A Graphical Data Visualisation Approach to Assessing Associations and Variations of the Impact of COVID-19

MWITONDI, Kassim, RAMAMURTHY, Anandi <http://orcid.org/0000-0002-4830-9029> and GUMBER, Anil <http://orcid.org/0000-0002-8621-6966>

A Graphical Data Visualisation Approach to Assessing Associations and Variations of the Impact of COVID-19

Kassim S. Mwitondi[1], Anandi Ramamurthy[2], and Anil Gumber[3]

[1]Sheffield Hallam University, Department of Computing (k.mwitondi@shu.ac.uk)
[2]Sheffield Hallam University, Dept of Media Arts & Communication (a.ramamurthy@shu.ac.uk)
[3]Sheffield Hallam University, Faculty of Health and Human Wellbeing (a.gumber@shu.ac.uk)

Abstract

The COVID–19 pandemic, like other global phenomena such as climate change, food security and democracy and human rights, arise from within confined geographical and legislative boundaries. Tackling such challenges entails robust and sustainable collaborative initiatives–a natural challenge, given the geo–political, socio–economic and cultural variations across countries and regions. Any initiatives to address the challenges inevitably evolve around data modelling–typically, uncovering the associations and variations across geo–political and socio–economic and cultural variations of our societies. This view aligns with the non–orthogonal nature of the 17 Sustainable Development Goals (SDG), initiated by the United Nations in 2015. We present a data–driven approach to assessing the associations and variations of the impact of the COVID–19 pandemic in the UK, across ethnic lines of selected health and care workers. The main motivation is that the exercise cuts across multiple SDGs and as such it highlights the impact of different data attributes and how they interact. While individual clinicians have tracked deaths through the press and social media to gain an understanding of which groups of health workers were particularly affected, gaps remain as to how the distribution of the impact of the pandemic actually was. This paper seeks to paint the overall picture of which health care workers were impacted based on those data attributes. Data was obtained from a nation–wide survey involving 380 responses from health and care workers on a range of demographic characteristics such as age, sex, ethnicity, job role and personal views on how they felt in different situations. The emerging graphical patterns provide insights into the way the health and care workers staff were impacted along those attributes. Our findings provide insights into what happens within a particular domain–country, sector or an individual SDG. Its is expected that the study will highlights the impact of the events on other SDGs and promote further collaboration.

Key Words: *Association Rules, BAME, Chi-Square Distribution, Correspondence Analysis, Data Visualisation, Sustainable Development Goals*

1 Introduction

COVID–19 invariably impacted societies across the globe in many ways–death rates, business slow down, economic effect, education impact, mental health, food security, human rights, etc, [1] all hinging on the 17 Sustainable Development Goals (SDG), a global initiative launched by the United Nations in 2015 [2]. The SDGs present a set of highly correlated data attributes spanning across the entire spectrum of human existence [3, 4]. Understanding those interactions and dynamics, goes through a number of interdisciplinary phases–from problem definition right through the solution. In other words, it cuts across multiple SDGs and, as such, it highlights the impact of different data attributes and how they interact. This work hinges more specifically on SDG#16–that seeks to promote just, peaceful and inclusive societies. There is enough empirical evidence to show that it plays a pivotal role in invigorating development in other SDGs. The impact of COVID–19 on societies has been widely studied since the pandemic and, in the last couple of years, we have seen a number of reports and research publications on the foregoing attributes and more [5, 6]. While the decisions on how to deal with the pandemic varied across geographical and legislative boundaries, everywhere, exposed to the brunt of the pandemic, were front line workers. There has since been a substantial public interest and concern that this should be understood. This interest led individual clinicians to try to track deaths through the press and social media to gain an understanding of which groups of health workers were particularly affected [7, 8].

Although the Public Health England (PHE) report [9] on the impact of BAME groups makes clear that there are disproportionate deaths in the wider population, there is currently no analysis of a consolidated list of all staff deaths in the health and social care sector. A lancet study identified higher risk of COVID–19 infections amongst health care workers. There are differing opinions regarding the extent of disproportionate deaths amongst ethnic minority staff in this sector. To date there is no overall assessment of deaths of health care workers by the key demographic characteristics such as age, sex, ethnicity and job role. Concerns over disproportionate impact of COVID–19 demand that we should assess such data in order to be able to consider the differing conclusion of the PHE Report on BAME deaths and the analysis by the Commission on Race and Ethnic Disparities [10]. The commission focused on areas including poverty, education, employment, health and the criminal justice system. While its commissioning was triggered by the tragic death of George Floyd in the US, at the hands of the Police, it coincided with the impact the pandemic was having on societies across the UK. It is therefore interesting to try and find out what we can learn by analysing the overall impact of health care worker deaths by ethnicity, gender, age and geographical location.

