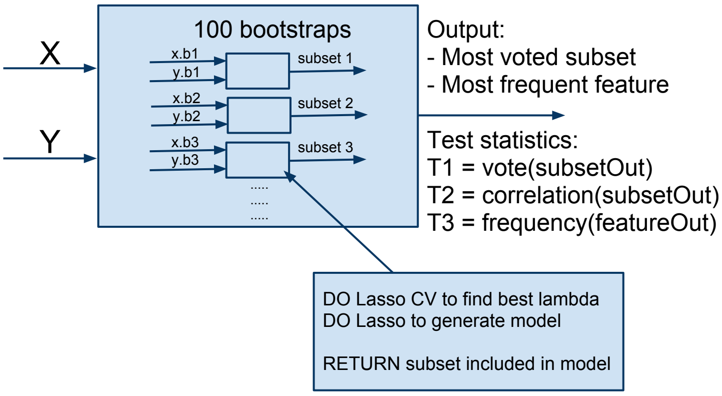


Figure 20: Lasso test design

We can define the most frequent subset as:

$$f^* = argmax_{f \subseteq F}(vote(B, f)), \tag{37}$$
$$\tag{38}$$

and then formally the test statistics are as follows:

$$T_1 = vote(f^*) \tag{39}$$
$$T_2 = correlation(f^*) \tag{40}$$
$$T_3 = max_{f \in F}(frequency(B, f)) \tag{41}$$

where F and B are the set of features and bootstraps respectively, and

$$vote(B, f) = \#b \in B : (lasso(b) = f) \tag{42}$$

The significance thresholds t_σ are found with permutation testing. The test (figure 20) is repeated many times with permuted labels, giving a set of T values. The value of T where it is greater than 95% of these is the significance threshold (with $p = 0.05$):

$$0.05 \geq \frac{\#p \in P : T(B_p) \geq t_\sigma}{\#p \in P}, \tag{43}$$

where P is the set of permutations, such that under H_0:

$$p(T(B) \geq t_\sigma) \leq 0.05 \tag{44}$$

Results: Feature Subsets Tables 13 and 14 show the most voted subsets in 100 bootstraps. Permutation testing found significance thresholds for T_1 of 7 for both labels. 3 significant models were found for GAL but none for WVS. The GAL significant subsets always contained log-health-expenditure, income-distribution, health-expenditure and 2 of either mortality-rate, gender-labour-ratio, and proportion-women-parliament.

All GAL significant feature subsets contained: log-health-expenditure, income-distribution and health-expenditure

mortality-rate, gender-labour-ratio, and proportion-women-parliament were each found in 2 of the 3 significant GAL subsets

Lasso with WVS generated few repeated subsets with 88 different subsets for 100 bootstraps. Compared with GAL these results are much more unstable and this may be because WVS has a smaller less varied sample of countries. The most voted model is however quite prominent with a vote of 5 in comparison to votes of 1 or 2 for all other subsets found. This model contains: life-expectancy, child-immunization, log-health-expenditure, mammal-threatened, income-distribution, and health-expenditure.

The most voted WVS model also contains the three features found in all significant GAL models: log-health-expenditure, income-distribution and health-expenditure

The significance thresholds for correlations of the most voted subsets (T_2) were 0.3284 and 0.4369 for GAL and WVS respectively. The correlations of all subsets were higher and thus significant, with mean correlations of 0.8620 (GAL) and 0.7660 (WVS).

All subsets found are significantly predictive of LS (compared to random case)

Results: Individual Features Tables 13 and 14 show the feature frequencies in the models generated by lasso in the 100 bootstraps. A conservative approach to discarding features is taken due to the high dependency between variables. For instance, given two variables v_1 and v_2 that are part of a feature set on which lasso is performed, and where $v_1 \sim v_2$. Due to the random selection of instances in the bootstraps and the small differences between v_1 and v_2 we can expect that their vote will be shared approximately equally between them. Hence the votes of each are reduced and this could potentially cause two significant variables to fall outside of the significant range.

	EC3	HE1	HE2	LOG HE3	HE4	EN2	EQ2	EQ3	FR1	ED2	CL2	LI2	CR2	DI1	LOG DI2	HE3	DI2	vote
Models	0	1	0	1	1	0	1	0	0	0	0	0	0	1	0	1	0	9*
	0	1	0	1	1	0	0	1	0	0	0	0	0	1	0	1	0	8*
	0	1	0	1	0	0	1	1	0	0	0	0	0	1	0	1	0	7*
	0	1	1	1	0	0	0	1	0	0	0	0	0	1	0	1	0	6
	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	1	0	4
	0	1	0	1	0	0	0	1	1	0	0	0	0	1	0	1	0	4
	0	1	0	1	1	0	1	1	0	0	0	0	0	0	0	1	0	4
	0	1	0	1	1	1	0	1	0	0	0	0	0	1	0	0	0	3
	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	1	0	2
	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0	1	0	2
	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	1	0	2
	0	1	0	1	0	0	1	1	0	0	0	0	0	1	0	0	0	2
	0	1	0	1	0	0	1	1	1	0	0	0	0	1	0	0	0	2
	0	1	0	1	1	0	0	0	1	0	0	0	0	1	0	1	0	2
	0	1	0	1	1	0	0	1	0	1	0	0	0	1	0	0	0	2
	0	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	2
	0	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	2
	0	1	0	1	1	0	1	1	0	1	0	0	0	0	0	0	0	2
	0	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	2
	. . . & 32 with vote 1																	1
Tot Feat freq (max 100)	0	100	14	100	45	12	50	68	18	9	2	8	0	64	0	70	3	0

Table 13: Unique feature sets for GAL (constrained to model size 6, 100 bootstraps)

	EC3	HE1	HE2	LOG HE3	HE4	EN2	EQ2	EQ3	FR1	ED2	CL2	LI2	CR2	DI1	LOG DI2	HE3	DI2	vote
Models	0	1	1	1	0	1	0	0	0	0	0	0	0	1	0	1	0	5
	0	0	1	1	0	1	0	0	0	0	0	1	0	1	0	1	0	2
	0	1	0	1	0	0	0	0	0	0	1	1	0	1	0	1	0	2
	0	1	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	2
	0	1	0	1	0	0	0	1	0	0	1	1	0	1	0	0	0	2
	0	1	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	2
	1	1	0	1	0	1	0	0	0	0	0	0	0	1	0	1	0	2
	1	1	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0	2
	1	1	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	2
	. . . & 79 with vote 1																	1
Tot Feat freq (max 100)	38	59	29	97	7	57	5	26	17	14	25	34	15	90	1	42	0	0

Table 14: Unique feature sets for WVS (constrained to model size 6, 100 bootstraps)

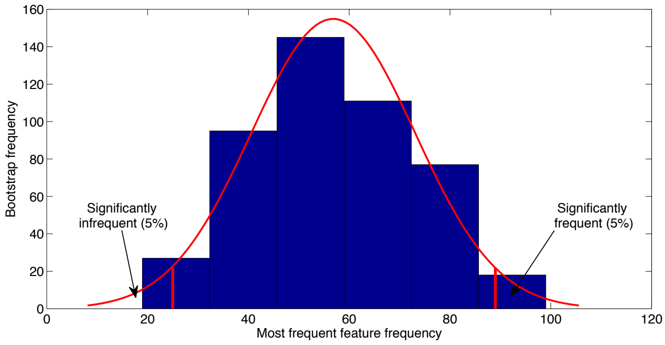


Figure 21: Feature frequency distribution illustration

Therefore, in addition to the 95% threshold we also take a 5% threshold to find features that are significantly *infrequent*, to aid us in discarding irrelevant features. The two thresholds help assess features with high correlation and dependencies, as illustrated in figure 21. A minimum threshold value will help us discard those features that have significantly little contribution to the models, such that the coefficient is often reduced to zero (significantly more often than the random case). Table 15 shows these three types of results.

	EC3	HE1	HE2	LOG HE3	HE4	EN2	EQ2	EQ3	FR1	ED2	CL2	LI2	CR2	DI1	LOG DI2	HE3	DI2	$t_{0.05}$	$t_{0.95}$
GAL	✗	✓	✗	✓	-	✗	-	-	✗	✗	✗	✗	✗	-	✗	-	✗	32	89.8
WVS	-	-	✗	✓	✗	-	✗	✗	✗	✗	✗	-	✗	✓	✗	-	✗	33	89.25

Table 15: Feature significance in lasso models: significantly infrequent (✗), intermediate (-) and significantly frequent (✓) (and threshold values)

7 features are significantly infrequent for both labels and can be discarded from the feature set (if both surveys find them infrequent). 3 are significantly frequent for at least one label; life-expectancy, log-health-expenditure and income-distribution. Log-health-expenditure dominates the results, used in all models for GAL and 97% for WVS. GAL and WVS have some different results. They both have two significant features but the second differs, being child-immunisation and income-distribution for GAL and WVS respectively (the survey not finding significance also did not find the feature significantly infrequent either and so this is not a big inconsistency).

Significantly frequent features: life-expectancy, log-health-expenditure, income-distribution

Notable features: employment, mortality-rate, mammals-threatened, gender-labour-ratio, proportion-women-parliament, population-growth, health-expenditure

Significantly infrequent features: child-immunisation, time-start-business, pupil-teacher-ratio, light, homicide, log-population-over-65, population-over-65

The results indicate a set of core features present consistently in the models, and others that are interchangeable and hence often present

Sample Models We can gain some insight from these generated models, and the type of contribution of each feature, indicated by the sign and size of the coefficient values. Examples equations 45 and 46 show two models for the top ranked subsets of GAL and WVS respectively.

$$\begin{aligned} LS_{GAL} \;=\; & 0.7815 \;+\; 0.7232\; life-expectancy \;+\; 0.1569\; log-health-expenditure \;-0.1143\; mortality-rate \\ & -0.1588\; gender-labour-ratio \;+\; 0.2694\; income-distribution \;+\; 0.3291\; health-expenditure \end{aligned} \tag{45}$$

$$\begin{aligned} LS_{WVS} \;=\; & 2.5259 + \; 0.1902\; life-expectancy \;+\; 0.0906\; child-immunisation \;+\; 0.3218\; log-health-expenditure \\ & +\; 0.1525\; mammals-threatened \;+\; 0.4660\; income-distribution \;+\; 0.1968\; health-expenditure \end{aligned} \tag{46}$$

All health variables included in the models have an intuitive sign. For instance, mortality-rate has a negative coefficient because mortality rate is negatively correlated with LS. However, the remaining variables show unexpected coefficients with respect to the concept each is representing. These variables are income-distribution, gender-labour-ratio and mammals-threatened.

The income-distribution coefficient is positive indicating a higher value is associated with higher LS. However, this is unexpected because high income-distribution means greater *inequality* of income. At first thought, this may be due to confounding with GDP but the initial correlations are not consistent with this. LS is correlated strongly and positively with GDP but GDP is correlated negatively with income-distribution.

In fact, work by Verme [50] details how previous results of the relationship between LS and relative income have been inconsistent, finding both negative and positive correlations. It is noted how the type of data might affect the results such as using small datasets, particular noting cross-country studies. Hence, perhaps income-distribution may be more appropriate as an indicator when people are entities and the relationship between LS and income equality can be directly analysed. One suggested explanation of a positive coefficient is the affect of relative income on social mobility.[50] A greater degree of income inequality relates to more potential to 'climb the ladder', to have ambition and progress in life, which improves LS.

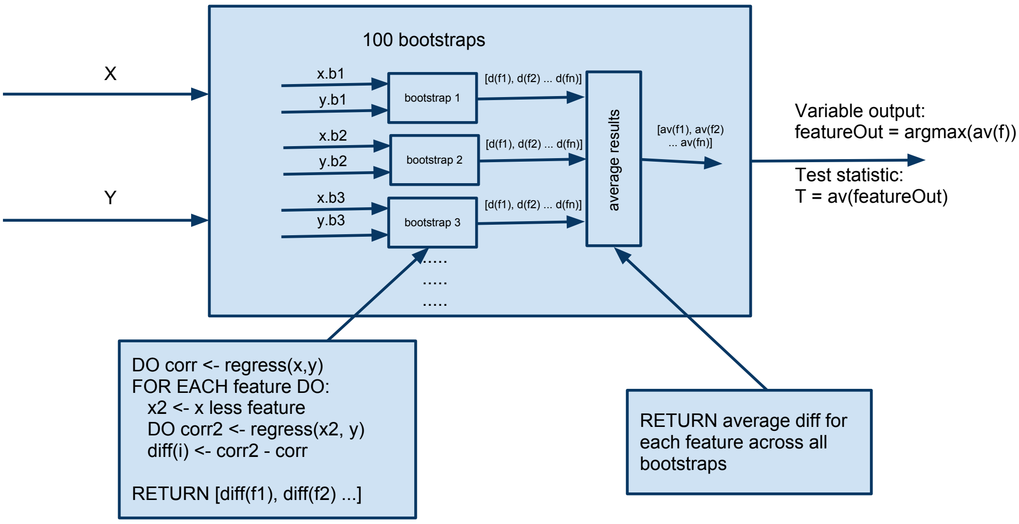
The variable mammals-threatened contributes positively to the WVS model. However, with hindsight it is unclear what this variable is representing and therefore will be discarded. A higher gender-labour-ratio value corresponds with a more equal gender employment, but the GAL model includes gender-labour-ratio with a small but negative coefficient. However, we showed with PCA analysis in section 5.3 that gender equality has a convex correlation with life satisfaction and therefore this may be the reason why the coefficient is not intuitive.

