No Pain, No Gain: The Challenges and Rewards of Analysing Handwritten Records

Antriksh Dhand
Supervised by Dr. Jenkin & Professor Keay

School of Medical Science
The University of Sydney

May 17th, 2024

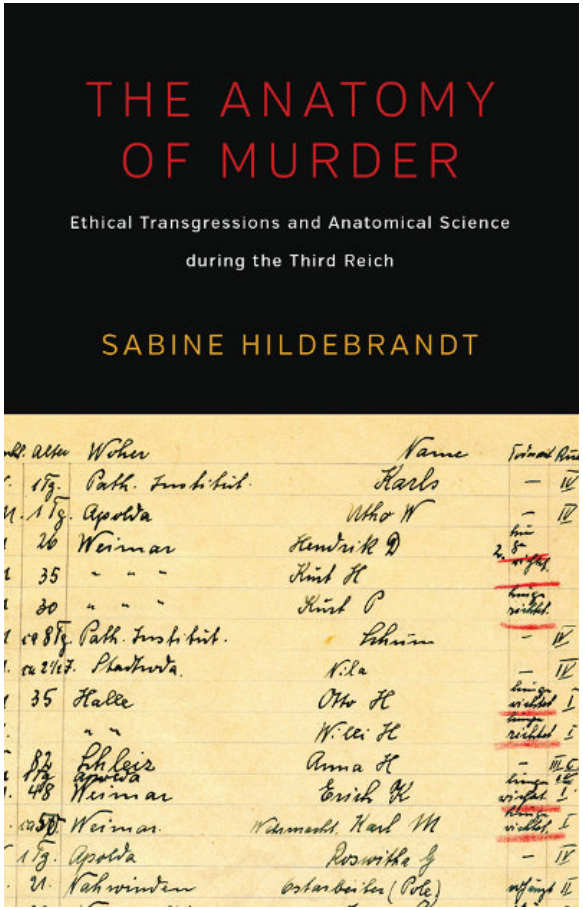
Table of Contents

Table of Contents

Why analyse historical records?

- Importance of preserving history: preventing historical data from decaying unutilised.
- Learning about societal trends in the past through their records.
- Applying learnings from the past to modern society.

See: Hildebrandt (2016), Garton (1988)



Introduction to the Anatomy Registers project

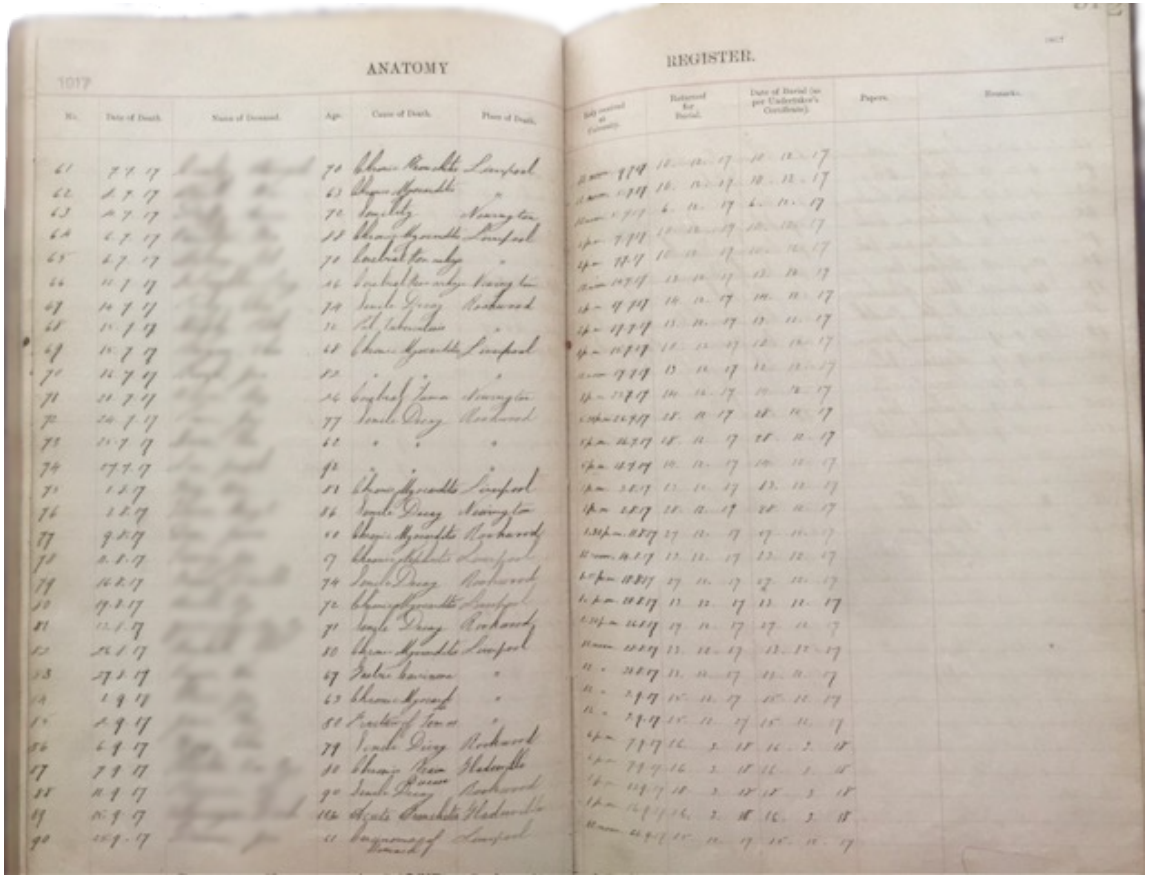


Figure: The primary data in this research project: the Anatomy Registers

Introduction to the dataset

13 attributes taken from the anatomy registers:

- Year
- ID
- Sex
- Age
- Cause of death
- Place of death
- Time of reception
- Date of death
- Date of reception
- Date of return for sepulture
- Date of burial or cremation
- Retention time
- Cemetery