This paper presents a data–driven approach to assessing the associations and variations of the impact of the COVID–19 pandemic in the UK, across ethnic lines of sampled health and care workers. It seeks to paint the overall picture of how health care workers were impacted based on those data attributes. Data was obtained from a nation–wide survey involving 380 responses from health and care workers on a range of demographic characteristics such as age, gender, ethnicity, job role and personal views on how they felt in different situations. Equipped with the combination of domain knowledge, data, tools and skills, we can focus on decoding and interpreting different sources of data across the SDG spectrum [11, 12], including gaining insights into how the COVID–19 battle was fought. This study does not seek to identify individuals but to understand the overall picture of which health care workers died by ethnicity, age, sex, and geographic location. It presents not only a substantial public interest and concern, but also a moral imperative to understand in order to recognise the impacts on particular communities in order to ensure that these impacts and losses are recognised and addressed and that disproportionate impact can be avoided as much as possible in the future.

1.1 Motivation and Analysis Setup

The paper was motivated by several factors, all revolving around knowledge extraction from data, with a focus on the complex interactions of the SDGs and the dynamics of their inherent data attributes. The coincidence of the impact of COVID–19 on our society and the commissioning of the Commission on Race and Ethnic Disparities report [10] on ethnic disparities and inequality in the UK, was a key motivating factor. The motivation was amplified by the fact that both phenomena hinge on SDGs–each of which is a potential source of Big Data [13], from which we seek uncover potentially useful knowledge. Gaining access to knowledge hidden in the data attributes is associated with issues around data randomness [14], thus our analysis setup is two-fold, soft and technical. The former relates to the legitimate interest from both health care workers and within the wider society to understand which groups of health care workers suffered loss of life due to COVID–19, while the latter relates to the methodological approach to uncovering such information. In both cases, such information remains buried deep in data attributes, some of which are yet to be formally identified, which we can do by focusing on identification of subtle, meaningful data.

1.2 Problem Space and Objectives

Zhang and Zhang [15] presented a framework of quotient space theory of problem solving in which they describe problem space "...as a triplet, including the universe, its structure and attributes." We describe our problem space in line with this description, and in juxtaposition with the foregoing motivation and analysis setup. Our universe is the entire set of the 17 SDGs, the structure is the embodiment of their interactions and the attributes are triggers of attainment of the targets. These triggers are typically unknown; they are largely buried in data and in many forms of grey literature [4]–hence, they need to be uncovered. Generally, uncovering them amounts to striking a balance between two concepts associated with data randomness–masking and swamping [16, 17, 14]. The former describes a situation where meaningful data is missed and the latter arises when random data is accepted as meaningful.

In the context of the foregoing motivation and analysis setup, interest is in how random or systematic the impact of COVID–19 was on our society. While individual clinicians have tracked deaths through the press and social media to gain an understanding of which groups of health workers were particularly affected [7, 8], gaps remain as to how the distribution of the impact of the pandemic actually was. Our contribution to filling such gaps derives from an approach that is underpinned by underlying societal demographics on the one hand and some theoretical aspects of data science

and their manifestation (data visualisation) on the other. We seek to answer the following question: **How equitable was the battle fought by health and care workers during the pandemic?** We set the following objectives.

1. To explore relevant literature on key socio-demographic conditions
2. To collect COVID–19 related data attributes
3. To prepare data for modelling
4. To carry out EDA and further visualisation for pattern recognition
5. To interpret emerging patterns
6. To make recommendations to relevant bodies...

The problem and objectives, defined above, align with communicating research findings to both technical and non-technical audience. In the next exposition, we outline the methods adopted for attaining this goal.

2 Methodology

This section presents the implementation phases–from data acquisition to implementation strategy. The overall framework blends quantitative and qualitative approaches in addressing the foregoing problem space via the set objectives. Its outcome, detailed in Section 3, are based on the current sample and the adopted methods.

2.1 Data Sources

Data was obtained from a nation–wide survey involving 380 responses from health and care workers on a range of demographic characteristics such as age, gender, ethnicity, job role and personal views on how they felt in different situations. The survey was open to all health and social care workers. They were all informed that it was anonymous and that its purpose was to create a manifesto for change. They were also advised that participation was voluntary, they could withdraw anytime, before pressing the submit button, and their data would be deleted from the study.

Variables	**Description and Relevance**
BVA	Verbal abuse from staff **before** the pandemic
BPA	Physical abuse from staff **before** the pandemic
BPP	Prevented from progressing **before** the pandemic
BES	Excessive scrutiny and/or punishment **before** the pandemic
BOH	Any other harassment or exclusions **before** the pandemic
DVA	Verbal abuse from staff **during** the pandemic
DPA	Physical abuse from staff **during** the pandemic
DPP	Prevented from progressing **during** the pandemic
DES	Excessive scrutiny and/or punishment **during** the pandemic
DOH	Any other harrassments or exclusions **during** the pandemic
PPE	Distribution of PPE or Physically unsuitable PPE
Risk	Had access to risk assessment or had reasonable adjustment following risk assessment
COVIDEnvironment	Worked in COVID–19 environment
Challenged	Whether they challenged discrimination
Treated	Whether they felt unfairly treated
ImpactCode	Whether they feel COVID–19 had an impact on them
ImmigStatVuln	Fear due to immigration status vulnerability
Ethnicity	Ethnicity of respondent
Age	Age of respondent
Gender	Gender of respondent

Table 1: Recorded variables representing the health and care work experiences before and during the pandemic

The data attributes, as compiled from the survey questions are presented in Table 1, representing the pre and during pandemic experiences of health and care workers of different ethnicities, age and gender. As noted above, the data attributes were collected from different locations, recorded on the basis of employment type and setting. The setting, geographical and employer type distribution of the data sources are graphically illustrated in Figure 1.