The unexpected coefficients highlight the restrictions of a linear model. Although the model has a high correlation it is constructed in a way which is contradictory to previous work and our own intuition. Lasso is an effective method for generating models, but these models are still only simple linear models and this is limiting. Further analysis with decision trees will be useful to investigate non linear patterns in the data.

Some features (e.g. mammals-threatened) have unexpected coefficients and one possible cause is confounding with economic success (GDP-per-capita)

8.2 Least Squares Leave One Feature Out

We use least squares to provide an independent alternative approach to finding key features, by analysing the effect on correlation values when removing a feature in turn from the feature set.



```
DO corr <- regress(x,y)
FOR EACH feature DO:
  x2 <- x less feature
  DO corr2 <- regress(x2, y)
  diff(i) <- corr2 - corr

RETURN [diff(f1), diff(f2) ...]
```

Figure 22: Least squares test design

Methodology Figure 22 shows the test design where the difference in correlation is calculated for each bootstrap with and without each feature. These values are averaged for each feature, and the features are output, ranked by average reduction of correlation. The first feature therefore causes the biggest reduction in correlation when removed from the feature set. The test statistic is the average correlation difference:

$$T = max_{f \in F}(av(f)), \tag{47}$$

where function av computes the average difference for a feature over all bootstraps. The significance threshold t_σ (with $p = 0.05$) is found by performing the test many times with permuted labels and finding T such that 95% of the results on random data give a smaller test statistic.

Results Tables 23(a) and 24(a) show the results of these tests, with the features ranked by test statistic. The permutation tests found significance thresholds of 0.0592 and 0.0571 for GAL and WVS respectively. Therefore, the results are not significant compared to the random case. However this is likely to be because the features are highly related and thus removing one feature and another feature can in effect represent much of the variance of the removed variable.

Tables 23(b) and 24(b) show t-test results of comparisons between the set of bootstrap outputs of each feature. This shows most are significant indicating the feature rankings found have significant differences between the results of each feature (such that the ranking is stable and not

Feature	Corr diff
LOG HE3	0.028
DI1	0.027
HE2	0.026
HE1	0.025
EQ3	0.021
LOG DI2	0.020
HE4	0.014
CL2	0.011
CR2	0.011
EQ2	0.011
HE3	0.011
LI2	0.010
EC3	0.010
FR1	0.010
ED2	0.010
EN2	0.010
DI2	0.008

(a) Correlation difference ranking

	LOG HE3	DI1	HE2	HE1	EQ3	LOG DI2	HE4	CL2	CR2	EQ2	HE3	LI2	EC3	FR1	ED2	EN2	DI2
LOG HE3	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DI1	1	-	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
HE2	1	0	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1
HE1	1	1	1	-	1	1	1	1	1	1	1	1	1	1	1	1	1
EQ3	1	1	1	1	-	1	1	1	1	1	1	1	1	1	1	1	1
LOG DI2	1	1	1	1	1	-	1	1	1	1	1	1	1	1	1	1	1
HE4	1	1	1	1	1	1	-	1	1	1	1	1	1	1	1	1	1
CL2	1	1	1	1	1	1	1	-	0	0	1	1	1	1	1	1	1
CR2	1	1	1	1	1	1	1	0	-	0	1	1	1	1	1	1	1
EQ2	1	1	1	1	1	1	1	0	0	-	1	1	1	1	1	1	1
HE3	1	1	1	1	1	1	1	1	1	1	-	1	1	1	1	1	1
LI2	1	1	1	1	1	1	1	1	1	1	1	-	0	0	1	1	1
EC3	1	1	1	1	1	1	1	1	1	1	1	0	-	0	1	1	1
FR1	1	1	1	1	1	1	1	1	1	1	1	0	0	-	1	1	1
ED2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-	0	1
EN2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	-	1
DI2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-

(b) Significance of difference between correlation differences of each feature (GAL)

Figure 23: GAL least squares results

Feature	Corr Diffs
LOG HE3	0.052
EN2	0.032
DI1	0.029
ED2	0.026
HE2	0.026
CR2	0.026
HE4	0.024
HE1	0.023
LOG DI2	0.022
EC3	0.022
FR1	0.021
EQ3	0.020
HE3	0.020
CL2	0.018
LI2	0.018
EQ2	0.017
DI2	0.012

(a) Correlation difference ranking

	LOG HE3	EN2	DI1	ED2	HE2	CR2	HE4	HE1	LOG DI2	EC3	FR1	EQ3	HE3	CL2	LI2	EQ2	DI2
LOG HE3	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
EN2	1	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
DI1	1	1	-	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ED2	1	1	1	-	0	0	1	1	1	1	1	1	1	1	1	1	1
HE2	1	1	1	0	-	0	1	1	1	1	1	1	1	1	1	1	1
CR2	1	1	1	0	0	-	1	1	1	1	1	1	1	1	1	1	1
HE4	1	1	1	1	1	1	-	0	1	1	1	1	1	1	1	1	1
HE1	1	1	1	1	1	1	0	-	0	1	1	1	1	1	1	1	1
LOG DI2	1	1	1	1	1	1	1	0	-	0	1	1	1	1	1	1	1
EC3	1	1	1	1	1	1	1	1	0	-	0	1	1	1	1	1	1
FR1	1	1	1	1	1	1	1	1	1	0	-	0	1	1	1	1	1
EQ3	1	1	1	1	1	1	1	1	1	1	0	-	0	1	1	1	1
HE3	1	1	1	1	1	1	1	1	1	1	1	0	-	1	1	1	1
CL2	1	1	1	1	1	1	1	1	1	1	1	1	1	-	0	1	1
LI2	1	1	1	1	1	1	1	1	1	1	1	1	1	0	-	1	1
EQ2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-	1
DI2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-

(b) Significance of difference between correlation differences of each feature (WVS)

Figure 24: WVS least squares results

trivial). For example, log-health-expenditure is the first ranked feature and on removal caused a reduction of correlation significantly different from the other variables.

log-health-expenditure is the top ranked feature for both labels, causing a significantly larger reduction in correlation if excluded

8.3 Lasso / Least Squares Results Comparison

Feature selection using lasso and least squares has found some similar results. Firstly the features found to be significant with lasso are near the top of the least squares rankings. In particular log-health-expenditure is ranked first for both labels. Income-distribution was ranked second and third in least squares and was present in 0.64 and 0.9 of the lasso models, for GAL and WVS respectively. Life-expectancy is ranked fourth and eighth which is consistent with the lasso results where this variable was found in 1.0 and 0.59 of lasso models for GAL and WVS respectively.

Both tests find log-health-expenditure, income-distribution and life-expectancy to be important features

With respect to the highly infrequent features in the lasso's model, least squares also found population-over-65 to be very infrequent.

population-over-65 has been found significantly less important than the other features in predicting LS

8.4 Results Summary

LS predictors Health variables have shown particular prominence in the results, particularly health-expenditure and life-expectancy. These variables were key in tests of all three methods, and were the only variables used as internal nodes in the decision trees. Health expenditure is quite a general term and therefore this is now removed from the feature set as more direct notions of health care quality are preferred.

The variables proportion-women-parliament and gender-labour-ratio, which were intended to represent gender equality, were present in the significant feature subsets of the GAL lasso. It is interesting to note how the top two feature subsets contained one of these, thus perhaps they represent the same concept and lasso chose between them. This is also consistent with the fact that these features were neither significantly frequent of significantly infrequent (table 15).

The feature income-distribution was consistently a key component of the models with a positive coefficient, which infers that income inequality indicates LS. However, the reason for this is unclear, and previous work has shown conflicting results on this topic.

Methods Lasso and least squared gave some similar results but also many differences. Lasso is 'fairer' with respect to variable selection (see section 3.3.5) and therefore is more powerful and informative than the greedy approach of least squares. Least squares provides a useful comparison for the lasso results but more importance will be given to the lasso results as it is able to cope with dependent variables better and so is more reliable. The models trees are very short due to the regression models at the leaves and therefore it may be more informative to use regression trees. Lasso and least squares are restrictive because they produce linear models, and hence we will investigate further machine learning approaches to find non linear patterns.

Features discarded The features discarded from this stage and reasons:

- health-expenditure, log-health-expenditure: Too general (replaced with more direct health indicators)
- GDP-growth, GDP-per-capita: Economic variables
- population-urban, population-density, C02-emissions, mammals-threatened: Likely confounding with GDP
- population-over-65, log-population-over-65, homicide, light,pupil-teacher-ratio, time-start-business: Significantly infrequent in lasso tests

9 Feature Selection (2)

9.1 Improved Feature Set

Table 16 shows the updated features set, including prominent features from the work of section 8 and also new features (further details in appendix A, table 13.3).

	Topic	Feature	Code	
Existing	HEALTH	Life expectancy %	life-expectancy	HE1
		Immunization %	child-immunisation	HE2
		Mortality rate	mortality-rate	HE4
	EQUALITY	Ratio of gender labor participation rate	gender-labour-ratio	EQ2
		Proportion of seats held by women in national parliaments	proportion-women-parliament	EQ3
		Distribution of family income - Gini index	income-distribution	DI1
	LIFE	Employment to population ratio	employment	EC3
		Population growth	population-growth	LI2
New	HEALTH	Hospital beds	hospital-beds	BH1
	FREEDOM	% of the largest religion or religious brand	percent-largest-religion	BF1
		Number of unique incidents of religious conflict	religious-confilict	BF2
		Freedom of the world	freedom	BF3
	CLIMATE	Mean temperature	temp-mean	BC1
		Min temperature	temp-min	BC2
		Max temperature	temp-max	BC3
	EDUCATION	School enrolment (primary gross)	primary-education-enrolment	BE1
		School enrolment (secondary gross)	secondary-education-enrolment	BE2
		Literacy rate	adult-literacy-rate	BE3
	CONFLICT	Military expenditure (relative to health exp)	military-expenditure	BP2
	EQUALITY	Secondary education (% female)	education-gender	BQ1

Table 16: Feature set 2

The additional features were chosen based on previous research and also the results found so far. Climate has been found to affect happiness such as work described in section 2. The results for pupil-teacher-ratio were surprising and hence this is investigated further by using different education variables. Freedom is an important subject but only one variable, time-start-business, has been used to represent this which was perhaps not highly representative of freedom. Religious indicators are also included to represent religious freedom / equality, as well as a generic freedom index.

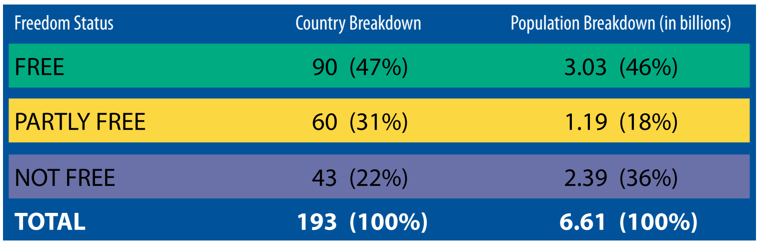
Freedom Index [28] We prefer to use primary sources but freedom is very difficult to quantify, and therefore in this case we use a freedom index[22]. This is derived from survey data and covers the following freedom areas; politics, expression and belief, religion, academic, associational and organizational rights, law and personal. The index values range from 1 to 7 representing highly free and not free respectively, and figure 25 shows the freedom of the world.

