Integrated Research Infrastructure

for the Social Sciences

Work Package 3

Technical Report 1

Tomasz Zajac, Michael Rigby, Angela Ryan, Matthias Kubler, Denise Clague, Jonathan Corcoran, Wojtek Tomaszewski

31 August 2022

Table of contents

1. Introduction

The Integrated Research Infrastructure for the Social Sciences (IRISS) Project aims to address the fragmentation of the Australian social science research infrastructure. One of the identified barriers hindering social research in Australia is the lack of dataset integrating information on people, places, time, and space. The IRISS Work Package 3 seeks to address this issue by developing and piloting a "proof-of-concept" data integration service called GeoSocial, which will allow researchers to enhance people-centred survey data with spatially structured data capturing information on places where these people live. Associated Work Package 4 Demonstrator 1 will showcase the features of an example integrated dataset. This report outlines the overall design of the service and its functionalities.

The service design has been informed by a series of discussions and consultations with social researchers within the IRISS project team, including during the AURIN-ISSR workshop on the 16th of August 2022, as well as reviews of existing studies using integrated geosocial data and data documentation. Furthermore, while designing the service, we needed to consider legal issues, including existing data governance frameworks that regulate access to and the use of data that is being integrated. Finally, we needed to take into account the available technical solutions that can be deployed within the project timeline. To a large extent, this report builds on our previous documents outlining some of these issues, which are presented in the appendices:

- Appendix 1: User requirements note
- Appendix 2: Data requirements note
- Appendix 3: Most downloaded ADA surveys
- Appendix 4: Previous work with HILDA involving spatial analysis

These appendices are referred to at the relevant points in this technical report.

2. Solution Development

As this is an early, "proof-of-concept" stage of the service development, the scope of the project needs to be limited. The currently developed version of the GeoSocial service will cater to relatively advanced users who require individual-level data for their analysis but are not able to, or prefer not, to integrate data on their own. This could include individuals lacking the technical skills necessary for performing data integration, not having sufficient knowledge of the data (e.g., not knowing what geographical identifiers should be used in data linkage), or simply preferring to save time by using readily available integration tools. For those users, the service will offer an easy-to-use, trustworthy, transparent, and reproducible solution for integrating data from Australian major longitudinal survey with data on places. The product targeting these relatively advanced users will have a form of editable scripts integrating data in the user's local computer environment.

The GeoSocial service will be developed using an iterative approach to software development using Agile methods that allows outputs from different stages, comprising several activities, to feed back into the design and development of the solution, catering for flexibility in functionality under time and cost constraints. The stages and activities comprising this workflow are described in Figure 1 with a high-level diagram illustrating the relationships between elements presented in Figure 8 (Appendix A). Further information about each of these stages will described in Sections of this report.

1. Planning
2. Iterative development
    - Understand researchers, their contexts and needs
    - Identify, categorise and prioritise user requirements
    - Design, develop and increment solutions
    - Evaluate solution against requirements
3. Completion

Figure 1: Iterative approach to software development

Following initial planning, the first round of development commenced with the capture of preliminary user data through contextual inquiry processes (e.g., researcher interviews, reviews of past projects and data audits), which were synthesised and analysed to produce various outputs, including preliminary user profiles, contextual information and needs. These outputs were then used to define the specific problems to be targeted by GeoSocial with high-level requirements for the solution identified, categorised and prioritised (further details relating to some of the key design considerations are described in Section 3).

The formal solution design phase of the project (expected to be performed September – December 2022) will layout a subset of the requirements for the purposes of developing a demonstrator of the GeoSocial service that will be evaluated against its design specification and internal testing. To incorporate learnings from this process, the software development process will then iterate over previous activities to ensure feedback is included thereby allowing the solution to be incremented towards the design and development of the project's operational pilot (expected by 30 June 2022). Such progression is expected to see greater functionality, usability and user experience, resulting in increments in technology readiness level from a demonstrator (level 2-3) to operational prototype (level 6-7) (Figure 2).

