Application of Business Intelligence Techniques using SAS on Open Data: Analysing Health Inequality in English Regions

Neha Thakkar, Ah-Lian Kor, Sanela Lazarevski
School of Computing, Creative Technologies, and Engineering,
Leeds Beckett University, Leeds, UK
E-mail: {A.Kor,S.Lazarevski}@leedsbeckett.ac.uk

***Abstract*—Health inequality is a widely reported problem. There is an existing body of work that links health inequality and geographical location. This means that one might be more disadvantageous health-wise if one was born in one region compared to another. Existing health inequality related work in various developed and developing countries rely on population census or survey data. Effective conclusions drawn require large scale data with multiple parameters. There is a new phenomenon in countries (e.g. the UK), where governments are opening up citizen-centric data for transparency purposes and to facilitate data-informed policy making. There are many health organisations, including NHS and sister organisations (e.g. HSCIC), which participate in this drive to open up data. These health-related datasets can be exploited health inequality analytics. This work presents a novel approach of analysing health inequality in English regions solely based on open data. A methodological and systematic approach grounded in CRISP-DM methodology is adhered to for the analyses of the datasets. The analysis utilises a well-cited work on health inequality in children and the corresponding parameters such as Preterm birth, Low birth weight, Infant mortality, Excessive weight in children, Breastfeeding prevalence and Children in poverty. An authority in health datasets, called Public Health Outcomes (PHO) Framework, is chosen as a data source that contains data with these parameters. The analysis is carried out using various SAS data mining techniques such as clustering, and time series analysis. The results show the presence of health inequality in English regions. The work clearly identifies the English regions on the right and wrong side of the divide. The policy and future work recommendations based on these findings are articulated in this research. The work presented in this paper is novel as it applies SAS based BI techniques to analyse health inequality for children in the UK solely based on open data.**

Keywords*—*SAS; BI techniques; open health data; data mining; health inequality

I. Introduction

Inequality is a widely reported problem in modern day societies. González [7] focuses on the regional divide in the UK and notes that inequality affects policy decisions in the country. Thomson et al. [21] have investigated on tensions in public health policy which involves health inequalities. The World Bank and overseas development institute McKay [16] have defined inequality broadly in terms of living standards. According to this interpretation, inequality is measured against living standards which encompass income equality, health, education, crime and housing standards (ibid, [20]). Health inequality is defined as "difference in people or groups due to social, biological, geographical or other factors" [2]. This research is anchored on this narrow interpretation of health inequality within a "geographical" context. The focus on the exact geographical context is based on datasets availability. This leads to the consideration of health inequality for children in England.

Existing health inequality analytics work is limited because most of them merely provide census or survey results [4], [17]-[19]. Such work predominantly relies on sampling techniques which raises the issue of totality and coverage of such sampled datasets. Bottlenecks relating to health inequality analytics are data access and the availability of 'sufficiently large data'. The emerging "Open Data" phenomenon addresses such bottlenecks. The UK government's Open Data Initiative[1] is part of the government's transparency and accountability initiative. Central and local UK government datasets are made available through data.gov.uk and organisations such as the Open Data Institute[2] helps drive this initiative. The initiative supports the release of citizen-related data by government agencies (e.g. health, facilities, crime events, council and government expenses, etc.). The open-sourced datasets could be used by the general public and businesses to perform various data analyses [8]. Additionally, it provides a means for citizen engagement and used as a vehicle for transparency and efficiency [10]. The Health & Social Care information centre (HSCIC[3]) is the national provider of health and social care related data. HSCIC datasets are primarily based on prescriptions and care for various diseases across England and Wales NHS. There are other contributors of health datasets in data.gov.uk, such as the estates and spending data, as well as hospital safety data.

[1] https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential
[2] https://theodi.org/
[3] https://www.gov.uk/government/organisations/health-and-social-care-information-centre

A. Aims and Objectives

SAS is one of the most popular and powerful data analytics software in the market.[4] It provides data mining and analytics capabilities over large datasets. Hence, the triangulation of health inequality use cases, access to open-sourced health datasets, and SAS data analytics capabilities lead to this project's aim: "*to study how SAS techniques can be applied to analyse possible health inequalities between various regions of England based on open data*". When phrased in layman's terms, it is as follows: "*Will you be at disadvantage from health perspective if you were born in one region of the England compared to other region?*" The following objectives support the achievement of this aim:

- Identify possible health inequality use cases in England regions.
- Select relevant open sourced datasets to be used as use cases.
- Employ SAS techniques (e.g. cluster and time series analyses) to unravel health inequalities amongst England regions.
- Application of SAS techniques on selected dataset and representation of results.

B. Potential Application

The sole purpose of the open data movement is to support and more importantly, influence government's policy making. Health inequality use cases presented in this paper could influence policy-making due to the following reasons:

- This research is based on the use of SAS data analytics on non-disputed government data sources. The research findings could be exploited to educate the general public.
- Government policy makers could obtain useful insight for making evidence-based policies. Several recommendations are made based on the findings.

