Geocoding regional and remote ‘poor quality’ address records with confidence

Miro Palfy
Statistical Analyst, SA NT DataLink

The Australian Government provides financial support to SA NT DataLink through the National Collaborative Research Infrastructure Strategy (NCRIS).

Who is SA NT DataLink?

Government of South Australia
Department for Education

Government of South Australia
SA Health

Government of South Australia
Department of Human Services

Northern Territory Government

University of South Australia

SA·NT DataLink

Flinders University

The University of Adelaide Australia

Cancer Council Beat Cancer Project

Improving lives through ground-breaking research
Funded by Cancer Council SA's Beat Cancer Project on behalf of its donors and the South Australian Department of Health

SAHMRI
South Australian Health & Medical Research Institute

hcasa
Health Consumers Alliance of SA Inc.

SA NT DataLink is an unincorporated Joint Venture established by the Joint Venture Consortium Partners, and administered by the University of South Australia on behalf of the Partners.

What does SA NT DataLink do?

SA NT DataLink is data linkage infrastructure bringing together unit record level records on health and human services across jurisdictions, government agencies and organisations, for approved:

- Policy analysis
- Program / Operational Evaluation
- Academic Research

What is Data Linkage?

- Secondary use of administrative data
- Typically population based longitudinal data collected for another purpose
- Linkage may take place across data sets in a single domain or across domains
- Also known as ‘Record Linkage'

Why is Data Linkage Important?

- Invaluable tool for population wide health & humans services research and evaluation of outcomes
- Unbiased picture of the entire population
- Cost-effective relative to other data collection mechanisms
- Enables studies to be done that could not otherwise be performed.

What data does SA NT DataLink link?

Birth Records
States/Territories

Child Protection Data
States/Territories

Public Housing and Homelessness Records
States/Territories

Mental Health Records
States/Territories

Death Records
States/Territories

CRADLE

GRAVE

Child Development Records
States/Territories

Education Records
States/Territories
Non-Government

Criminal Justice
States/Territories

Public Hospital Records
States/Territories
Non-Government

Centrelink Records
Australian Government

Health Condition Registries
States/Territories

SA NT DataLink linked data collections

Health data

- SA Public Hospital Separations
- SA Public Hospital Emergency Dept.
- SA Dental Service (Titanium)
- SA Perinatal (by baby)
- SA Perinatal (by mother)
- SA Child Health Check
- SA Public Mental Health Services
- SA Cervical Screening
- Drug and Alcohol Services SA
- NT Perinatal (Trends) - by baby
- NT Perinatal (Trends) - by mother
- NT Inpatient Activity
- NT Emergency Department
- NT Primary Health Care Collection
- NT Public Hospital Pharmacy

Social Data

- SA Child Protection
- SA Electoral Roll
- SA Youth Justice
- Housing SA - Public housing
- Housing SA - Homelessness to Home
- NT Child Protection
- NT Public Housing (Urban)

National Registries

- Australia and New Zealand Dialysis and Transplant Registry (ANZDATA)
- Australian Orthopaedic Association National Joint Replacement Registry (AOANJRR)

Registries

- SA Cancer Registry
- SA Birth Registry (by baby)
- SA Birth Registry (by mother)
- SA Birth Registry (by co-parent)
- SA Death Registry
- NT Cancer Registry
- NT Birth Registry
- NT Death Registry
- NT Immunisation Register

Education Data

- SA Public School Enrolments Census
- SA NAPLAN
- SA Running Records
- SA English as an Additional Language
- Australian Early Development Census (AEDC SA & NT)
- NT Student Activity
- NT NAPLAN (Public Schools)
- NT NAPLAN (Catholic and Christian)

Geocoding at SA NT DataLink

- Increased demand for geocoding from researchers requiring more accurate addresses for innovative and important spatial analysis
  - Service Planning
  - Evaluation and monitoring of activities and outcomes

- Data custodians also request geocoding services for their organisation's address data to be validated after data collection
- Varying levels of address quality across administrative datasets (from Emergency Department to Electoral Roll).
- Techniques and strategies to manage the risk of re-identification

Factors making geocoding challenging?