Introduction to the Anatomy Registers project

Year Code	ID	Sex	Age	Date of death	TIME OF RECEPTION	DATE OF RECEPTION	DATE OF RETURN FOR SEPULTURE	DATE OF BURIAL OR CREMATION	@ # or ^	RETENTION TIME		Extra	CAUSE OF DEATH	PLACE OF DEATH	PLACE OF DEATH CATEGORY CODE	SOURCE CODE	DONOR REGISTRATION	DONOR STATUS CODED	CEMETERY	Year Code	Age < 20 or Missing	Missing Age
1883	1	M	71	8/5/1883		11/5/1883	13/6/1883	13/6/1883		33			Diarrhoea & Broncholitis	Callan Park Hospital	100	12		13	Rookwood	1883	0	0
1883	2	M	78	13/5/1883		15/5/1883	19/5/1883	19/5/1883		4			?Diarrhoea & Debility	Sydney Hospital	23	11		11	Rookwood	1883	0	0
1883	3	M	38	22/5/1883		24/5/1883	16/10/1883	16/10/1883		145			General Paralysis	Gladesville Mental Hospital	3	12		13	Rookwood	1883	0	0
1883	4	M	60	7/7/1883		08/07/1883	16/10/1883	16/10/1883		100			Psoriasis & Debility	Sydney Hospital	23	11		11	Rookwood	1883	0	0
1883	5	M	52	10/7/1883		13/7/1883	16/10/1883	16/10/1883		95			Internal Injuries	Sydney Hospital	23	11		11	Rookwood	1883	0	0
1883	6	M	38	27/7/1883		28/07/1883	16/10/1883	16/10/1883		80			Phthisis (Tuberculosis)	Sydney Hospital	23	11		11		1883	0	0
1883	7	M	57	7/10/1883		8/10/1883	18/10/1883	18/10/1883		10			Effusion on brian	George Street Asylum Parramatta	11	12		12	Rookwood	1883	0	0
1883	8	M	64	13/10/1883		14/10/1883	28/12/1883	28/12/1883		75			Bronchitis	George Street Asylum Parramatta	11	12		12	Rookwood	1883	0	0
1883	9	M	32	28/10/1883		5/11/1883	13/11/1883	13/11/1883		8			Syncope from diarrhoea - probably intestinal obstruction	Temora, NSW	44			4	Rookwood	1883	0	0
1883	10	M	62	4/10/1883		5/10/1883	18/12/1883	18/12/1883		74			Diarrhoea	George Street Asylum Parramatta	11	12		12	Rookwood	1883	0	0
1884	11	M	55	22/3/1884		22/3/1884	13/5/1884	30/05/1884		53			Valvular disease of the heart	Macquarie Street Asylum	12	12		12	Rookwood	1884	0	0
1884	12	M	79	3/4/1884		5/4/1884	14/6/1884	14/06/1884		70			Senile decay	Macquarie Street Asylum	12	12		12	Rookwood	1884	0	0
1884	13	M	85	17/4/1884		17/4/1884	14/6/1884	14/06/1884		58			Senile decay	Macquarie Street Asylum	12	12		12	Rookwood	1884	0	0
1884	14	M	64	16/5/1884		16/5/1884	14/7/1884	14/07/1884		59			Paralysis	George Street Asylum Parramatta	11	12		12	Rookwood	1884	0	0
1884	15	M	84	4/6/1884		4/6/1884	1/8/1884			59			Paralysis	Macquarie Street Asylum	12	12		12		1884	0	0
1884	16	M	64	23/6/1884		23/6/1884	02/09/1884	03/09/1884		64			Bronchitis	George Street Asylum Parramatta	11	12		12	Rookwood	1884	0	0
1884	17	M	72	14/7/1884		14/7/1884	1/8/1884			18			Senile decay	George Street Asylum Parramatta	11	12		12		1884	0	0
1884	18	M	72	17/7/1884		17/7/1884	24/09/1884			69			Senile decay	George Street Asylum Parramatta	11	12		12		1884	0	0
1884	19	F	76	18/7/1884		18/7/1884	23/09/1884			67			Old age, jaundice	Hyde Park Asylum	10	12		12		1884	0	0
1884	20	M	64	31/07/1884		1/8/1884	02/09/1884	03/09/1884		32			-	George Street Asylum Parramatta	11	12		12	Rookwood	1884	0	0
1884	21	M	55	02/09/1884		2/9/1884	12/11/1884			61				George Street Asylum Parrama	11	12		12		1884	0	0
1884	22	M	63	4/10/1884		4/10/1884	29/12/1884			86	missing place	3		GEORGE STREET ASYLUM	11	12		12		1884	0	0
1884	23	M	25	30/11/1884		30/11/1884	13/1/1885	15/01/1885		45				SYDNEY HOSPITAL	23	11		11	Rookwood	1884	0	0
1884	24	M	20	2/12/1884		2/12/1884	13/1/1885	15/01/1885		33			-	GLADESVILLE MENTAL HOSPITAL	3	13		13	Rookwood	1884	0	0
1885	25	M	20	12/1/1885		12/1/1885	8/3/1885			55				Sydney Hospital	23	11		11	ROOKWOOD	1885	0	0
1885	26	M	55	23/3/1885		23/3/1885	9/5/1885			47				GLADESVILLE MENTAL HOSPITAL	3	13		13	ROOKWOOD	1885	0	0
1885	27	M	52	22/3/1885		23/3/1885	20/5/1885			58				GLADESVILLE MENTAL HOSPITAL	3	13		13	ROOKWOOD	1885	0	0
1885	28	M	85	26/3/1885		26/3/1885	9/5/1885			44				LIVERPOOL ASYLUM	14	12		12	ROOKWOOD	1885	0	0
1885	29	M	45	14/4/1885		14/4/1885	25/6/1885			72				LIVERPOOL ASYLUM	14	12		12		1885	0	0
1885	30	M	56	21/4/1885		21/4/1885	25/6/1885			65				GLADESVILLE MENTAL ASYLUM	3	13		13		1885	0	0
1885	31	M	83	22/4/1885		22/4/1885	4/7/1885			73				PARRAMATTA HOSPITAL FOR THE INSANE	5	13		13		1885	0	0

Figure: From cursive to Calibri – the first few rows of Dr. Jenkin's transcribed data

Introduction to the dataset

- The dataset contained 7609 records spanning 101 years from 1883 to 1983.
- The format was typical of most manually-collected datasets.
- No cleaning had been conducted on the dataset, other than basic checks during entry.
- All data was de-identified in line with USYD's Human Research Ethics Committee approval.[1]

[1] *"Investigation of historical, demographic and medical information in cadaver records held by the Discipline of Anatomy."* Project 2017/898

The scope for SCDL3991

Original project aim

Compare cause of death with place of death to uncover trends in regards to how and where people died.

Updated project aim

Improve the data quality of the Asset Register dataset and bring it into a state useable for analysis.

Table of Contents

Ensuring data completeness

Data completeness

The extent to which all required or expected records and data fields are present in a dataset.

Potential solutions:

- Delete the missing entry
- Replace with a statistical summary (mean/median)
- Impute the missing value based on context

Ensuring data completeness

```
1 data.isna().sum()
```

```
year 0
id 0
sex 5
age 10
death_date 2436
reception_date 0
return_date_sepulture 1680
burial_date 4408
death_place 6
place_code 0
```

Solution

Replaced missing date of death with the date the body reached USYD

Ensuring data consistency

Data consistency

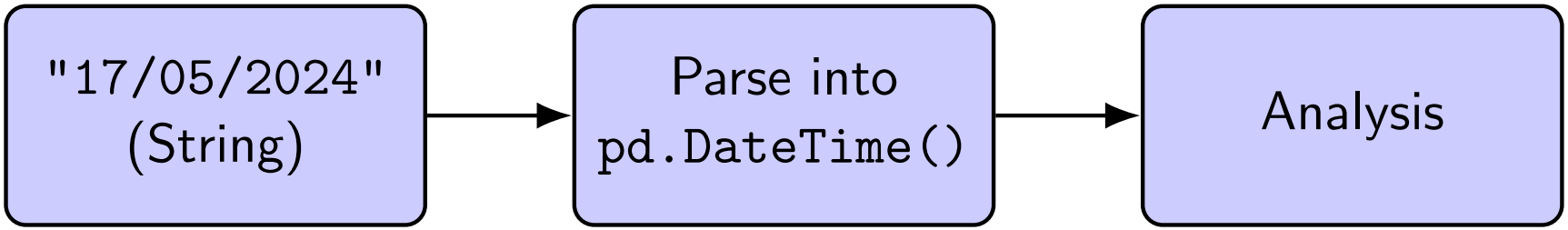
Data adheres to a set of predefined rules and formats, ensuring uniformity in the representation of information.

i.e. The representation of the data is actually useable (programmatically or in the dataset itself).

Ensuring data consistency

Formatting dates correctly

Check to see if our 4 date columns are in the right format: `death_date`, `reception_date`, `return_date_sepulture` and `burial_date`.



Ensuring data consistency

Formatting dates correctly

Match dates in the format DD/MM/YYYY

```
"^(0?[0-9]|[12]\d|3[01])(0?[0-9]|1[0-2])(000\d|00\d{2}
|0\d{3}|1\d{3}|2000)$"
```

- **Day Component:** `"^(0?[0-9]|[12]\d|3[01])"`
  - `"0?[0-9]"`: Matches a single digit day from 01 to 09
  - `"[12]\d"`: Matches any two-digit day from 10 to 29
  - `"3[01]"`: Matches the days 30 or 31
- **Month Component:** `"(0?[0-9]|1[0-2])"`
  - `"0?[0-9]"`: Matches a single digit month from 01 to 09
  - `"1[0-2]"`: Matches the months 10, 11, and 12
- **Year Component:** `"(000\d|00\d{2}|0\d{3}|1\d{3}|2000)"`
  - Iteratively matches any year between 0000 and 2000

Ensuring data consistency

Formatting dates correctly

```
1 data[~data["death_date"].str.match(regex)]
```

Some notable findings from running the regex matcher on our dataset:

Ensuring data consistency

Formatting dates correctly

```
1 data[~data["death_date"].str.match(regex)]
```

Some notable findings from running the regex matcher on our dataset:

- `id=5044, death_date = "20/11/9161"`

Ensuring data consistency

Formatting dates correctly

```
1 data[~data["death_date"].str.match(regex)]
```

Some notable findings from running the regex matcher on our dataset:

- `id=5044, death_date = "20/11/9161"`
- `id=5014, death_date = "26/87/1961"`

Ensuring data consistency

Formatting dates correctly

```
1 data[~data["death_date"].str.match(regex)]
```

Some notable findings from running the regex matcher on our dataset:

- `id=5044, death_date = "20/11/9161"`
- `id=5014, death_date = "26/87/1961"`
- `id=5424, death_date = "31/06/1964"`

Ensuring data validity

Data validity

The data accurately represents the real-world values it is supposed to depict.

i.e. The values in the dataset make sense. This often involves checking the values of one attribute against another.

Ensuring data validity

Checking the relationship between dates

```
# Entries with difference greater than 10 days
data[(data["reception_date"] - data["death_date"]).dt.days
     > 10]

# Entries where death_date is after reception_date
data[data["death_date"] > data["reception_date"]]
```

id	death_date	reception_date
1560	26/03/1913	27/03/1916
5036	23/10/1961	23/10/1951

In reality, there were dozens of such entries.

Introduction to data transformation

Data transformation involves transforming some parts of the dataset in order to make performing statistical analysis on it more feasible.

Often, unstructured data (such as textual data) benefits the most from such data cleaning.

