Identifying and analysis of indicators of community strength

Christopher Chapman: 760 426

Domain: Community

How do we identify strong communities and their indicators? This report seeks to identify through correlation, the factors that can be used to identify and quantify strong communities, allowing for focused analysis on its benefits in all areas including mental and physical health and crime. This is a useful question for local governments, who can use this information to target issues in their communities. Through promoting highly correlated activities, communities will benefit through greater community engagement, and reap the benefits.

The following two data sets were used:

- **LGA11 Community Strength - Torren University of Australia (csv)**
  Contains modelled estimates of community strengths for the year 2011. Data is divided by Local Government areas, containing information about volunteering in the last 12 months, receiving or giving support and percentage of people that feel safe walking at night.
  Columns used: area_code, cs_cul_syn_4_uci_8_10_11_10, cs_vol_syn_4_uci_8_10_11_10, cs_gsup_syn_4_uci_8_10_11_10, cs_saf_syn_4_uci_8_10_11_10
  Link: https://data.aurin.org.au/dataset/tua-phidu-lga11-communitystrength-lga2011

- **Local government Area (LGA) profiles data 2011 – Government of Victoria – Department of Health (csv)**
  Contains annually updated Local Government Profiles containing a broad range of data. Used a subset of the data set, analysing relevant family, behavioural and community data.
  Link: https://data.aurin.org.au/dataset/vic-govt-doh-health-lga-profilesdata2011-lga
  Columns used: area_code, perc_membership_groups, perc_parents_schools, perc_multiculturalism.

Initial transformations:

- Cleaned up the information so that it would be easier to process and visualise, removing whitespace from column headings, changing numerical inputs from string to double or integer types.
- Removed unreliable data as defined by the data sets by RRMSE, the relative square root mean square error, removing all unreliable data, (those with a value of 2). This meant that a majority of the "disagree or strongly disagree with acceptance of other cultures" column data was unusable, and despite some really interesting cultural information to be gained from this, I had to cut it due to unreliability.
- Evaluated whether any of the columns used had outliers, using the IQR (inter quartile range) * 3 to identify and remove any major outliers. Through this evaluation, I found no outliers. However, initially, when my question had a focus on community and offences, this method was used to aggregate Melbourne, an LGA with an immensely high offence rate.
- Restricted the datasets to only relevant data, prioritising data that was measured per 100 people and removing columns containing data not concerning the proposed question.

- Increased clarity of column headers, renaming used columns in LGA11 Community Strength to human understandable names.
- Used LGA codes as a key, sorting sets in ascending order in terms of LGA codes for comparison of datasets.
- Removed rows with inconsistent LGAs, in particular unincorporated Vic, which only contained insufficient data and had an extremely small sample size.
- Keeping consistency in the column titles and allowing for some operations to be applied to both sets, renaming 'area_code' in community strength to 'lga_code'.

Results

Identifying strong communities and indicators:

A way to identify strong communities is through looking at correlating factors. I looked at the Pearson values between 7 areas listed below: (figure 1)

Keys:
Support -> Received support in times of crisis per 100
Vol -> Volunteered in the last 12 months
Members-> % of people in membership groups
Safety -> % of people that feel safe walking alone in the local area at night
Parents -> % of parents that are involved in their local schools
Services -> % of population who believes that there are good facilities and services in the LGA like shops, schools and libraries.
Multi-cult -> % believe that multiculturalism makes life better

Pearson coefficients - r values: (figure 1)

	Support	Vol	Members	Safety	Parents	Services	Multi-cult
Support		0.839	0.688	0.809	0.260	-0.094	-0.389
Vol	0.839		0.690	0.925	0.225	-0.406	-0.630
Members	0.688	0.690		0.580	0.369	-0.270	-0.389
Safety	0.809	0.925	0.580		0.209	-0.430	-0.592
Parent inv	0.260	0.225	0.369	0.209		-0.234	-0.169
Services	-0.094	-0.406	-0.270	-0.430	-0.234		0.754
Mutli-cult	-0.389	-0.630	-0.389	-0.592	-0.169	0.754	