Variability in geocoding tools

- Wide range of geocoding products and commercial services available
- Algorithms used by geocoding tools not transparent
- Many geocoding tools have regional focus / bias

1/1 Unknown road

1/1 UNKNOWN ROAD, ACACIA HILLS, NT, 0822

1/1 UNKNOWN ROAD, ALDINGA, SA, 5173

1/1 UNKNOWN ROAD, ARKAROOLA, SA, 5732

Data quality

- Quality of address data is domain dependant with little or poor validation at record creation, especially for legacy datasets
- Poor address quality directly impacts the geocoding success and utility

- High cost of manual review to improve address data at input as well as for output validation
- High-quality research requires at least 90% of records to be consistently and accurately geocoded

Cardiac ARIA research project

- Cardiac **Accessibility and Remoteness Index of Australia** (ARIA): index to measure accessibility to medical care for cardiac emergency patients

- Objective: analyse Cardiac events and the accessibility to Cardiac services to identify potential mismatch between access to and need for services in SA and NT

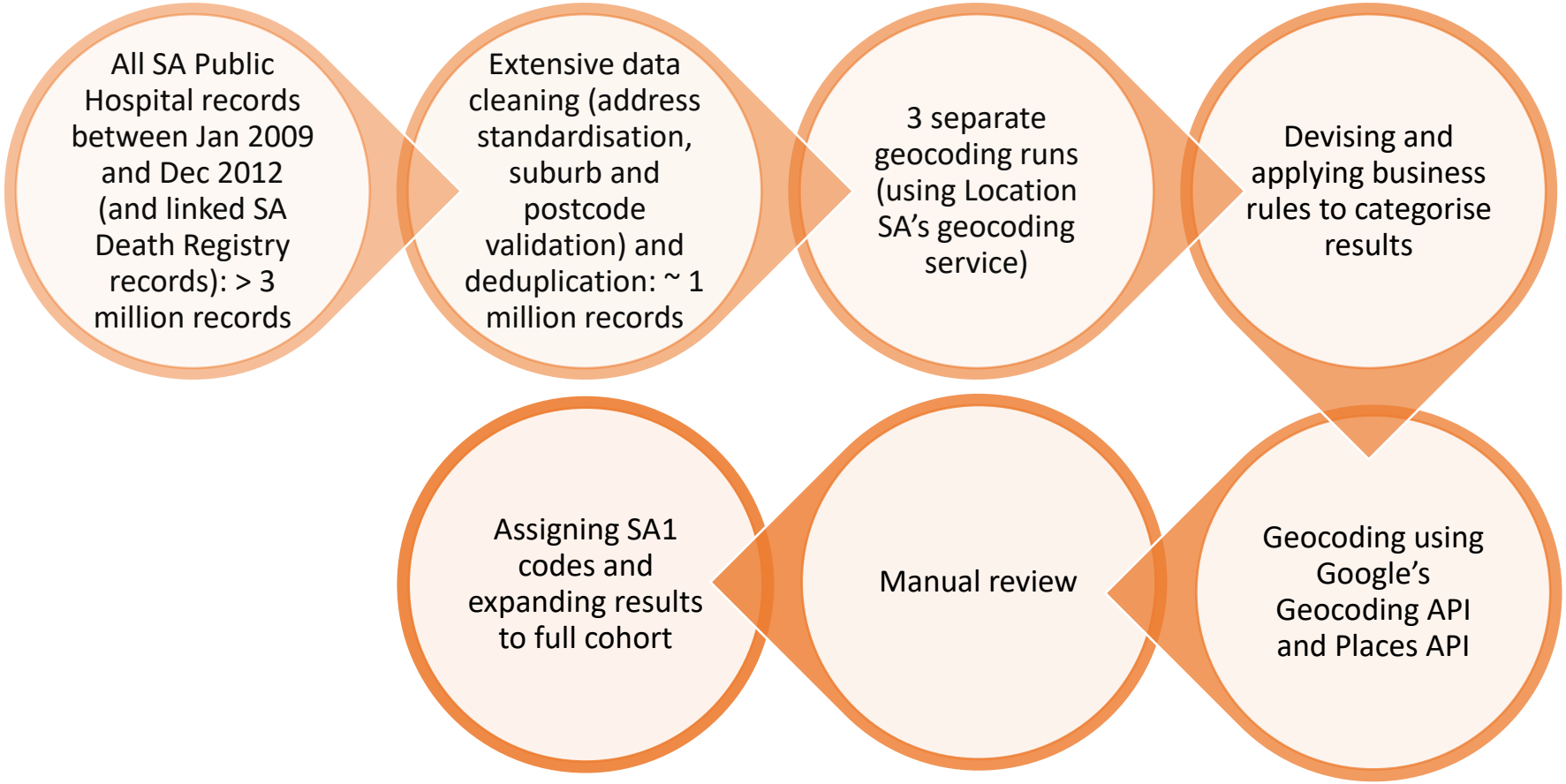
- Measures two sub-indices
  1. The time from 000 call for cardiac emergency until arrival at medical facility
  2. Access to basic services
     (family doctor, pharmacy, cardiac rehabilitation and pathology services)