A range of responses in death_place

Some practitioners took extreme care and precision when entering in the place of death:

- "Former" and "new" names
  - `st george district hospital formerly 301 queen st concord west`
  - `netherleigh private hospital now eastern suburbs private hospital chapel st randwick`
- Compound locations
  - `hornsby hospital from oatlands nursing home dundas`
  - `9 redmyre rd strathfield and royal north shore hospital`

A range of responses in `death_place`

Some practioners were on the other end of detail:

- Minimally descriptive entries
  - `at home`
  - `donor`
  - `morgue`
- Suburbs only
  - `orange`
  - `waverley`

A range of responses in death_place

And some practioners were in-between:

- `royal prince alfred hospital camperdown page pavilion`
- `royal prince alfred hospital`
- `rpa`

A range of responses in death_place

Some people passed away in odd spots:

- in omnibus kingslangly rd greenwich
- at sea ss ellinis
- in the road guildford rd guildford

Where do you even begin to analyse such varied data?

Categorisation of death_place institution

Category	**Classification criteria**
Private residence	A private address (commencing with a street, flat, lot or unit number)
Aged care	Locations containing "Nursing", "Convalescent", "Aged Care", ...
Hospice	Locations containing "Hospice" + some known hospices
Private hospital	Locations containing "Private Hospital" or "surgery"
Public hospital	All hospitals without "private" in their name, including district hospitals
Public asylum	Currently coded 12 + three special cases (Liverpool, George St, Rockwood)
Public mental asylum	Currently coded 13 + locations containing "mental asylum" or "psychiatric"
Aged 2 and under	Currently coded 70
Morgues	Currently coded 6 + locations containing "morgue"
Other	Other

Extracting suburbs from `death_place`

Another transformation we could perform using the place of death attribute is to categorise it based on the suburb of where the patient died.

lot 611 20 augusta st casula	→	Casula
linwood nursing home 87 bowden st ryde		Ryde
howard avenue dee why		Dee Why
5b699 military rd mosman		Mosman
sydney adventist hospital 185 fox valley rd wahroonga		Wahroonga
leumeah nursing home 284 castle hill rd castle hill		Castle Hill
broughton vale rd berry		Berry
mana house hospital 17 pacific highway wahroonga		Wahroonga
mayfair nursing home marrickville rd marrickville		Marrickville
macquarie lodge 171 wollongong st arncliffe		Arncliffe
bankstown hospital eldridge rd bankstown		Bankstown
the prince of wales hospital high st randwick		Randwick
59 watkins rd wangi wangi		Wangi Wangi
illawarra hospital		Wollongong

Extracting suburbs from `death_place`

We utilised a NSW Government dataset containing the geographical boundaries of all localities in New South Wales.

suburb	geometry
Aarons Pass	`POLYGON (...)`
Abbotsbury	`POLYGON (...)`
Abbotsford	`POLYGON (...)`
Abercrombie	`POLYGON (...)`
Abercrombie River	`POLYGON (...)`
...	...

Extracting suburbs from death_place

```
regex = r'\b(' + '|'.join(suburbs) + r')\b'
matches = donors["death_place"].str.extractall(regex)
```

Some entries were matched with multiple suburbs.

id	suburb_1	suburb_2	suburb_3
...	...	...	...
7601	ettrick	ashbury	-
7602	lidcombe	-	-
7603	castle hill	-	-
7604	rockdale	woodford	banksia