Figure 2: Technology Readiness Level (TRL) descriptions (Source: TWI-Global)

While this early version of the system is purposefully limited in scope, it is designed with possible future extensions in mind. We map possible future developments in the last section of this report. With the future extensions, the service will be able to offer, among others, a wider range of data, a more flexible data integration process, and some analytic solutions. By doing so, the service will progressively lower the bar in terms of necessary skills and will become accessible for a broader pool of researchers interested in studying geographically enhanced survey data.

3. Towards Design

The first step towards designing the GeoSocial service requires iterating and building on user data captured earlier in the project to define input parameters for the design stage. Key elements of this section include the user profile, data and interactions discussed below.

3.1 User profile for the GeoSocial Service

An effective user-centred design process relies on a well-established understanding of users, their contexts and needs. As highlighted in Section 1, preliminary outputs from the user data capture process (including early analysis of user skills presented in Appendix 1) have identified two target user profiles for development of the GeoSocial service that are summarised below:

Social science researcher: Advanced user Confident with using Python and/or R for data wrangling, integration, and analysis Good understanding of geospatial data Needs to integrate longitudinal and geospatial data for analysis Supports other social science researchers

Social science researcher: Mid-level user Confident with understanding and tweaking R scripts Experienced in the use of Stata software Limited understanding of geospatial data Needs to integrate longitudinal and geospatial data for analysis May consult with data science researchers to achieve goals

The user profiles may undergo minor updates as the project progresses and will be critical to assist the design and testing processes of the project's service design phase.

3.2 Data

The GeoSocial service aims to facilitate social science research by providing researchers with survey data enhanced with information about places. The long-term goal is to offer the service for a large pool of surveys available through the Australian Data Archive (ADA), for which geographical information is available. The geographical data will initially come from AURIN's repositories but could be expanded in the future to include other data sources, and other types of data. In the initial (proof-of-concept) phase, the service will be used to demonstrate the utility of the service and pilot its functionality. This will be achieved by using the service to produce a Demonstrator data set (as per Work Package 4) to illustrate the service capabilities and the added value of spatially integrated survey data.

The selection of datasets for the Demonstrator was preceded by data audits reviewing metadata and assessing the current usage of various datasets. The first data audit focused on the ADA data collection. It concluded that the Household, Income and Labour Dynamics in Australia (HILDA) study with 33,924 downloads[1] was by far the most often downloaded survey available in ADA. Results for selected other surveys are presented in Appendix 3. HILDA makes a good example dataset for the service prototype for one more reason. It consists of multiple waves which allows the demonstration of the service's capability for temporal data integration.

The second audit ranked AURIN datasets by their usage base. Together with the review of previous HILDA-based studies focused on geographical factors (see Appendix 4), the audit will inform the final selection of geographical datasets that can be used with the service, including for the purpose of building the WP4 Demonstrator 1 data set

The review of survey documentation (see Appendix 2) recommended that the service focuses on enhancing individual-level survey data by adding information about places (i.e., performing individual-area linkage by using geospatial area codes). Even for HILDA, which is the largest survey in the ADA collection, the sampling design allows reliable and unbiased estimation only for very large areas, e.g., states, which limits its utility for research. Many smaller geographical areas (e.g., SA2s) might be represented by very few observations or not be represented at all in the data. Furthermore, these observations might be geographically clustered and unrepresentative of their area. For this reason, terms and conditions of HILDA access explicitly prevent publishing area-level estimates for areas more granular than broad regions within state (e.g., reporting at an SA2 level is not permitted under the standard HILDA Terms and Conditions of access). This issue is likely to be even more pronounced in the case of smaller surveys.

3.3 Data Flows

The key datasets described above in Section 2.2, namely those from HILDA and the ABS, are hosted by various organisations that may have different governance frameworks and access protocols. As these will impact on user interaction relating to any data integration solution, understanding this dependency is essential for the design of the GeoSocial service demonstrator and operational pilot.