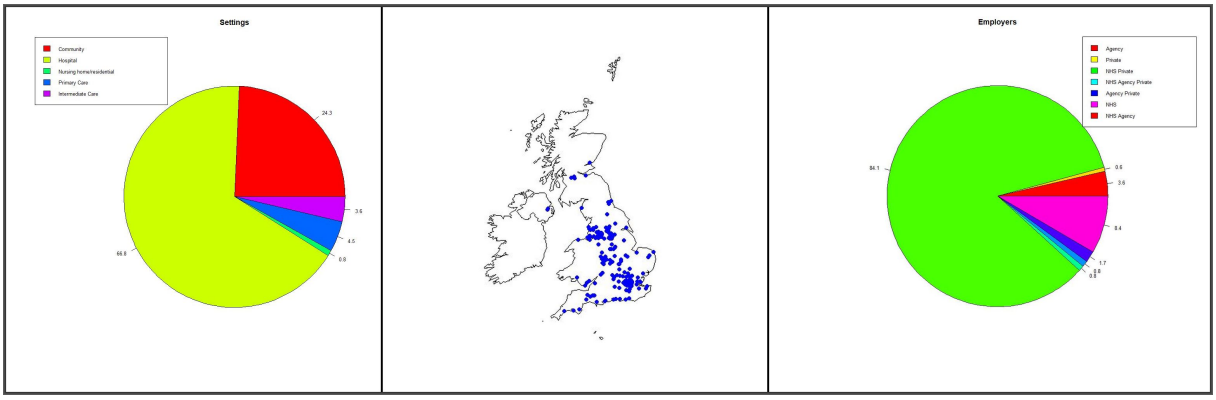

Figure 1: Graphical illustration of the distribution of the data sources

The variable **ImpactCode** in Table 1 provides different levels of impact, described in Table 2. It is reasonable to assume that these individual responses derive naturally from the staff's working experience during the pandemic. We can therefore use the variable, in combination with others, to assess equitability of the COVID–19 battle.

Impact Level	**Description**
X	Not impacted at all by the pandemic. Working conditions hadn't significantly changed
A	Mental Health
B	Sick Leave
C	Difficult to do job
D	Leave job
E	Mental Health and Sick Leave
F	Mental Health and Difficult to do job
G	Mental Health and Leave job
H	Sick Leave and Difficult to do job
I	Sick Leave and Leave job
J	Difficult to do job and Leave job
K	Mental Health, Sick Leave and Difficult to do job
L	Mental Health, Sick Leave and Leave job
M	Mental Health, Difficult to do job and Leave job
N	Sick Leave, Difficult to do job and Leave job
O	Mental Health, Sick Leave, Difficult to do job and Leave job

Table 2: Different levels of impact health and care workers felt the pandemic had on them

2.2 Modelling Strategy

Real–life data attributes take different forms–typically, numeric and categorical, the analyses which require different strategies. The data attributes in Table 1 are all categorical and as such, they take on different levels such as age groups, gender, or any other response such as whether they felt they were unfairly treated. We divide the modelling strategy in two parts–exploring associations among the data attributes and predictive modelling. The former provides insights into the underlying distribution of the features and reveals how different variables interact. We shall deploy two standard methods, Chi–Distribution [18] and Correspondence Analysis [19].

2.2.1 Measuring Associations Among Data Attributes

A good way to interpret some of the questionnaire findings is to use a two–way contingency table, for measuring associations, which it does by using the signed contribution to the Pearson's correlation, defined as

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k} \tag{1}$$

where O_k and E_k are the observed and expected counts respectively. In other words, for a two-way contingency table, Pearson's contribution to Chi–Square criterion for each data point i, j is the average of the differences

$$d_{ij} = \frac{f_{i,j} - e_{i,j}}{\sqrt{(e_{i,j})}} \tag{2}$$

where f_{ij} and e_{ij} are the observed and expected counts. Equation 1 represents a test of independence, to determine whether two categorical variables are statistically related or not. Establishing association between two attributes of interest, such as ethnicity & risk exposure, may imply that the battle for COVID–19 was not equitable. The test creates a two–way table, sorting the data according to the variables being tested and "test the hypothesis that there is no relationship between them". This is done by comparing the "actual counts" from the sample data with the "expected counts", given that the null hypothesis of no relationship is true, and it is defined as

$$\text{Expected Count} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Total Count of the sample}} \tag{3}$$

Equation 3 is an "average" measure of the opinion of people in a sample. It is the only way we can generalise our findings on how they feel, and we draw attention to the associated issues of data randomness [14].