Military Expenditure We investigate any relationship of military size with LS. The original variable is military expenditure as a percentage of GDP. The models generated included this feature with negative coefficients. However, this may be due to confounding with GDP. This variable is investigating the relationship:

$$\alpha \frac{MilExp}{GDP} = LS. \tag{48}$$

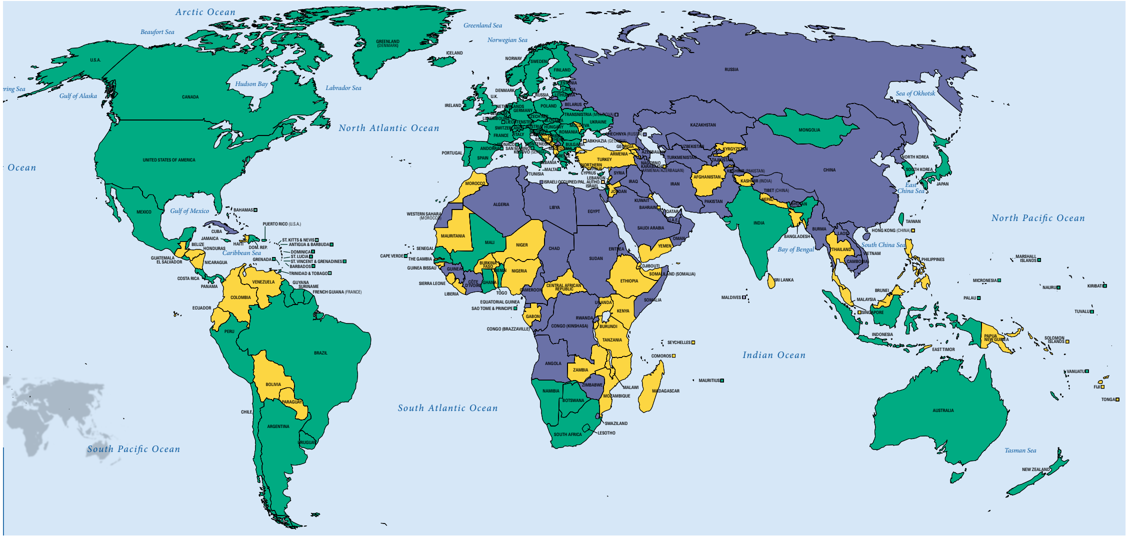
Assuming that $MilExp$ stays constant then α would be strongly negative because GDP-per-capita has a strong positive correlation with LS. Therefore, where the covariance with LS of military expenditure is small compared to that of GDP, using this variable relative to GDP is in fact just incorporating a measure of GDP in the models, rather than military expenditure as intended. Therefore this variable is adapted to be relative to health expenditure instead. It can now be thought of as a measure of the investment in military relative to the investment in health care. As is shown in the subsequent tests this variable gives different results to the original.

[22] produced by Freedom House [28]



Freedom Status	Country Breakdown	Population Breakdown (in billions)
FREE	90 (47%)	3.03 (46%)
PARTLY FREE	60 (31%)	1.19 (18%)
NOT FREE	43 (22%)	2.39 (36%)
TOTAL	**193 (100%)**	**6.61 (100%)**

(a) Map Key & Statistics

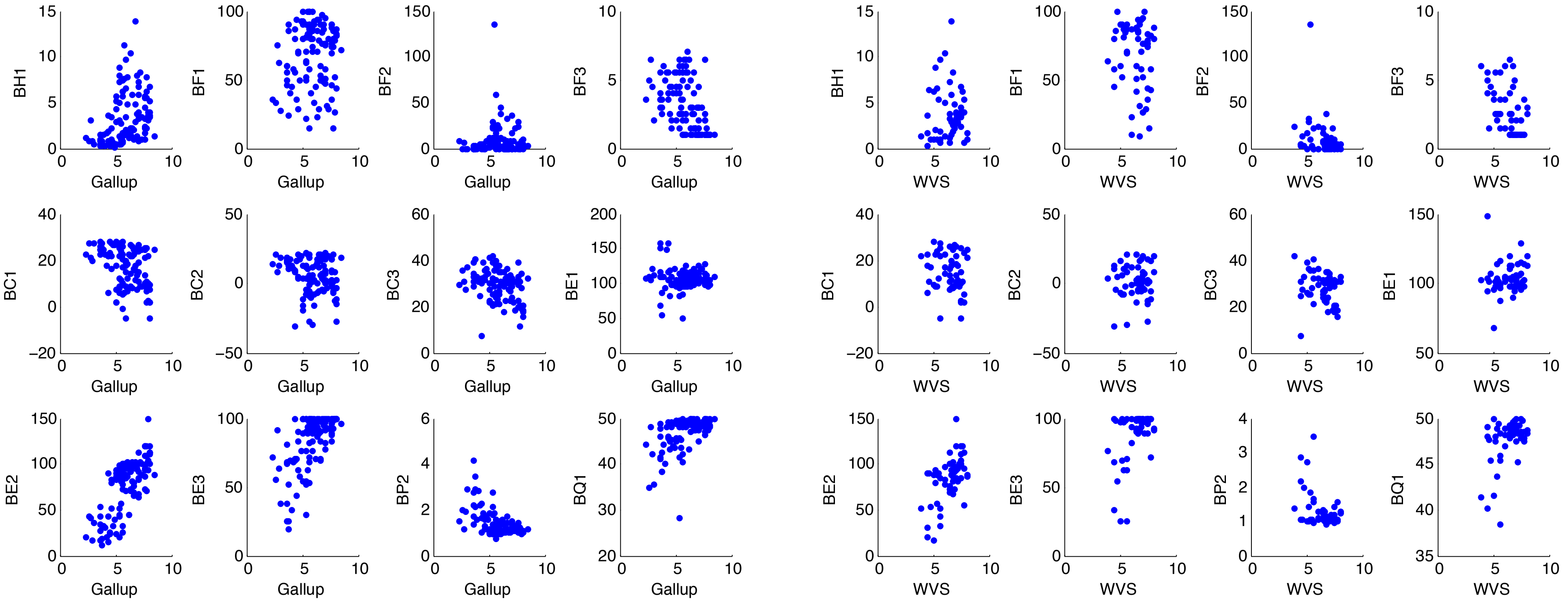


(b) World Map

Figure 25: Freedom House world map [28]

9.2 Initial Analysis and Correlations

The new features (figure 26) do not show many clear relationships on visual inspection. Freedom appears linear (and negative) but quite 'fat', with alot of variance at each LS value. Secondary enrolment does show a linear and positive correlation with LS. The difference between primary and secondary school enrolment is quite striking, where primary-education-enrolment has much less variance with LS. Also, the countries with lower LS appear to have greater variance of primary-education-enrolment, both above and below 100, whereas those with high LS are clustered much tighter around the mean (100%). These enrolment variables are gross enrolment figures, relative to the population that officially corresponds to that education level. Hence primary enrolment may be above 100% for countries of lower LS because there are older members of the population taking primary level education.



(a) GAL against new features

(b) WVS against new features

Figure 26: Label / feature correlations

9.3 Finding Significant Features: Decision Trees

Regression trees are used for feature selection to generate larger trees than that of model trees. REPTree in Weka is used which used reduced error pruning to improve the stability of the tree. The WVS tree, shown in figure 28 is the same as was generated previously. The GAL tree (figure 27) however is much larger and uses 6 different features as internal nodes.

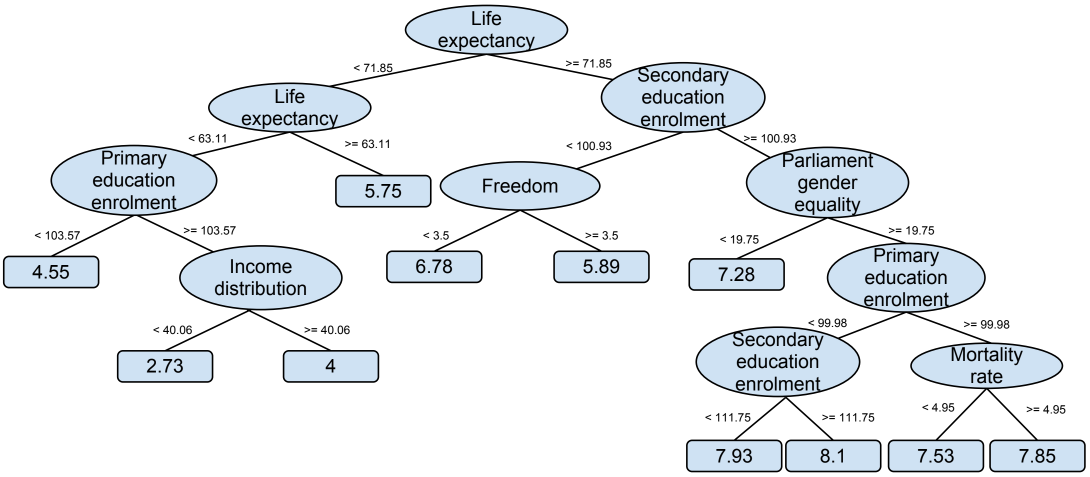


Figure 27: Decision tree: All features & GAL Label

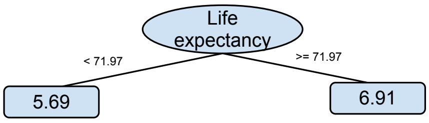


Figure 28: Decision tree - all features & WVS Label

Features selected by decision tree: mortality-rate, life-expectancy, income-distribution, primary-education-enrolment, secondary-education-enrolment, proportion-women-parliament

9.4 Finding Significant Features: Lasso

Lasso significance tests are repeated using the feature set of table 16. Table 17 shows the significance of each feature. 4 features are significant with at least one label, and life-expectancy is significant for both. Again some of the signs are not consistent with our knowledge of the features, such as income-distribution.

Significantly frequent: life-expectancy, proportion-women-parliament, freedom, secondary-education-enrolment

9.5 Assessing Significant Features

The following subset has been found to be predictive of LS using either decision trees or lasso for at least one label:

{life-expectancy, mortality-rate, proportion-women-parliament, freedom, secondary-education-enrolment, income-distribution, primary-education-enrolment}

	GAL		WVS	
Feature	Frequency	Signif	Frequency	Signif
life-expectancy	100	✓	86	✓
child-immunization	23	✗	10	✗
mortality-rate	6	✗	1	✗
gender-labour-ratio	22	✗	1	✗
proportion-women-parliament	81	✓	35	-
income-distribution	44	-	76	-
employment	2	✗	32	-
population-growth	6	✗	38	-
hospital-beds	3	✗	10	✗
percent-largest-religion	40	-	4	✗
religious-conflict	0	✗	15	
freedom	66	-	88	✓
temp-mean	1	✗	0	✗
temp-min	17	✗	13	✗
temp-max	3	✗	4	✗
primary-education-enrolment	6	✗	11	✗
secondary-education-enrolment	88	✓	71	-
adult-literacy-rate	0	✗	16	✗
military-expenditure	19	✗	7	✗
education-gender	34	-	36	-
$t_{0.05}$	31		31	
$t_{0.95}$	87		85	

Table 17: Lasso results

We compare this subset with GDP-per-capita as done previously using 10 fold cross validation for lasso and decision trees. This test is comparing a single feature with a subset of features, hence the latter has a higher VC dimension. This means is can fit the data more closely during training. The use of 10 fold CV is highly important here where the correlation is calculated on new test data such that there is no advantage given to the larger feature set. If this feature set overfits the training data it will perform badly on the test data. 10 sets of tests were performed to increase the reliability of the results (the folds are randomly chosen each time).

		Lasso		DT	
Test #	Test	GAL Mean	WVS Mean	GAL Mean	WVS Mean
1	GDP-per-capita only	0.6633	0.5337	0.7584	0.4732
2	log GDP-per-capita only	0.7764	0.5927	0.7735	0.5691
3	log GDP-per-capita & GDP-growth	0.7750	0.5907	0.7647	0.5302
4	subset*	0.8553	0.7786	0.8612	0.7456
	t-test p-val (1 & 4)	1.7575×10^{-20}	6.4391×10^{-9}	4.5530×10^{-09}	4.8387×10^{-11}
	t-test p-val (2 & 4)	1.7760×10^{-07}	1.2577×10^{-5}	7.9029×10^{-10}	3.3353×10^{-07}
	t-test p-val (3 & 4)	1.4042×10^{-07}	8.6039×10^{-7}	2.1372×10^{-10}	2.2592×10^{-09}

Table 18: Lasso correlations & significance

Table 18 shows the mean correlations and the p-values of t-tests comparing the significant subset(4) with variations of economic variables. Our feature subset gave significantly higher correlation than all economic variable sets tested. This was the case for both labels and for both lasso and DTs.