e.g. `"rockdale nursing home 22 woodford st banksia"`

Table of Contents

A Geographical Analysis of Donor Data

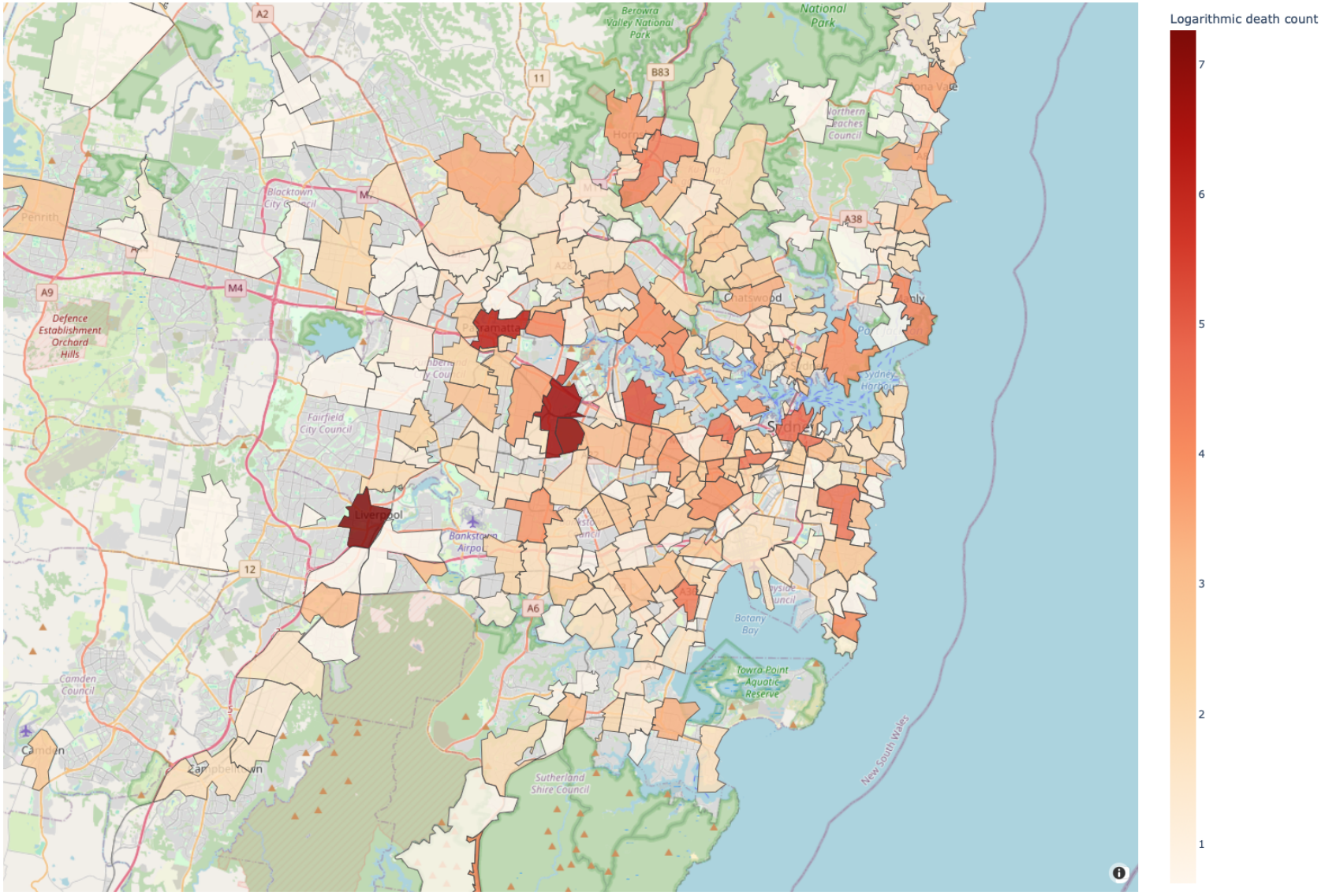
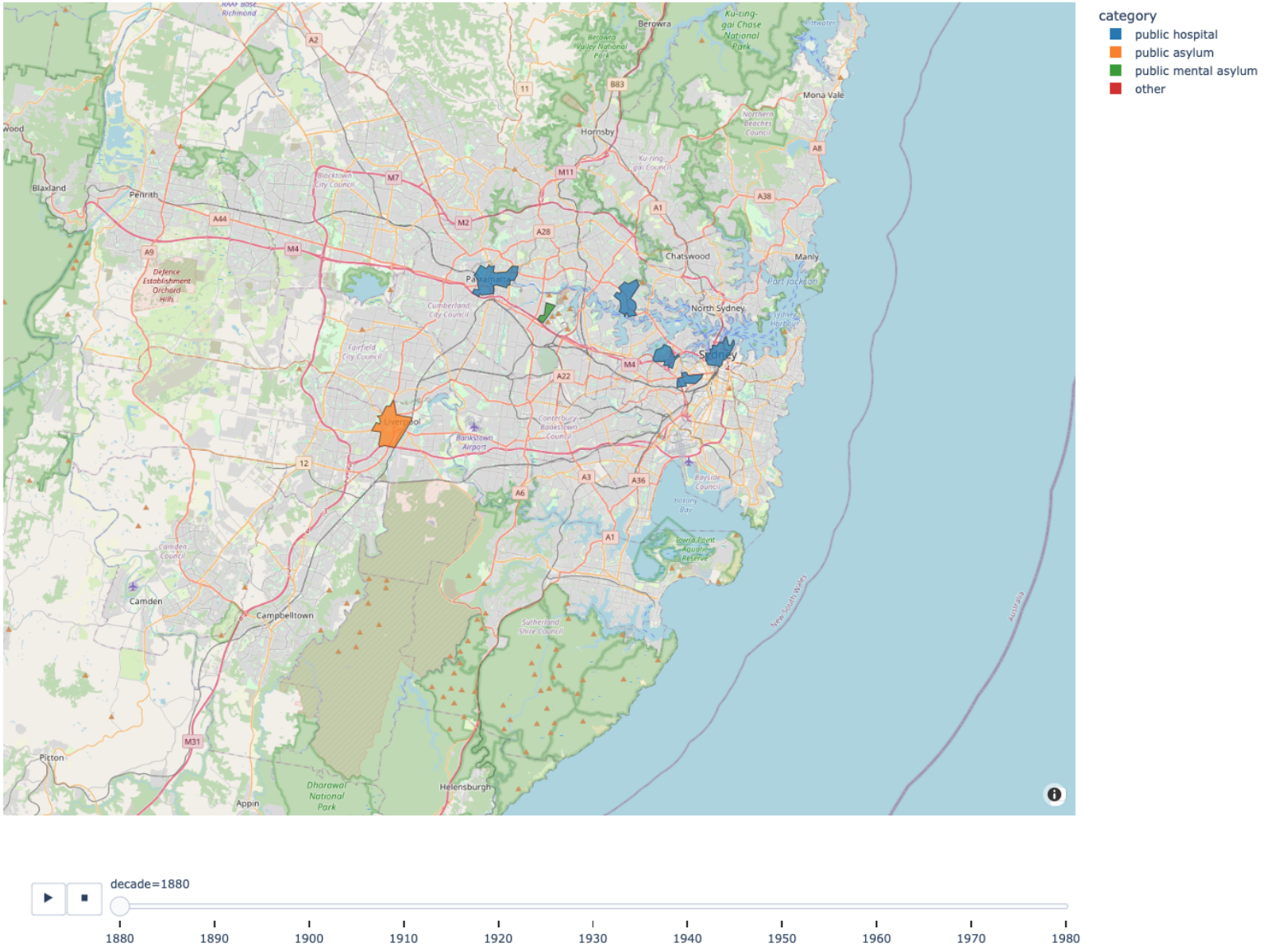
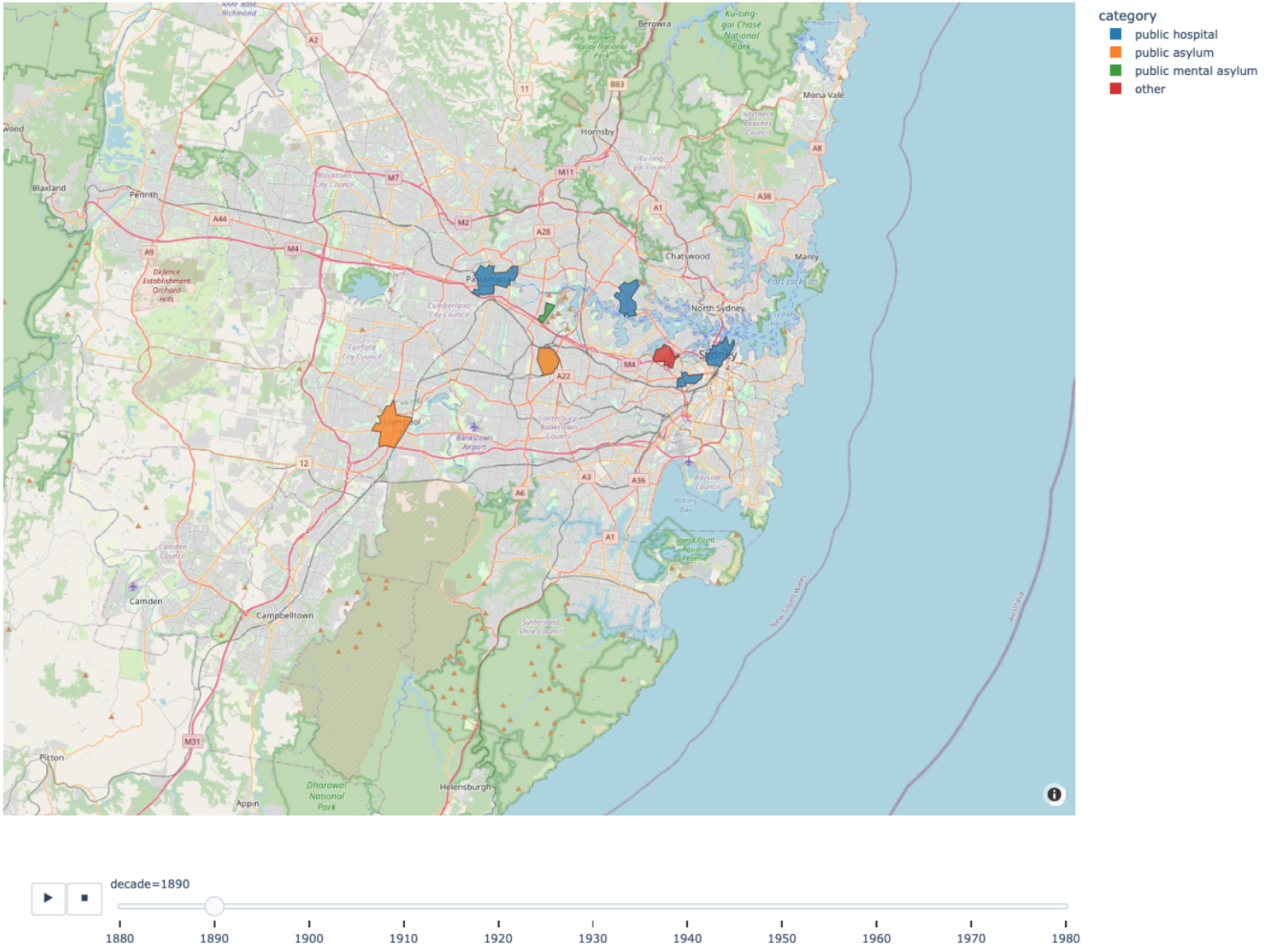