2.2.2 Correspondence Analysis

Correspondence analysis [19], is a useful visualisation technique, particularly for uncovering associations between different categories of selected data attributes in a two–way contingency table. A typical question, in this case, is whether we can establish association between some row elements and some column elements. Its mechanics are designed to generate orthogonal components, with maximisation of variation in the data in mind. Let $\mathcal{M}$ denote a two–way matrix of size r $\times$ c, where r is the number of rows and c the number of columns, then we can compute a set of weights

$$\mathrm{w}_r = \frac{1}{\mathrm{c}_{\mathcal{M}}}\mathcal{M}\mathbf{1} = \mathrm{diag}\left[\frac{1}{\sqrt{\mathrm{w}_r}}\right] \quad \text{and} \quad \mathrm{w}_c = \frac{1}{\mathrm{c}_{\mathcal{M}}}\mathbf{1}^T\mathcal{M} = \mathrm{diag}\left[\frac{1}{\sqrt{\mathrm{w}_c}}\right] \tag{4}$$

where $\mathbf{1}$ is a univariate vector of ones and $\mathrm{c}_{\mathcal{M}}$ is the sum of the contents of $\mathcal{M}$ across rows and columns, expressed as

$$\mathrm{c}_{\mathcal{M}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\mathcal{M}_{ij} \tag{5}$$

From Equations 4 and 5, we can obtain the same two–way table transformed into proportions, $\Pi = \frac{\mathcal{M}}{\mathrm{c}_{\mathcal{M}}}$. Correspondence Analysis is an extension of Principal Component Analysis [20] and, as defined above, it presents mechanics quite similar to those of the Chi-Square, except that it provides more intuitive graphical visualisation.

2.2.3 Predictive Modelling

For predictive modelling, the categorical data attributes in Table 1 need to be transformed into numerical features and one of the most common approach is the creation of dummy variables–also known as one-hot encoding. The approach creates as many new dummy variables as there are levels in every variable. One–hot encoding makes it possible to carry out supervised modelling based on probability proportions in each variable. The major downside is the high cardinality, as a 5–level variable, say, generates 5 variables, increasing data dimension and the risk of over-fitting [21].

We can use the variables in Table 1 to predict the impact COVID–19 had on the respondees or their capability to challeng discrimination. Given labelled data, we take a Bayesian approach that allows to use existing *prior* knowledge

to learn more about the data and generate new *(posterior)*. Basically, we are interested in the *(posterior)* probability of a particular event, e.g., belonging to the k^{th} class given evidence in the data

$$f\left(y=k|x\right)=\frac{f\left(y|x\right)\pi_k}{\int\limits_{-\infty}^{\infty}f\left(y|x\right)\pi_k\mathrm{d}y}\propto\frac{f_k\left(x\right)\pi_k}{\sum_{k=1}^{K}f_k\left(x\right)\pi_k}\propto P\left(y|x\right)=\frac{P\left(x|y\right)P\left(y\right)}{P\left(x\right)} \tag{6}$$

where $f\left(x\right)$ and π_k represent the data density and prior class priors, proportional to $P\left(x|y\right)$ and $P\left(y\right)$ respectively. In real–life applications both the densities and priors are estimated from data. The analyses in Section 3 are based on Random Forests [22], a well–documented ensemble learning method for aggregating multiple decision trees[23], based on the original mechanics of decision trees[24]. Its predictions, assuming multiple trees $m>1$ and individual weight functions w_j, gives the equivalent of Equation 7 as follows

$$\hat{f}\left(\phi\right)=\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n}(x_i,x')w_i f\left(\phi\right)_i=\sum_{i=1}^{n}\left[\frac{1}{m}\sum_{j=1}^{m}(x_i,x')w_i\right]f\left(\phi\right)_i \tag{7}$$

In both cases we have to deal with the swamping and masking effects[16, 17]. The estimates can be obtained in various ways of for sampling from probability distributions, as long as the density function can be evaluated.

3 Analyses and Discussions

The findings of this research are presented in...

N			
Chi-square contribution			
N / Row Total			
N / Col Total			
N / Table Total			

Total Observations in Table: 358

survey$ImmigrationStatusVulnerability	survey$Challenged: No	Yes	Row Total
N/A	90	114	204
	0.658	0.448	
	0.441	0.559	0.570
	0.621	0.535	
	0.251	0.318	
No	25	49	74
	0.825	0.561	
	0.338	0.662	0.207
	0.172	0.230	
	0.070	0.137	
No,N/A	1	3	4
	0.237	0.162	
	0.250	0.750	0.011
	0.007	0.014	
	0.003	0.008	
Yes	29	47	76
	0.103	0.070	
	0.382	0.618	0.212
	0.200	0.221	
	0.081	0.131	
Column Total	145	213	358
	0.405	0.595	