II. Literature Review

Undeniably, health inequality is an age old problem. Work dated back to the 1930s [22] reveals evidence of health inequality in the UK. Although there is a lack of consensus over how health inequality is measured [2], there is universal agreement that it does exist. This literature review focuses on the following: exploitation of open data for policy making; need for measuring health inequality for evidence-based policy making; existing approaches to measure health inequality.

A. Use of Open Data for Policy Making

Some of the benefits of Open Data are: transparency and accountability, public service improvement, promote innovation and increase economic value, empowering citizens, and improved efficiency.[5,6] If data analytics are conducted correctly, then there is a huge potential for policy making.[7] According to Gurstein [8], the UK and Canada are among some of the countries that are at the forefront of Open Data Initiative (see[8,9,10]) where citizen related data is made public [8]. The UK government issues a code of practice[11] to provide guidance on the release and re-use of open sourced data. Currently, the number of datasets released in data.gov.uk is 42,891. There are numerous examples for the use of open data in policy making. In California, USA, they conduct an annual health survey, called California Health Interview Survey (CHIS)[12]. One of the countries uses this open-sourced survey data to successfully argue against building another truck stop by one of the country roads. The argument is based on the evidence that in the California state, the county has the highest overall asthma symptoms prevalence (ibid).

B. Need for Tackling Health Inequality Using Evidence-Based Policy

Existing work highlights major tension among various components of health policy making (Thomson et al., 2005). The authors argue that health inequality exists due to biased policies favouring only well-off and well-engaged patients. They advocate evidence-based policy making and emphasise health inequality monitoring with an appropriate right policy to tackle it. This study supports this line of research and contributes by showing how the mix of open data and SAS can be used for the monitoring of health inequality in children. This work is timely and beneficial due to the following great concern of reduced budget: "NHS is expected to transfer healthcare funding away from younger, more deprived areas to older, more affluent ones" [9].

C. Health Inequality Measurements

Various possible parameters can be used to measure health inequality. Work by [13] establishes a link between income inequality and health, where inequality is linked to mortality rates. Similarly, Wilkinson and Pickett [23] establish a link between inequality and health and social problems. The widely cited work by [2], defines health inequality as: "Any change in the distribution of health that keeps the mean level of health the same but involves a sick person getting healthier and a healthy person getting sicker is registered as a reduction in inequality in health irrespective of the socioeconomic status of the persons concerned." There has been a recent trend that points to using geographical location while measuring the health inequality. For example, work utilising cross-country comparisons [22]. Curtis and colleague [5] consider how ideas and evidence concerning geographical health variation are used in discourses relating to health inequalities. In particular, the authors in the context of Britain find that place has some significance on health inequality, i.e. if you live in an area of high health inequality then there is a higher chance of a

[4] Elliott, A.C. & Woodward, W.A., 2010. *SAS essentials: a guide to mastering SAS for research*, John Wiley & Sons.

[5] http://opendatatoolkit.worldbank.org/en/starting.html

[6] https://www.publications.parliament.uk/pa/cm201314/cmselect/cmpubadm/564/564.pdf

[7] https://www.civilserviceworld.com/articles/opinion/big-data-get-it-right-and-benefits-policy-making-could-be-huge

[8] http://data.gov.uk/

[9] http://www.data.gov/

[10] https://openparliament.ca

[11] https://data.gov.uk/consultation/code-of-practice

[12] http://healthpolicy.ucla.edu/chis/Pages/default.aspx

negative impact on your health. Another similar yet important study investigates the effect of environment amnesties in France. For example, green spaces and its effect on health inequality and they find correlation between the two [14].

This study is built on several existing research [3], [9], [11], [15] in this area that highlights health inequality parameters when it comes to children. It is built on infant mortality statistics parameters that are important for measuring health inequality in children: Infant mortality; Low birth weight; Excessive weight in children; Breastfeeding prevalence; and Children in poverty. These parameters are linked to socio-economic and geographical/environmental situation of women and family. It is important to analyse health inequality with respect to children as there is a proven link between adulthood diseases and childhood circumstances [6]. Hence, when the dataset is searched for this, support for the aforementioned five criteria in dataset is an important factor for selection of datasets for this study.

III. Methodology

A. *CRISP-DM Methodology*

The CRISP-DM [24] consists of six steps shown in Fig. 1.

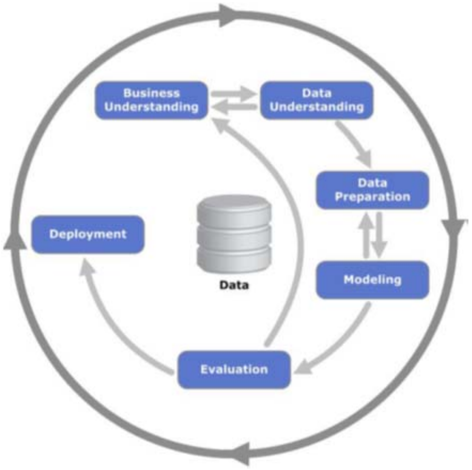

Fig. 1. CRISP-DM methodology steps.[13]

The six steps are:

1) ***Business understanding*****:** Understand what exactly we are searching for, patterns, analysis etc. and then build use cases, models or hypothesis.

