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

Ah-Lian Kor, Neha Thakkar, Sanela Lazarevski
School of Computing, Creative Technologies, and Engineering,
Leeds Beckett University, Leeds, UK
(e-mail:{A.Kor,}@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 disadvantage 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. This 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

1. INTRODUCTION

Inequality is a widely reported problem in modern day societies. González (2010) focuses on the regional divide in the UK and notes that inequality affects policy decisions in the country. The World Bank and overseas development institute (McKay, 2002) 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 (McKay, 2002; Thakkar, 2015). Health inequality is defined as "*difference in people or groups due to social, biological, geographical or other factors*" (Bartley 2004). 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 (e.g. Phillimore et al., 1994; Ross et al. 2000; Sacker et al. 2000; Currie, 2008). 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 is 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 (Gurstein, 2011). Additionally, it provides a means for citizen engagement and used as a vehicle for transparency and efficiency (Huijboom & Den Broek, 2011). 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

1.1 Aims and Objectives

SAS is one of the most popular and powerful data analytics software in the market (Elliott & Woodward 2010). 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.

1.2 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.

2. LITERATURE REVIEW

Health inequality
https://www.publications.parliament.uk/pa/cm200809/cmselect/cmhealth/286/286.pdf

Undeniably, health inequality is an age old problem. Work dated back to the 1930s (Wagstaff, et al. 1991) reveals evidence of health inequality in the UK. Although there is a lack of consensus over how health inequality is measured (Bartley, 2004), 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.

2.1 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, improved efficiency[4,5]. If data analytics are conducted correctly, then there is a huge potential for policy making[6].

According to Gurstein (2011), the UK and Canada are among some of the countries that are at the forefront of Open Data Initiative (see[7,8,9]) where citizen related data is made public (Davies, 2010, Gurstein, 2011). The UK government issue a code of practice[10] 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)[11]. One of the counties, 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).

2.2 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.

http://www.who.int/rpc/meetings/en/Hunter_Killoran_Report.pdf

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2724448/

https://www.shu.ac.uk/~/media/home/research/hccj/files/reports/tacklinghealthinequalitiesreport.pdf

http://jech.bmj.com/content/58/10/811

https://www.rcpe.ac.uk/sites/default/files/smith_1.pdf

The work carried out in this project 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. My work is part of the vital work in this area especially there is a great concern that reduced budget in "NHS is expected to transfer healthcare funding away from younger, more deprived areas to older, more affluent ones" (Hargreaves et al. 2013).

2.3 Health Inequality Measurements

[4] http://opendatatoolkit.worldbank.org/en/starting.html
[5] https://www.publications.parliament.uk/pa/cm201314/cmselect/cmpubadm/564/564.pdf

[6] https://www.civilserviceworld.com/articles/opinion/big-data-get-it-right-and-benefits-policy-making-could-be-huge
[7] http://data.gov.uk/
[8] http://www.data.gov/
[9] https://openparliament.ca
[10] https://data.gov.uk/consultation/code-of-practice
[11] http://healthpolicy.ucla.edu/chis/Pages/default.aspx

A framework for measuring health inequality
http://www.who.int/healthinfo/paper05.pdf

A framework for measuring health inequity
https://www.ncbi.nlm.nih.gov/pubmed/16020649

Measuring Health Inequalities
http://www.yhpho.org.uk/resource/view.aspx?RID=9957

Measures of health inequalities: part 1
http://jech.bmj.com/content/58/10/858

Measures of health inequalities: part 2
http://jech.bmj.com/content/58/11/900

A three-stage approach to measuring health inequalities and inequities
https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-014-0098-y

Expert Review and Proposals for Measurement of Health Inequalities in the European Union
http://ec.europa.eu/health/sites/health/files/social_determinants/docs/full_quantos_en.pdf

Health inequalities and population health
https://www.nice.org.uk/advice/lgb4/chapter/introduction

Defining and measuring health inequality: an approach based on the distribution of health expectancy
http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S0042-96862000000100005

There are various possible parameters that can be used to measure health inequality. Work by (Kawachi et al., 1997) establish link between income inequality and health, where inequality is linked to mortality rates. Similarly, (Wilkinson and Pickett, 2010) found link between inequality and health and social problems. The widely cited work by (Bartley, 2004), define 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 geography while measuring the health inequality. For example, work utilising cross-country comparisons (Wagstaff et al. 1991). Authors in (Curtis & Jones ,1998) considers 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 area of high health inequality that has higher chances of having impact on your health. In another important research similar to this, authors study the effect of environment amnesties in France, for example, green spaces and its effect on health inequality and they find correlation between the two (Kihal-Talantikite, et al. 2013).

My work uses existing research in this area that highlights the health inequality parameters when it comes to children, including few significant works (Mathews et al. 2012) (Calling, et al. 2011) (Janevic, et al. 2010) (Hargreaves, et al. 2013) infant mortality statistics that lists following parameters important for measuring health inequality for children:

- Preterm birth
- Low birth weight
- Infant mortality
- Excessive weight in children
- Breastfeeding prevalence and
- Children in poverty

They link these criteria 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 (Diderichsen, et al. 2012). Hence, when the dataset is searched for the project, support for the aforementioned six criteria in dataset will be an important factor for selection of datasets for the project.

3. PROGRAMME

3.1 Research Aim and Objectives

4. METHODOLOGY

5. UNDERLYING COMPONENTS FOR THE PROPOSED CLOUD SERVICE ECOSYSTEM

5.1 Energy Software

5.2 Programming Model, Program Construction, and Incorporation of Energy Metrics

5.3 Service Composition, Deployment, Operation, and Evaluation

5.5 Environmental Impact

6. CONCLUSIONS

REFERENCES

[1]. Agarwal, V., Chafle, G., Mittal, S., and Srivastava, B. (2008). Understanding approaches for web service composition and execution, In Proceedings of the 1st Bangalore Annual Compute Conference (Bangalore, India, January 18 - 20, 2008); COMPUTE '08. ACM, New York, NY.

[2]. Beloglazov, A., and Buyya, R. (2010). Energy Efficient Allocation of Virtual Machines in Cloud Data Centers. Proceedings of the 10th IEEE/ACM

International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2010), Melbourne, Australia, May 17-20, 2010.

[3]. Dong, W. L., and Jiao, L. (2008). QoS-Aware Web Service Composition Based on SLA, Proceedings of the Fourth International Conference on Natural Computation,Vol.5,pp.247-251.

[4]. Mittinen, A. P., and Nurminen, J. K. (2010.). Energy efficiency of mobile clients in cloud computing, Proceedings of HotCloud'10, USENIX Association Berkeley, CA, USA.

[5]. Srikantaiah, S., Kansal, A., and Zhao, F. (2008). Energy Aware Consolidation for Cloud Computing, Proceedings of the 2008 Conference on Power Aware Computing and Systems(HotPower'08). Berkeley, CA, USA.

[6]. Srivastava, B., and Koehler, J. (2003). Web Service Composition - Current Solutions and Open Problems, ICAPS 2003 Workshop on Planning for Web Services, pp. 28-35,

[7]. Yan, J., et. al (2007). Autonomous service level agreement negotiation for service composition provision, Future Generation Computer Systems, Volume 23, Issue 6, pp.748-759.