N			
Chi-square contribution			
N / Row Total			
N / Col Total			
N / Table Total			

Total Observations in Table: 358

survey$ImmigrationStatusVulnerability	survey$Race: Colour	White	Row Total
N/A	165	39	204
	0.629	3.876	
	0.809	0.191	0.570
	0.536	0.780	
	0.461	0.109	
No	65	9	74
	0.028	0.172	
	0.878	0.122	0.207
	0.211	0.180	
	0.182	0.025	
No,N/A	3	1	4
	0.057	0.349	
	0.750	0.250	0.011
	0.010	0.020	
	0.008	0.003	
Yes	75	1	76
	1.414	8.709	
	0.987	0.013	0.212
	0.244	0.020	
	0.209	0.003	
Column Total	308	50	358
	0.860	0.140	

Figure 2: Responses touching on immigration status

We looked at other cross tabulations of the variables in Table 1 including the relationships between the distribution of PPE and exposure to risk as well as gender and risk exposure. Figure 3

N	
Chi-square contribution	
N / Row Total	
N / Col Total	
N / Table Total	

Total Observations in Table: 358

survey$PPE	survey$Risk No	Yes	Row Total
No	218	66	284
	2.298	5.181	
	0.768	0.232	0.793
	0.879	0.600	
	0.609	0.184	
Yes	30	44	74
	8.819	19.883	
	0.405	0.595	0.207
	0.121	0.400	
	0.084	0.123	
Column Total	248	110	358
	0.693	0.307	

N	
Chi-square contribution	
N / Row Total	
N / Col Total	
N / Table Total	

Total Observations in Table: 358

survey$Gender	survey$Risk No	Yes	Row Total
Female	189	84	273
	0.000	0.000	
	0.692	0.308	0.763
	0.762	0.764	
	0.528	0.235	
Male	57	26	83
	0.004	0.010	
	0.687	0.313	0.232
	0.230	0.236	
	0.159	0.073	
Other	2	0	2
	0.273	0.615	
	1.000	0.000	0.006
	0.008	0.000	
	0.006	0.000	
Column Total	248	110	358
	0.693	0.307	

Figure 3: Responses touching on the overall risk exposure

Next, we generate association plots indicating deviations from independence of rows and columns in a 2-dimensional contingency table. The left hand side panel of Figure 4 represents associations between working in COVID–19 environment and ethnicity, while the right hand side panel exhibits the relationship between gender and risk exposure. In both plots, each cell is represented by a rectangle that is assigned height proportional to the signed contribution to the Pearson's correlation in Equation 1 and width proportional to the square root of the residuals, such that the area of the box is proportional to the difference in observed and expected frequencies.