2) ***Data understanding*****:** Select right datasets and often done in conjunction with 1). For the purpose of this study, Public Health Outcomes, Framework data[14] based on various indicators including health expectancy, infant mortality, etc. have been selected because PHO datasets are extremely detailed and rich dataset. Another reason is the availability of measurement parameters that concern children (e.g. infant mortality, excess weight in children, children in poverty, low birth weight, breastfeeding prevalence and note the definition of these parameters are given in Table 1. Further the datasets constrain the regions to the following nine in the England: South West Region; North East Region; South East Region; London Region; East of England Region; North West Region; Yorkshire and Humber Region; East Midland Region; West Midland Region.

3) ***Data preparation*****:** Prepare data for the next step (e.g. cleansing, aggregation, etc.). There are very few missing data and the missing data in this study are imputed using the mean of the observed variable.[15] The data is in excel file (.xlsx) format and SAS provides the facility to export excel files directly. As part of the scope, only data related to the five factors are extracted (i.e. infant mortality, low birth weight, breastfeeding prevalence, excess weight in children, and children in poverty).

4) ***Modelling*****:** Model data to discover patterns using statistical, data mining or machine learning techniques. In this study, the modelling phase involves (details are found in Box 1): descriptive as well as inferential statistics analysis; clustering.

5) ***Evaluation*****:** Evaluate and validate the model built in (iv). The aim is to produce findings related to pattern of inequality within the selected regions.

6) ***Deployment*****:** Apply model in use cases. Recommendations formulated based on findings to inform policy making at local or central government level.

TABLE I. Definition of Parameters

Criteria	Definitions	Source
Infant Mortality	"Infant mortality is an indicator of the reflects the relationship between causes of infant mortality and upstream determines of population health such as economic, social and environmental conditions."	Office for National Statistics (ONS)
Low Birth Weight	"Live births with a recorded birth weight under 2500g and a gestational age of at least 37 complete weeks as a percentage of all live births with recorded births weight and a gestational age of a least 37 complete weeks."	Public Health Outcome Framework
Breast feeding prevalence	"This is the percentage of infants that are totally or partially breastfed at age 6-8 weeks."	Public Health Outcome Framework
Excess weight in children	No definition available in the source document	Public Health Outcome Framework
Children in poverty	"the percentage of dependent children aged under 20 in relative poverty (living in households where income is less than 60 per cent of median household income before housing costs)"	Public Health England

[13] Wirth, R. & Hipp, J., 2000. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. pp. 29–39.

[14] http://www.phoutcomes.info/

[15] Scheffer, J., 2002. Dealing with missing data.

BOX I. MODELLING TECHNIQUES EMPLOYED FOR THE STUDY

For each criterion (i.e. parameter):

1. Aggregate the values for each of the cities in the region to arrive a MEAN (and STDDEV) value for a region;

2. Find the difference between various regions and whether it is statistically significant using ANOVA (PROC ANOVA in SAS). In particular, this will highlight the best performing regions and how they are compared to other regions.

3. Clustering analysis is considered to group regions together based on similarity features using decision trees.

B. Implementation Details

Data Mining Tools: Statistical Analysis Software (SAS enterprise guide 7.1) is chosen because: it supports large number of datasets; predominantly used by UK Business users and government. However, the drawbacks of SAS are: it is expensive because it is a proprietary software though it is provided free for academic use.

C. SAS BI Implementation Pipeline and Setups

1) ***Library setup and data upload*****:** Create a library object called "INEQLITY" using the code:
LIBNAME INEQLITY "/home/n.thakkar9755/sasuser.v94";

The library is placed under auto execute so that it prompts for execution of library each time, when the user logs in. The excel file named premature mortality and healthcare.xlsx (the one downloaded from original source) is modified using filter and divided into five different datasheets, namely, *infant mortality, breastfeeding prevalence, low birth weight, excess weight, and children in poverty*. Subsequently, the excel files for the datasets are uploaded to INEQLITY. See the code in Box 2. The child inequality name is given to the xlsx file.

BOX II. SAS UPLOAD OF AN EXCEL FILE

```
File>Import Data>source Data>Dissertation Data>Premature
mortality and health care-data> child_inequality.xlsx
```

Following this, datasets are created. For example, datasets "Infant_mortality_2011" is created (i.e. using the code in Box 3) from the child_inequality.xlsx file. Following is the code used to produce this dataset and a sample of the dataset infant_mortality_2011 is depicted in Table 2.

