DATA2901 Group Assignment

Harry Breden

Contents

THE UNIVERSITY OF
SYDNEY

1 Dataset Description

1.1 SA2 Regions

This dataset provides a collection of Australian Statistical Geography Standard digital boundaries of Statistical Area Level 2 (SA2) regions. These regions provide a geofencing of Australia and its territories, with each of the 2,473 regions representing "a community that interacts together socially and economically" (ABS, 2021). They have an average population of approximately 10,000 people, but are generally larger and less populous in regional areas. All SA2 regions have a parent SA3, SA4, Greater Capital City Area, and State/Territory. These are larger statistical areas, and are also included in the dataset.

The data provided is contained in 6 files. The main shapefile (.shp), encodes the actual geographical Polygons and MultiPolygons of the SA2 regions. There is also a dBASE file (.dbf) which contains atribute data such as the code, name, area etc. of each SA2 geometry. The 4 other files provide metadata and scheme information about how these 2 files interact and should be interpreted.

The dataset was obtained from the ABS's Digital Boundary Files website and imported into a GeoDataFrame using Fiona. Most of the information was deemed obsolete for this project. The dataset was pre-processed by filtering down to SA2's with parent Greater Capital City Area equal to Greater Sydney. Only 4 attributes per SA2 were kept; their name, code, area (in square kilometers) and geometry (as a MultiPolygon in Well Known Text).

1.2 School Catchments

Similarly to the SA2 Regions, the data set contains digital boundaries of school catchment zones that geofence NSW. Each child in NSW is provided with a public school to attend based on their residential address. The school catchment within this dataset provides an indication of what school one would attend if their primary address was within a particular Polyogon or Multiploygon.

The data is partitioned into primary, secondary and future schools. Each partition has a shapefile and dBASE file (where the main data is stored), as well metadata and scheme information similar to that in 1.1.

The dataset was obtained from the NSW's Department of Education School Intake Zones website and imported into a GeoDataFrame using Fiona. Most of the information was again deemed unnecessary for this project. The dataset was pre-processed keeping only 2 attributes; the school ID and its geometry (as a MultiPolygon in WKT).

1.3 Businesses

This dataset provides counts of businesses in each SA2 region based on their size and industry. The data is provided in a single Comma Separated Values (.csv) file, and was obtainted from the ABS's Counts of Australian Businesses website. To preprocess this data, it was imported into a Pandas DataFrame. The industry code was converted to an integer index and a mapping table was created matching industry code and name. The 3 attributes kept per SA2 region were industry code, SA2 code and total count of businesses. The data was also filtered to SA2 regions within Greater Sydney.

1.4 Income

This dataset contains information about the income of residents within each SA2 region. It was sourced from the ABS's Personal Income in Australia website. It was provided as a CSV file, and was preprocessed in Pandas by dropping the SA2 name column and only keeping SA2 entries in Greater Sydney. The final data included SA2 codes, the amount of earners and their median age, as well as median and mean income ($AUD).

1.5 Population

This dataset was sourced from ABS's Regional Population by Age and Sex website and provides data for the populations of SA2s. It was provided as a CSV, with information on population bucketed to 5 year age ranges. The necessary data kept was the total population, and population between the age of 0 and 19 years old of all SA2s in Greater Sydney. This was preprecessed in Pandas.

1.6 Polling Places

This datatset provides geographic information on where polling places are within NSW from 2019. It was provided as a CSV file from AURIN's AEC Federal Election Polling Places website. Much of the data was deemed irrelevant to this project. In order to preprocess the data, all columns apart from the Polling Place ID, longitude and latitude were kept. The longitude and latitude were then converted into Point geometries in WKT.

1.7 Public Transport Stops

This dataset contains information on the public transport stops within NSW. The data provided by NSW Transports' Open Data Hub Timetables Complete GTFS website was a plain text file. The plain text file was, however, in a CSV-like structure and was able to be read into Pandas. Again, much of the data was not necessary for this project. In order to only consider "stops" and with accordance to the General Transit Feed Specification, the data was filtered to only have missing/empty Location Types (indicating a stop). The longitude and latitude were then converted into Point geometries in WKT, with all other columns apart from Stop ID being dropped.

