A Geo-temporal Analysis of the Relationship Between Fertility Rates and World Development

Jakob Brown

Abstract—The present report presents a visual analytics approach to exploring the relationships between several World Development Indictaors (WDIs), as well as how they have changed over time in different parts of the world. The analysis, which entails geo-temporal analysis, choropleth mapping of k-means clutstering outputs, and a linear regression predictive model, focuses especially on babies per woman as an indicator, in the interest of acsertaining how significant an influence it is in the overall progress and advancement of a society.

1 Problem Statement

The UN's world development indicators (WDI) [18] contain records on the majority of countries in the world, as far back as 1800. This data provides both a generous landscape and time-scape over which these indicators can be displayed and analysed to see how they have changed and relate to each other.

The following research questions are posed:

1. Can correlations between babies per woman and other development indicators be established and evidenced through matching trends geo-temporally?
2. What parts of the world are most in need of further development in terms of Human Development Index?
3. To what extent can a credible prediction model be made to predict overall development using babies per woman as the focal predictive element?

These questions aim to add to the body of research that posits high unwanted fertility rates a key hindrance on human development [3]. Through geo-temporal analysis methods, explanations of high fertility rates can be inferred, which – as has been highlighted by research [14] - are of great value to the progress of human development everywhere through shaping of policy and intervention funding, at a national and local level of any country. The data, when combined together, provides a uniquely broad scale, encompassing the whole world and the whole of the past two centuries for some variables. This substantial breadth, used correctly, can give clarity to the big questions of where and when development has surged forward, where it can still be improved, and how that further development can be facilitated.

2 State of the Art

Much prior visualisation research has been conducted in to investigating a single variable over both space and time, or multiple variables over the course of both space and time [9]. The nature of the present study, however, requires the effective observation of the relationships between multiple thematic variables over both time and space. State-of-the-art techniques that allow for this level of comparison are crucial to gleaning the trends and patterns that the present research questions demand as answers [9].

One pioneer of such techniques was the late Hans Rosling, whose visualisations of world demographics also made great use of the WDIs, championing the visual encoding of multivariate data to allow for both time and space to be represented simultaneously in one visualisation [11]. Use of size, colour, labelling, mapping, and animation gave ample communicatory channels through which these dimensions could be conveyed, giving valuable style and attractiveness to the visualisations which often aided his storytelling purpose [10]. These accessible visualisation tools have proved extremely effective in communication of information to their audience, regardless of the domain knowledge level of the audience members [12]. Consequently, they have been useful in debunking harmful and ill-informed narratives about parts of the world in terms of poverty and development.

There are however apparent trade-offs between style and analytical power [8], as well as limitations to using animation for temporal movement with regards to more analytical purposes [17]. One of the most popular means of gaining information on broad geo-temporal spectrums are choropleth mapping, visually embedding variable information into the map's feature's fill colour [9][4]. Multiples can be made to differ across variables or time, working better in communicating correlations to users than animations or single-map variations [9]. 3-Dimensional layering can also be done [16]; however, this requires more interaction from the user, necessitating more elaborate navigation methods that can hinder insight perception [13].

The referenced research aids the present study in its approach: The identical subject matter of Rosling et al's (2005) work lends itself well to providing a strong state-of-the-art method for the effective visualisation of the WDIs over time and space. The visual encoding of variables in particular enables the relevant observation of relationships and correlations over time and space, and will be key in informing the input of the predictive modelling component of the study and determining its ultimate accuracy. Other critical evaluations of this work [9] and alternative techniques proposed [9][4] will tailor the present approach, by using more apposite techniques where others are not as appropriate for present purposes. This includes opting for single graphs when

juxtaposing time, and multiple graphs when focussing on space at one single point in time [9]. The finding that time intervals and spatial regions yield generally better and quicker understanding of variable relationships than inclusion of every single time and region [9] will also be honoured in the present study's visualisations by grouping by space and time where suitable.