BOX III. CREATION OF DATASET

```
data infant_mortality_2011;
set ineqlity.child_inequality;
keep Indicator 'Time Period'n Value
'Parent Code'n 'Parent Name'n;
if 'Time period'n eq '2011';
run;
```

TABLE II. DATASET INFANT_MORTALITY_2011

	Indicator	**Time Period**	**Parent Code**	**Parent Name**	**Value**
1	4.01 Infant Mortality	2011	E12000001	North East Region	26857654421
2	4.01 Infant Mortality	2011	E12000002	North West Region	28991509629
3	4.01 Infant Mortality	2011	E12000003	Yorkshire and Humber Region	36078845477
4	4.01 Infant Mortality	2011	E12000004	East Midlands Region	47353760446

Similar procedure is repeated for other datasets relating to low birth weight, breastfeeding prevalence, etc.

2) ***Conduct Analyses on the datasets*****:** Details of the various techniques are shown in Table 3.

TABLE III. DATA ANALYSES

Technique	**Description**
Statistical Analysis	Basic statistical analysis such as Mean and Standard deviation.
Anova PROC ANOVA	ANOVA is used to compare the statistical difference among the mean values of three or more different groups.[16]
Clustering Analysis PROC CLUSTER	Cluster analysis divides data objects into different groups based only on the information found in the data that describes the objects and their relationships.[17]

Various types of clustering procedures are available in SAS (e.g. proc CLUSTER, proc FASTCLUS, proc MODECLUS, proc VARCLUS, and proc TREE).[18] The Procedure CLUSTER is used in this study because of the following advantages (ibid): easier to use; can produce as many clusters as possible in one run; produces an output which could be fed into the Procedure TREE to produce dendogram.

3) ***SAS Visualisation*****:** The visualisation technique used in presenting the results is found in Table 4.

TABLE IV. VISUALISATION TECHNIQUE

Technique	**Description**
Bar chart with PROC GCHART	The bar chart is one of the easiest and most frequently used graph.

IV. RESULTS

To reiterate, this study considers an investigation on the five parameters (i.e. infant mortality, low birth weight, breastfeeding prevalence, excess weight in children, children in poverty) for six regions of England (South West, North East, South East, London, East of England, north west,

[16] Wallenstein, S., Zucker, C.L. & Fleiss, J.L., 1980. Some statistical methods useful in circulation research. Circulation Research, 47(1), pp.1–9.

[17] Kaufman, L. & Rousseeuw, P.J., 2009. Finding groups in data: an introduction to cluster analysis, John Wiley & Sons.

[18] Elliott, A.C. & Woodward, W.A., 2010. *SAS essentials: a guide to mastering SAS for research*, John Wiley & Sons.

Yorkshire and Humber, East midlands, and West midlands). However, details of the analysis of only one parameter (see Box 1) are discussed followed by a summary of the overall analysis.

A. Descriptive Statistical Analysis for Low Birth Weight

The coding to generate the results in Table 5 and Fig. 2 is found in Thakkar (2015). Results reveal that for the year 2011, South East region has the lowest average birth weight (i.e. 2.47) while London has the highest value of 3.13. The four regions that are not performing well for Low birth weight are: South East, South West, East of England, and North East regions.

TABLE V. LOW WEIGHT AT BIRTH REPORT FOR YEAR 2011

Obs	Region_Name	Mean_value	Std_deviation_value
1	South East region	2.47	0.49
2	South West region	2.51	0.37
3	East of England region	2.64	0.91
4	North East region	2.65	0.58
5	North West region	2.90	0.71
6	East Midlands region	2.96	0.77
7	Yorkshire and the Humber region	2.97	0.77
8	West Midlands region	3.06	0.81
9	London region	3.13	0.65

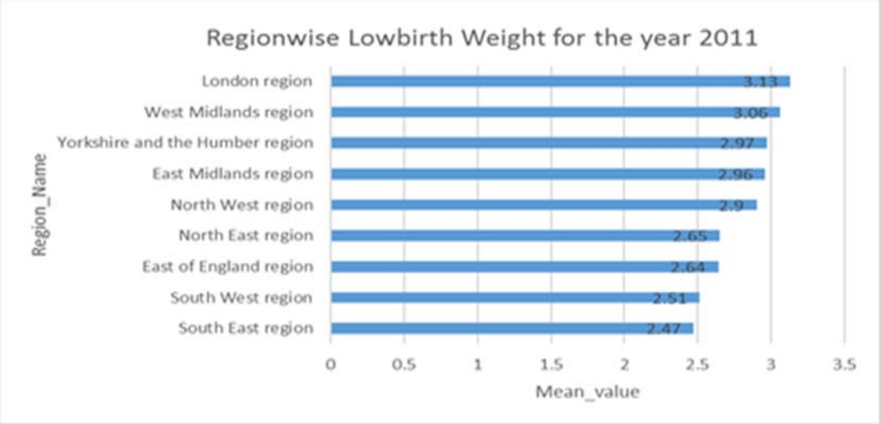

Fig. 2. Results for low birth weight.