1.8 Childcare Facilities

This dataset provides locations of childcare facilities within NSW in 2022. The data was downloaded using a web scraping function from the DataNSW Early Childhood Education and Care Provider Locations website and converted into a CSV file for easier manipulation using Pandas. The dataset includes attributes associated with each childcare facility, such as its ID, provider type, provider name, address, suburb, state, postcode, latitude, and longitude. Only the 'ID' column and the latitude and longitude coordinates were retained from the original dataset. The 'ID' column was renamed to 'childcare_id', and the latitude and longitude were used to create Point geometries in WKT using the GeoPandas library.

1.9 NSW Points of Interest

This dataset provides Points of Interets (POI) within NSW. A POI is a feature, service or activity that someone might want to visit and/or utilise. They are grouped into the following categories; Community, Education, Recreation, Transportation, Utility, Hydrography, Physiography and Place. A specific subset of POIs were investigated, namely libraries. The data was provided as a GeoJSON file downloaded from the NSW Spatial Data Portal POI website. The GeoJSON file format is a standardised format to transfer geospatial data as well as attributes associated with the geometries. The collection of POIs was filtered in GeoPandas by finding the POI type to be library. The Point geometries were also converted to WKT, and the ID column was renamed to align with the filtering to just libraries.

2 Database Description

A normalised scheme was produced to store the data sets outlined in 1. Below is a database diagram:

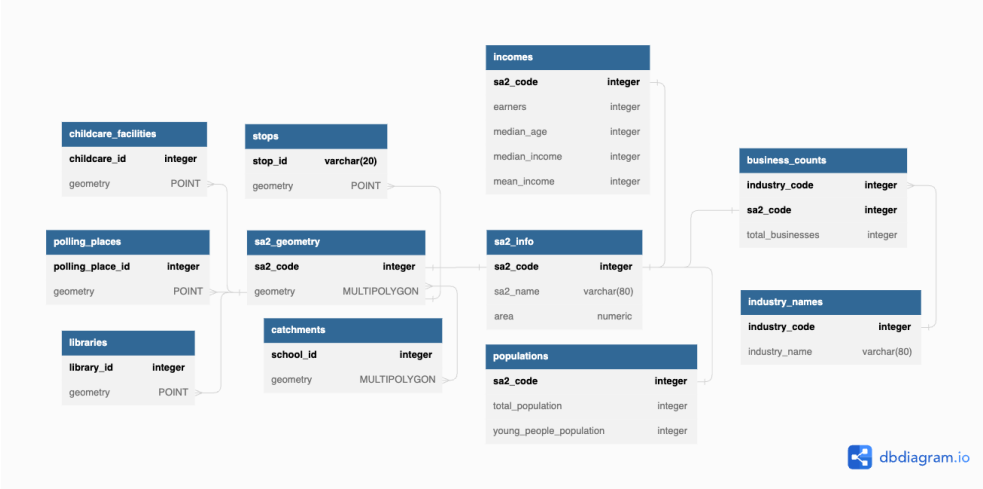

Figure 1: Database Diagram

Primary keys are in bold, and foreign keys are linked. Note that the foreign keys linking geometries were not added into the database explicitly, but rather provide a visual intuition of the relationship they have to the SA2 geometry. GIST indexes were also generated on all geometry columns, as well as those automatically generated on primary key columns. A composite primary key of Industry Code and SA2 Code was used within the Business Counts table.

3 Score Analysis

3.1 Scoring Function Rationale

The assignment brief initially proposed a scoring function that computes the sum of z-scores of each attribute and applies a logistic function. While this approach provides a degree of standardisation and normalisation, it embodies certain limitations that the refined scoring function addresses.