[1] Data as of 20 April 2022.

Figure 3 illustrates a typical scenario in which a researcher seeks to obtain data from sources that may have different request and access requirements, from establishing a data sharing agreement and/or signing a confidentiality deed poll then registering for the service, to downloading the data using an open connection or a secure API. In this figure, HILDA would have the most restrictions requiring Confidentiality Deed Poll, service registration and token steps, while others, such as the ABS, are open and without controls to access the data.

Functional requirements for the GeoSocial service are described in detail in Section 3.

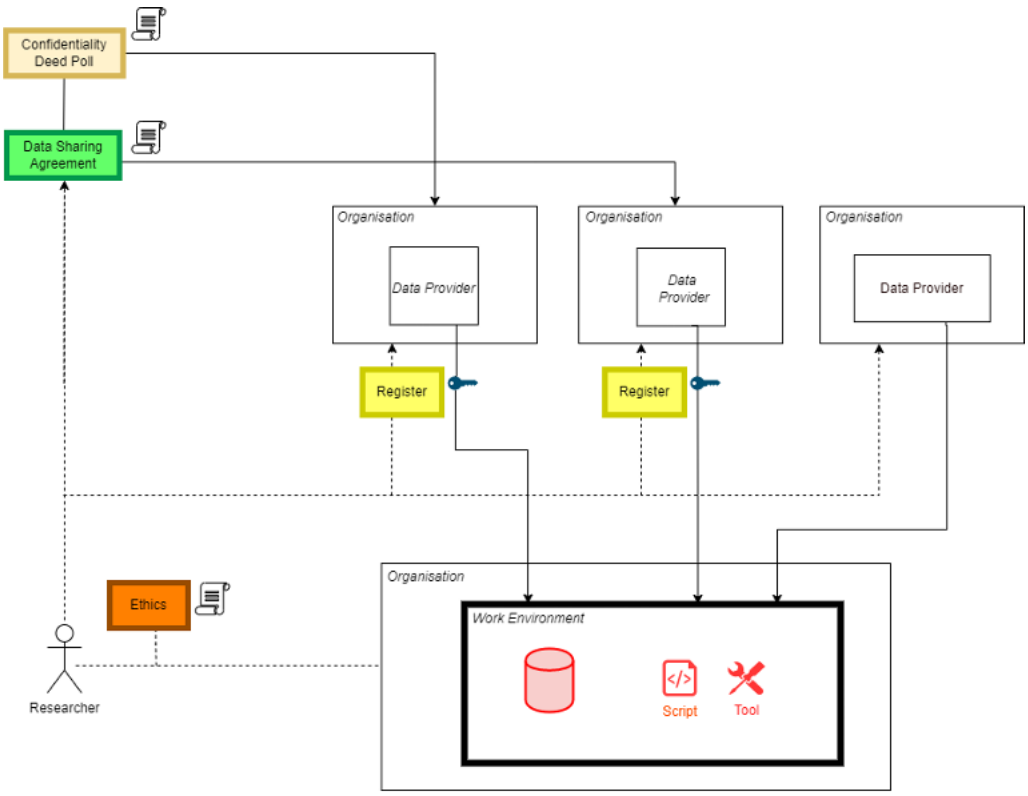

Figure 3: Example scenario of a researcher's needs to source and access data from different sources

4. Preliminary Service Design

This section describes the preliminary design of the GeoSocial service, drawing on the user data captured so far, and the inputs and considerations presented in Section 3. It sketches the high-level solution that will form a foundation for the service design phase that will be performed in September – December 2022.

4.1 Translating user needs into prioritised requirements

User needs and high-level requirements, including those from Section 2, were refined, before collaboratively defining functional requirements and their prioritisation using the Agile method MoSCoW in which requirements were categorised according to being ‘Must have’, ‘Should have’, ‘Could have’, ‘Won’t have’. These were then added to project’s prioritised requirements list (PRL) that will inform both service design and development phases.