3 Properties of the Data

The array of the UN's WDIs used in the present study was obtained from the website of the Gapminder foundation, founded by the aforementioned Rosling. The data "is a compilation of relevant, high-quality, and internationally comparable statistics about global development and the fight against poverty." [18]

Gapminder collated each indicator from the World Bank into a data file, with each country/economy as a row, and each year as a column. This format was modified such that a single year column was made with multiple data points of each country in the data, each representing a single year. This would be a far more conducive format for time-series analysis.

Because some variables did not have records as far back in time as others, and some indicators data were ostensibly sparse, containing many missing or bogus values, several separate versions of the data were made for each computational task of the research. The variables that were full enough to feature in the time series analysis were: babies per woman; life expectancy; child mortality (death of children <= 5 years old per 1000 births); population; GDP per capita. Human Development Index (HDI) (a statistical representation of average health, education and living standard of a country's populace) only goes back to 1990, so could not be included in the time-series analysis, it was however deemed the most fitting dependent variable for the predictive modelling component, as a sufficient representation of overall development of a country by measure of quality of life. Finally, data sets which contained each feature individually were also made for visual juxtaposition of clustering by different variables.

Data distribution was addressed via plots containing looped histograms on each variable (*figure 1)*. These graphs allowed clear identification of skewness or kurtosis amongst the indicators.

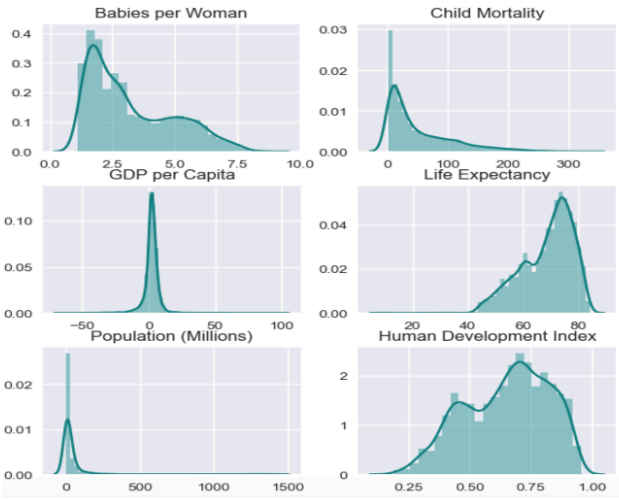

Figure 1: Histogram plot of data distribution.

Of these variables, several had non-Gaussian distributions. This was remedied by log transforming the variables and iterating the distribution plots to see which benefitted in terms of distribution (*figure 2*). Those that were distributed more normally on a logarithmic scale were kept, and those that were not were reverted to their original distribution.

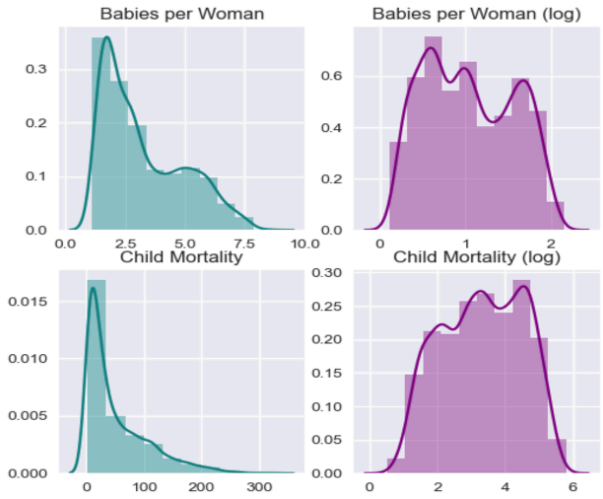

Figure 2: Histogram display of pre- and post-log transformation of features.

Other modifications made to the data include the imputation of certain missing variables. Lack of domain expertise meant that minimal data alteration was preferred, however the life expectancy metric was deemed too important an indicator to omit from the time-series analysis, and so null values present in smaller countries in earlier 19th century were imputed with the mean of the two data points for the same country that were closest in year either side of the missing value, e.g. Antigua's 1806 life expectancy was calculated as the mean of Antigua's 1805 and 1807 life expectancy. This replicated the trend of steady increase in life expectancy that was observed in the data as a whole, and so would have retained a high degree of verisimilitude. In

cases of multiple consecutive missing values, the same principle was applied by creating evenly spaced intervals between the last and next available data point.