> The feature set {life-expectancy, mortality-rate, proportion-women-parliament, freedom, secondary-education-enrolment, income-distribution, primary-education-enrolment} can predict LS significantly better that economic variables using Lasso and DTs

Comparing the lasso correlations with the previous results of table 12, they are marginally reduced. However, the difference is small and not significant, with p = 0.4451 and 0.2456 for GAL and WVS respectively. The first feature set contained 17 variables but the one used now has just 7, which shows some variables used initially contribute a small degree to the predictive ability of a LS model.

The feature set {life-expectancy, mortality-rate, proportion-women-parliament, freedom, secondary-education-enrolment, income-distribution, primary-education-enrolment} can predict LS as well as the 17 original features

The GAL and WVS models are highly consistent. The most frequent model includes the same features for both labels, and each coefficient has the same sign. There are also similarities with regard to the coefficient magnitude, such as having large coefficients for life-expectancy and income-distribution for both labels, although the life-expectancy coefficient seems quite a bit larger for GAL. This may be due to the sample where GAL has many more poorer countries with low health care which may make it more sensitive to changes of life expectancy. Primary-education-enrolment has a negative coefficient and this is quite unexpected, especially as secondary-education-enrolment uses a positive coefficient.

9.6 Frequent Subsets

Lasso was run again using just the 7 key features and GAL label in order to find frequent subsets. There are 7! = 5040 possible subsets. However over 100 bootstraps lasso used just 17. The significance threshold for subset vote is 13 at a 5% level, and table 19 show the significant subsets found.

5 features are present in both frequent subsets. Mortality-rate is not included in either and additionally it's coefficients are not stable as the sign varies in the results, whereas all other variables are only ever included with the same sign. The votes show that the first subset has a much lower p-value, whereas $p \sim 0.05$ for the second (because the vote equals the threshold). The only difference is the inclusion of primary-education-enrolment and this indicates this variable should be included in the model.

	life-expectancy	mortality-rate	proportion-women-parl	income-distribution	freedom	primary-education-enrol	secondary-education-enrol	vote
subset 1	1	0	1	1	1	1	1	43
subset 2	1	0	1	1	1	0	1	13

Table 19: Lasso frequent subsets

Significant features subsets:
{life-expectancy, proportion-women-parliament, income-distribution, freedom, primary-education-enrolment, secondary-education-enrolment}

{life-expectancy, proportion-women-parliament, income-distribution, freedom, secondary-education-enrolment}

10 Best Prediction

Linear and nonlinear methods are used to find the optimal prediction of LS. We test with both the whole feature set (of stage 2) and our key feature set to provide a comparison. The GAL LS label is used as it provides better coverage, and again 10 fold cross validation is used to prevent overfitting.

10.1 Optimising SVM

SVM's have several parameters that must be optimised for each dataset (See section 7 for more details). A grid search was used to optimise the loss and cost values. Two kernels were tried; RBF and poly. RBF is a gaussian kernel and hence the gamma parameter was optimised also. This was done after the grid search using the optimal loss and cost values, trying gamma values at 0.5 intervals. This gives a good but non optimal result as the gamma is not optimised in conjunction with the other parameters. However the optimal correlation was very near to that found with the original gamma value (that used in the grid search).

10.2 Results

Table 20 shows the p-values of t-tests comparing the results of 10 times 10 fold cross validation. This was done using linear and nonlinear learners and also using the significant features and whole feature set to provide comparisons. These show a significantly better correlation for the key feature set, compared with the whole feature set for both the model tree and SVM. The lasso predictions did not significantly improve using only the key features, which indicates lasso is effective at feature selection.

Using key features significantly improved the best prediction found for model trees and SVMs

		Key Features				All Features			
		Model tree	RBF	Poly	Lasso	Model tree	RBF	Poly	Lasso
Key	Model tree	NaN	0.0197	0.0027	0.5150				
	RBF	0.0197	NaN	0.0152	0.9996				
	Poly	0.0027	0.0152	NaN	0.7917				
	Lasso	0.5150	0.9996	0.7917	NaN				
All	modeltree	0.0000	0.0000	0.0000	0.0000	NaN	0.1115	0.2627	0.0000
	RBF	0.0000	0.0000	0.0000	0.0000	0.1115	NaN	0.0006	0.0001
	poly	0.0000	0.0000	0.0000	0.0000	0.2627	0.0006	NaN	0.0000
	lasso	0.0476	0.2207	0.3852	0.2136	0.0000	0.0001	0.0000	NaN
mean correlation		0.8605	0.8503	0.8458	0.8504	0.7519	0.7648	0.7420	0.8326

Table 20: Prediction results comparison: p-values

Best mean correlation: 0.86 (model tree)

The best correlation was found for model trees with a correlation of 0.86. However, this was not significantly better than lasso which indicates a highly linear relationship such that LS can be represented using a linear model.

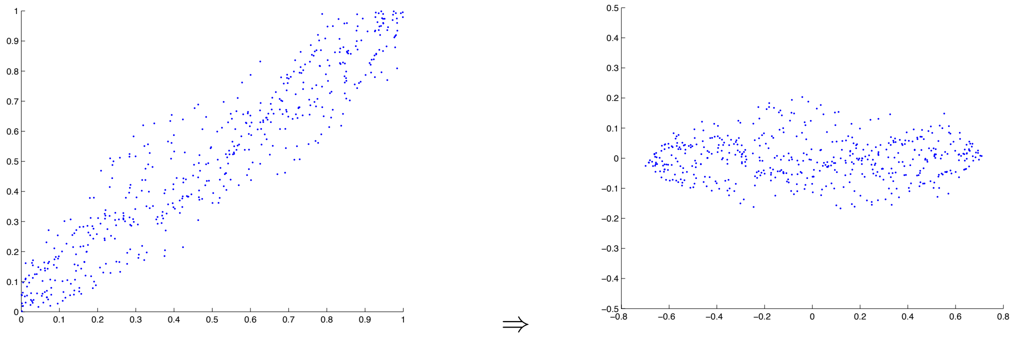
The relationship between the features and LS is highly linear

Using the large feature set gave significantly worse performance. For instance, a t-test comparing the two feature sets for model trees gave a p-value of 2.1212×10^{-16}. Decision trees naturally perform features selection, but can still be affected by irrelevant attributes as lower down the tree the number of instances at each node reduces and the chance of choosing an irrelevant attribute to split the data increases. Also, model trees use a linear model at the leaves which irrelevant attributes can affect. Lasso appears much more robust to irrelevant features, with no significant difference in performance between the two feature sets.

Lasso is robust to irrelevant features

10.3 DT & PCA

PCA can be an effective transformation method when using decision trees because the variance of the principal components are orthogonal to the axes (because the axes are by definition the direction of highest variance in the data). Decision trees split the data into 'boxes' with boundaries that are orthogonal to the axes and therefore principal components may be more easily split during tree construction.



(a) Original data

(c) Principal components

Figure 29: PCA demonstration

10.3.1 Results

		Model			Regression		
		Key features	PCs	Both	Key features	PCs	Both
Model	Key	-	5.63×10^{-8}	9.85×10^{-2}	6.98×10^{-5}	6.35×10^{-7}	2.77×10^{-5}
	PCs		-	2.78×10^{-7}	8.52×10^{-1}	5.75×10^{-3}	3.4428×10^{-1}
	Both			-	5.47×10^{-4}	3.25×10^{-7}	1.44×10^{-4}
Regression	Key				-	1.86×10^{-2}	2.33×10^{-1}
	PCs					-	5.04×10^{-2}
	Both						-
		0.8574	0.8179	0.8521	0.8157	0.7771	0.8057

Table 21: PCA decision tree results (10 x 10 fold Correlation)

The key features performed the same as using both principal components and key features alone. The results of table 21 show that using PCs alone gave significantly worse results. Using only key features gave the highest mean correlation across the folds. This is not significantly better (p = 0.0986) than using the combined features, which is expected due to the feature selection capabilities of trees (where if the PCs do not split the data well they can simply be ignored). However, the p-value is significant with a 1% level and this again supports the results of section 10.2 where using only key features improved results.

The correlations of PCs against key features shown in table 22 indicate from which features the principal components are composed. PC1 is clearly representing health and education, where

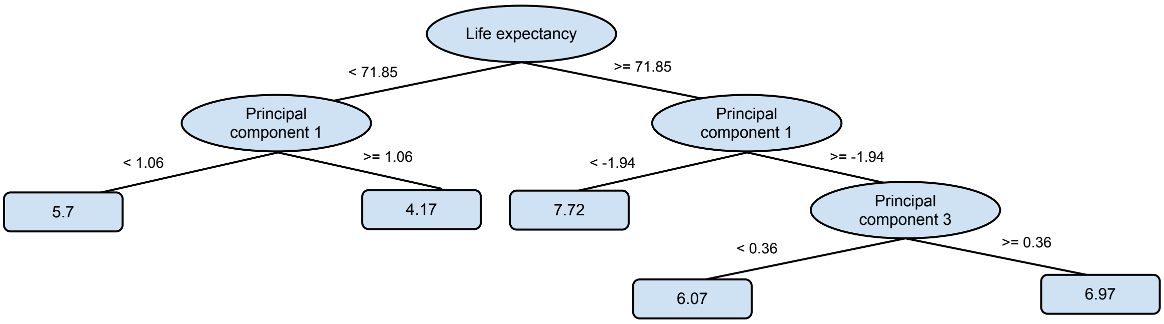


Figure 30: Decision tree with principal components & key features

		life-exp	mort-rate	prop-wom-parl	income-dist	sec-edu-enrol	freedom	prim-edu-enrol	GAL
Corr	PC1	-0.9473	0.9274	-0.2033	0.4433	-0.9309	0.6402	0.0926	-0.8047
	PC2	-0.1111	0.1715	0.9109	0.1981	-0.0923	-0.3568	0.2116	0.0761
	PC3	0.0683	-0.1075	-0.2728	0.8163	0.1023	-0.4013	0.2815	0.1702
	PC4	0.1653	-0.2826	0.2446	0.2604	0.1361	0.5229	0.2522	0.1971
	PC5	-0.0515	0.0384	0.1505	0.3900	-0.0675	0.0156	1.0000	-0.0694
P-val	PC1	0.0000	0.0000	0.0324	0.0000	0.0000	0.0000	0.3336	0.0000
	PC2	0.2459	0.0719	0.0000	0.0372	0.3351	0.0001	0.0258	0.4270
	PC3	0.4765	0.2614	0.0038	0.0000	0.2855	0.0000	0.0028	0.0741
	PC4	0.0830	0.0027	0.0097	0.0058	0.1544	0.0000	0.0076	0.0381
	PC5	0.5916	0.6893	0.1148	0.0000	0.4815	0.8711	0.0000	0.4691

Table 22: Correlations of key features against principal components

high values correspond to lower LS, lower life-expectancy, higher mortality-rate, lower secondary-education-enrolment, better income equality, and lower freedom. The decision tree infers countries with lower PC1 values have a higher LS. Here income-distribution is related negatively with LS such that a greater income equality relates with higher LS which differs from our earlier results.

The decision tree also uses PC3. This correlates with income-distribution, proportion-women-parliament, freedom and primary-education-enrolment. A high PC3 value corresponds with greater income inequality, lower proportion-women-parliament, higher freedom, and higher primary-education-enrolment. Income-distribution has the same sign for PC3 and PC1 but these PCs have opposite relationships with LS. PC3 is used to split those countries with the highest living standards (according to PC1 and life expectancy), where the lower PC3 values are related to lower LS for these countries. Therefore, for these countries the decision tree is inferring that more equal income-distribution relates with lower LS values. This indicates that the relationship between income-distribution and LS appears to differ when LS reaches a certain level. Income *equality* may be important for LS in poorer countries to improve the standards of lives. However once quality of life reaches a certain level, income *inequality* may relate positively with LS to provide opportunities, aspirations and social mobility.

11 Graphical Models

Bayesian networks are described in section 3.3.8. Our aim is to use BBNs to assess the relationships between the key features we have identified.

11.1 Discretizing

Bayesian networks require nominal data and so the data is converted from the original numeric features. Discretizing using equal width caused some bins to hold very few instances, which may cause more unreliable results because it is easily affected by each of these instances. Therefore equal frequency was used to give approximately the same number of instances in each bin in the network. The bins are slightly different sizes because the number of instances cannot be divided exactly into 4 bins. Some of the variables have only a small number of values such as freedom where the values are between 0 and 7 in 0.5 intervals. This means that the number of instances in the bins is more variable, because of the groups of intervals with identical freedom values.