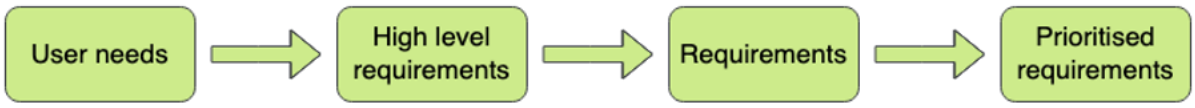

Figure 4: Workshop method to translate user needs into prioritised requirements

In total 49 functional requirements were identified within the initial PRL (refer Appendix B). to take forward into the service design phase. It is important to note that the uneven distribution of requirements across priority categories shows a significant number of ‘Must haves’ that will require careful management to ensure development remains on track (Figure 5).

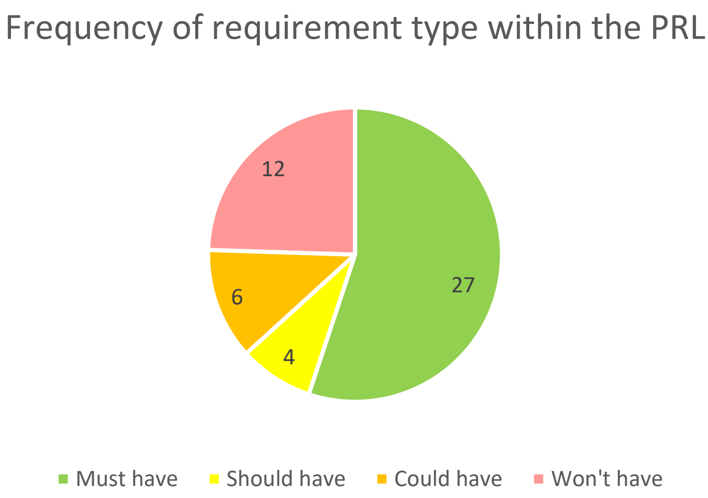

Figure 5 Prioritisation of functional requirements within the PRL following the ISSR-AURIN Workshop 16 August 2022 (see Appendix R for the full list of requirements):

4.2 User experience and usability goals

A user's experience of the GeoSocial service is paramount of the solution's design process. User experience (UX) can be defined as a person's perceptions and responses resulting from use (or expected use) of a service (ISO, 2019). Preliminary UX goals identified include:

Satisfying - Gives me what I need

Supportive - Lowers the barrier to entry

Informative – Gives additional information about data integration and possible datasets

Helpful – Provides information on how to use the service

Motivation – Provides an option to learn/adopt a programmatic way of working

Usability goals differ from UX goals in that they are used to describe the service itself and how it might help a user obtain outputs and achieve outcomes. Preliminary usability goals can be described in terms of the service's usefulness (Sharp et al., 2019) and have been identified as:

Effective (e.g. successfully integrates data according to the user's input)

Efficient (e.g. encourages interaction flow and saves time)

Utility (e.g. provides the functionality a user needs)

Learnable (e.g. allows a user to find and repeat tasks in a natural way)

Memorable (e.g. follows design norms)

Safe (e.g. provides suitable warning messages)

Both UX and usability goals will be used to guide the formal design phase of the GeoSocial service.

4.3 Use cases

To aid design and planning activities towards the service demonstrator and operational pilot, use case diagrams were created at different levels using the 'Must have' requirements only. The purpose of this was to help identify a minimum viable product (MVP) that is feasible, without logical gaps in user interaction, functionality, or feedback, across different levels. This process will be an important activity to differentiate between the demonstrator and operational pilot solutions that will be performed within the project's solution design phase (expected be performed September – December 2022). Preliminary use cases are presented in Figure 6 and Figure 7 below.