4 Analysis

4.1 Approach

In this section, the process through which the research questions are to be answered will be outlined. This will be organised in terms of the computational operations carried out for each task and the visual components that will be produced to accompany them. The purposes of these actions will also be stated, as well as how they interact with human judgement of the present author (*figure 3*).

In order to answer research question 1 and 2, the state-of-the-art approach of geo-temporal visualisation will be employed to allow for the observation of trends in indicators through both time and space. First through time, data pre-processing previously outlined will format the data set, plus feature engineering of time interval features of half century and century. Moreover, a continent feature will be made through merging the data with Geopandas' world map GeoDataFrame. Following that, individual Tableau figures will be created in order to produce simple, clear graphs displaying all relevant features by harnessing the graphs' physical features such as size and colour, and the newly made spatial and temporal groupings. Iterating through visualisations of different combinations of variables will provide visual description, from which any relationships between variables and the strength of those relationships can be inferred through human perception. Disparities between regions' development over time can also be explored, and explanation can be extrapolated through human reasoning interacting with previous literature on world development.

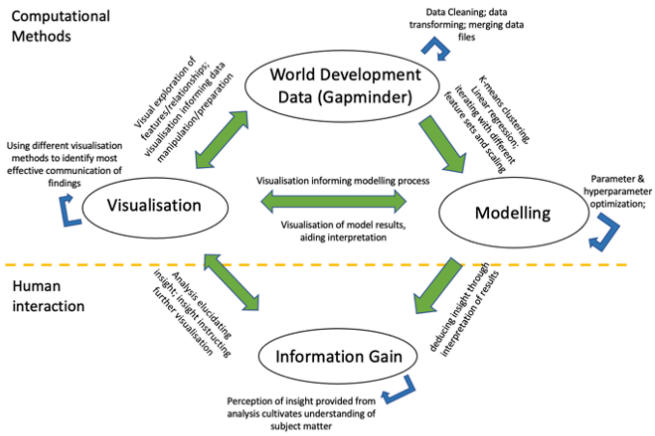

Figure 3: Workflow Diagram.

A k-means clustering will then be used to cluster the world's countries based on multiple WDIs in present day. Different numbers of clusters will be iteratively tested and visualised via silhouette analysis and PCA scatter plotting to ascertain the optimum number. Choropleth mapping will then be used to create separate maps, visually encoded by the clustering of the two main indicators being analysed: babies per woman and HDI. This mapping will allow for visual corroboration of the areas in different stages of development with regards to these indicators posited by the time-series analysis. Furthermore, any detail lost by the continental grouping will be clarified, further refining human understanding. By using multiple graphs, a correlation between the two variables can be drawn or refuted. Furthermore, focussing on the present day will give clearer insight into what countries are in need of development in the future, useful knowledge for future work on world development.

Finally, to answer research question 3, two linear regression models will be made with HDI as the dependent variable: one a univariate model with babies per woman as the sole independent variable, and another multiple linear regression model with several WDIs as input. The features selected for this model will be determined through a combination of PCA analysis and a correlation matrix of all the variables. Following standardization and normalization, an array of train/test split sizes will be iterated over to determine the optimal split for model performance. Performance for each will be visualised with scatterplots and histograms of residual distribution, from which human judgement will be needed to infer reliability of results or lack thereof.

4.2 Process

The time-series analysis was conducted using the previously made version of the data containing all of the variables that had data extending far back enough in time. Visualisations were created using different scales, including by year, by half century and by century. In general, average statistics were used for each variable, aggregated by continent to prevent visual noise: there were too many countries to display them all individually. Averages were preferred to summed totals in general, as they neutralised the effects of countries with much larger populations distorting the data, leading to misrepresentation for an entire continent. The yearly analysis was found to provide the most detail and the best scale of the when the most change occurred over the 220 years up to today. The 2020 data was dropped as it was unclear whether the records for the most recent year were in fact empirical records yet, or UN forecasts for the indicators, which were made for every year up to 2100 for many of the indicators.