The primary limitation of the initial scoring function is its assumption of equal importance for all metrics. Real-world data seldom adheres to such uniform significance across variables. To rectify this, the refined scoring function introduces a set of weights for each metric. These weights were determined using Principal Component Analysis (PCA), a machine learning procedure that identifies the most significant direction of variance in the data (see Machine Learning). As a result, the weights reflect the relative importance of each metric in explaining the overall variance in the data.

Additionally, the refined scoring function addresses the issue of outliers, which the initial function could be sensitive to. To mitigate the potential adverse impact of outliers, the z-scores in the refined scoring function are capped at -3 and 3. This statistical decision is grounded in the properties of a standard normal distribution, where approximately 99.7% of the data lies within three standard deviations from the mean. This is also a standard method to mitigate against outliers.

In the context of assessing the 'well-resourced' status of a region, the selection of additional datasets plays a crucial role. The inclusion of 'Childcare Facilities' and 'Libraries' in the refined scoring function is a testament to this understanding.

The selection of childcare facilities as an additional dataset extends the scope of the assessment to include the provision of services that directly impact the welfare of families, particularly those with young children. By factoring in the availability of childcare facilities, the refined scoring function provides a more comprehensive perspective on the region's resources.

Similarly, libraries serve as hubs of learning, community engagement, and social inclusion. Their presence is indicative of a region's commitment to education, literacy, and community development. Incorporating libraries as an additional metric into the scoring function enables a broader evaluation of a region's resources, taking into account educational and social infrastructure, not just commercial or health-related amenities.

The density of polling places, transport stops, libraries and childcare facilities were calculated by dividing the counts of these facilities by the respective area. Additionally, the density of schools was computed by counting the number of schools within each area and dividing it by the young people population and area. The densities of retail and health services were also calculated by dividing the total number of businesses by the total population and area.

The refined scoring function is defined as follows:

Let M denote the set of metrics: retail, health, stops, polls, schools, libraries, childcare facilities. For each category $X \in M$, calculate the capped z-score Z'_X:

$$Z'_X = \begin{cases} 3 & \text{if } Z_X > 3 \\ -3 & \text{if } Z_X < -3 \\ Z_X & \text{otherwise} \end{cases}$$

Then, calculate the weighted sum of these capped z-scores, S:

$$S = \sum_{X \in M} w_X \cdot Z'_X$$

Finally, calculate the final score by standardising S:

$$\text{Final score} = \frac{S - \mu_S}{\sigma_S}$$

3.2 Results

Table 1: Descriptive Statistics Table

Metric	Min	Q1	Median	Q3	Max
Retail	-0.539	-0.426	-0.264	0.018	3.000
Health	-0.620	-0.499	-0.332	0.069	3.000
Stops	-1.722	-0.742	-0.014	0.710	3.000
Polls	-0.519	-0.366	-0.207	0.068	3.000
Schools	-0.414	-0.341	-0.262	-0.023	3.000
Libraries	-0.516	-0.516	-0.516	0.209	3.000
Childcare	-1.013	-0.701	-0.230	0.450	3.000
Final Score	-1.205	-0.701	-0.172	0.419	4.582

Table 1 provides a comprehensive view of the distribution of various metrics across SA2 areas. All metrics share a maximum value of 3, representing the highest z-score. This shows that there are areas that significantly outperform the mean in these categories.

However, the minimum value of these metrics varies considerably. For example, the 'Stops' metric reveals the most negative minimum z-score, indicating that some areas are considerably underperforming with the availability of public transport. In contrast, the 'Libraries' metric has the least variation, with its 25th percentile, median, and 75th percentile values all equal to -0.516, suggesting a shortage of libraries across most areas.

The third quartile (Q3) values highlight the distribution among higher-scoring regions. The 'Childcare' metric, for example, has a Q3 value of 0.450. This indicates that even in high-performing areas, childcare facilities tend to be closer to the mean than those in top-performing regions for other metrics.

Looking at the final scores, there's a substantial range in the level of resources across regions, as indicated by the minimum and maximum values of approximately -1.21 and 4.59, respectively. The 25th, median, and 75th percentiles (-0.701, -0.172, and 0.419, respectively) further illuminate this distribution. The negative median score suggests that more than half of the regions have z-scores below the mean, indicating a skewed distribution. The distribution of these scores further reinforces this interpretation, with a right-skewed distribution indicating a concentration of lower-scoring areas and fewer high-scoring areas.