- Requirement to accurately geocode >90% (SA1) of patient address data collected by public hospitals to reliably calculate distance between patient and service locations

- Research study for the Northern Territory and South Australia led by Professor Robyn Clark, Senior Clinician, from Flinders University.

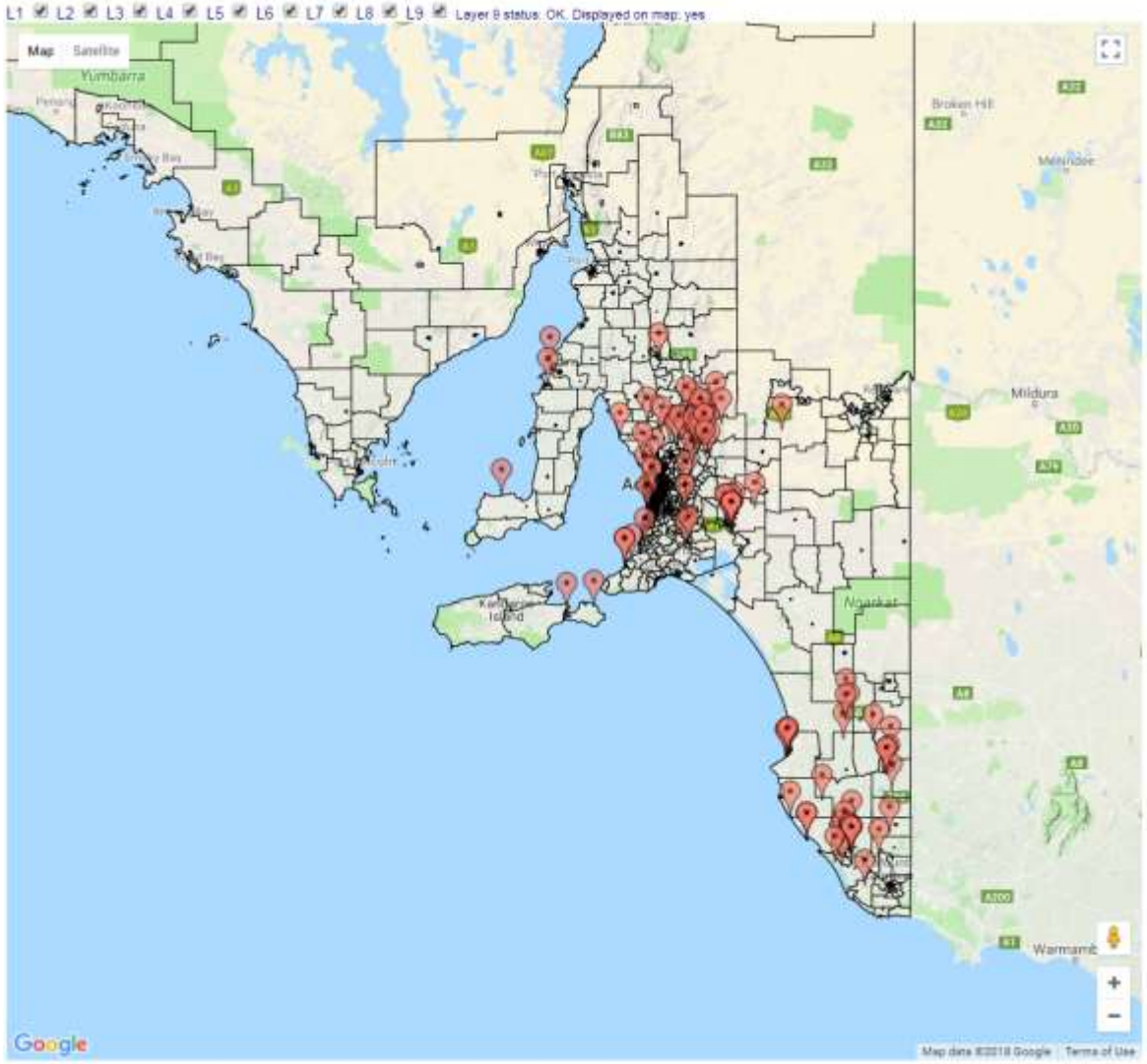
Geocoding Cardiac ARIA: proposed workflow