The number of bins was chosen by experimentation, where the class distribution of the bins was visually compared. A balance is needed between simplicity and loss of information. A small number of bins is preferred such that the degree of freedom of the network is kept low. However, there needs to be a sufficient number of bins to capture relationships between the variables, as reducing the number of bins reduces the information content of the network. We found 4 bins to be appropriate (named VLOW, LOW, HIGH, VHIGH respectively).

11.2 Network Structure

There are several methods for automatically generating networks, but these are not able to infer causality and therefore the network structure is built manually. The probabilities are then calculated using estimates of the probabilities (see section 3.3.8).

The networks use the key features, and additionally GDP-per-capita is included. Figure 31(a) shows a basic structure, which we are confident represents the variable relationships. This structure represents the notion that GDP affects the levels of other variables and these in turn impact on LS. We investigate some alterations to this basic network, shown in figures 31(b), 31(c) and 31(d). These have one or two additional edges connecting GDP-per-capita to proportion-women-parliament, freedom and also to LS directly.

11.2.1 Metrics used for assessment

We use % correct, ROC curve area and degrees of freedom to assess the networks, to look at the performance of each structure.

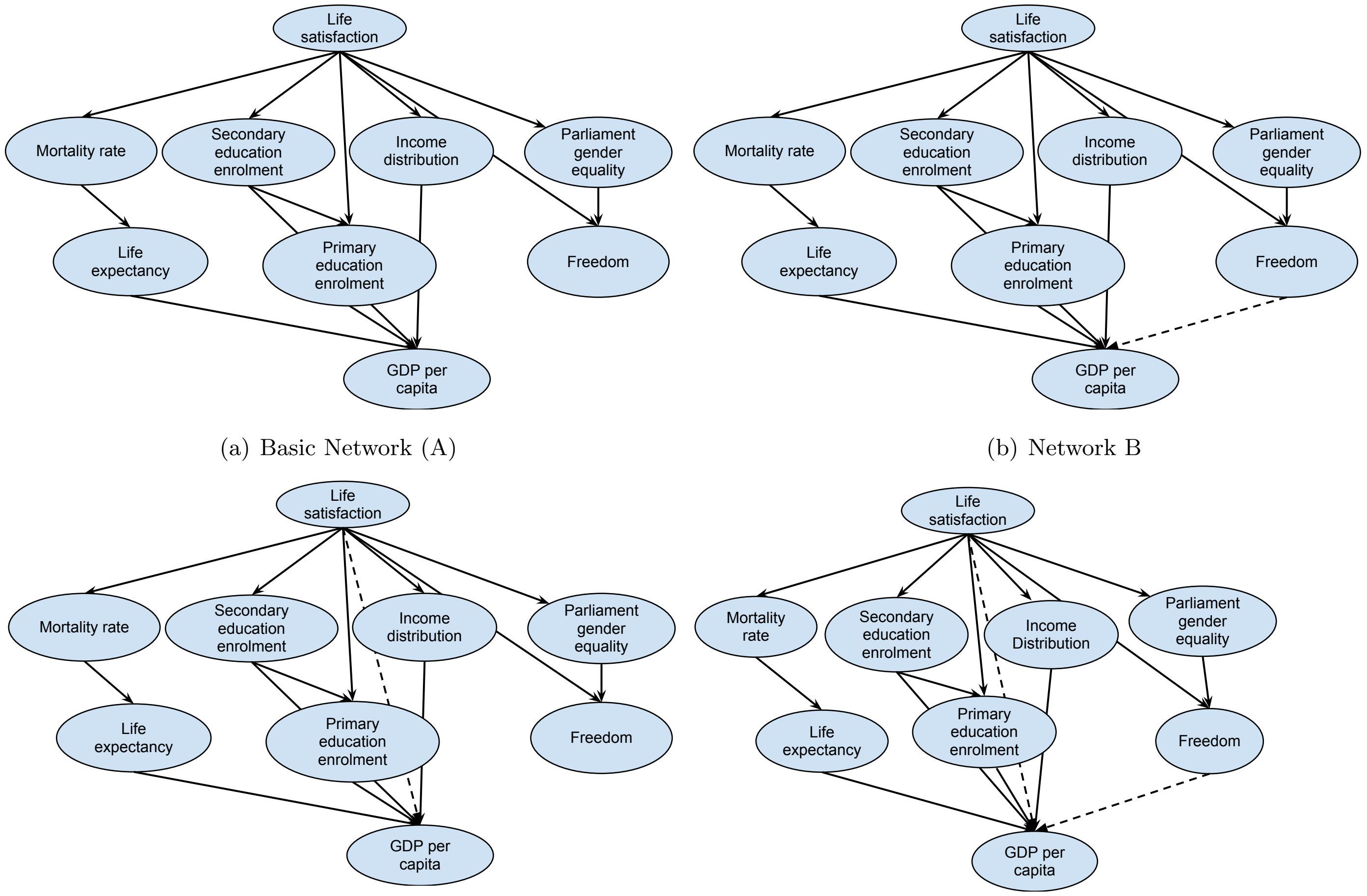
Degrees of Freedom The degrees of freedom (DF) are the number of parameters that can change in a model. This is an important consideration as models with higher DF are more easily able to fit to a pattern.

The DF of a network is the number of probabilities in the probability tables of each node. These are probabilities of each value of this node conditional on each combination of parent node values. For instance, one probability of node mortality-rate is $p(mortality - rate = VLOW \mid LS = VLOW)$. The number of values in a network is given by:

Structure (fig 31)	# edges	Degrees of Freedom
A	13	1236
B	14	4308
C	14	4308
D	15	16596

Table 23: BBN details

$$V = \sum_{i=1}^{n}(K_i \cdot (\prod_{X_j \epsilon parent(X_i)} K_j)), \tag{49}$$



(a) Basic Network (A)

(b) Network B

(c) Network C

(d) Network D

Figure 31: Network structures

where K_i is the number of discretised values of node (feature) i. Therefore the number of variables at each node grows exponentially in the number of edges leading to that node.[23]

Receiver Operating Characteristic (ROC) Curves A ROC curve is a graph of false positive rate against true positive rate. The area underneath (AUC) indicates how well the test instances are classified. This combines both the rate the class is correctly identified (true positive) and the extent that instances are classified as this class incorrectly (false positive). With respect to this work, a ROC curve is generated for each band of LS, and figure 32 shows the ROC curve for VLOW LS and the network in figure 31(a).

Specifically, for a ROC curve of class c and given two random instances i_1 and i_2 where $c(i_1) = c$ and $c(i_2) = \neg c$, having $p_1 = p(c \mid i_1)$ and $p_2 = p(c \mid i_2)$. The AUC corresponds to the probability that $p_1 > p_2$.[21] An instance that is misclassified but only just such as if $p(c \mid i_1)$ is only slightly smaller than $p(\neg c \mid i_1)$, will reduce the area, but by less than if the difference in these probabilities is much greater (given that the points of other classes lie between these values). In this sense this value can be seen to quantify the distribution of values in a confusion matrix into a single value.

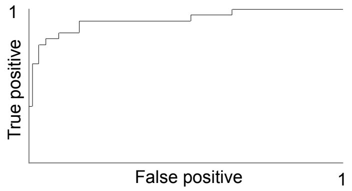


Figure 32: ROC curve (basic structure low LS)

Each ROC curve corresponds to a class value (discretised LS value), and here we use the average AUC (across LS values) as a measure of how well LS is predicted by each network. A value of 1 indicates no false positives and 100% true positives (and 0.5 is as good as the random case). Hence this metric is a useful measure as it indicates class prediction relative to other

[23] For our network all nodes have 4 values. None: $4, 1 : 4^2 = 16, 2 : 4^3 = 64, 3 : 4^4 = 256, 4 : 4^5 = 1024, 5 : 4^6 = 4096, 6 : 4^7 = 16384$

instances, rather than the actual prediction values. [21] Instances ranked in the correct order (according to the ratio of $p(c \mid i)$ and $p(\neg c \mid i)$ such that $\frac{p(c|i)}{p(\neg c|i)}$ is always larger for instances of class c) will give an optimal AUC, but in this case it is still feasible that the classifications are incorrect. For instance, if no instances are classified as $VLOW$ but also if the $VLOW$ instances are ranked higher such that they are more likely to belong to this class than the other instances, then the ROC curve will be optimal. As such we combine use of this measure with others such as the % correct to provide an overview of the effectiveness of each network.

In addition, weighted ROC area is assessed. This takes into account the class distribution of the test instances of each fold[24], and this gives more importance to the results of more frequent classes.

11.2.2 Results

Table 24 shows the affect of structural changes to the original network (figure 31(a)) which are shown in figures 31(b),31(c) and 31(d). These results compare 10 sets of 10-fold cross-validation tests (100 result values per network)[25]. Firstly, the proportion correct is surprisingly low (approximately 58%). This may be because the numeric data is discretised which loses much of the information is contains. Confusion matrices shows the magnitude of the errors, such as that of graph A (table 25) where mistakes of smaller magnitude are more common. For instance, no countries with an extreme LS value (VLOW or VHIGH) were classified as the opposite extreme value.

		P-value				H-value				Mean
		A	B	C	D	A	B	C	D	
% correct	A	-	Same	0.0014	0.9863	-	0	1	0	58.4817
	B	Same	-	0.0014	0.9863	0	-	1	0	58.4817
	C	0.0014	0.0014	-	0.0003	1	1	-	1	56.8433
	D	0.9863	0.9863	0.0003	-	0	0	1	-	58.4740
ROC Area	A	-	0.3197	0.0009	0.1939	-	0	1	0	0.7047
	B	0.3197	-	0.0008	0.1760	0	-	1	0	0.7049
	C	0.0009	0.0008	-	0.0165	1	1	-	1	0.6895
	D	0.1939	0.1760	0.0165	-	0	0	1	-	0.7007
Weighted ROC Area	A	-	0.3197	0.1562	0.0395	-	0	0	1	0.7966
	B	0.3197	-	0.1733	0.0341	0	-	0	1	0.7965
	C	0.1562	0.1733	-	0.0050	0	0	-	1	0.7938
	D	0.0395	0.0341	0.0050	-	1	1	1	-	0.7993

Table 24: BBN t-test results comparison of networks of figure 31

The t-test result sets for the 3 measured values are first considered independently. The % correct was highest for graph A, although this was not significant. In fact A and B showed identical results and no significant difference was found between A,B and D. Graph B gave the largest ROC area, marginally larger than A. Again, no significant difference between A,B and D. The weighted ROC area results did however find a significant improvement, for graph D when compared to all other graphs.

	(-inf-4.95]	(4.95-5.85]	(5.85-7.05]	(7.05-inf)
(-inf-4.95]	22	3	1	0
(4.95-5.85]	4	14	6	3
(5.85-7.05]	1	5	14	8
(7.05-inf)	0	3	7	20
ROC area	0.936	0.686	0.721	0.827

Table 25: Confusion matrix of graph A

Therefore, all three tests showed graph C performs significantly worse than the other network, with respect to the % correct and weighted and unweighted ROC curve areas. It also has a larger

[24] A weighted average of ROC areas of each class, uses the confusion matrix to find the weights, see class method weka.classifiers.Evaluation.weightedAreaUnderROC()

[25] The division of instance to fold is random

degree of freedom than graph A and hence graph A is preferred for this reason also. With respect to graphs A,B and D, only weighted ROC curve area gives a significant difference, and this is in favour of graph D. However with $p = 0.0395$ for graph A versus graph D this is significant at a 5% level but not at a 1% significance level. These results are less than definitive, and it is only when considering model complexity that we prefer one graph over the others. Graph D has 16596 degrees of freedom, which is much larger compared to 1236 of graph A. This means that D is much more prone to overfitting as there are many more tunable parameters in the model. This reason leads us to believe it is reasonable to accept graph A as an appropriate graph compared to the three variations considered. Also, according to Occam's razor when two models perform equally well then the simpler one should be preferred.

A degree of improvement was expected when connecting GDP-per-capita directly with LS. There are several possible justifications for this representation, firstly that this does indeed show that GDP-per-capita has a direct impact on LS. Alternatively, this may have occurred if there are other relationships not captured in our network involving variables we have not considered. The feature set considered is small compared with the potentially relevant features which may impact on LS, which is likely to be a large number of variables. Therefore, these results are in line with our expectations.

Connecting GDP-per-capita directly with LS shows no performance improvement of the network

12 Data Visualisation

Chernoff faces (CF) are a method for visualisation of multivariate data.[12] Each facial feature corresponds to a feature in the data set and changes shape or size according to the feature value (relative to the values of the other data points). Facial expression is closely associated with happiness and hence LS, and as such it is an ideal representation. Satisfied countries can be easily identified using facial expression as well as differences between particular facial features, and hence trends and anomalies are easy to identify. Two star visualisations have been generated for a comparison (figure 33(c) and 33(d)) and these are clearly much less intuitive.