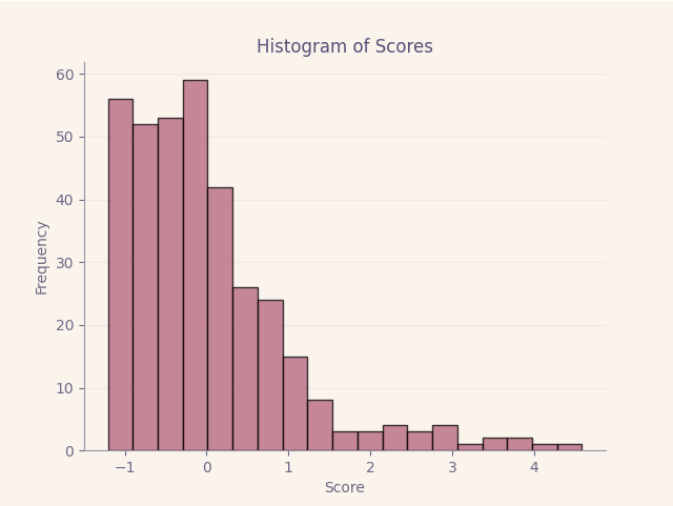

Figure 2

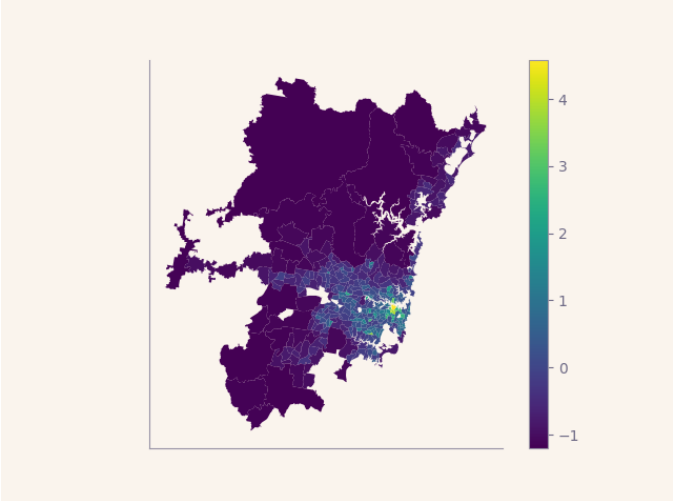

Figure 3: Darker shades indicate lower well-resourced scores, while lighter shades represent higher scores.

The heatmap and 3D kepler.gl visualisation [1] underscore the concentration of resources in the Central Business District

[1] The data may take up to 30 seconds to load. It is recommended to have only one layer open at a time and they can be hidden and unhidden on the left. High and Colour both indicate the score of a SA2

(CBD) and inner-city regions. This finding aligns with the principle of agglomeration in urban economics, where businesses and services (such as retail stores, health facilities, schools, and childcare facilities) tend to locate close to each other and to the population centres to benefit from proximity to customers, labour markets, and complementary services.[1]

Table 2: Top 10 Scoring Areas

SA2 Area	Retail	Health	Stops	Polls	Schools	Libraries	Childcare	Score
Sydney (North) - Millers Point	3.0	3.0	2.81	3.0	0.81	3.0	3.0	4.58
Ultimo	3.0	3.0	1.03	0.42	3.0	3.0	3.0	4.01
Surry Hills	3.0	3.0	1.42	1.87	2.41	3.0	1.32	3.92
Sydney (South) - Haymarket	3.0	2.99	3.0	3.0	1.60	-0.52	1.83	3.78
Hurstville - Central	2.58	2.62	2.52	3.0	1.29	-0.52	2.75	3.61
Darlinghurst	3.0	3.0	2.42	3.0	3.0	-0.52	-0.12	3.46
North Sydney - Lavender Bay	1.62	3.0	2.48	0.57	0.12	2.05	3.0	3.22
Chippendale	3.0	2.84	0.06	0.61	3.0	-0.52	2.30	2.90
Pyrmont	2.42	2.12	1.97	0.04	0.67	3.0	1.47	2.90
Double Bay - Darling Point	2.33	3.0	1.62	0.49	0.67	2.63	0.47	2.82