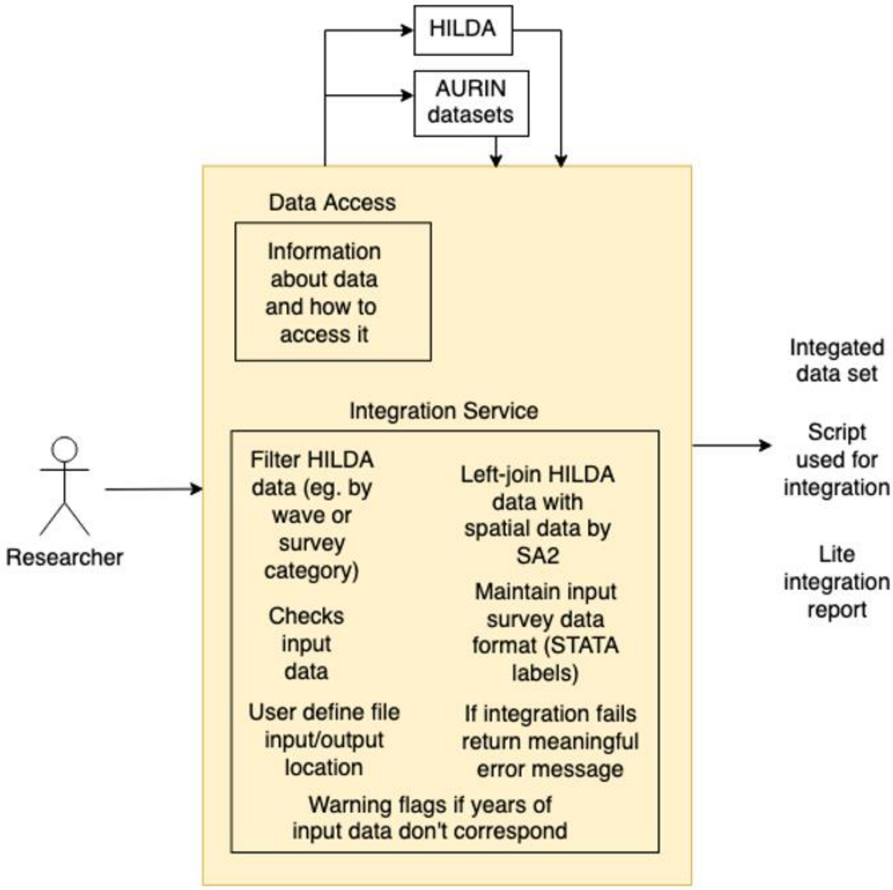

Figure 6: High level use case diagram of the GeoSocial service (must have requirements only)