Geocoding Cardiac ARIA: actual workflow



Google geocoding API

- 2,500 address lookups per day (free service); 100,000 per day (paid, $0.50USD/1000 records)
- Filtering results through region biasing and defining a bounding box
- Alternatively, Google Places API can be used when looking up establishments, geographic locations or prominent points of interest - proved very useful when dealing with incomplete communal addresses (e.g. nursing homes etc.)



Google geocoding API

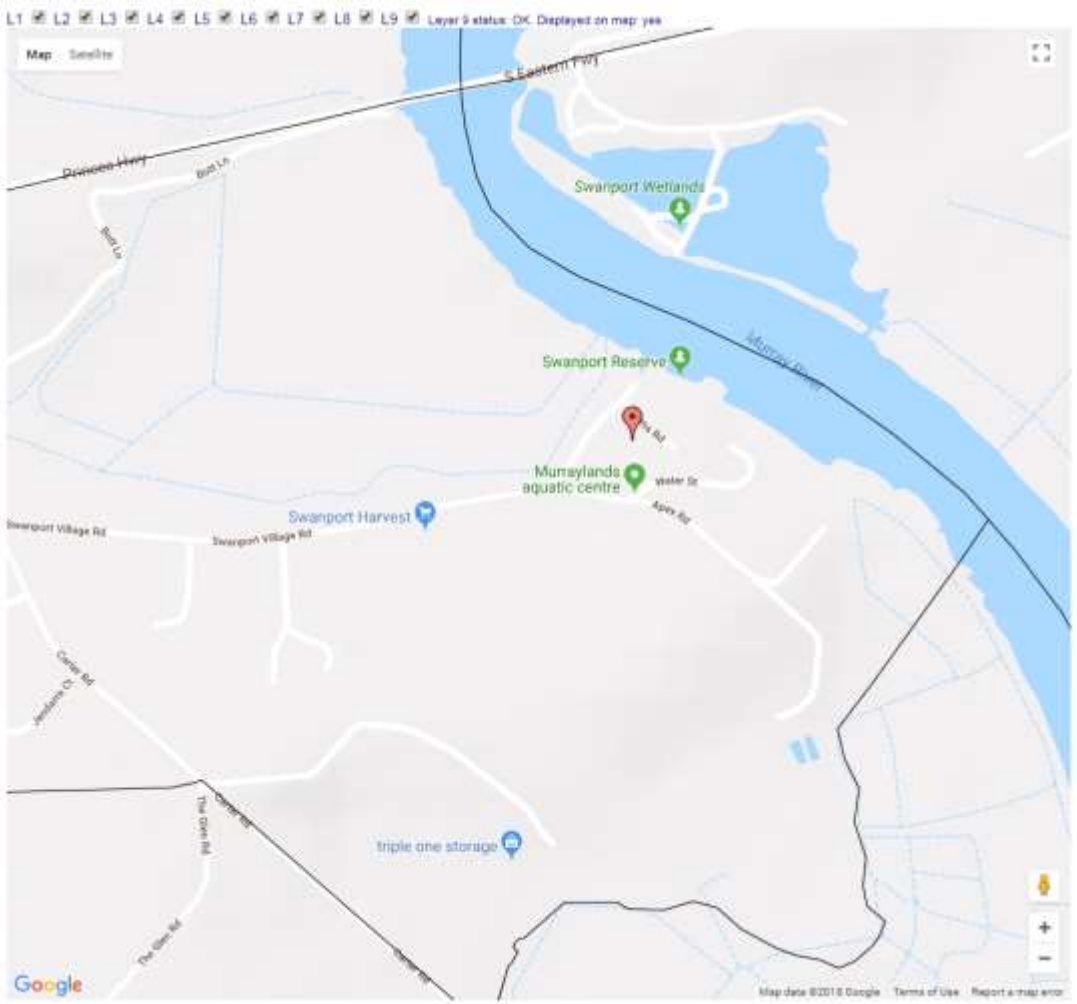
- Batched process using Python and JavaScript
- HTTPS requests, output in JSON (or XML)
- If successful, result may contain one or more locations consisting of multiple fields used to describe location properties and match quality
- As Google collects all data, dummy addresses were added to cohort for obfuscation