This principle is further supported by the findings from the analysis of Table 2. Sydney (North) - Millers Point emerges as the leading region with a score of 4.58, boasting exceptional access to Retail, Health, Polls, Libraries, and Childcare resources. Notably, Ultimo, Surry Hills, and Darlinghurst also demonstrate superior access to Retail and Health resources. However, the Libraries metric reveals a deficit in certain areas like Sydney (South) - Haymarket, Hurstville - Central, Darlinghurst, and Chippendale. Despite their high overall performance, North Sydney - Lavender Bay and Double Bay - Darling Point show room for improvement in the Schools and Childcare sectors respectively. Furthermore, with the exception of Chippendale and Pyrmont, the top-scoring regions display strong public transportation networks. Interestingly, Darlinghurst, despite its high overall score, exhibits a deficiency in childcare resources. These observations underline the necessity of comprehensive evaluations when considering resource distribution, as high-performing areas can still display specific deficiencies.

The negative Pearson's correlation coefficient of -0.654 between the z-score and distance from the CBD also corroborates this urban concentration pattern. As one moves away from the city centre, the level of resources across the seven metrics tends to decrease. This pattern reflects the influence of urban structure and commuting patterns on the location of services. For instance, transport stops are likely more abundant in and around the CBD to facilitate commuting.

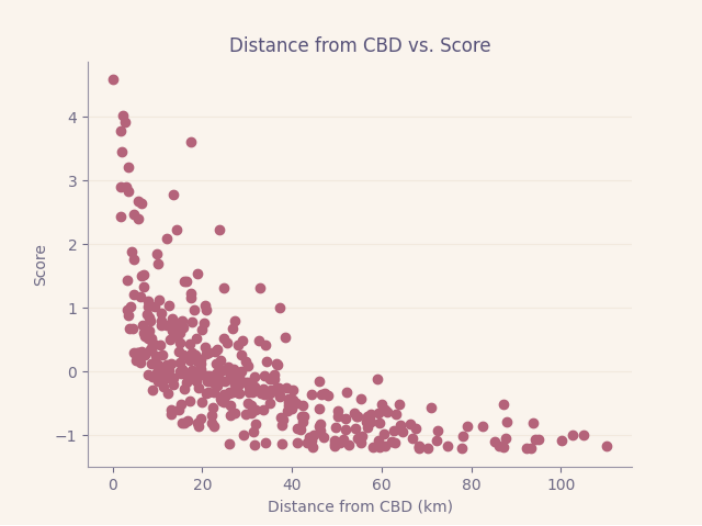

Figure 4

Further, the average z-score for areas topologically adjacent to the top 10 scoring areas is 1.778. A two-tailed t-test revealed that this average is significantly higher than the average z-score for Greater Sydney. This suggests that areas surrounding the top 10 scoring regions also tend to be well-resourced. This could indicate a clustering effect, where

well-resourced areas are often surrounded by other well-resourced areas. This spatial autocorrelation could be due to a variety of factors such as economic forces, urban planning decisions, or historical development patterns.

While this study offers significant insights, several limitations need to be considered. The metrics utilised to compute the z-scores, though comprehensive, do not capture all facets of an area's resources. Important amenities such as parks, cultural institutions, and social services have not been included in the computation of the final score, possibly presenting an incomplete picture of resource richness.

Moreover, these metrics focus on the quantity of resources rather than their quality or accessibility. A higher quantity of a certain resource, such as schools, may not necessarily equate to better accessibility or quality. For instance, an area with many schools may score high in the 'schools' metric, but if these schools are of poor quality, or if they're not easily accessible to all students, then the high score may not accurately reflect the actual level of resources in the area.