As can be seen from figure 4, clear relationships can be observed between avg. babies per woman and many other variables, including avg. child mortality, avg. life expectancy, year, and continent. As time has passed, babies per woman have decreased everywhere in the world, which has also saw a drastic reduction in premature death in children, and significant increase in life expectancy. The change in centurion average appears to have accelerated for the 21st century, suggesting that most of the change occurred in recent times.

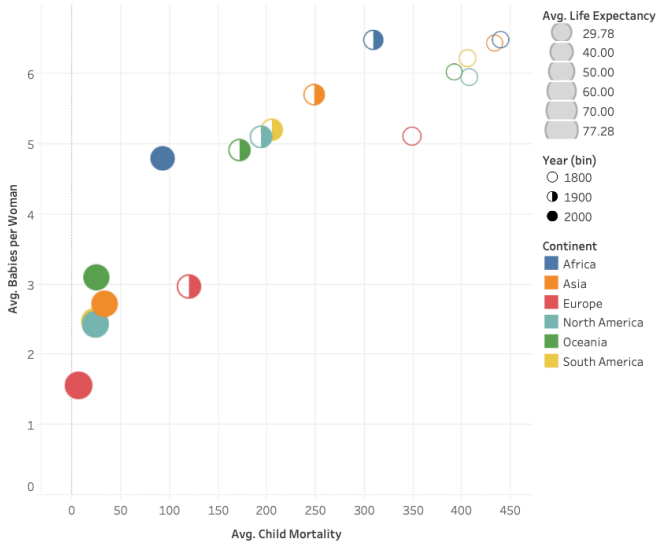

Figure 4: Avg. Babies per Woman Against Avg. Child Mortality, Filtered by Century and Avg. Life Expectancy, Aggregated by Continent.

When looking closer, on a year-by-year scale (*figure 5*), the rate of change throughout time becomes even clearer. Birth rates appeared to take a sharp decline everywhere around the world in the 1960s, in synchrony with child mortality rates. This is likely due to the invention and release of the contraceptive pill in 1960, suggesting that the reason or high birth rates in many instances was due to the lack of agency women had over their reproductive activity. Further visualisation saw similar surged improvement in the other WDIs around this time. This finding solidifies the relationship that babies per woman has with all world development indicators analysed, and gives credence to the theory posited by many that the empowerment of women through contraception is *the* method by which poverty can be

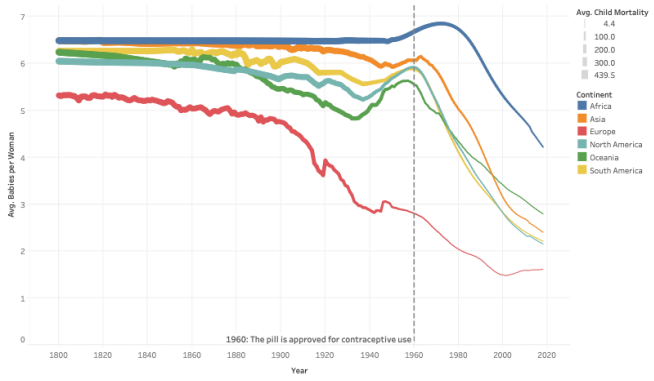

Figure 5: Avg. Babies per Woman Over Time, Filtered by Avg. Child Mortality, Aggregated by Continent.

eradicated [1][6][7]. This visualisation also highlights the effect that world events can have on world development indicators, with apparent slumps in babies per woman during both the World Wars, followed by 'baby booms' at their end. This time-series analysis would also suggest that Africa is – on average – further behind the rest of the world in terms of development by all indicators analysed. This finding was further scrutinised through the geographical mapping section.

The clustering data sets were prepared by slicing only the most present year (2019) from the data, and then creating two separate sets, one with babies per woman and without HDI, and the other vice versa. The data were normalised and standardised to prevent undue dominance of larger-scale variables.