Figure: A visual guide to donor contributions across NSW suburbs

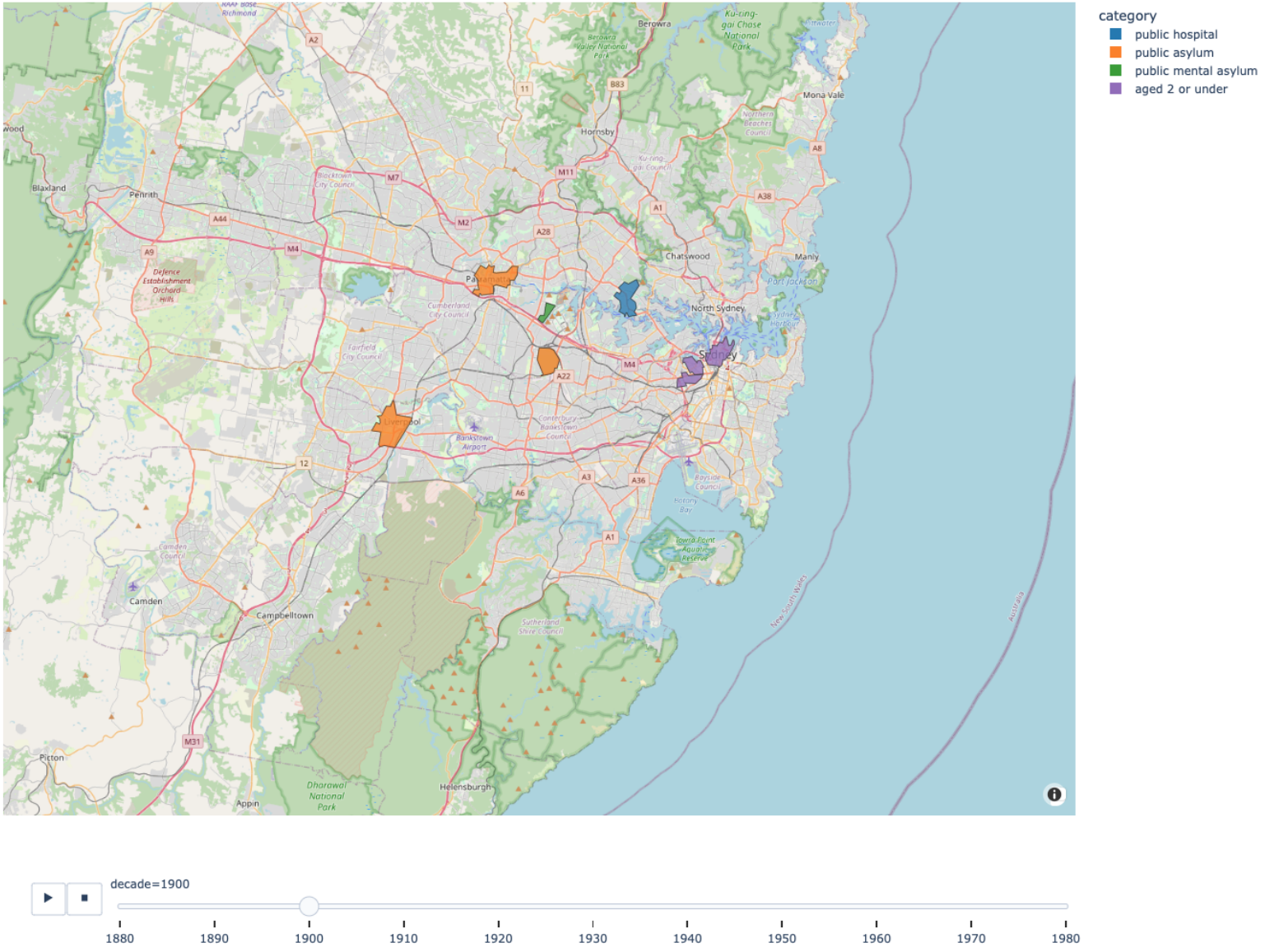
A Geographical Analysis of Donor Data



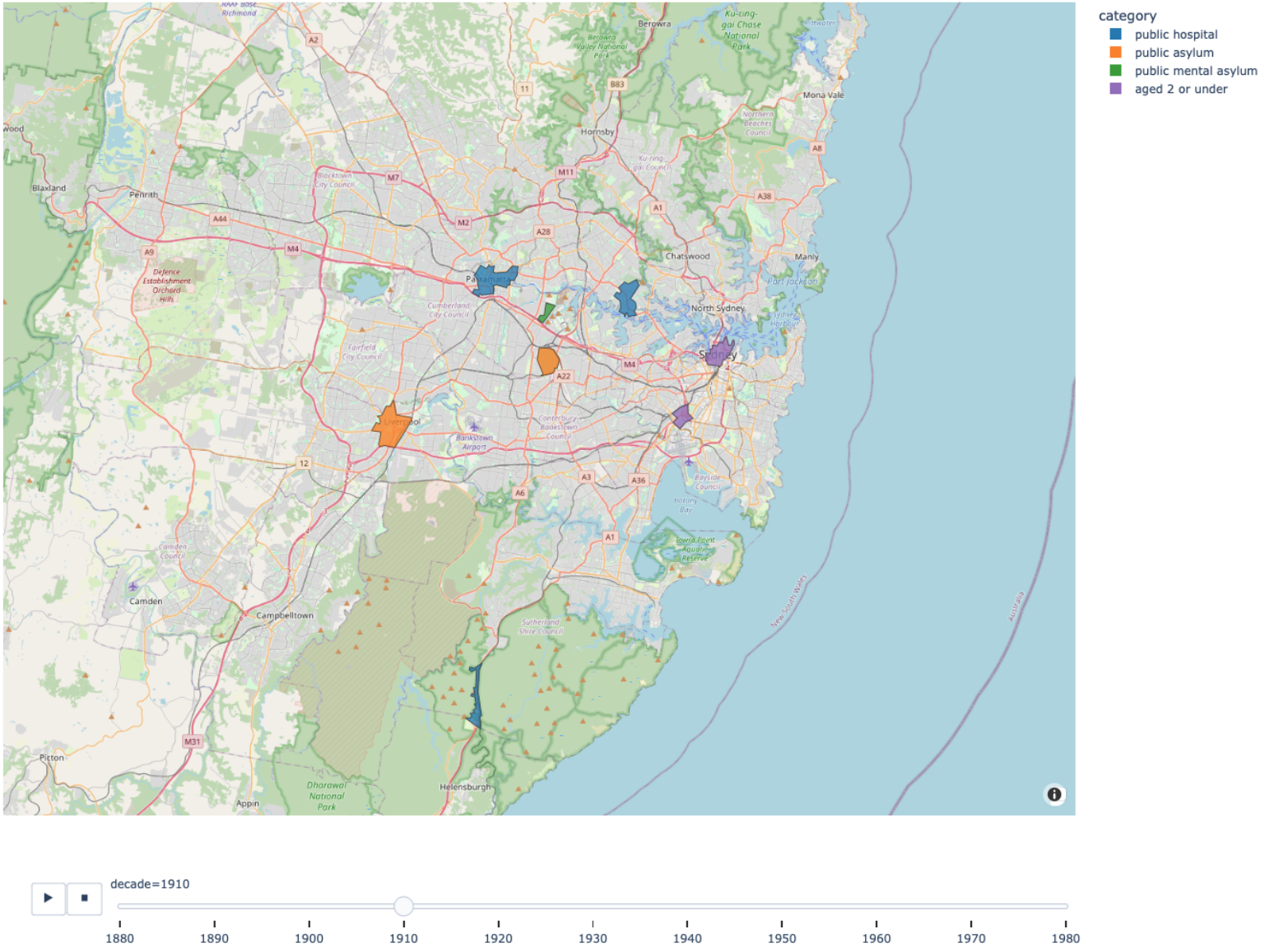
A Geographical Analysis of Donor Data



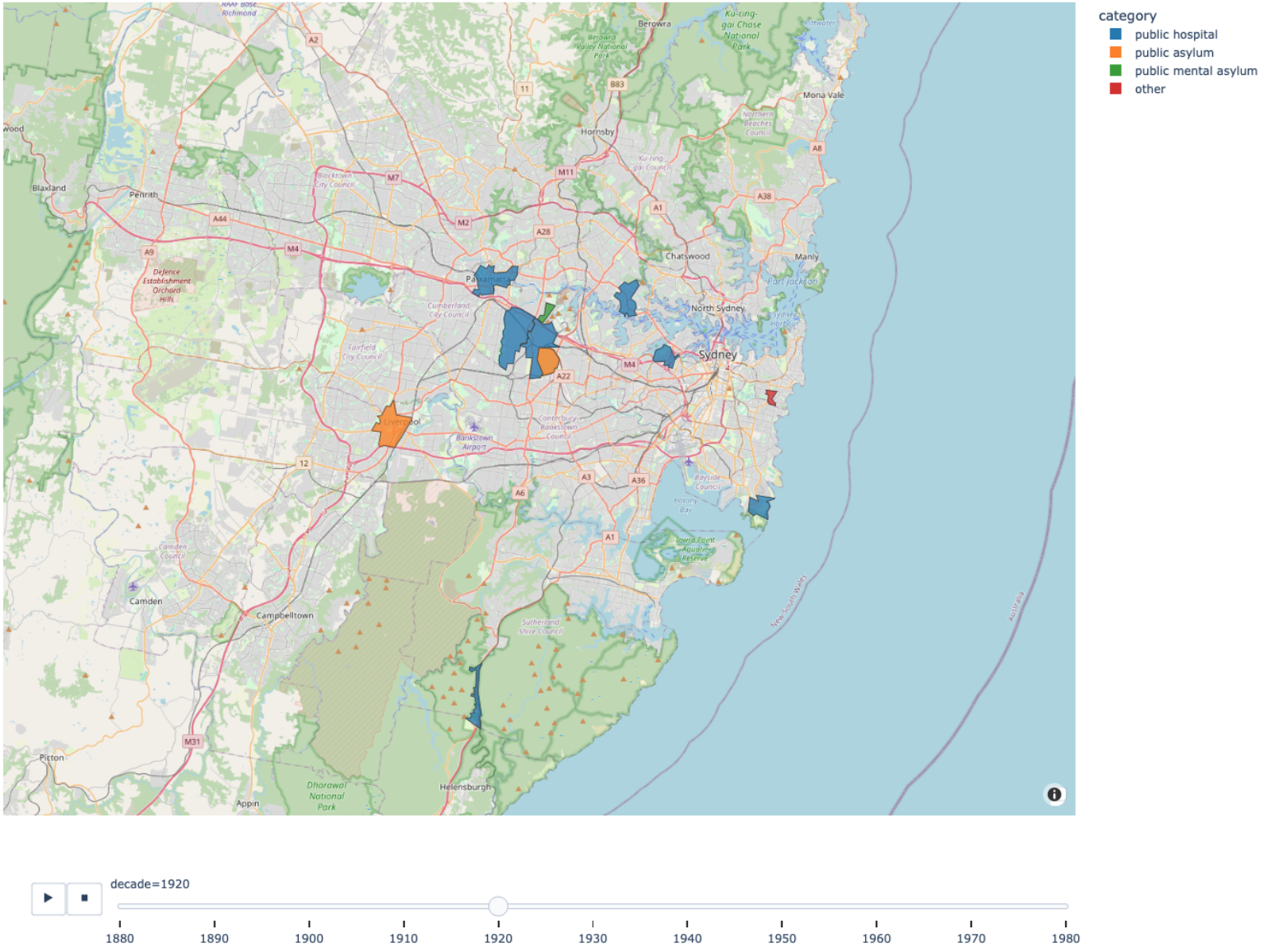
A Geographical Analysis of Donor Data



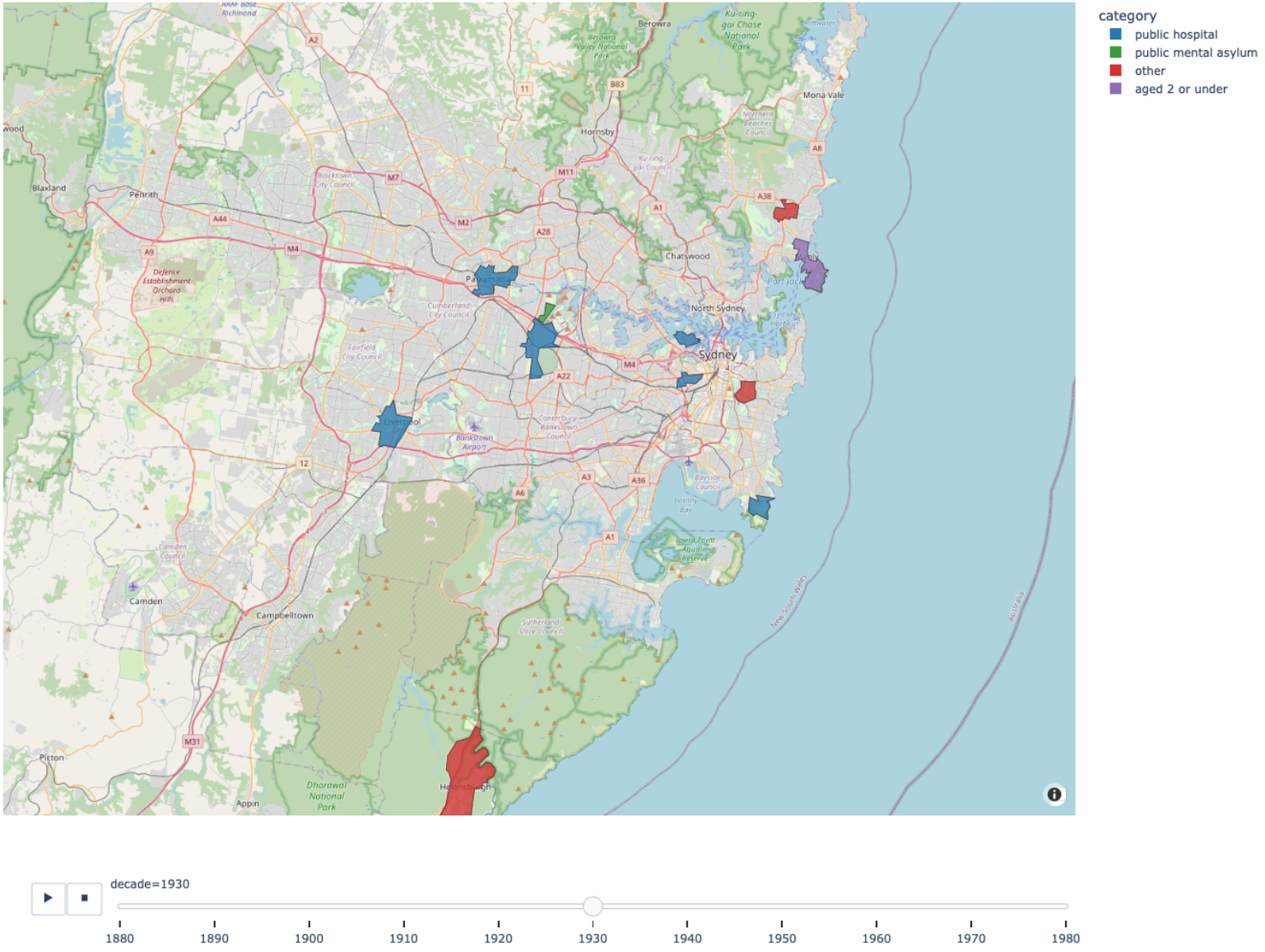
A Geographical Analysis of Donor Data



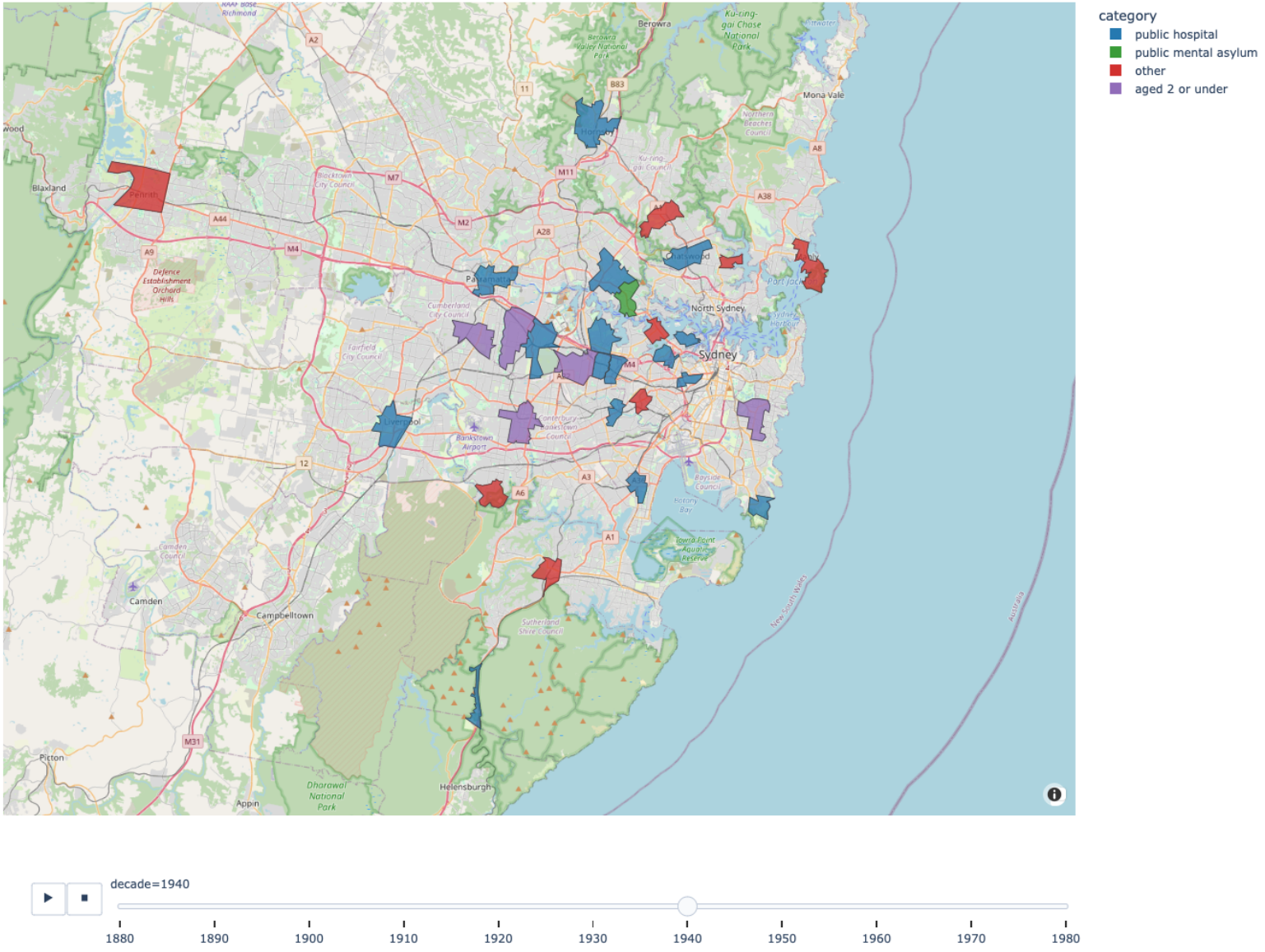
A Geographical Analysis of Donor Data



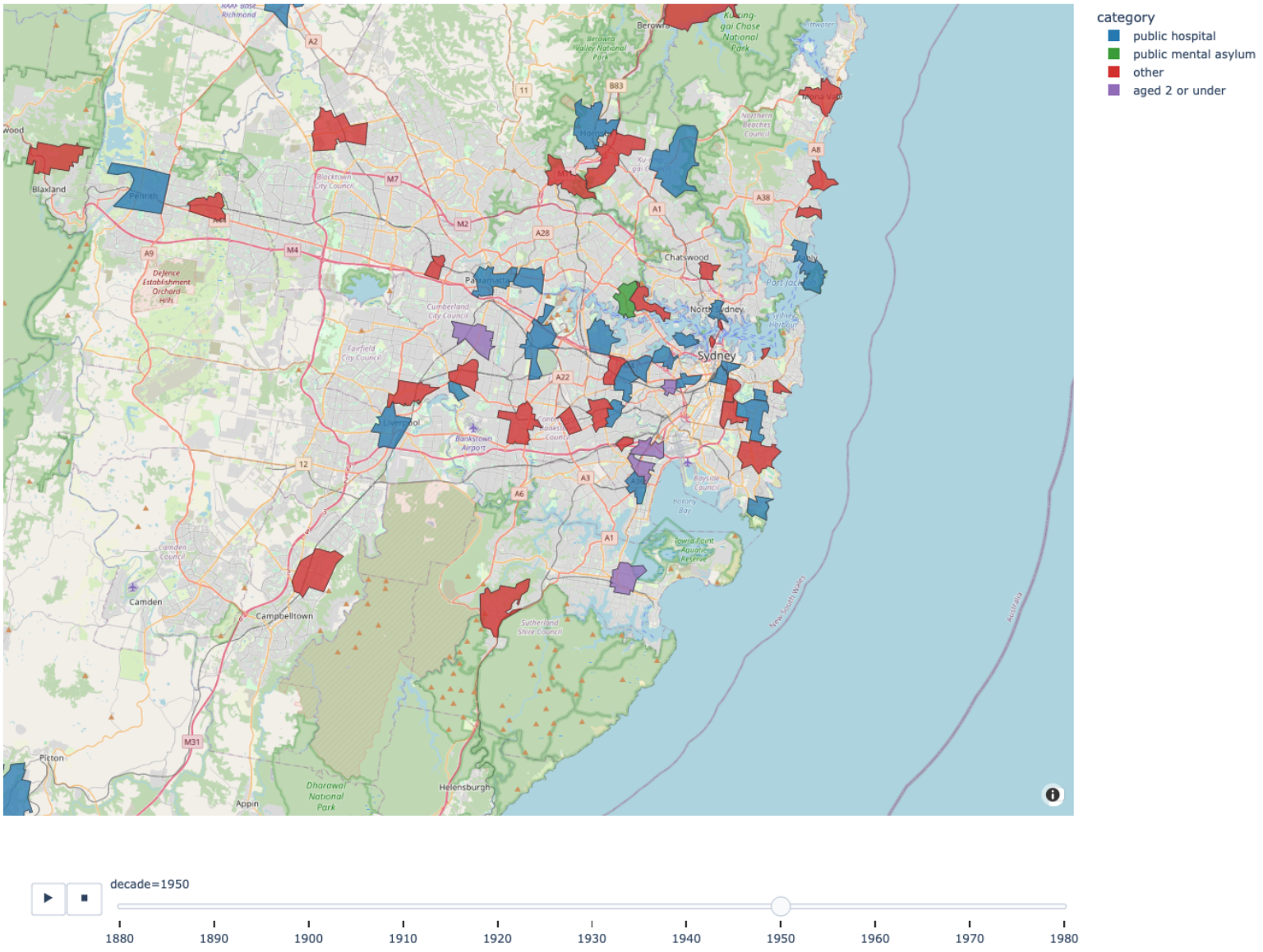
A Geographical Analysis of Donor Data



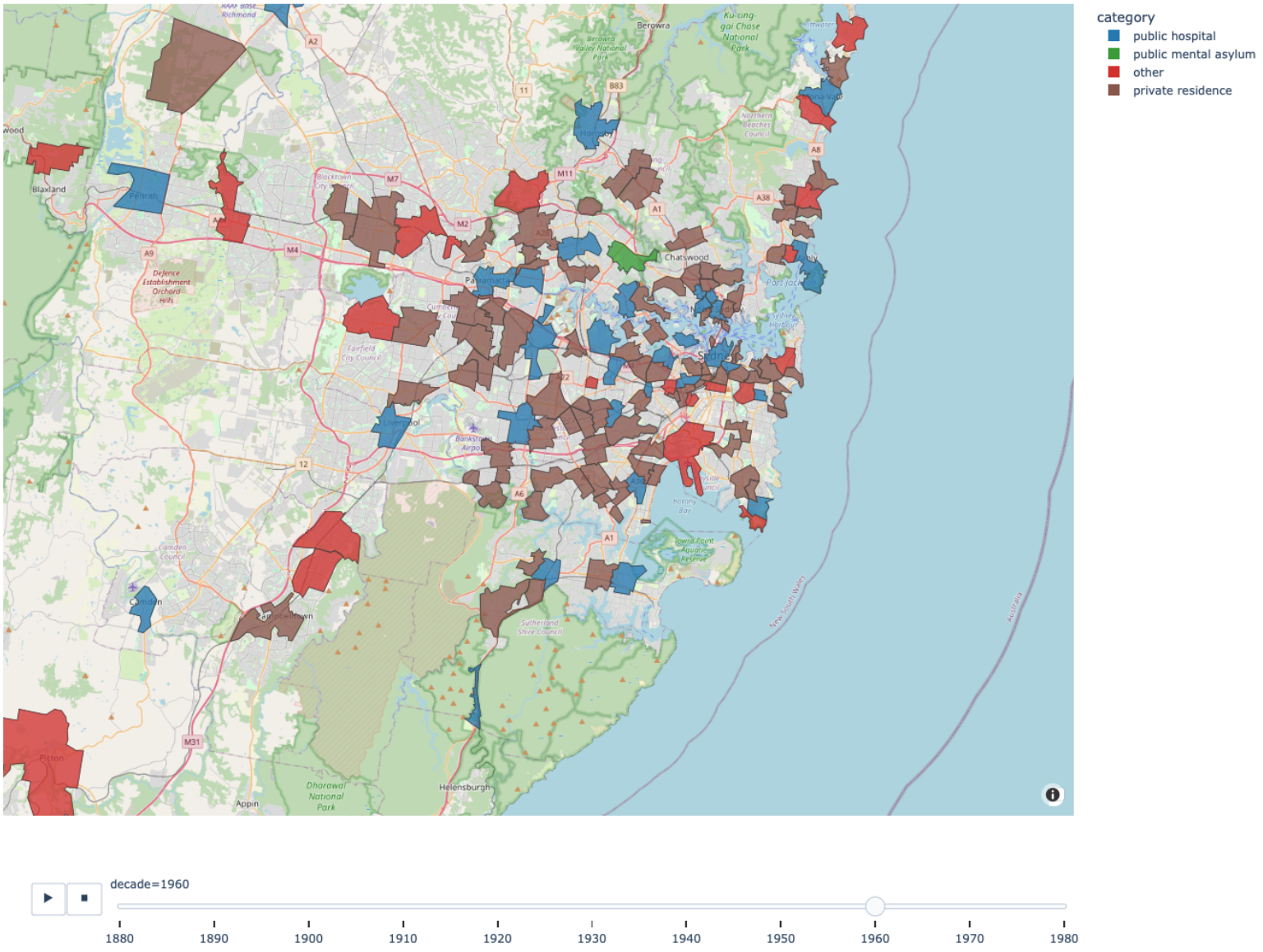
A Geographical Analysis of Donor Data



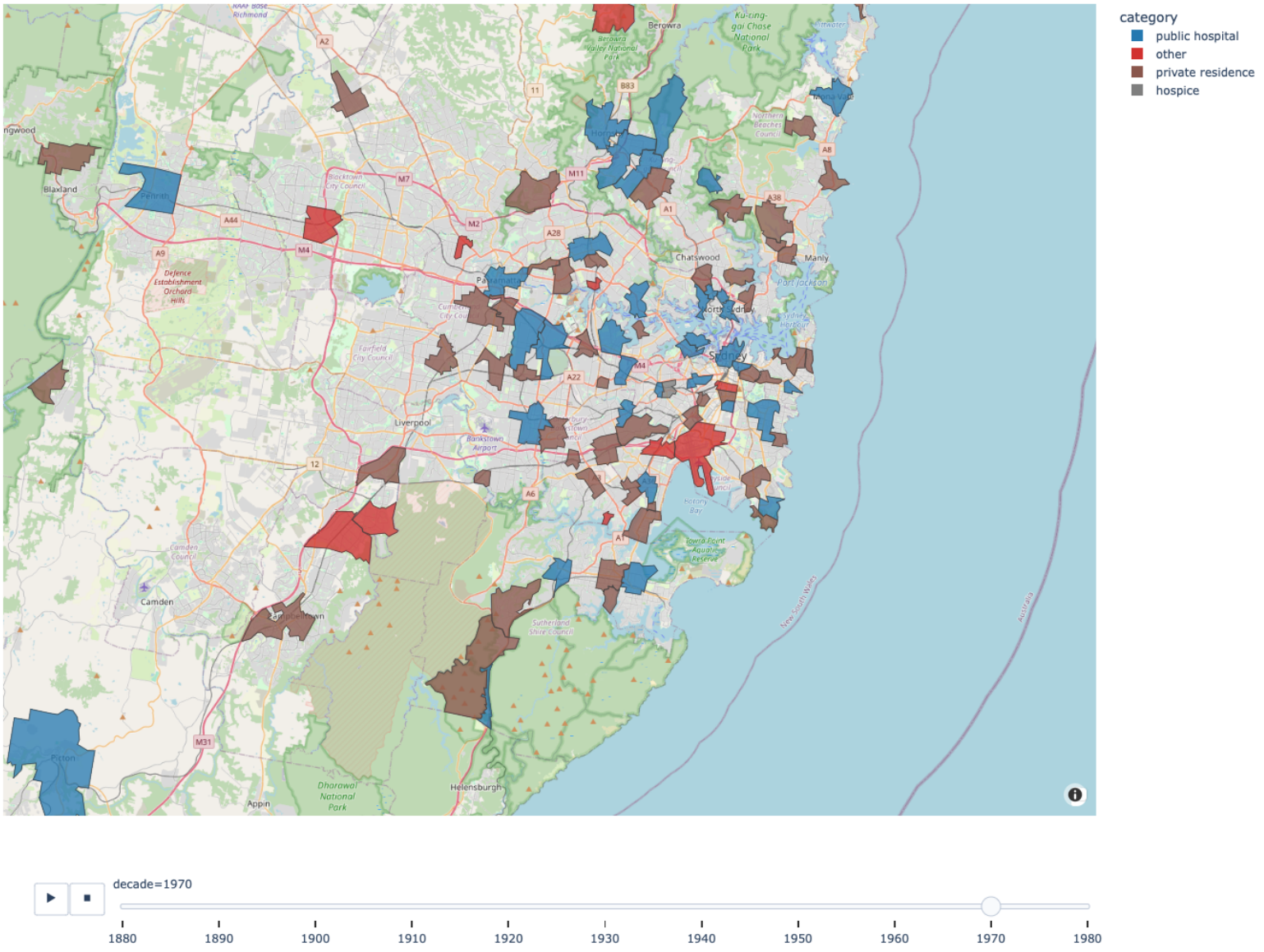
A Geographical Analysis of Donor Data



A Geographical Analysis of Donor Data



A Geographical Analysis of Donor Data



A Geographical Analysis of Donor Data

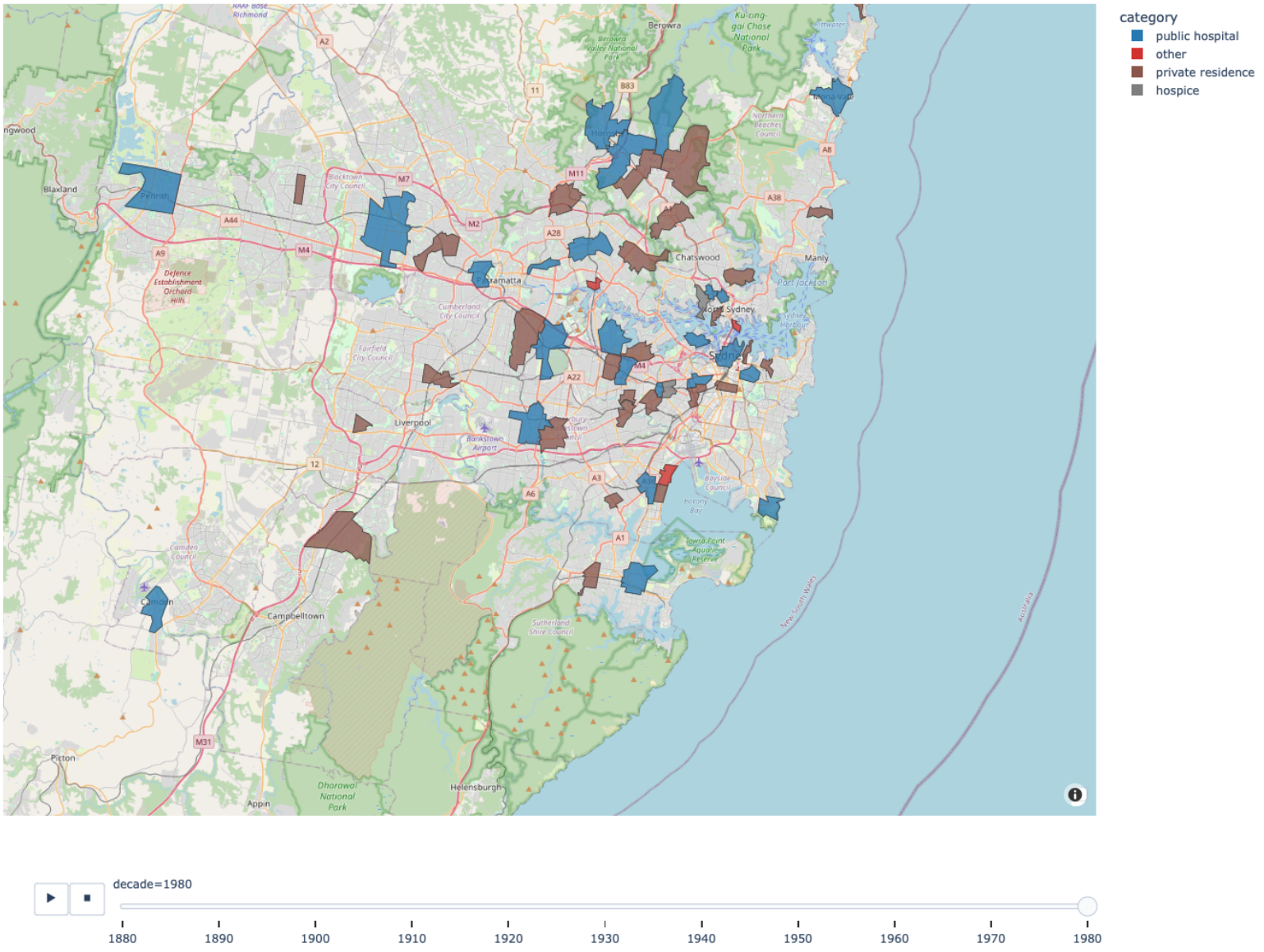


Table of Contents

Conclusion

- Working with hand-transcribed data poses unique challenges as a data analyst, requiring a rigorous and time-consuming data verification process.