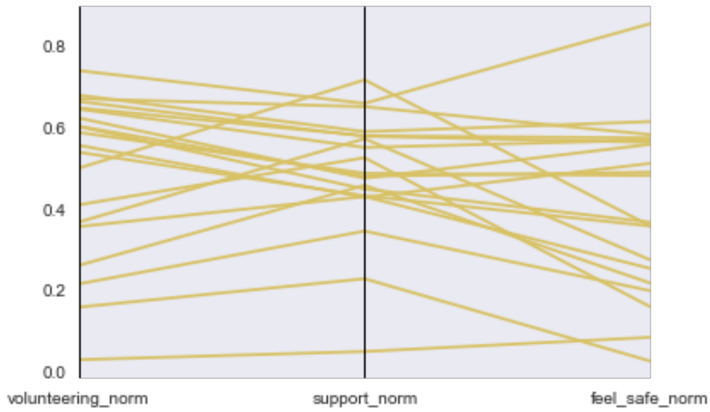
As seen, the Pearson values between support, volunteering and safety all have strong correlation of r > 0.8 (Jacob Cohen) between one another and there is a sizeable relationship between these sets and membership, with r>0.5. Although it is used as a reference for looking at stronger communities, parent involvement has a small correlation between 0.2 and 0.37, and is not as strong of an indictor. Surprisingly, community services showed a negative correlation, which could show a disconnect between people's beliefs or that all indicators are not currently present. Interestingly, there seems to be a strong negative trend between pro multi-culturalism communities and other community indicators, but is beyond the scope of this analysis.

However, Pearson coefficients can only show that elements are strongly connected, and thus do not prove causality.

Further similarity comparisons:

I further looked at the strongest correlated, support, volunteering and safety through parallel coordinates. I normalised support, volunteering and membership values to a scale from 0-1, comparing the first 20 LGAs in numerical order. Through this visualisation, it is easy see similar movement between the coordinates.

Parallel coordinates between normalised volunteering, support and safe values (figure 2)



Quantifying strong communities

Assuming that support, volunteering and safety are good indicators of a strong society, we can rank different LGAs. To find a basis to rank communities, I normalised the values of support, volunteering, and membership for each LGA, then taking the average for each (figure 2). I chose to use normalisation as major outliers have been removed, thus the normalised data is not present only in a small range, which can occur.

figure(3) – strong communities based off of normalised

$$\left(\sum_{i=0}^{3}\left(\frac{xi - min(x)}{max(x) - min(x)}\right)\right) \div 3$$

A problem that could occur is because this ranking system is only using three different indicators; it may not always give the best indication of a community's strength. However, a way to quantify indicators to get a scale is extremely important for further analysis.

Ranking communities:

Top 5 Ranks: figure (3) **Lowest 5 ranks: figure (4)**

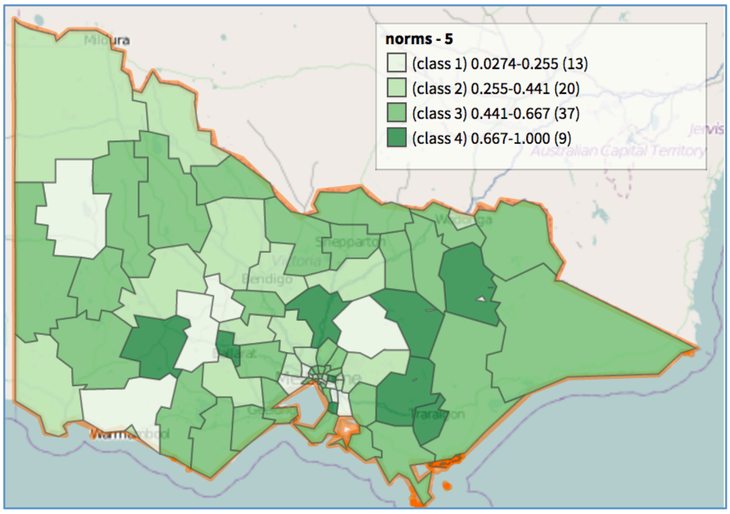
LGA code	LGA name		LGA code	LGA name
20110	Alpine		21180	Brimbank
22170	Frankston		21610	Casey
20570	Ballarat		22670	Greater Dandenong
23810	Latrobe		25490	Moyne
20830	Baw Baw		25990	Pyrenees

By creating a scale to measure strong communities, we are now able to look rank and look critically at what are the strengths of top ranked communities vs lower ranked communities.