An array cluster amounts from 2-6 were then looped over, performing a k-means clustering algorithm on the babies per woman data, creating a silhouette plot with an average silhouette value as a marker (*figure 6 (left)*), accompanied by a PCA transformation with 2 components in order to visualise each data point as a scatter plot alongside, coloured by the cluster they had been assigned (*figure 6 (right)*). These plots simultaneously displayed together allowed for simple evaluation of each number of clusters while juxtaposing with the others. The human judgement component involved several actions: checking the silhouette plot for any data points that yielded negative values, suggesting the country it belonged to had been assigned to the wrong cluster; observing the overall silhouette score, which gave the average value between 0 and 1 of how effectively the countries had been clustered; and authenticating the silhouette analysis through observing any overlap of clustered data points in the PCA scatter plot.

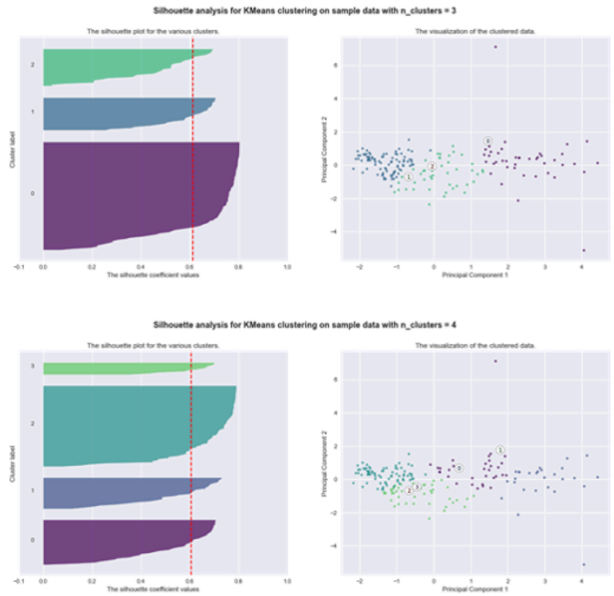

Figure 6: Silhouette and PCA Analysis of K-means Clustering Algorithm.

For 2 clusters, there were some negative values, suggesting that the algorithm was being forced to group together countries that did not belong together. This would also have given poor detail when it came to the mapping stage. The rest of the number of clusters had no negative values, but the silhouette score peaked at 3 clusters, tapering off as the number increased. The accompanying PCA plot also showed excellent segregation of clusters, thus, the

choropleth mapping was executed with 3 clusters for each variable.

Even after string matching the country names of the WDI data and the Geopandas world map GeoDataFrame, there were missing data for some geometric values, namely Greenland and some islands such as Taiwan. These were simply greyed out in the choropleth maps, by making their cluster value NaN.

Overall, the geographical mapping supported the relationship between babies per woman and the main development indicator, HDI (*figure 7*). The vast majority of the world appears to have been clustered in a similar way for both variables. It also substantiated previous findings in the time-series analysis that showed a great convergence of birth rate and other development indicators in all continents but Africa. Further detail elucidated from the mapping of all individual countries across multiple visuals includes the apparent localisation of the lowest development in Africa to the central-most countries. It can also be seen that lower HDI can be present without a high fertility rate, as in some parts of East and South-East Asia, as well as South and Central America. OECD countries are – unsurprisingly – consistently in the top cluster for babies per woman and HDI. The perspective the spatial mapping shows that areas still in need of development in terms of HDI are spread all over the world and are still very common, in spite of the drastic improvements that the time-series analysis highlighted.