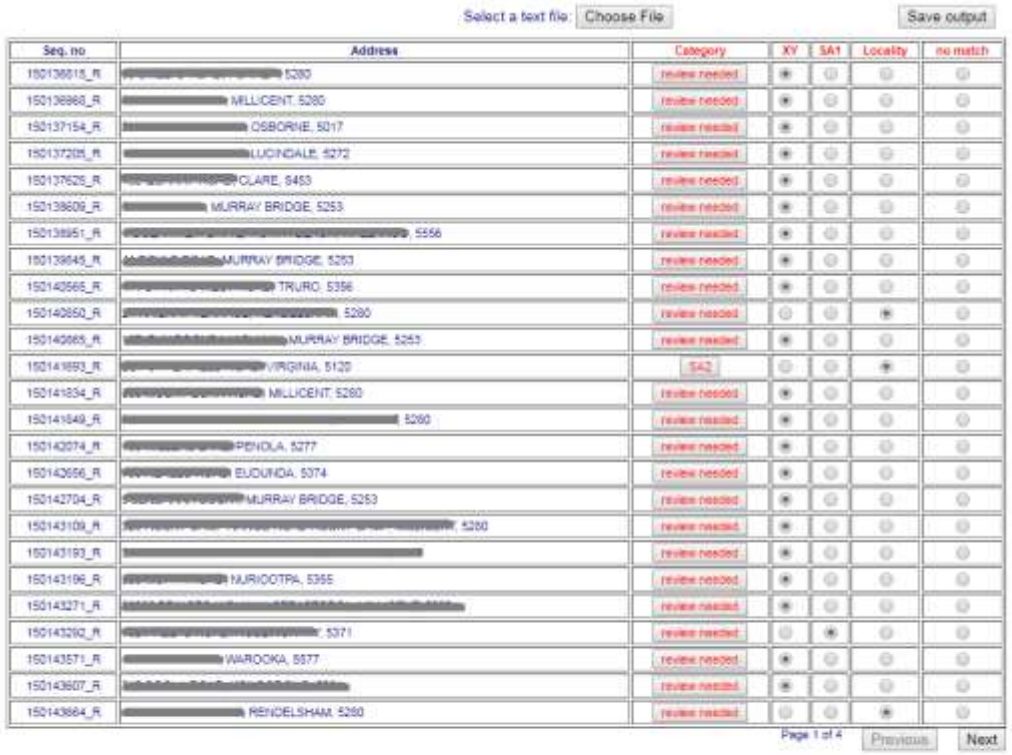
```
        # approximate area in square meters
        area = x_diff * y_diff * 12365161.32
        if area < area_threshold:
            res = False
    return res


def geocode_file(file_name, output, gkey=None, gtimeout=None, gret_timeout=60):
    # initialise geo-coder
    gmaps = googlemaps.Client(key=gkey,
                              timeout=gtimeout,
                              retry_timeout=gret_timeout)
    # read text addresses file
    i = 1
    current_batch = i
    delimiter = "#|#"
    f = open(file_name, 'r')
    out = open(output, 'w')
    res_list = list()
    header = list()
    for line in f:
        try:
            address = str.strip(line)
            print('Geo-coding ' + address)
            raw_list = gmaps.geocode(address)
            if len(raw_list) > 0:
                raw = raw_list[0]
            else:
                raw = dict()
            res_dict = undict("google_geo", raw)
            res_dict['original_address'] = unicode(address)
            res_dict['row_number'] = unicode(i)
            res_dict['geo_match'] = unicode(True)
            if notExact(raw):
                print('Exact match not found, trying \'places\' api...')
                res_dict['geo_match'] = unicode(False)
                try:
                    places_raw = gmaps.places(address)
                    if type(places_raw) == dict:
                        if 'results' in places_raw:
                            if len(places_raw['results']) > 0:
                                print('At least one result found using \'places\' api...')
                                place_raw = places_raw['results'][0]
                                res_dict.update(undict("google_places", place_raw))
                            else:
```