Visualisation through a heat map:

This geographical heat map was created using the normalised measurement created above, importing a csv containing the relevant normalised data for each LGA to Aurin. I used its functionality to create a cloropeth, with darker shades of green representing stronger communities based off the above classification. This is an easily readable map, showing explicitly where the stronger communities are, and where some have clustered. However, as the data is divided by LGAs, which are relatively large areas, this cannot be used to look at the detailed information about specific areas.

Figure 3 – Measuring community strength by LGA

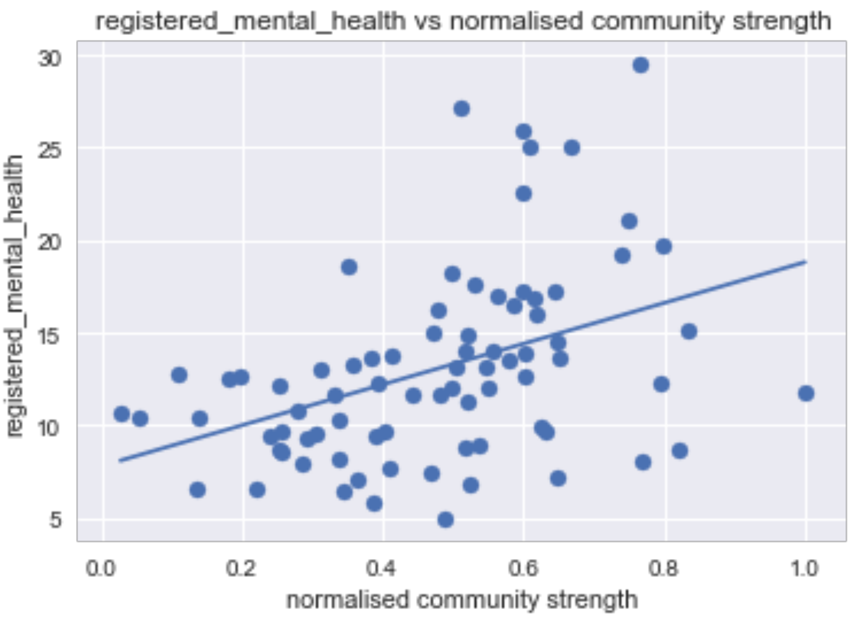


Value:

Through the processing of data, I was able to create a more legible code, changing the names of unclear columns, keeping consistency in column titles between datasets and sorting in terms of LGA codes. I made the datasets more python readable to allow for easier analysis, removing whitespace from column headings, changing numerical inputs to their required types. Unreliable data was removed from the data sets, avoiding all data with high uncertainty, and eliminated major outliers, allowing for more accurate calculation of correlation.

Using the quantified strength to find the benefits of strong communities

The usefulness of having a measure of strength shines when using it as a scale to compare against. An example that would interest local governments is using a scatter plot and line of best fit, to see comparative trends. For example, through a simple comparison against mental health, as seen below, there is a mid-strong correlation of r=0.418, which could lead to new and profound information



With just the raw data, it is difficult to understand or visualise which communities are strong. Through the figures below and creating a scale to measure community strength, correlations between communities and benefits such as better health become visible. Through processing an in particular visualisation, comparing and contrasting different LGAs is made much easier.

Challenges and reflections:

I was initially interested in other topics, including aspects of tourism, housing and child participation in sport. It was extremely frustrating being unable to find any data in the required formats, whether it was hidden behind a pay-wall or was private data and couldn't be accessed. I found that there was no clear-cut way of analysing what the aspects of community strength were. Despite finding very strong correlations, it is difficult to prove causality.

Question resolution:

Only by being able to quantify and find factors of strong communities can there be further analysis into the benefits of strong community. Through my proposed scale, local government bodies will be able to use rankings and normalised data in order to compare against community problems. An example of which was provided in the value section of the report.

Code:

I used some workshop code, which I referenced through comments within my code. These were used to make the similarity comparison, and in order to find the Pearson coefficients of each values. Except for some lines, which were found from online forums which were referenced in the code, the rest I created myself, whether reusing code from project 1 or made during this project.