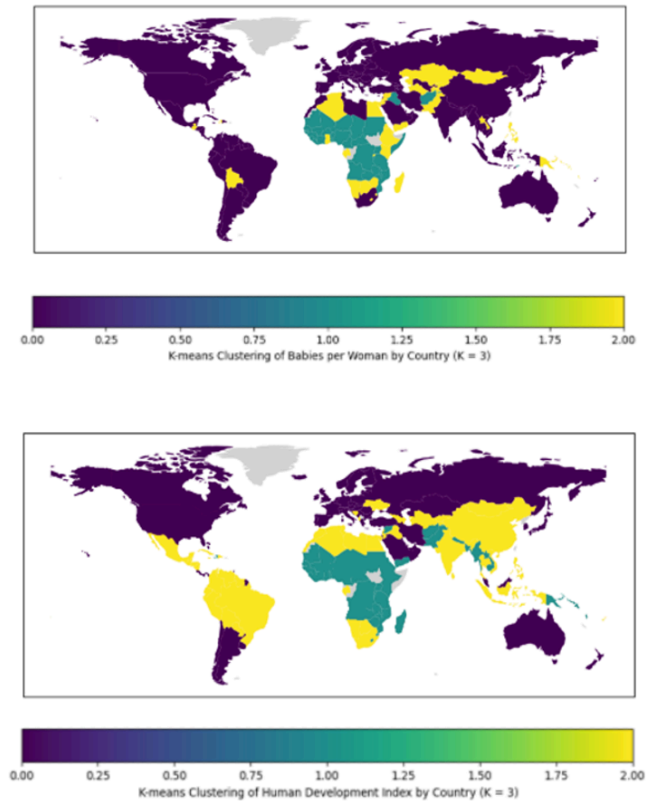

Figure 7: Choropleth World Maps, Clustered by Babies per Woman and HDI.

In preparation of the predictive modelling stage, a correlation matrix was made of all the variables available. (*figure 8*). This had to include HDI as the dependent variable, and so the data was modified yet again to only include the years 1990 to 2018, the most up to date data. Population was scaled to millions in anticipation of modelling.

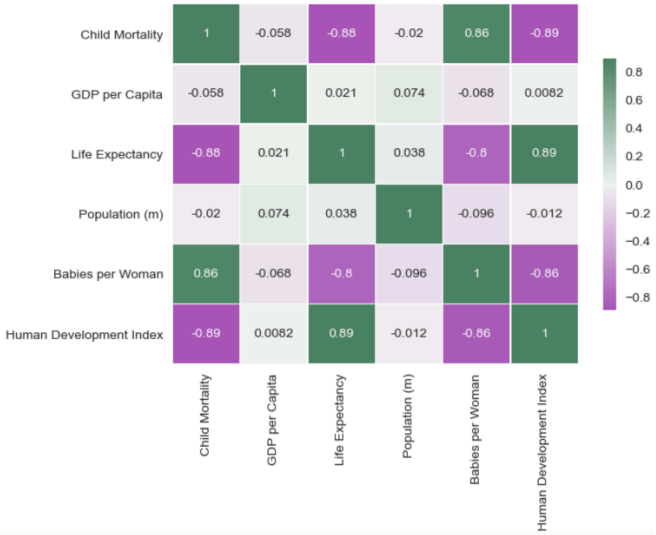

Figure 8: Correlation Matrix for All WDIs.

The matrix with correlation labelling showed clearly that child mortality, life expectancy, and babies per woman had strong associations with HDI, while GDP per capita and Population had almost no association. Analysis of the loadings of the PCA analysis from figure 6 showed also that the variance captured by these two variables was little to none. The second, far less influential principal component appeared to essentially be responsible for capturing some severe outliers. This component was found to predominantly be accounting for GDP per capita and population. They were therefore dropped, leaving three predictor variables for the multiple regression model. These indicators were also highly correlated together, threatening multicollinearity issues with the predictive model. The model was created nonetheless, assured that the intercorrelated variables would still yield greater predictive power.

After normalisation and standardisation scaling, a range of train/test splits were looped over (0.35 – 0.15 at intervals of 0.5) for each model, calculating R2 and mean squared error (MSE) as performance measures, as well as generating scatter plots of predicted values against actual values for the test sets (which were kept the same for each model through using the same random seed to ensure comparability). Regression lines were plotted through these scatter plots also (*figure 9*). Alpha and size of the markers were reduced to accommodate for the great number of data points being plotted.

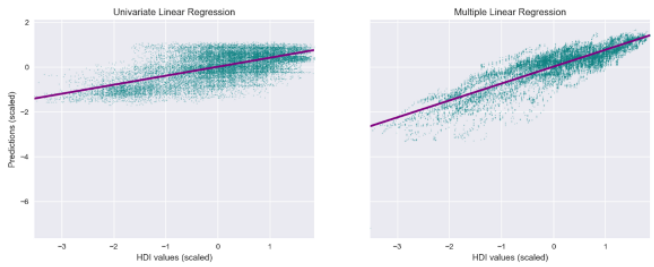

Figure 9: Regression Plots for Univariate and Multivariate Models (scaled).