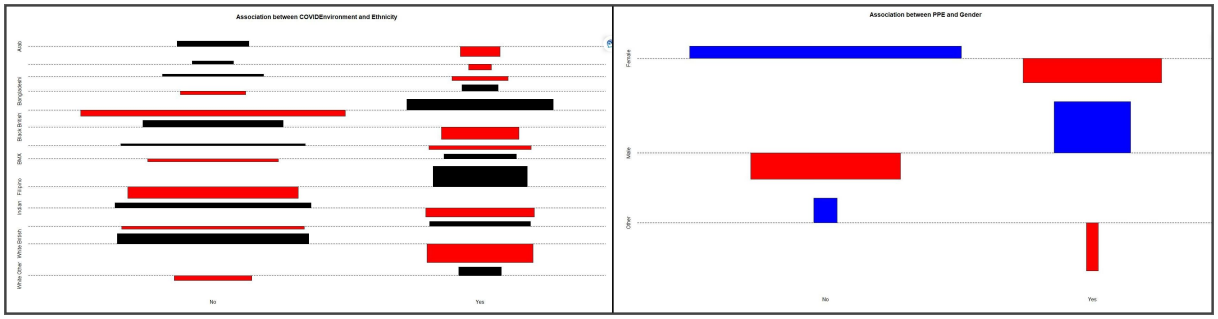

Figure 4: Associations between COVID-19 environment, Gender and Risk Exposure

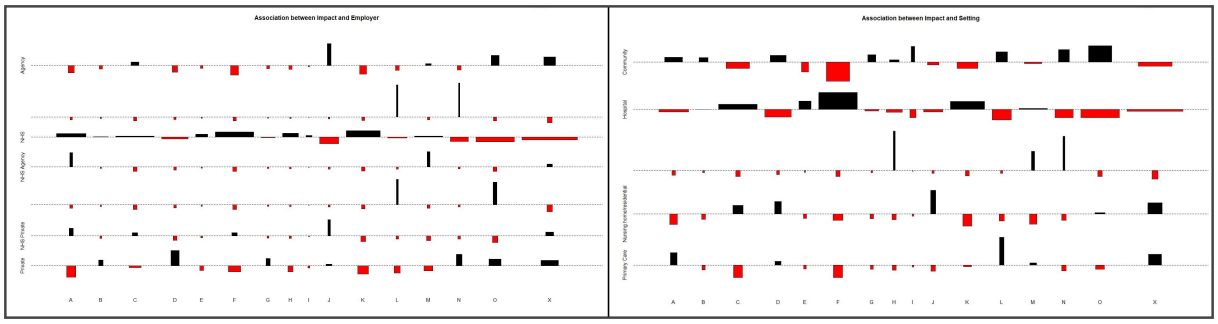

Figure 5: Associations impact and employer, settings

The rectangles in each row are positioned relative to a baseline indicating independence when the correlation is zero. If the observed frequency of a cell is greater than the expected one, the box rises above the baseline and is shaded in the colour specified by the investigator–black and blue, in this case; otherwise, the box falls below the baseline and is shaded in red. Here it can be seen that Other Asians, Bangladeshi and Other Whites, for instance, exhibited the lowest difference between observed and expected data. Notice that the closer the two parameters get, the smaller the correlation, with zero implying that the two attributes aren't related.

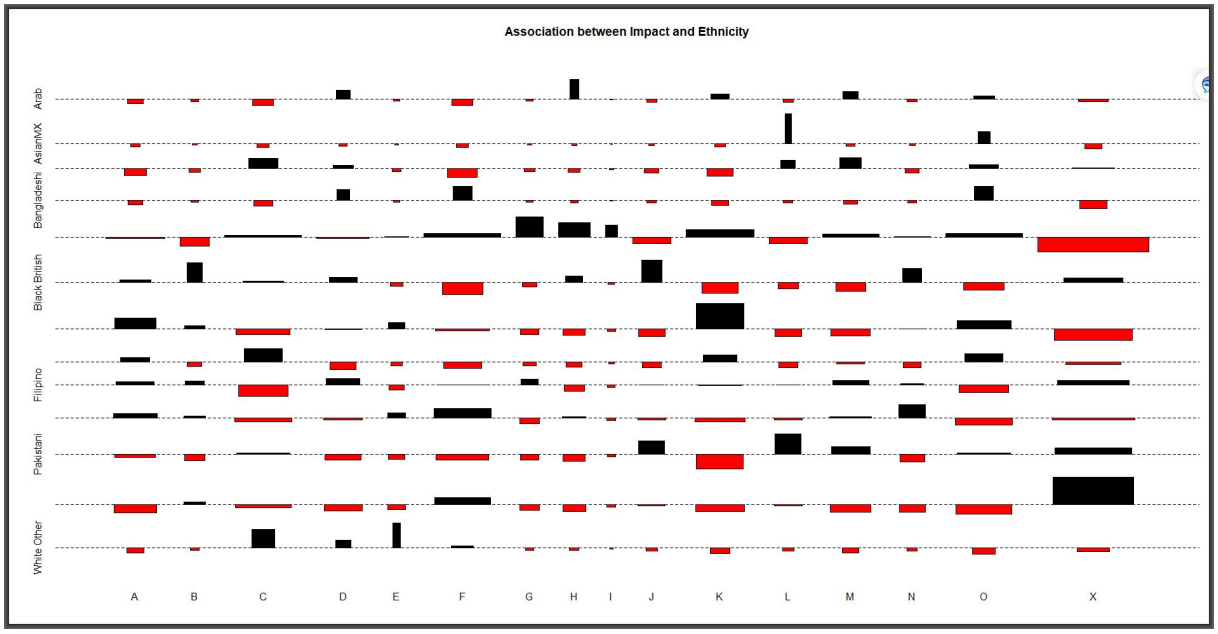

Figure 6: An illustrations of the impact of the pandemic on different ethnicities

Figure 6 exhibits how different ethnicities were impacted by the pandemic. In interpreting visual outputs of our analyses, we must always be mindful of the underlying theory and vice versa. There are many variants of inferential statistics and all of which can be presented in graphical and numeric formats. Understanding and being able to interpret inferential outputs is crucial.

3.1 Results from Correspondence Analysis

Correspondence analysis is a very useful method for understanding relationships among variables. In this case, we apply it to map experiences of health and care workers and their demographics. If specific features fall within close proximity on the same plot, that indicates a similarity which can improve working conditions.