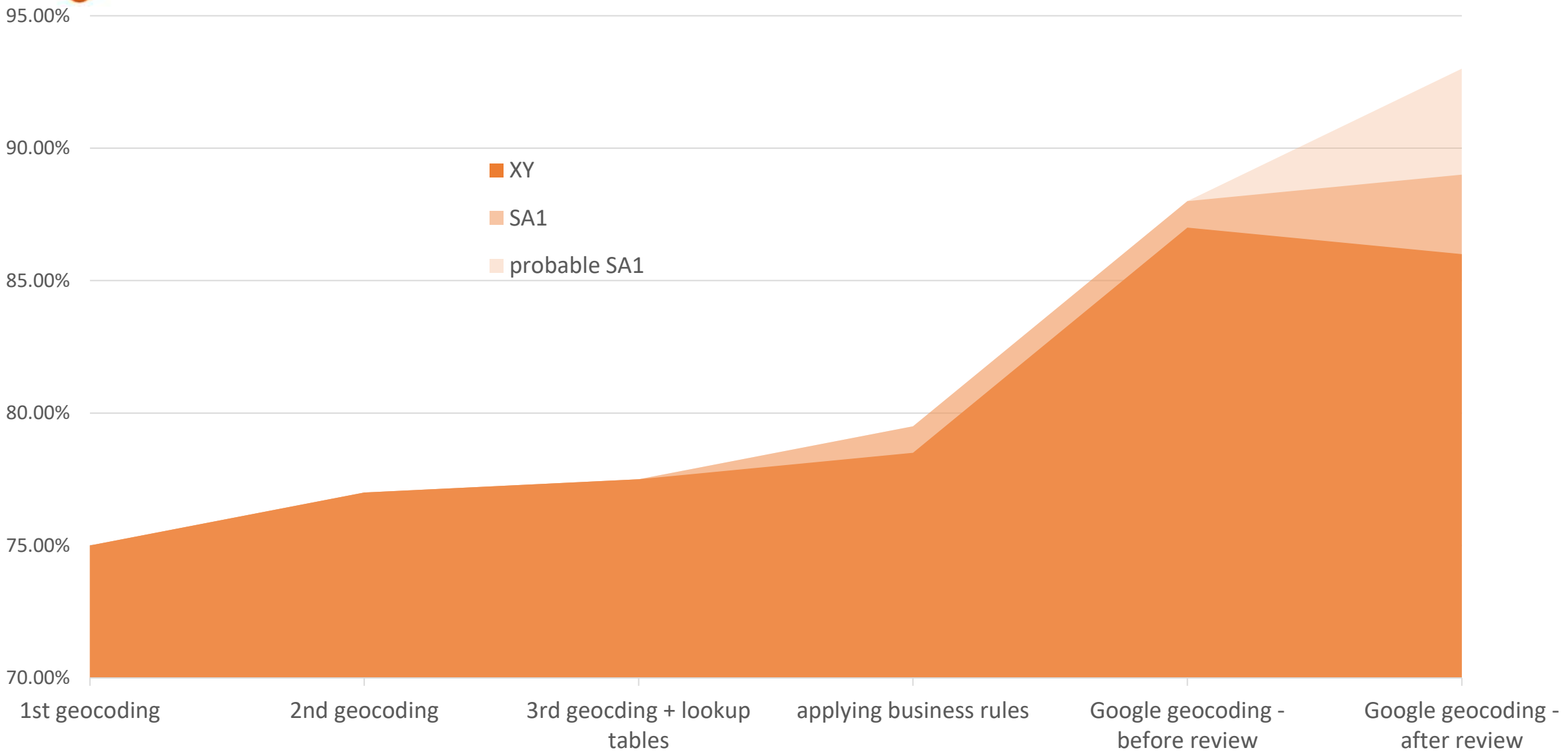
Google geocoding API

- Manual review in browser (JavaScript)





Match accuracy progression



Conclusion

- Variable data quality renders purely automated geocoding difficult
- 90% (SA1) data quality only achievable with significant manual review efforts
- Addresses ideally validated at point of data entry
- Update and use of GNAF required to improve address validation
- Geocoding outcomes vary more between tools for remote properties

acknowledge and thank the following contributors to this project

For further information please contact:
Miro Palfy
SA NT DataLink, Principal Statistical Analyst
miro.palfy@unisa.edu.au
+61 (8) 8302 1719
www.santdatalink.org.au

The Australian Government provides financial support to SA NT DataLink through the National Collaborative Research Infrastructure Strategy (NCRIS).