The optimal train/test split was found to be 0.35. The univariate model failed to perform to a successful standard (R2 score: 40.4%; MSE: 0.34), while the multivariate model performed extremely well (R2 score: 75.4%, MSE: 0.14). This contrast in performance can be seen in figure 9 from the more sparsely distributed values around the regression line in the univariate plot as compared alongside the multivariate, which has a more stringent, dense scatter. Residual plots on the multivariate model (*figure* 10) showed the distribution of residuals to be evenly spread in a Gaussian shape, though with a slight knew to the left. This generally implies that the error terms remain constant as the value of the predictor changes.

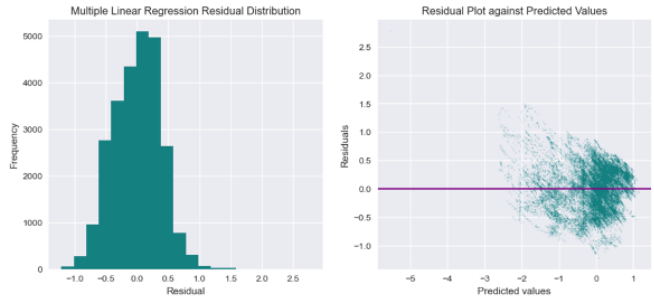

Figure 10: Residual Plots of Multivariate Linear Regression Model (scaled).

4.3 Results

RQ 1. Time-series analysis indicates a strong relationship between babies per woman and other indicators of world development. The general trend is still heading downwards, as world development trends upwards everywhere. The sharpest decline in babies per woman occurred as a result of making oral contraception publicly available to the masses. This suggests that areas where babies per woman is still high and development is still low may be so because they still have limited access to contraception or had delayed access.

RQ 2. Successful clustering and choropleth mapping saw that many places were still at different stages of development, with Central Africa being furthest behind, although sharply improving. Integrating the geo-temporal findings would allude to these areas having lower access to contraception for women. Future research should investigate in to the reasons for this such as religiosity and infrastructure [4][1], further instructing development work by governments and charities.

RQ 3. A successful linear regression model produced a 75.4% accuracy rate, with a corresponding residual distribution that suggested credible, reliable results. This could be used by governments or charities as a tool for monitoring development increases or decreases in a country through three relatively easily attainable statistics.

5 Critical Reflection

While the time-series analysis worked excellently in answering the research question at hand, it did reveal some drawbacks of the data. Many indicators appeared to have unnaturally stable values for some less developed countries across time. This can be seen with the exceptionally smooth, stable first part of the Africa line in figure 5, which suggests that the records kept are in many cases just estimates, particularly those that are older. Moreover, GDP per capita was found to have very erratic patterns across time, making it difficult to identify clear relationships with any of the other indicators. This was especially unexpected given the clear relationship between GDP per capita and other development indicators that have been drawn in other research [15]. The analysis is only as accurate as the data being analysed, and so as time passes and the now exceptional standards of data collection is maintained, this kind of long-term time-series work will only become more assured in its credibility and more complete in the metrics it has available for use. This will also hopefully result in filling the gaps in the world map for which data could not be obtained for this study.

While aggregating by continent improved the visual readability of the analysis, grouping together so many hugely varying countries and averaging their data inevitably misses detail, as the choropleth mapping only began to demonstrate. This highlights one of the still pervasive difficulties with geo-temporal visualisation when both the times and/or places being examined are myriad. More useful information such as differing explanations of high and low development rates are most certainly there to be discovered in case study analyses of the idiosyncrasies of different countries [2]. Gathering of enough individual instances at a national or community level may even lead to the 'nonspuriousness' required to infer causation of development from lower fertility rates [5].