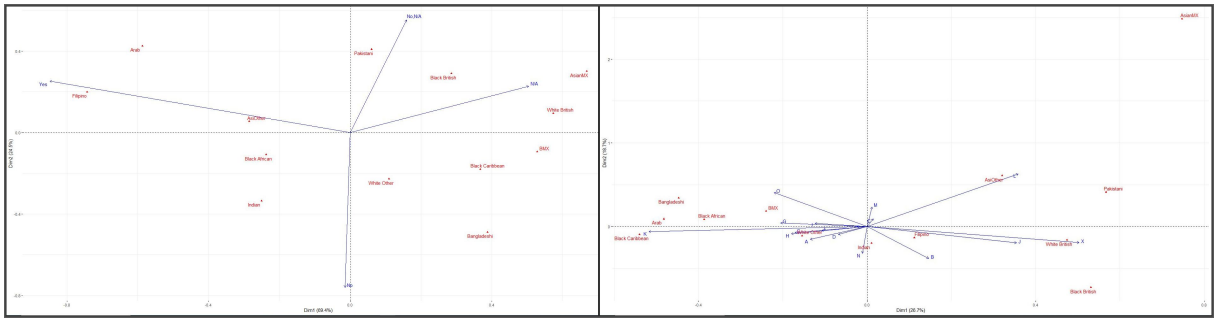

Figure 7: Graphical illustration of correspondence analysis

The two panels in Figure 7 represent relationships between immigration status vulnerability and ethnicity (on the left hand side) and the impact of the pandemic and ethnicity (on the right hand side). The simplest way to understand the plot is to recognise that both axes measure the levels of variation in the data, with the extreme left of the horizontal axis representing the negative measure and the extreme right, the most positive measure, similarly for the south and north directions of the vertical. In this particular application, we use the plots to try and answer questions such as:

1. Could the fight against COVID–19 have been fought in a more equitable way, than it did?
2. Are there lessons we can take from these results that could help improve our working conditions?
3. Was the study really representative of the feeling across the country?
4. Could employers have done better?
5. Are different settings really distinguishable for efficiency?

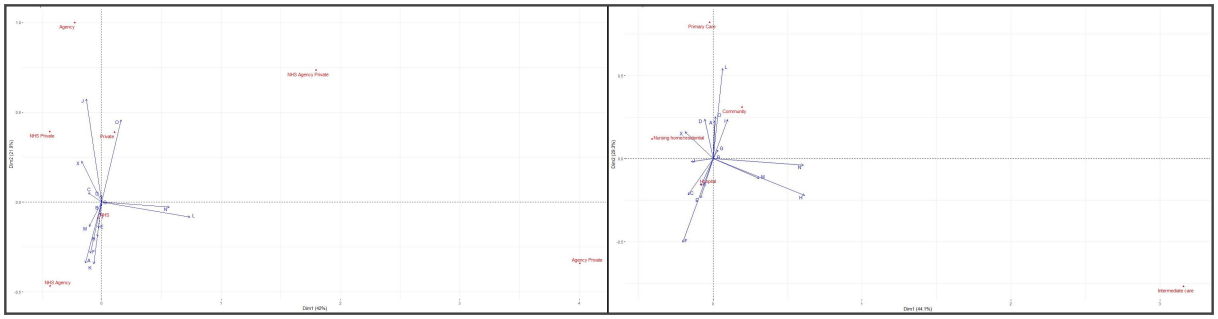

Figure 8: Exploring the impact based on employers and settings

The most interesting thing here is to know the contributing points to the solution provided by the method. Generally, we are placing two data attributes at a time, from Table 1, on the graph. Some of the relationship will be positive and some will be negative or non–existent. Being close to zero means no much relationship and these data points will be relatively unimportant to the interpretation. Being farther away from the origin means being more closely associated with the factors in the proximity This simple geometric interpretation of correspondence analysis makes it a really valuable technique in understanding relationships among variables, not provided by other graphical methods.

To identify the row and column points that are the most associated with the principal dimensions, you can use the function dimdesc() [in FactoMineR]. Row/column variables are sorted by their coordinates in the dimdesc() output.

`http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/113-ca-correspondence-analysis-in-r-essentials/`

4 Concluding Remarks

The 16^{th} SDG, as defined by the UN, is central to societal transformation. But even before the SDG launch in 2015, its role in social prosperity has always been known to man. The private sector call it "competition" and we all know, it is competition that has given us all the novel technologies we take for granted. No wonder, as the world gears up to even more advanced technologies. All that is needed is a political independent implementation of SDG #16. It is imperative to acknowledge that any of us can make or break our society. It is true of those already in power as it is for those in the opposition. With a full fledged SDG #16, a dysfunctional system will eventually fail. The election frauds and police brutality we so often hear nowadays are symptoms of failing systems that would do whatever it sustain the status quo. Such systems know only too well that they are not meeting peoples' expectations, but holding them to ransom. They are never confident in what they are doing, otherwise they would be open to meritocracy.

As the outputs are all based on statistical distribution theory, it is imperative to acknowledge that some outputs be difficult to interpret or may not agree with "current perceptions". It is our responsibility to make sense of what we generate and make them understandable to the audience we target. Quite often, some outputs are challenging because there is knowledge gap between the underlying theory and the graphics, say. We need to be mindful here that the concept of "theory" is not confined to statistics. The Chi–Square is a standard and widely used test in investigating associations. In this particular example, the Pearson's Chi-squared test returned a p-value of 0.2641, meaning it was not significant or that the null hypothesis could not be rejected. This p-value represents the probability of obtaining a chi-square as large or larger than that in the current experiment–i.e., the probability of deviations from what was "expected" being due to mere chance.