- Interesting results can be achieved through structuring data to determine relationships in the dataset.

Limitations

The unstructured attributes in the dataset reduces data quality and limits potential statistical analysis that may uncover interesting findings.

Future work

Further standardisation of place of death

As a result, it is important to spend more time cleaning and structuring the data and attributes to allow for complex analysis.

e.g. How can we programmatically consolidate similar entries into one?

mona vale district hospital
mona vale district hospital 1 coronation st mona vale
mona vale district hospital coronation parade mona vale
mona vale district hospital coronation st
mona vale district hospital coronation st mona vale
mona vale district hospital mona vale
mona vale hospital
mona vale hospital mona vale

Mona Vale Hospital

Future work

Further standardisation of place of death

We tried to achieve this using:

- String similarity techniques
  - Difficult to find the one best similarity cut-off
  - Either too harsh e.g. won't allow for absence of address
  - Or too lenient e.g. will group `balmain district hospital` and `bowral district hospital` together
- Simple ML approaches
  - Attempted to use KMeans clustering (unsupervised learning) to try and cluster the entries together
  - An issue was that we needed to specify the number of clusters, which is unknown

Resources

All code written for this project is available in a well-documented Jupyter Notebook, which can be accessed at github.com/antrikshdhand/SCDL3991-research.

Acknowledgements

I extend my deepest appreciation to my supervisors, Dr. Rebekah Jenkin and Professor Kevin Keay, for helping me see this project through from start to finish despite the many roadblocks in the middle. This project would not have been made possible without our extensive, and often tangential, discussions.

Thank You

Any questions?

Contact information

- Email: adha5655@uni.sydney.edu.au
- GitHub: github.com/antrikshdhand
- LinkedIn: linkedin.com/in/antrikshdhand

References and further reading

Rebekah A. Jenkin.

Altruism in Death. Historic and Contemporary Use of Mortal Remains in Anatomical Examination for Education and Research in Australia and New Zealand.

Doctoral dissertation, The University of Sydney, 2023.

Sabine Hildebrandt.

The Anatomy of Murder: Ethical Transgressions and Anatomical Science during the Third Reich.

New York: Berghahn, 2016. ISBN 978-1785330674.

References and further reading

Kostas Petrakis, Samaritakis Georgios et al.

Digitizing, Curating and Visualizing Archival Sources of Maritime History: the case of ship logbooks of the nineteenth and twentieth centuries.

Drassana: revista del Museu Marítim, 28, Mar. 2021, pp. 60-87.

Stephen Garton.

Medicine and Madness: A Social History of Insanity in New South Wales 1880-1940.

Sydney: New South Wales University Press, 1988. ISBN 0868401153.