The visual approach of producing arrays of plots by looping over multiple nested parameters such as number of clusters in k-means clustering, and train/test split and which variables used in predictive modelling proved excellent in its efficiency and its ease of comparison of different outputs. This was important in optimising several processes throughout the research, and was found to greatly reduce iterative workload of revisiting and altering data each time, as it could all be tested and compared in one fell swoop.

The best predictive model did exhibit some collinearity amongst predictor variables, meaning that overfitting may have taken place in results. However, as training and test R2 were

the same, and the only real purpose of the model was to work as a predictive tool rather than a relationship descriptor, it can be assumed that this did not have a negative effect on the desirability of results.

Table of word counts

Problem statement	250
State of the art	487
Properties of the data	494
Analysis: Approach	500
Analysis: Process	1500
Analysis: Results	200
Critical reflection	456

REFERENCES

[1] Bailey, M.J., 2006. More power to the pill: The impact of contraceptive freedom on women's life cycle labor supply. *The quarterly journal of economics*, *121*(1), pp.289-320.

[2] Dias, J.G. and de Oliveira, I.T., 2015. Multilevel effects of wealth on women's contraceptive use in Mozambique. *PloS one*, *10*(3), p.e0121758.

[3] Gillespie, d., Ahmed, s., Tsui, a. and Radloff, s., 2007. unwanted fertility among the poor: an inequity?. *bulletin of the world health organization*, *85*, pp.100-107.

[4] Guo, D., Chen, J., MacEachren, A.M. and Liao, K., 2006. A visualization system for space-time and multivariate patterns (vis-stamp). *IEEE transactions on visualization and computer graphics*, *12*(6), pp.1461-1474.

[5] Hitchens, c., 2012. the missionary position: mother teresa in theory and practice. hachette uk.

[6] Manandhar, D.S., Osrin, D., Shrestha, B.P., Mesko, N., Morrison, J., Tumbahangphe, K.M., Tamang, S., Thapa, S., Shrestha, D., Thapa, B. and Shrestha, J.R., 2004. effect of a participatory intervention with women's groups on birth outcomes in nepal: cluster-randomised controlled trial. *the lancet*, *364*(9438), pp.970-979.

[7] Milazzo, A. and Van de Walle, D., 2015. *women left behind? poverty and headship in africa*. the world bank.

[8] Moere, A.Y., Tomitsch, M., Wimmer, C., Christoph, B. and Grechenig, T., 2012. evaluating the effect of style in information visualization. *ieee transactions on visualization and computer graphics*, *18*(12), pp.2739-2748.

[9] Peña-Araya, V., Pietriga, E. and Bezerianos, A., 2019. a comparison of visualizations for identifying correlation over space and time. *ieee transactions on visualization and computer graphics*, *26*(1), pp.375-385.

[10] Rosling, h., 2006. debunking third-world myths with the best stats you've ever seen. ted.

[11] Rosling, h., 2007. visual technology unveils the beauty of statistics and swaps policy from dissemination to access. *statistical journal of the iaos: journal of the international association for official statistics*, *24*(1), pp.103-104.

[12] Rosling, H., Rosling, R.A. and Rosling, O., 2005. new software brings statistics beyond the eye. *statistics, knowledge and policy: key indicators to inform decision making. paris, france: oecd publishing*, pp.522-530.

[13] Shneiderman, B., 2003. Why not make interfaces better than 3d reality?. *IEEE Computer Graphics and Applications*, *23*(6), pp.12-15.

[14] Sonfield, A., Basstedt, K., Kavanaugh, M.L. and Anderson, R., 2013. the social and economic benefits of women's ability to determine whether and when to have children.

[15] Summers, L.H., 1994. investing in all the people: educating women in developing countries. the world bank.

[16] Tominski, C., Schulze-Wollgast, P. and Schumann, H., 2005, July. 3d information visualization for time dependent data on maps. In *Ninth International Conference on Information Visualisation (IV'05)* (pp. 175-181). IEEE.

[17] Tversky, B., Morrison, J.B. and Betrancourt, M., 2002. animation: can it facilitate?. *international journal of human-computer studies*, *57*(4), pp.247-262.

[18] World Bank, 2014. World development indicators.