Dr Chinakizwa's research is an interesting one and could be more so if it goes on to integrate its findings with those of other researchers in the field of education across the SADC region, at least. This is important because of two closely related reasons. One, the variation in conclusions reached by different studies, as a result of data randomness needs harmonisation. Two, our young scientists and researchers, as he correctly put it, are the decision makers of tomorrow. Tracking their experiences and perceptions in a spatio-temporal context is naturally ideal. The two generate highly valuable data that for transforming various aspects of of SDGs - particularly those revolving around SDG#4.

References

[1] Tommasel, A.; Diaz-Pace, A.; Godoy, D.; Rodriguez, J. M. Tracking the evolution of crisis processes and mental health on social media during the COVID-19 pandemic. *Behaviour & Information Technology* **2021**, *0*, 1–20.

[2] SDG, Sustainable Development Goals. 2015; `https://www.un.org/sustainabledevelopment/sustainable-development-goals/`.

[3] Mwitondi, K.; Munyakazi, I.; Gatsheni, B. Amenability of the United Nations Sustainable Development Goals to Big Data Modelling. *International Workshop on Data Science-Present and Future of Open Data and Open Science, 12-15 Nov 2018, Joint Support Centre for Data Science Research, Mishima Citizens Cultural Hall, Mishima, Shizuoka, Japan* **2018**,

[4] Mwitondi, K.; Munyakazi, I.; Gatsheni, B. A robust machine learning approach to SDG data segmentation. *Journal of Big Data* **2020**, *7*.

[5] IISD, COVID-19 Wreaking Havoc on Bangladesh's Poor: A Story of Food, Cash, and Health Crises. 2021; `https://sdg.iisd.org`.

[6] Newman, K. L.; Jeve, Y.; Majumder, P. Experiences and emotional strain of NHS frontline workers during the peak of the COVID-19 pandemic. *International Journal of Social Psychiatry* **2022**, *68*, 783–790.

[7] Miao, L.; Last, M.; Litvak, M. Tracking social media during the COVID-19 pandemic: The case study of lockdown in New York State. *Expert Systems with Applications* **2022**, *187*, 115797.

[8] Fernandez, G.; Maione, C.; Zaballa, K.; Bonnici, N.; Spitzberg, B. H.; Carter, J.; Yang, H.; McKew, J.; Bonora, F.; Ghodke, S. S.; Jin, C.; Ocampo, R. D.; Kepner, W.; Tsou, M. H. The Geography of Covid-19 Spread in Italy Using Social Media and Geospatial Data Analytics. *The International Journal of Intelligence, Security, and Public Affairs* **2021**, *23*, 228–258.

[9] PHE, Beyond the Data: Understanding the Impact of COVID-19 on BAME Groups. 2021; www.gov.uk/phe.

[10] CRED, Commission on Race and Ethnic Disparities: The Report. 2021; https://www.gov.uk/government/publications/the-report-of-the-commission-on-race-and-ethnic-disparities.

[11] Mwitondi, K.; Munyakazi, I.; Gatsheni, B. An Interdisciplinary Data-Driven Framework for Development Science. *DIRISA National Research Data Workshop, CSIR ICC, 19-21 June 2018, Pretoria, RSA* **2018**,

[12] Pearce, W.; Mahony, M.; Raman, S. Science advice for global challenges: Learning from trade-offs in the IPCC. *Environmental Science & Policy* **2018**, *80*, 125–131.

[13] Mwitondi, K. S.; Said, R. A. A data-based method for harmonising heterogeneous data modelling techniques across data mining applications. *Journal of Statistics Applications & Probability* **2013**, *2*, 293 – 305.

[14] Mwitondi, K. S.; Said, R. A. Dealing with Randomness and Concept Drift in Large Datasets. *Data* **2021**, *6*.

[15] Zhang, L.; Zhang, B. The Quotient Space Theory of Problem Solving. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Berlin, Heidelberg, 2003; pp 11–15.

[16] Lawrence, A. J. Deletion Influence and Masking in Regression. *Journal of the Royal Statistical Society. Series B (Methodological)* **1995**, *57*, 181–189.

[17] Bendre, S. M. Masking and swamping effects on tests for multiple outliers in normal sample. *Communications in Statistics - Theory and Methods* **1989**, *18*, 697–710.

[18] Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1900**, *50*, 157–175.

[19] Hirschfeld, H. O. A Connection between Correlation and Contingency. *Mathematical Proceedings of the Cambridge Philosophical Society* **1935**, *31*, 520–524.

[20] Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901**, *2*, 559–572.

[21] Mwitondi, K. S.; Moustafa, R. E.; Hadi, A. S. A Data-Driven Method for Selecting Optimal Models Based on Graphical Visualisation of Differences in Sequentially Fitted ROC Model Parameters. *Data Science Journal* **2013**, *12*, WDS247–WDS253.

[22] Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

[23] Breiman, L. Bagging Predictors. *Machine Learning* **1996**, *24*, 123–140.

[24] Breiman, L.; Friedman, J.; Stone, C.; Olshen, R. Classification and Regression Trees. *Atmospheric Environment* **1984**,