B. Inferential Statistical Analysis for Low Birth Weight

The SAS procedure Proc Anova is repeated to find out if there is statistical significant difference for low birth weight in different English regions. The null hypothesis and alternate hypothesis are assumed as follows:

H_0 (null hypothesis): $\mu_1 = \mu_2 = \ldots = \mu_9$, Means of Low weight at birth values of different regions are equal.

H_a (alternative hypothesis): $\mu_i \neq \mu_j$: where i and j are any pair of numbers from 1 to 9 and at least two low weight at birth mean values of different regions are not equal.

Tables 6 and 7 shows the ANOVA analysis results for Low Birth Weight and the nine different English regions. The level of confidence for the test, α, is 0.05. The p-value (< 0.0001) is less than α value and thus, the null hypothesis is rejected. This implies that the Low Birth Weight means for the nine regions and significantly different.

C. Clustering Analysis

The SAS Clustering technique facilitates grouping of objects (i.e. nine regions in this study) based on values (i.e. five parameters) or data that describe the objects and their relationships. The greater the similarity within a group and the greater the difference between groups, the better or more distinct is the clustering.[19] The procedure PROC CLUSTER is used for this analysis. The centroid method is used for analysis purpose and it is similar to the medoid method discussed in [12].

1) ***Data preparation*:** The mean value of each parameter (i.e. Infant mortality, Low birth weight, Excess weight in children, Breastfeeding prevalence and Children in poverty factors) is collated for each region (see Table 8). To simplify the coding each region is given a numeric identifier (i.e. 1-South West, 2-North East, 3-South East, 4-London, 5-East of England, 6-North West, 7-Yorkshire and Humberside, 8-East Midlands, 9-West Midlands).

TABLE VI. ONE-WAY ANOVA ANALYSIS OF LOWBIRTH WEIGHT FOR YEAR 2011

Comparisons significant at the 0.05 level are indicated by * **				
Region_Name Comparison	**Difference between means**	**Simultaneous 95% Confidence Limits**		**Sig.**
4-2	1.88	1.88	1.88	***
4-6	3.57	3.57	3.57	***
4-9	4.38	4.38	4.38	***
4-7	5.07	5.07	5.07	***
4-8	6.94	6.94	6.94	***
4-5	8.20	8.20	8.20	***
4-3	10.27	10.27	10.27	***
4-1	10.85	10.85	10.85	***

[19] Kaufman, L. & Rousseeuw, P.J., 2009. Finding groups in data: an introduction to cluster analysis, John Wiley & Sons.

TABLE VII. ONE WAY ANOVA FOR LOW BIRTH WEIGHT

The ANOVA Procedure					
Dependent Variable: lowbirth_weight					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr>F**
Model	8	217.9028444	27.2378556	2.16E15	<0001
Error	9	0.0000000	0.0000000		
Corrected Total	17	217.9028444			

TABLE VIII. MEANS INPUT FOR THE CLUSTER PROCESS (I.E. EVERY PARAMETER FOR EACH REGION)

Region Number	**Region Name**	**Infant Mortality**	**Excess Weight**	**Low Birth Weight**	**Breastfeeding Prevalence**	**Children in Poverty**
1	South West	3.18	22.62	2.51	49.05	15.36
2	North East	3.29	24.48	2.68	29.45	24.33
3	South East	3.43	20.86	2.47	52.23	15.94
4	London	3.66	23.13	3.13	68.72	26.21
5	East of England	3.85	22.22	2.64	45.92	18.01
6	North West	4.12	22.92	2.90	33.48	22.64
7	Yorkshire and Humberside	4.18	22.30	2.97	37.26	21.14
8	East Midlands	4.24	22.52	2.96	44.13	19.27
9	West Midlands	5.06	23.46	3.06	38.86	21.83

Feature Scaling**:** The values in Table 8 vary greatly which is inappropriate for clustering techniques that rely on similarity of values. Hence, values need to be normalised such that all values can be represented in the range [0 – 1]. This process is often referred as feature scaling [1]. In this particular situation, all the mean values are divided by its respective highest column value. This means that normalisation is applied to all the data values. The second feature scaling that is required is to carry out adjusting the semantics and numeric representation of the values. For all the parameters, the lower the value, the better the result except for the "Breastfeeding prevalence" criterion, where a lower value indicates a better result. To unify semantics, values for all parameters (except for Breastfeeding prevalence are subtracted from 1). As for the Breastfeeding prevalence, the normalised values are retained because the higher the value, the better it is. The normalised and scaled values are tabulated in Table 9.

Fig. 3 shows a sample SAS procedure for a 4-cluster solution for nine different regions. Fig. 4 and Table 11 shows the output of the clustering process with the regions being divided into 4 clusters and summarised in Table 10. The figure also represents the dendogram created by a procedure tree.