12.1 Label Value Inference

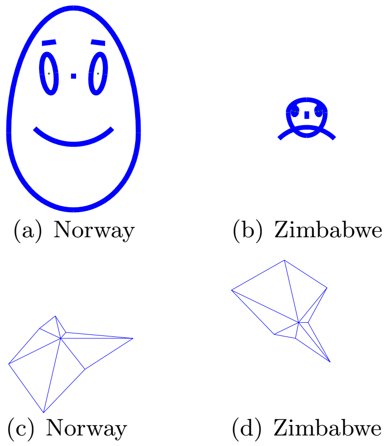
To maximise the data available, we infer GAL values from the WVS where the GAL values are missing, by regressing WVS against GAL. The resultant model is $GAL = 0.9525\, WVS + 0.1938$ with $p = 1.4054 \times 10^{-10}$ and a correlation of 0.6287. Therefore while the overall correlation is good there are some differences which may affect the country order for the two labels.

12.2 Face Generation

The faces for each country are shown in figure 27 and these results are ordered by LS such that the trends can be identified. Table 26 shows the key for the face features, chosen such that high LS relates to happier looking faces. For example, eyebrow angle is used to represent gender equality and hence proportion-women-parliament was negated such that angry eyes relates to poor equality.

Face feature	Data feature	Visual meaning
Face size	life-expectancy	large = high
Length of nose	mortality-rate	large = high
Eyebrow angle	proportion-women-parliament	unhappy = unequal
Eye: dist between	income-distribution	close = equal
Forehead shape	freedom	round = more free
Shape of jaw	primary-education-enrolment	round = high
Smile	secondary-education-enrolment	smile = high

Table 26: Chernoff face key


(a) Norway

(b) Zimbabwe

(c) Norway

(d) Zimbabwe

Figure 33: Chernoff and star examples

The similarities and differences are easily recognisable, now the features are represented visually. For example, two countries having very low and high LS values are Zimbabwe and Norway respectively, shown in figures 33(a) and 33(b). These show very different values for each feature. However we can see that primary education rate is very similar (shape of jaw line). Even though Zimbabwe's primary enrolment is high, the secondary enrolment is very low (sad smile). This highlights our findings that secondary education is more predictive of LS than primary education.

Health was shown to significantly predict LS and our CFs show this with both increasing face size (life expectancy) and decreasing nose size (mortality rate).

	life-exp	mort-rate	prop-women-parl	income-dist	sec-educ-enrol	freedom	prim-educ-enrol
Norway	80.74	3.50	36.10	25.79	112.38	1.0	98.25
Zimbabwe	44.21	93.40	15.20	50.10	40.99	6.5	103.60

Figure 34: Feature values for Zimbabwe and Norway

12.3 Interactive Results

A website[26] was built to provide an interactive visualisation of the results. The core technologies used are: Google maps, Google visualisations, HTML, JavaScript and JSON. The Chernoff faces were overlayed onto a map such that the LS and feature values can be easily analysed by global location. The standard map pointers are replaced with faces, which when clicked display charts of the feature values for this country next to the map (see figure 36). There is also a separate page where the correlations between features can be viewed where on mouse over the country details are displayed.

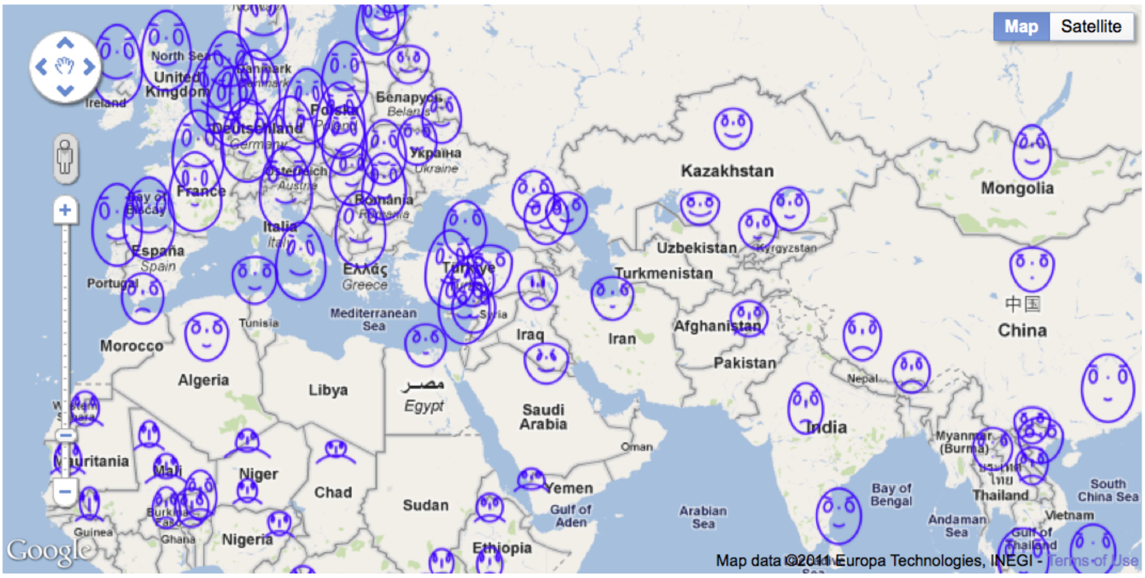


Figure 35: Website interactive happiness map

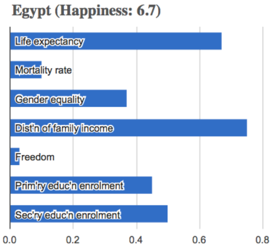


Figure 36: Example feature chart

[26] http://intelligentsystems.bristol.ac.uk/research/Millard/

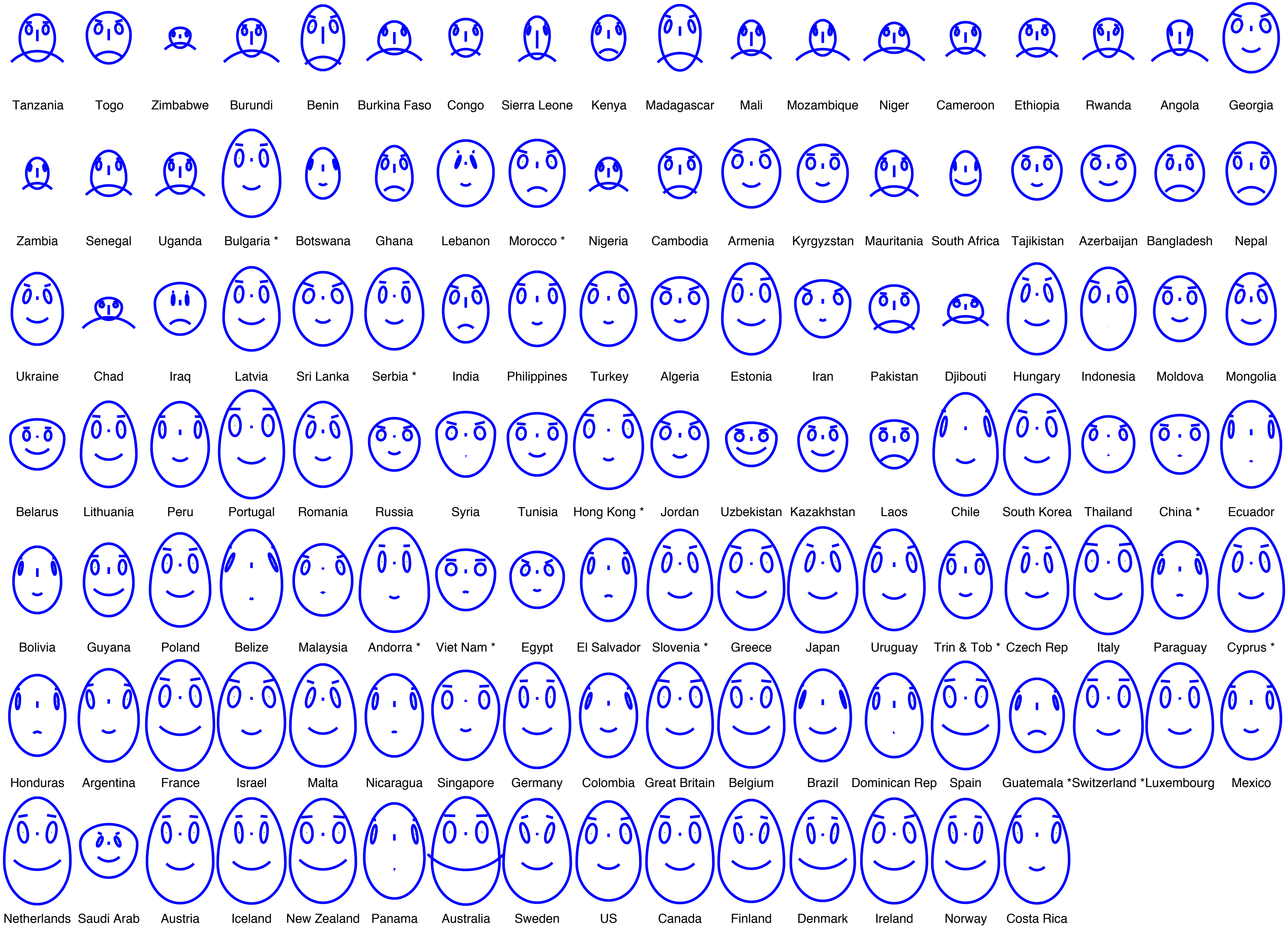


Table 27: Chernoff faces ordered by LS (* GAL value inferred from WVS)

13 Project Conclusions

Drawing conclusions It is important to first note the extent to which conclusions can be drawn from our results. We have found key features and models to represent and predict LS. However, justifying these results with common sense reasoning has limited worth as demonstrated by Paul Lazarsfield, a mathematician and Professor of Sociology. [2] Lazarsfield switched findings of military research looking at the stress of soldiers. The research found that education was positively related to the stress a soldier experiences, which could be justified by environmental differences such that farmers are more used to the tough conditions. However when switching the result, he showed that this could just as easily be justified by another reason, that education provides skills to cope with stress. This shows how we should only be guided by results rather than our intuition and common sense. [32]

13.1 Results Summary

13.1.1 Feature selection & construction

Variables were constructed such that they were most representative of the intended concept. For instance, latitude was used to represent light, and this was weighted by population to ensure it was representative of where people actually live.

Military expenditure was initially found to be significant. However, this variable was the amount relative to GDP per capita. Confounding was found to be the underlying reason for it's inclusion. This was determined by altering the variable to military expenditure per health expenditure, after which it was no longer found to be a significant feature (see section 9.1).

We found much sensitivity to choice of variable to represent a concept. For example, several education indicators were considered; primary enrolment, secondary enrolment, literacy rate and pupil teacher ratio. Literacy rate and pupil teacher ratio were not significant features, but the enrolment variables (particularly secondary) were prominent in the results.

Visualising the data indicated some variables have an exponential relationship with LS. Transforming these with logs was found to be beneficial, where these variables were often selected instead of the original (section 6.2).

13.1.2 Key features & feature subsets

The key features identified are:

- Health: life-expectancy, mortality-rate
- Education: primary-education-enrolment, secondary-education-enrolment
- Equality: proportion-women-parliament*, income-distribution
- Freedom: freedom index

* Additionally, PCA found an interesting relationship between gender equality and both LS and GDP per capita (section 5.3).

Although mortality rate was found to be significant as an individual feature, it was not present in the significant feature subsets. The following frequent subsets were found using lasso tests:

{life-expectancy, proportion-women-parliament, income-distribution, freedom, primary-education-enrolment, secondary-education-enrolment}

{life-expectancy, proportion-women-parliament, income-distribution, freedom, secondary-education-enrolment}

13.1.3 Key Models

Tests have shown that our models using the key features are significantly more predictive of LS than economic variables alone (section 7 and section 9.5). The best correlation used was 0.86 for model trees. There was no significant different between the results of linear and nonlinear methods, indicating the relationship of our feature set with LS is highly linear (section 10). A regression tree for the GAL label is shown in figure 37.