Similarly, the count of polling places, while indicative of civic engagement and infrastructure availability to some degree, may not provide a complete picture of a region's well-resourcedness. The count does not account for the effectiveness of these polling places, their manageability, or the quality of the voting process. Furthermore, polling places are used infrequently and thus may not represent a continuous resource compared to other metrics such as healthcare facilities or public transportation.

This focus on quantity over quality extends to all metrics used in this study and underscores the need for more comprehensive and nuanced assessments of resource distribution. Furthermore, while the analysis has provided a snapshot of the current distribution of resources, it does not shed light on how this distribution has evolved over time or how it might change in the future. Further research could explore these temporal dynamics, which could offer valuable insights for urban planning and policy-making.

4 Income Correlation Analysis

The Pearson's correlation coefficient of 0.293 indicates a weak positive relationship between the well-resourced score of each region and the median income. This suggests that although regions with higher scores tend to have higher median incomes, the relationship is not strong or deterministic, indicative of the influence of other socioeconomic and geographical factors.

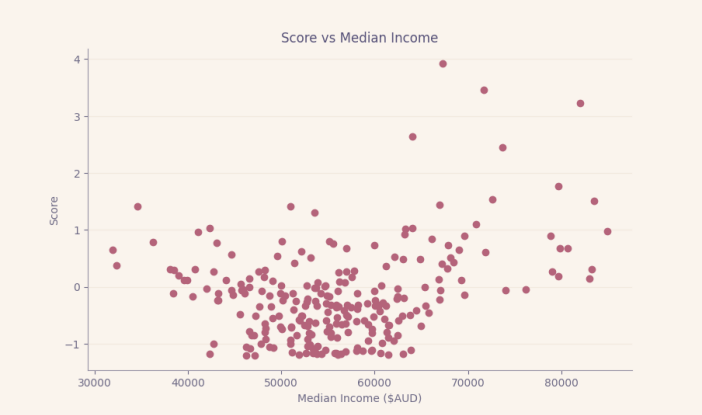

Figure 5

This correlation aligns with established research on urban geography within the Australian, and more specifically, Sydney context. It has been noted that income levels tend to be higher in inner-city areas compared to outer suburbs or rural regions, often referred to as the 'urban wage premium'.[3] These areas are typically better serviced with public amenities and have higher accessibility to resources, leading to higher well-resourced scores.

Conversely, as demonstrated previously, distance from the CBD and the well-resourced score is negatively correlated. This phenomenon, often termed 'spatial mismatch', arises from historical urban development patterns and contemporary planning policies, leading to a concentration of resources and higher incomes in inner-city areas.[3]

However, this pattern does not imply a strict deterministic relationship between income, resource distribution, and distance from the CBD. For example, government policies often aim to reduce resource disparities between inner-city and outer-

suburban regions, thereby ensuring a certain standard of service accessibility regardless of income levels or geographical location.[2]

5 Advanced Deliverables

5.1 Rank Based Scoring

This alternative methodology, using ranks instead of z-scores, provides a robust scoring system that is less sensitive to outliers and skewed distributions in the underlying metrics. For each metric, the region with the highest value is given a rank of 1, the region with the second highest value is given a rank of 2, and so forth. Thus, the rank of a region for a specific metric is a measure of how that region compares to all other regions in terms of that metric. The final score for a region is the sum of the ranks for all metrics.

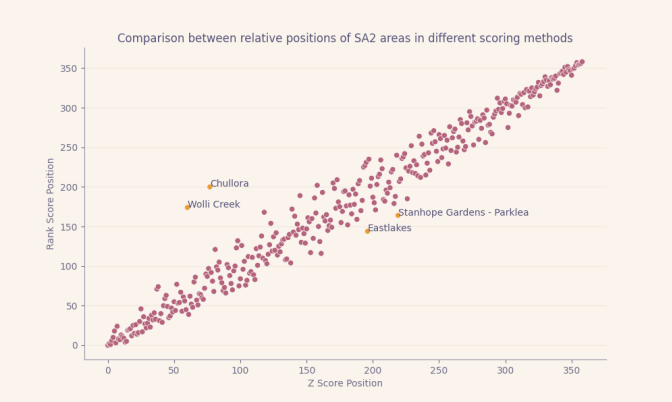

Figure 6: Outliers, defined as areas where the difference in positions is greater than 50, are highlighted and annotated.