TABLE IX. NORMALISED AND SCALED VALUES

Region Number	**Region Name**	**Infant Mortality**	**Excess Weight**	**Low Birth Weight**	**Breastfeeding Prevalence**	**Children in Poverty**
1	South West	0.37	0.07	0.20	0.71	0.41
2	North East	0.35	0.00	0.14	0.43	0.07
3	South East	0.32	0.15	0.21	0.76	0.39
4	London	0.28	0.06	0.00	1.00	0.00
5	East of England	0.24	0.09	0.16	0.67	0.31
6	North West	0.19	0.06	0.05	0.49	0.14
7	Yorkshire and Humberside	0.17	0.83	0.05	0.54	0.19
8	East Midlands	0.16	0.83	0.05	0.64	0.26
9	West Midlands	0.00	0.83	0.02	0.64	0.17

TABLE X. 4-CLUSTERS OF REGIONS

Cluster	Region/s
Cluster #1	North West, Yorkshire and Humber, East Midlands, West Midlands
Cluster #2	South West, South East and East of England
Cluster #3	North East
Cluster #4	London

V. DISCUSSION OF FINDINGS AND RECOMMENDATIONS

The case used for this data analytics has one primary central aim:

"To study how SAS techniques can be applied to analyse possible health inequalities between various regions of England based on open data". It is translated to the following question: *"Will you be at disadvantage from health perspective if you were born in one region of the England compared to other region?"*

A. Summary of Findings

The findings reveal that as per 2011 data snapshot:

- "Health inequality exists in English regions".
- "West Midland, East Midland, Yorkshire and Humber and North West are in the wrong side of the health inequality in England".
- "South East, South West, North East and East of England are on the right side of the health inequality in England".
- "London has neutral results in terms of health inequality".

B. Recommendations

There are two types of recommendations. Firstly, what policy actions can be taken based on the findings and secondly, what further research can be carried out.

1) Policy recommendations: Findings presented in this paper can be used as an evidence of health inequality that should feed into increased resource allocation in the health deprived areas.

2) Health promotions in schools.

3) Active NHS's role in tackling health inequality.

4) Promote nutrition awareness campaign.

5) Specific health inequalities initiatives: Government could tackle health inequality by targeting on worst performing areas reported in Table 5.

```
data forclustering_2011;
input Region_name
       Infant_mortality
       excess_weight
       lowbirth_weight
       breastfeeding_prevalence
       children_in_poverty@@;
/*
       1 = "South West region"
       2= "North East region"
       3= South East region;
       4= London region;
       5=East of England region;
       6=North West region;
       7=Yorkshire and the Humber region;
       8=East Midlands region;
       9=West Midlands region ;
*/
datalines;
1 0.37 0.07 0.20 0.71 0.41
2 0.35 0.00 0.14 0.43 0.07
3 0.32 0.15 0.21 0.76 0.39
4 0.28 0.06 0.00 1.00 0.00
5 0.24 0.09 0.16 0.67 0.31
6 0.19 0.84 0.07 0.49 0.14
7 0.17 0.83 0.05 0.54 0.19
8 0.16 0.83 0.05 0.64 0.26
9 0.00 0.83 0.02 0.64 0.17
;

proc cluster noeigen simple method=centroid
rmsstd rsquare nonorm out=tree;
id Region_name;
var Infant_mortality excess_weight
lowbirth_weight breastfeeding_prevelence
children_in_poverty;
run;

proc tree =tree out=clus3 nclusters=4;
id Region_name;
copy Infant_mortality excess_weight
lowbirth_weight breastfeeding_prevelence
children_in_poverty;

proc sort; by cluster;
proc print; by cluster;
var Region_name Infant_mortality
excess_weight lowbirth_weight
breastfeeding_prevelence
children_in_poverty;
title2 '4-cluster solution';
run;
```

Fig. 3. SAS Procedure for a 4-Cluster solution for the nine regions.

TABLE XI. CLUSTER SOLUTIONS FOR THE NINE REGIONS (PART 1)

CLUSTER=1						
Obs	**Region_name**	**Infant_mortality**	**excess_weight**	**lowbirth_weight**	**breastfeeding_prevalence**	**children_in_poverty**
1	6	0.19	0.84	0.07	0.49	0.14
2	7	0.17	0.83	0.05	0.54	0.19
3	8	0.16	0.83	0.05	0.64	0.26
4	9	0.00	0.83	0.02	0.64	0.17
CLUSTER=2						
Obs	**Region_name**	**Infant_mortality**	**excess_weight**	**lowbirth_weight**	**breastfeeding_prevalence**	**children_in_poverty**
5	1	0.37	0.07	0.20	0.71	0.41
6	3	0.32	0.15	0.21	0.76	0.39
7	5	0.24	0.09	0.16	0.67	0.31
CLUSTER=3						
Obs	**Region_name**	**Infant_mortality**	**excess_weight**	**lowbirth_weight**	**breastfeeding_prevalence**	**children_in_poverty**
8	2	0.35	0.00	0.14	0.43	0.07
CLUSTER=4						
Obs	**Region_name**	**Infant_mortality**	**excess_weight**	**lowbirth_weight**	**breastfeeding_prevalence**	**children_in_poverty**
9	4	0.28	0.06	0.00	1.00	0.00