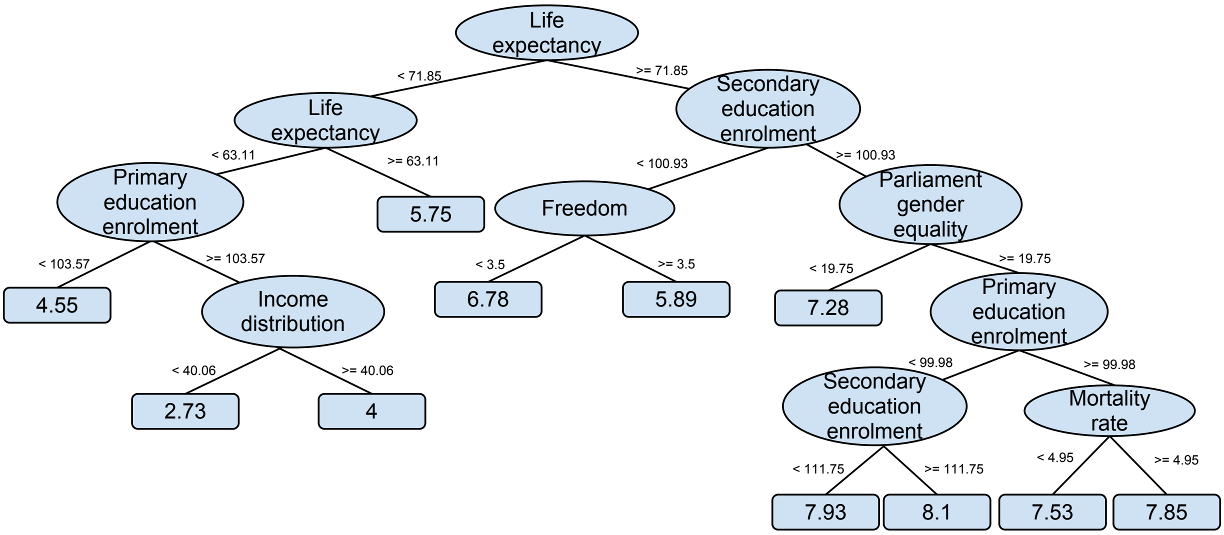


Figure 37: Decision tree: Key features & GAL label (not standardised)

The models generated using lasso and least squares consist of coefficient values that were often against intuition. For instance, primary education enrolment has a small negative coefficient. This is to be expected as only using a small number of features are used and thus this is only a 'good' but less than perfect model of life satisfaction. Therefore some compromises are built into the generated models to account for this as best as possible.

13.1.4 Graphical Models

Tests with Bayes networks investigated slight variations and found no significant improvement in prediction when connecting GDP per capita directly with life satisfaction. Additionally, the performance was poor with a mean prediction of 59 %. This is likely to be due to discretisation where to keep the degrees of freedom low a small number of 'buckets' was used, which causes much of the information in the data to be lost.

13.1.5 Consistency of results & data quality

Tests results for WVS and GAL found some similar results but also some differences. For instance, models for WVS preferred freedom whereas proportion-women-parliament was used for GAL. Life expectancy however was a consistently significant feature throughout this work for both labels.

The availability of a suitable ground truth is a constraint of this project, where there are differences between the sources that are not ideal (see section 4). In brief, WVS is too small and missing many countries with low LS, whilst GAL is larger but possibly skewed by question positioning. Additionally, the variables used as features included some missing values and these were inferred. This involved both using alternative sources and also imputation using the k-nearest neighbour method. Imputation (section 6.1) did not give good results and was used only where reliable alternative sources could not be found. The reliability of alternative sources is a concern, as small differences between the sources can affect the data. Therefore, future work in this area will improve with better data availability and coverage.

13.1.6 Results visualisations

Chernoff faces provide an effective visualisation of our key feature set. They have shown to be an excellent method to visualise multivariate data where patterns and anomalies can be easily identified. Facial expression was a particularly appropriate and intuitive representation for this work, corresponding to life satisfaction.

13.2 Assessing progress

The main objective was to predict happiness using a feature set excluding economic variables. Results surpassed expectations, having identified a feature set significantly *more* predictive than economic variables. Structural tests that identified no direct relationship between GDP and life satisfaction further support the notion proposed: "Economic factors are not goals in themselves but a means for acquiring other types of social goals". [25]

We have proven GDP to be inadequate as a measure of progress and shown the opportunity and value of an alternative concise but highly informative variable set. GDP by definition[27] is driven by consumption, but this is not the case for our feature set. Therefore, a measure of progress incorporating these could help lead us to a happier, more sustainable and planet friendly future.

13.3 Areas of further work

Our work found an interesting relationship between gender equality and GDP / life satisfaction and this needs further investigation. The two variables represent gender equality only (labour participation and equality in parliament), and hence more equality variables needs investigating to look at other types of equality.

An interesting area of future work would be to learn using more structured data to allow information that cannot easily be expressed in typical attribute value learners. For instance, Tilde is an ILP learning system able to learn decision trees (including regression) using first order logic. This may be of much use to investigate relationships and structure in the world such as political allies and country neighbours. Our results indicate close relationships between countries, such as the PCA graph (figure 14) where nearby countries have similar values of some variables. It would be interesting to investigate the extent to which a country's LS is determined by their own actions or heavily constrained by other properties such as the countries they border.

Finally, the relationship between income distribution and LS is an open question, and the uncertainty behind this is highlighted by our conflicting results regarding this. The constraining factor regarding research into this is the lack of available data.

[27] Gross Domestic Product (GDP) is the value of the goods and services produced by a country in a year. Formally, GDP is: The sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. [1]

14 Appendix A: Data Sources

	Source	Year	# respondents	# countries	Age	Question & answer	Answer scale	Standardisation
World Value Survey(WVS)	Primary	2008	Mostly > 1000	55	16+, 18+	All things considered, how satisfied are you with your life as a whole these days? (V22) 1 (Dissatisfied), 2, 3, 4, 5, 6, 7, 8, 9, 10 (Satisfied), -1 Dont know, -2 No answer, -3 Not applicable, -4 Not asked in survey, -5 Missing; Unknown	1 - 10	**Rescaled from 1-10 to 0-10
Happy Planet Index(HPI)	Secondary (Gallup)	2006	Mostly > 1000	143(112 Gallup life sat'n)	15+	All things considered, how satisfied are you with your life as a whole these days 0 to 10, where 0 is dissatisfied and 10 is satisfied	0 - 10	
OECD (Positive experience index)	Secondary (Gallup)	2006 -08	Mostly > 1000	35	15+	Positive experience index (combines 6 questions from Gallup; well rested, treated with respect, chose how time was spent, proud of something you did, learnt or did something interesting, enjoyment ladder- of-life questions, which ask respondents to rate their life from the worst (0) to the best (10) level, and refer to the share of people who rate their life (today and in the future) at step 7 or higher.	0 - 10 (but given as %)	Rescaled - divide by 10
OECD (Ladder of life)	Secondary (Gallup)	2006 -08	Mostly > 1000	35	15+	Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?	0 - 10	
European Social Survey (ESS)	Primary	2008	1500 or 800 if population < 2 million	26	15+	Taking all things together, how happy would you say you are? Extremely unhappy 00, 01 02 03 04 05 06 07 08 09, Extremely happy 10, (Dont know) 88	0 - 10	
Eurobarometer	Primary	2008	1000	31	15+	On the whole, are you very satisfied, fairly satisfied, not very satisfied or not at all satisfied with the life you lead? Very satisfied, Fairly satisfied, Not very satisfied, Not at all satisfied, DK	% satisfied	Rescale - divide by 10
European Value Survey(EVS)	Primary	2008	1500	48	18+	Variable v66: how satisfied are you with your life -5 (other missing), -4 (question not asked), -3 (nap), -2 (na), -1 (dk), 1 (dissatisfied), 2, 3, 4, 5, 6, 7, 8, 9, 10 (satisfied)	1 10 (-5 - -1 invalid)	**Rescaled from 1-10 to 0-10
Latinobarometre	Primary	2007	1000 - 1200	18	18 (16 Brazil & Nicaragua)	Q1ST.A In general, would you say you are satisfied with your life? Would you say you are....? 1 (Very satisfied), 2 (Fairly satisfied), 3 (not very satisfied), 4 (Not satisfied at all), 0 NK	1 4 (reversed*)	***Rescaled and reversed

Table 28: Surveys Descriptions: Surveys considered as happiness label (**rescale(x) = (x-1)*10/9 ***rescale(x) = 10 - (x-1)*10/3)

Id	Type	Code	Year	Feature	Source	# missing values	URL
EC1	ECONOMIC	EC1-GDPGROWTH-2008	2008	GDP growth (annual %)	World Bank	2	http://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG
EC3	ECONOMIC	EC3-EMPLOY-RATIO-2008	2008	Employment to population ratio, 15+, total (%)	World Bank	2	http://data.worldbank.org/indicator/SL.EMP.TOTL.SP.ZS
EC4	ECONOMIC	EC4-GDP-PER-CAPITA-2008	2008	GDP per capita (current US$)	World Bank	2	http://data.worldbank.org/indicator/NY.GDP.PCAP.CD
HE1	HEALTH	HE1-LIFEEXP-2008	2008	Life expectancy at birth, total (years)	World Bank	1	http://data.worldbank.org/indicator/SP.DYN.LE00.IN
HE2	HEALTH	HE2-IMMUN-08	2008	Immunization, DPT (% of children ages 12-23 months)	World Bank	1	http://data.worldbank.org/indicator/SH.IMM.IDPT
HE3	HEALTH	HE3-HEALTH-EXP-08	2008	Health expenditure per capita (current US$)	World Bank	2	http://data.worldbank.org/indicator/SH.XPD.PCAP
HE4	HEALTH	HE4-MORT-RATE-08	2008	Mortality rate, under-5 (per 1,000)	World Bank	1	http://data.worldbank.org/indicator/SH.DYN.MORT
EN1	ENVIRON	EN1-CO2EMIS-2007	2007	CO2 emissions (metric tons per capita)	World Bank	1	http://data.worldbank.org/indicator/EN.ATM.CO2E.PC
EN2	ENVIRON	EN1-MAM-THREAT-08	2008	Mammal species, threatened	World Bank	0	http://data.worldbank.org/indicator/EN.MAM.THRD.NO
EQ2	EQUALITY	EQ2-LABOR-RATIO-GEN	2008	Ratio of gender labor participation rate HIGHER=MORE EQUAL	World Bank	2	http://data.worldbank.org/indicator/SL.TLF.CACT.FE.ZS http://data.worldbank.org/indicator/SL.TLF.CACT.MA.ZS
EQ3	EQUALITY	EQ3-PARL-WOM-08	2008	Proportion of seats held by women in national parliaments (%)	World Bank	3	http://data.worldbank.org/indicator/SG.GEN.PARL.ZS
FR1	FREEDOM	FR1-TIME-START-BUS-2008	2008	Time required to start a business (days)	World Bank	2	http://data.worldbank.org/indicator/IC.REG.DURS
ED2	EDUCATION	ED2-PUP-TEA-RAT-08	2008	Pupil-teacher ratio, primary	World Bank	30	http://data.worldbank.org/indicator/SE.PRM.ENRL.TC.ZS
CL2	CLIMATE	CL2-LAT-WEIGHTED	2010	Latitude weighted average by population (indicator for light)	World Gazetier, geohive, wolfram alpha	2	
LI2	LIFE-STYLE	LI2-POP-GROW-08	2008	Population growth (annual %)	World Bank	0	http://data.worldbank.org/indicator/SP.POP.GROW
LI3	LIFE-STYLE	LI3-POP-URBAN-08	2008	Urban population (% of total)	World Bank	0	http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS
LI4	LIFE-STYLE	LI4-POP-DENS	2008	Population density (people per sq. km of land area)	World Bank	0	http://data.worldbank.org/indicator/EN.POP.DNST
CR2	CRIME	CR2-INTENT-HOMO-03-08	2003-2008	Intentional homicide, rate per 100,000 population	UN DATA	1	`http://www.unodc.org/unodc/en/data-and-analysis/homicide.html`
DI1	DISTRIB'N	DI1-FAM-INCOME	1975 - 2011	Distribution of family income - Gini index	CIA World Factbook HIGHER=MORE UNEQUAL	7	`https://www.cia.gov/library/publications/the-world-factbook/rankorder/2172rank.html`
DI2	DISTRIB'N	DI2-AGE-65-2008	2008	Age distribution	U.S. Census Bureau	0	`http://www.census.gov/ipc/www/idb/groups.php`

Table 29: Features (1)