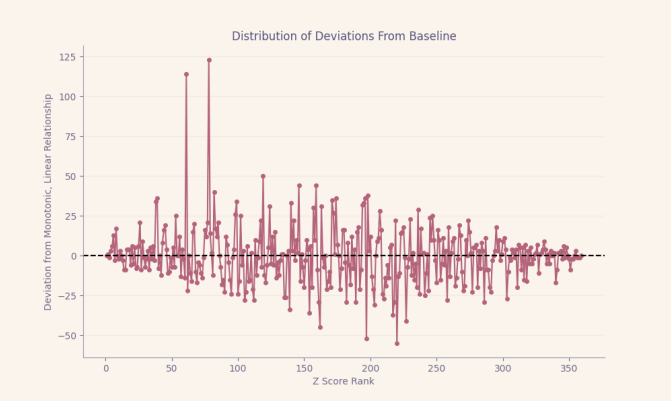

Figure 7: The dashed line represents a baseline if the scoring methods assigned the same rank. The x-axis is the rank assigned by the original scoring method. The y-axis represents the difference in the relative positions of each area between the scoring methods.

A Spearman's Rank Correlation coefficient of 0.985 indicates a strong positive relationship between the two sets of rankings. This suggests that despite the different methodologies, both scoring methods generally agree on the relative positions of the SA2 regions in terms of how well-resourced they are. This high correlation signifies that both methods are likely capturing the same underlying patterns in the data, despite their different approaches.

The comparative analysis of the two methodologies also revealed a distinct pattern in the degree of divergence between their results. The spread is not uniform across the range of scores (Figure 7). Instead, it appears to peak in the middle of the rankings, while being much tighter at both ends. This pattern suggests that both methodologies are particularly consistent in their evaluation of regions that are either very well-resourced or poorly resourced. This consistency likely stems from the distinctiveness of these areas in terms of their resource profiles, which makes them stand out regardless of the scoring method employed.

It is noteworthy that four areas differed in positions by more than 50 places. This discrepancy might arise due to the different sensitivities of the two methods to outliers and extreme values. The z-score method may be heavily influenced by extreme values in the metrics, causing some areas to have much higher or lower scores than the median. On the other hand, the ranking method is resistant to outliers, as it only considers the relative positions of the metric values, not their magnitudes. Nevertheless, the high degree of agreement between the two methods highlights the effectiveness of the decision to trim the z-scores of each metric at -3 and 3 in the original methodology. This z-score trimming is a method of handling outliers and it is designed to reduce the impact of extreme values on the overall score. The fact that the rank-based method, which is inherently resistant to outliers, produced similar results, supports the notion that the trimming strategy succeeded in its objective.

5.2 Machine Learning

Princple Componenet Analysis (PCA) was used within the data workflow to weight the z-scores of each component such to maximise the variance in the scores. This meant our score was able to explain the most varience in the data compared to any other linear combination of the z-score components, allowing for a more complete indication of a SA2's well-resourcefulness.

PCA is an unsupervised machine learning model which allows for the dimensionality reduction of complex data sets whilst preserving the maximum amount of information between features. By creating a score for each SA2, this is essentially reducing the dimensionality of the dataset to 1-dimension. As the dataset was to be restrict the to 1 dimension, the first principle component vector was calculated, and its components used as weights for the scoring metric. This first principle component vector is the unit eigenvector of the maximal eigenvalue of the co-variance matrix. The derivation of this is provided in the Appendix (7.3). In the case of this dataset, this eigenvector was computed using SKLearn using z-score normalisation, and the weights were automatically used within the SQL query to calculate the score for each SA2.