Note: 1 South West, 2 North East, 3 South East, 4 London, 5 East of England, 6 North West, 7 Yorkshire and Humberside, 8 East Midlands, 9 West Midlands

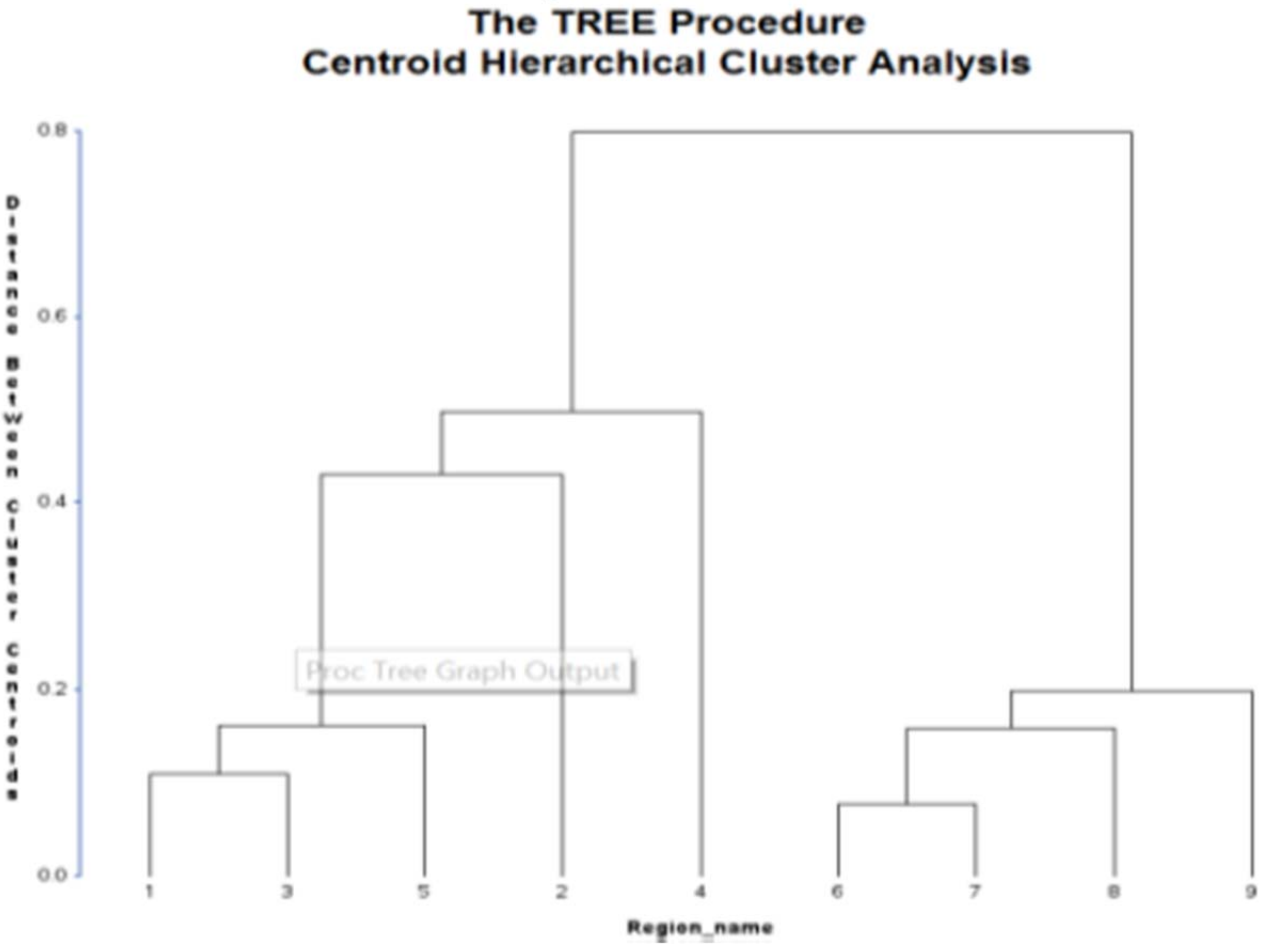

Fig. 4. 4-Cluster solution for the nine regions.