Id	Type	Code	Year	Feature	Source	URL	Missing values
BH1	HEALTH	B-HE-1-HOSP	avg 06-08 (and earlier if missing)	Hospital beds (per 1,000 people)	World Bank	`http://data.worldbank.org/indicator/SH.MED.BEDS.ZS`	Avg 2006 - 2009 & missing: use earlier year
BF1	FREEDOM	BF1-REL-PROP	08	percentage of the largest religion or religious brand (LG1PCT08)	thearda	`http://www.thearda.com/Archive/Files/Downloads/IRF2008_DL2.asp`	1: United States (51.3%), vietnam (9.3%) both from `https://www.cia.gov/library/publications/the-world-factbook/geos/us.html`, china (42.5%) from World Christian Database
BF2	FREEDOM	BF2-REL-CONF	08	Number of unique incidents of religious conflict (CONFL#08)	thearda	`http://www.thearda.com/Archive/Files/Downloads/IRF2008_DL2.asp`	1: United States - given mean value (7.18)
BF3	FREEDOM	BF3-WOR-FREE	08	Freedom of the World 2008 LOW = MORE FREE	Freedom House	`http://freedomhouse.org/template.cfm?page=351&ana_page=341&year=2008`	-
BC1	CLIMATE	BCL1-MEAN-TEMP	1961-1990	Mean Temperature	Tyndall Centre	`http://www.cru.uea.ac.uk/~timm/cty/obs/TYN_CY_1_1.html`	-
BC2	CLIMATE	BCL2-MIN-TEMP	1961-1990	Min Temperature	Tyndall Centre	`http://www.cru.uea.ac.uk/~timm/cty/obs/TYN_CY_1_1.html`	-
BC3	CLIMATE	BCL3-MAX-TEMP	1961-1990	Max Temperature	Tyndall Centre	`http://www.cru.uea.ac.uk/~timm/cty/obs/TYN_CY_1_1.html`	-
BQ1	EQUALITY	BQ1-EDU-GENDER	Average 2005 - 2010 (and earlier if missing)	Secondary education, pupils (% female). Transformed to gender equality: 50 - abs(x - 50)	World Bank	`http://data.worldbank.org/indicator/SE.SEC.ENRL.FE.ZS`	
BE1	EDUCATION	BE1-GROSS-PRI	Average 2005 - 2010 (and earlier if missing)	School enrolment, primary (% gross)	World Bank	`http://data.worldbank.org/indicator/SE.PRM.ENRR`	missing: Singapore (from Singapore Year of Statistics 2011 [46](263.906/209.1 $*$ 100))
BE2	EDUCATION	BE2-GROSS-SEC	Average 2005 - 2010 (and earlier if missing)	School enrolment, secondary (% net)	World Bank	`http://data.worldbank.org/indicator/SE.SEC.ENRR`	missing 1: Singapore (Singapore Yearbook of Statistics 2011 [46] (214.388/(125.3 + 134) = 82.67))
BE3	EDUCATION	BE2-LIT-RATE	Average 2000 - 2010 (and earlier if missing)	Literacy rate	World Bank	`http://data.worldbank.org/indicator/SE.ADT.LITR.ZS`	missing 8: Serbia (`http://stats.uis.unesco.org/unesco/TableViewer/tableView.aspx`), Andorra, australia, austria, belguim, canada, Czech Republic, denmark (`https://www.cia.gov/library/publications/the-world-factbook/fields/2103.html`
BP1	CONFLICT	BP1-MIL-EXP	2000 (and earlier if missing)	Military expenditure (% of GDP)	World Bank	`http://data.worldbank.org/indicator/MS.MIL.XPND.GD.ZS`	missing 3: costa rica (cia world factbook `https://www.cia.gov/library/publications/the-world-factbook/fields/2034.html`), andorra (no official military so given the mean value 2.04), Hong Kong (defense is responsibility of china - therefore given the same value as china)
BP2	CONFLICT	BP2-MIL-EXP	2000 (and earlier if missing)	BP1-MIL-EXP / Health expenditure, total (% of GDP)	World Bank	health expenditure: `http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS?page=2`	missing: hong kong [35] 59,661 / 1,234,964 = 4.83%

Table 30: Features (2)

References

[1] World bank: Gdp. `http://data.worldbank.org/indicator/NY.GDP.PCAP.CD/countries`.

[2] *Biographical Memoirs.* Number v. 56 in Biographical Memoirs. National Academies Press, 1986.

[3] Saamah Abdallah, Sam Thompson, Juliet Michaelson, Nic Marks, and Nicola Steuer. The happy planet index 2.0: Why good lives dont have to cost the earth. NEF (the new economics foundation), June 2009.

[4] Edgar Acuña and Caroline Rodriguez. The treatment of missing values and its effect on classifier accuracy. In David Banks, Leanna House, Frederick R. McMorris, Phipps Arabie, and Wolfgang Gaul, editors, *Classification, Clustering, and Data Mining Applications*, volume 0 of *Studies in Classification, Data Analysis, and Knowledge Organization*, pages 639–647. Springer Berlin Heidelberg, 2004. 10.1007/978-3-642-17103-1_60.

[5] Gustavo Batista and Maria Carolina Monard. A study of k-nearest neighbour as an imputation method. In *Soft Computing Systems: Design, Management and Applications*, Second International Conference on Hybrid Intelligent Systems, Santiago, Chile, pages 251–260. IOS Press, 2002.

[6] C.M. Bishop. *Pattern recognition and machine learning.* Information science and statistics. Springer, 2006.

[7] Christian Bjornskov. How comparable are the gallup world poll life satisfaction data? *Journal of Happiness Studies*, 11(1):41–60, 2010.

[8] David Blanchflower and Andrew J. Oswald. Hypertension and happiness across nations. The warwick economics research paper series (twerps), University of Warwick, Department of Economics, 2007.

[9] Washington's Blog. Inequality in america is worse than in egypt, tunisia or yemen. http://www.globalresearch.ca/index.php?context=va&aid=22999. Gini Index Map.

[10] Robert Burbidge and Bernard Buxton. B.f.: An introduction to support vector machines for data mining. In *Keynote Papers, Young OR12, University of Nottingham, Operational Research Society, Operational Research Society*, pages 3–15, 2001.

[11] David Cameron. Prime minister speech on wellbeing. Transcript, November 2010. http://www.number10.gov.uk/news/speeches-and-transcripts/2010/11/pm-speech-on-well-being-57569 [accessed April 23, 2011].

[12] Herman Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368, 1973.

[13] Andrew E. Clark, Paul Frijters, Michael A. Shields, Andrew E. Clark, Paul Frijters, Michael A. Shields, and Richie Davidson. Income and happiness: Evidence, explanations and economic implications, 2006.

[14] Nello Cristianini and John Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods.* Cambridge University Press, New York, NY, USA, 2000.

[15] Tijl de Bie. Pattern analysis and statistical learning, university of bristol (lecture notes). 2011.

[16] Angus Deaton. Income, health, and well-being around the world: Evidence from the gallup world poll. *The Journal of Economic Perspectives*, 22(2):53–72, 2008.

[17] R. A. Easterlin. Does economic growth improve the human lot? some empirical evidence. In Paul A. David and Melvin W. Reder, editors, *Nations and Households in Economic Growth: Essays in Honor of Moses Abramowitz.* Academic Press, 1974.

[18] Richard A. Easterlin, Laura Angelescu A. McVey, Malgorzata Switek, Onnicha Sawangfa, and Jacqueline Smith S. Zweig. The happiness-income paradox revisited. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52):22463–22468, December 2010.

[19] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[20] David V. Glidden Eric Vittinghoff, Stephen C. Shiboski and Charles E. McCulloch. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models.* Statistics for biology and health. Springer, 2005.

[21] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861 – 874, 2006. ROC Analysis in Pattern Recognition.

[22] Eibe Frank, Yong Wang, Stuart Inglis, Geoffrey Holmes, and Ian H. Witten. Using model trees for classification. *Machine Learning*, 32:63–76, 1998. 10.1023/A:1007421302149.

[23] Bruno S. Frey. Genes, economics, and happiness. CREMA Working Paper Series 2010-01, Center for Research in Economics, Management and the Arts (CREMA), January 2010.

[24] Joseph L. Gastwirth. Estimation of the lorenz curve and gini index. *The Review of Economics and Statistics*, 54(3):306–316, August 1972.

[25] P. Gundelach and S Kreiner. Happiness and life satisfaction in advanced european countries. *Cross-Cultural Research*, 38(4):359–386, November 2004.

[26] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[27] T. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer series in statistics. Springer, 2009.

[28] Freedom House. Freedom in the world 2008. `http://freedomhouse.org/template.cfm?page=363&year=2008`, 2008.

[29] Eduardo R. Hruschka, Estevam R. Hruschka Jr., and Nelson F. F. Ebecken. Evaluating a nearest-neighbor method to substitute continuous missing values. In Tamás D. Gedeon and Lance Chun Che Fung, editors, *AI 2003: Advances in Artificial Intelligence*, volume 2903 of *Lecture Notes in Computer Science*, pages 723–734. Springer Berlin / Heidelberg, 2003. 10.1007/978-3-540-24581-0_62.

[30] RD Lane, EM Reiman, GL Ahern, GE Schwartz, and RJ Davidson. Neuroanatomical correlates of happiness, sadness, and disgust. *The American journal of psychiatry*, 154(7), 07 1997.

[31] R Layard. Why subjective well-being should be the measure of progress. Available at http://www.oecdworldforum2009.org, 2009.

[32] PAUL F. LAZARSFELD. The american solidier an expository review. *Public Opinion Quarterly*, 13(3):377–404, 1949.

[33] David T. Lykken and Auke Tellegen. Happiness is a stochastic phenomenon. *Psychological Science*, 7(3):186–189, May 1996.

[34] Guy Mayraz, Gert G. Wagner, and Jürgen Schupp. Life satisfaction and relative income: Perceptions and evidence. IZA Discussion Papers 4390, Institute for the Study of Labor (IZA), 2009.

[35] Ms S.Y.YUE Miss Eva LIU. Health care expenditure and financing in hong kong, 1998.

[36] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[37] Tom Mitchell. Decision tree learning. http://www-2.cs.cmu.edu/ tom/mlbook-chapter-slides.html (Accessed).

[38] P.N. Mukherji and C. Sengupta. *Indigeneity and universality in social science: a South Asian response*. Sage Publications, 2004.

[39] Andrew J. Oswald and Nattavudh Powdthavee. Does happiness adapt? a longitudinal study of disability with implications for economists and judges. *Journal of Public Economics*, 92(5-6):1061 – 1077, 2008.

[40] Timo Partonen and Jouko Lnnqvist. Seasonal affective disorder. *The Lancet*, 352(9137):1369 – 1374, 1998.

[41] Ross J. Quinlan. Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, Singapore, 1992. World Scientific.

[42] Katrin Rehdanz and David James Maddison. Climate and happiness. Working Papers FNU-20, Research unit Sustainability and Global Change, Hamburg University, 2003.

[43] Remco Bouckaert Remco. Bayesian network classifiers in weka. Working paper series 14/2004, University of Waikato, Hamilton, New Zealand, 2004.

[44] Leora N. Rosen, Steven D. Targum, Michael Terman, Michael J. Bryant, Howard Hoffman, Siegfried F. Kasper, Joelle R. Hamovit, John P. Docherty, Betty Welch, and Norman E. Rosenthal. Prevalence of seasonal affective disorder at four latitudes. *Psychiatry Research*, 31(2):131 – 144, 1990.

[45] Ed Sandvik, Ed Diener, and Larry Seidlitz. Subjective well-being: The convergence and stability of self-report and non-self-report measures. In Ed Diener, editor, *Assessing Well-Being*, volume 39 of *Social Indicators Research Series*, pages 119–138. Springer Netherlands, 2009. 10.1007/978-90-481-2354-4_6.

[46] Singapore. Yearbook of statistics, singapore, 2011.

[47] Alois Stutzer and Bruno S. Frey. Does marriage make people happy, or do happy people get married? *Journal of Socio-Economics*, 35(2):326 – 347, 2006. The Socio-Economics of Happiness.

[48] R. Tibshirani. A simple explanation of the lasso and least angle regression. *Journal*, Year.

[49] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[50] Paolo Verme. Life satisfaction and income inequality. *Review of Income and Wealth*, 57(1):111–127, 2011.

[51] Y. Wang and I. H. Witten. Induction of model trees for predicting continuous classes. In *Poster papers of the 9th European Conference on Machine Learning*. Springer, 1997.

[52] A White. A global projection of subjective well-being: The first published map of world happiness 2006. Available at http://news.bbc.co.uk/1/shared/bsp/hi/pdfs/28_07_06_happiness_map.pdf [accessed March 2011].

[53] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.