The results of the PCA can be seen the Appendix (7.1). Retail and Health accounted for the largest varience in the dataset, being weighted 0.45 and 0.44 respectively. Schools and Libraries were the least significant attributes, being weighted 0.31 and 0.32. This may be due to retail and health services driving economic activity in SA2s, with high levels of these attributes requiring other resources to be provided to the SA2 in order to meet demand.

The eigenvector and eigenvalues were calculated using a full singular value decomposition of the co-variance matrix, and in some sense be considered 'accurate' because of this. However, there is no true accuracy measure for this algorithm as it does not classify the features.

6 Bibliography

References

[1] Masahisa Fujita and Jacques-Francois Thisse. *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth*. Cambridge University Press, 2002. DOI: 10.1017/CBO9780511805660.

[2] Bill Randolph and Alan Tice. *Suburbanizing Disadvantage in Australian Cities: Sociospatial Change in an Era of Neoliberalism*. Taylor & Francis, 2014. URL: https://www.tandfonline.com/doi/full/10.1111/juaf.12108.

[3] Sahar Sarkar. "The Scaling of Income Distribution in Australia: Possible Relationships between Urban Allometry, City Size, and Economic Inequality". In: *Environment and Planning B: Urban Analytics and City Science* 45.4 (2016), pp. 603–622. DOI: 10.1177/0265813516676488.

7 Appendix

7.1 PCA Metric Weights

The following weighings for each metric were calculated using PCA:

$$\begin{aligned} w_{\text{retail}} &= 0.45075328648827917 \\ w_{\text{health}} &= 0.4382001185821598 \\ w_{\text{stops}} &= 0.3542459899107671 \\ w_{\text{polls}} &= 0.36725888233772297 \\ w_{\text{schools}} &= 0.31043876860537345 \\ w_{\text{libraries}} &= 0.3236294063075545 \\ w_{\text{childcare facilities}} &= 0.37858235707779997 \end{aligned}$$

7.2 Spearman's Rank Correlation

The Spearman rank correlation coefficient is a non-parametric measure of rank correlation. It assesses how well the relationship between two variables can be described using a monotonic function. A Spearman correlation of 1 results when each of the variables is a perfect monotone function of the other, implying that as one variable increases, the other does too, and vice versa. The formula for Spearman's rank correlation coefficient (ρ) is expressed as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where:

- d_i represents the difference in ranks for each pair of observations.
- n denotes the number of paired observations.

7.3 Derivation of PCA eigenvector

In order to preserve as much information fron the dataset when projecting onto a 1-D line, we need to find a vector $\mathbf{u}$ that points in the general direction of all the SA2s z-scores. More formally, we want this vector $\mathbf{u}$ that maximises the length of all feature vectors $\mathbf{x}_i$ when projected onto $\mathbf{u})^2$, ie. wanting to maximise:

$$\sum_i (\mathbf{x}_i \cdot \mathbf{u})^2$$

However, we also have the restriction that $\mathbf{u}$ should be unit length, ie. $\mathbf{u} \cdot \mathbf{u} = 1$. Using matrices and the Lagrange Multiplier method of optimisation problems, we find with C being the co-variance matrix of the feature vectors:

$$\begin{aligned} 0 &= \nabla \left[\sum_i (\mathbf{x}_i \cdot \mathbf{u})^2 - \lambda(\mathbf{u} \cdot \mathbf{u} - 1) \right] \\ &= \nabla \left[\sum_i \mathbf{u}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{u} - \lambda(\mathbf{u}^T \mathbf{u} - 1) \right] \\ &= \nabla \left[\mathbf{u}^T C \mathbf{u} - \lambda(\mathbf{u}^T \mathbf{u} - 1) \right] \\ &= 2C\mathbf{u} - 2\lambda\mathbf{u} \\ \implies C\mathbf{u} &= \lambda\mathbf{u} \end{aligned}$$

This gives λ an eigenvector of C and in order to maximise information, we choose the largest eigenvector. We also find $\mathbf{u}$ the unit eigenvector corresponding to this eigenvector, as claimed.