1 South West, 2 North East, 3 South East, 4 London, 5 East of England, 6 North West, 7 Yorkshire and Humberside, 8 East Midlands, 9 West Midlands

VI. Conclusion

Health inequality is an age old problem. This study aims investigate health inequality based on open data (from Public Health Outcomes Framework) and using SAS as analytics tool. PHO dataset support all the five analytical parameters required for this work. A survey for relevant methodology for data analytics has been conducted. For the purpose of this study, the CRISP-DM methodology has been selected because it is very systematic. Using its recommended steps, data is prepared for modelling. Modelling using various SAS techniques includes the following: Mean and Standard deviation, Anova, and Clustering analysis. Applying means, standard deviation, and Clustering reveal a very interesting pattern from the data. The clustering results validate initial observations which are aligned with the summary of findings listed in Section 6.1.

References

[1] Aksoy, S. & Haralick, R.M., 2001. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern recognition letters*, 22(5), pp.563–582.

[2] Bartley, M., 2004. Health inequality: An introduction to theories, concepts and methods.

[3] Calling, S. et al., 2011. Socioeconomic inequalities and infant mortality of 46,470 preterm infants born in Sweden between 1992 and 2006. *Paediatric and perinatal epidemiology*, 25(4), pp.357–65. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21649678 [Accessed September 16, 2015].

[4] Currie, C., 2008. *Inequalities in young people's health: HBSC international report from the 2005/2006 Survey*, World Health Organization.

[5] Curtis, S. & Jones, I.R., 1998. Is there a place for geography in the analysis of health inequality? *Sociology of health & illness*. Available at: http://onlinelibrary.wiley.com/doi/10.1111/1467-9566.00123/pdf [Accessed September 16, 2015].

[6] Diderichsen, F. et al., 2012. Health inequality--determinants and policies. *Scandinavian journal of public health*, 40(8 Suppl), pp.12–105. Available at: http://sjp.sagepub.com/content/40/8_suppl/12.short [Accessed September 14, 2015].

[7] González, S., 2010. The North/South divide in Italy and England: Discursive construction of regional inequality. *European Urban and Regional Studies*.

[8] Gurstein, M.B., 2011. Open data: Empowering the empowered or effective data use for everyone? *First Monday*, 16(2).

[9] Hargreaves, D.S., Djafari Marbini, A. & Viner, R.M., 2013. Inequality trends in health and future health risk among English children and young people, 1999-2009. *Archives of disease in childhood*, 98(11), pp.850–5. Available at: http://adc.bmj.com/content/early/2013/05/17/archdischild-2012-303403.abstract [Accessed September 16, 2015]. 69

[10] Huijboom, N. & den Broek, T., 2011. Open data: an international comparison of strategies. *European journal of ePractice*, 12(1), pp.4–16.

[11] Janevic, T. et al., 2010. Neighborhood deprivation and adverse birth outcomes among diverse ethnic groups. *Annals of epidemiology*, 20(6), pp.445–51. Available at: http://www.sciencedirect.com/science/article/pii/S1047279710000566 [Accessed September 16, 2015].

[12] Kaufman, L. & Rousseeuw, P.J., 2009. *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons.

[13] Kawachi, I., et. al. (1997). Social capital, income inequality, and mortality, Am J Public Health, v.87(9); Sep 1997, PMC1380975

[14] Kihal-Talantikite, W. et al., 2013. Green space, social inequalities and neonatal mortality in France. *BMC pregnancy and childbirth*, 13(1), p.191.

[15] Matthews, T.J., MacDorman, M.F. & others, 2012. Infant mortality statistics from the 2008 period linked birth/infant death data set. *National vital statistics reports*, 60(5).

[16] McKay, A. (2002). Defining and Measuring Inequality, url: https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/3804.pdf, accessed date: 30/4/2017.

[17] Phillimore, P., Beattie, A. & Townsend, P., 1994. Widening inequality of health in northern England, 1981-91. *Bmj*, 308(6937), pp.1125–1128.

[18] Ross, N.A. et al., 2000. Relation between income inequality and mortality in Canada and in the United States: cross sectional assessment using census data and vital statistics. *BMj*, 320(7239), pp.898–902.

[19] Sacker, A. et al., 2000. Comparing health inequality in men and women: prospective study of mortality 1986-96. *BMJ*, 320(7245), pp.1303–1307.

[20] Thakkar, N. (2015). Application of Business Intelligence Techniques using SAS on Open Data: Analysing Health Inequality in English Regions using SAS and Open Data, MSc BI unpublished dissertation, Leeds Beckett University, Leeds, UK.

[21] Thomson, R., Murtagh, M. & Khaw, F.M., 2005. Tensions in public health policy: patient engagement, evidence-based public health and health inequalities. *Quality and Safety in Health Care*, 14(6), pp.398–400.

[22] Wagstaff, A., Paci, P. & Van Doorslaer, E., 1991. On the measurement of inequalities in health. *Social science & medicine*, 33(5), pp.545–557.

[23] Wilkinson, R. and Pickett, K. (2009b) 'Income Inequality and Social Dysfunction', *Annual Review of Sociology*, 35, pp. 493–511

[24] Wirth, R. & Hipp, J., 2000. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*. pp. 29–39.