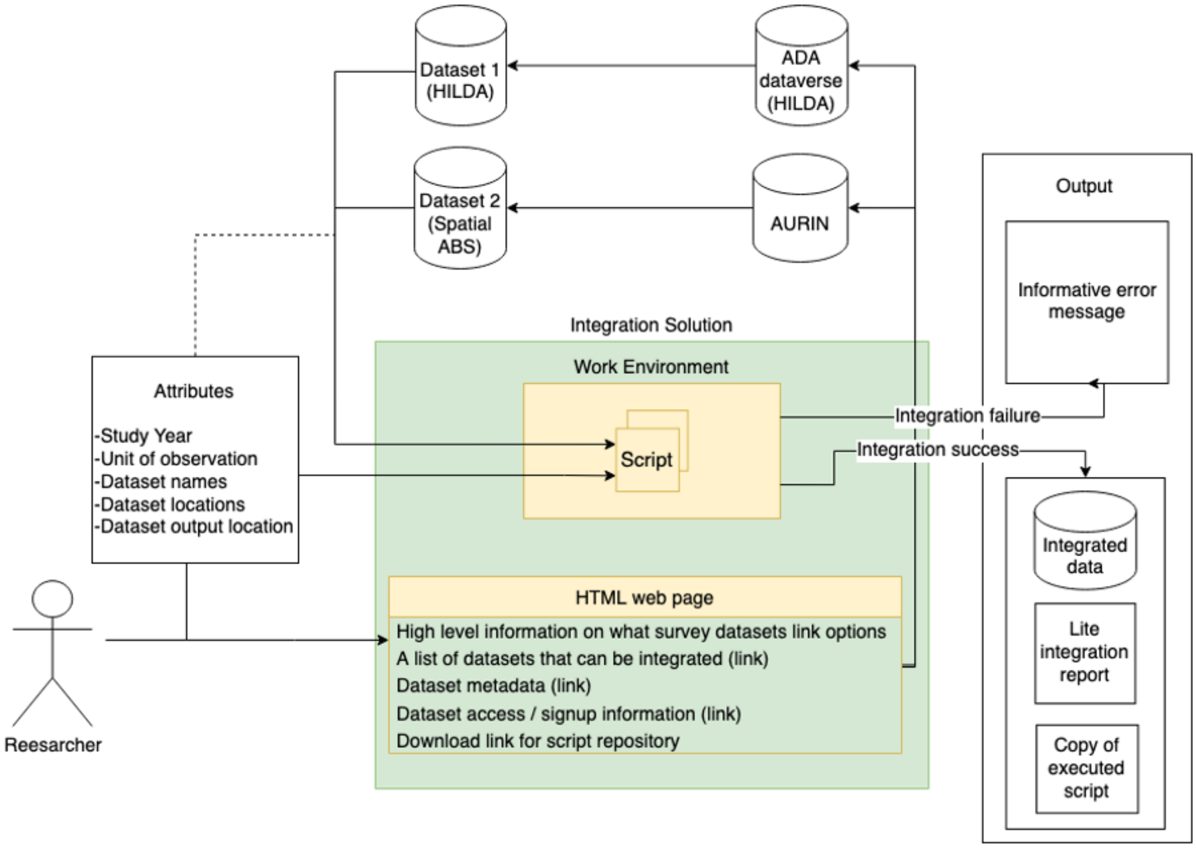

Figure 7: Lower-level use case diagram of the GeoSocial service showing inputs, solution components and outputs (must have requirements only)

4.4 Key features for the GeoSocial service

The following feature definitions have been created for the Geosocial service drawing on requirements from the PRL (Must haves only):

GeoSocial service

The solution for Work Package 3 of the IRISS project, comprising inputs, integration solution and outputs

Integration Solution

The core component of the GeoSocial service, comprising script repository, and HTML web page

Script repository

A suite of scripts to perform data integration written in the R language

HTML web page

A reference page providing background information about the GeoSocial service and links to data that could be used

4.5 Other Considerations

The following other design considerations have been identified for the GeoSocial service.

User experience and usability goals

The GeoSocial service will be designed to meet specific goals expanded from those listed in Section 4.2.

User interface

The GeoSocial service is planned to be expanded with a graphic user interface (GUI) in the operational pilot.

Security

The GeoSocial service will not be designed with authentication/authorisation or other access mechanisms. The data integration scripts will run within a working environment that is assumed secured.

Programming language

Scripts used within the GeoSocial service are planned to be written in the R programming language per the user profiles defined in Section 2.

Future expansion

This iteration of the design will only be for HILDA survey data and a selected number of AURIN datasets, however there is an intention for the service to be expanded to more survey datasets and more spatially enabled datasets in the future. The design of this iteration of the solution will allow for additional datasets to be added in the future.

4.6 FAIR4RS

To meet user needs, address requirements and make the GeoSocial service demonstrator valuable to the research community, the service design will seek to adopt the FAIR Principles for Research Software (FAIR4RS) defined under findable, accessible, interoperable, and reusable categories in Table 1.

Table 1: FAIR Principles for Research Software (Chue Hong et al., 2022)

Findable: Software, and its associated metadata, is easy for both humans and machines to find
F1. Software is assigned a globally unique and persistent identifier.
F1.1. Components of the software representing levels of granularity are assigned distinct identifiers.
F1.2. Different versions of the software are assigned distinct identifiers.
F2. Software is described with rich metadata.
F3. Metadata clearly and explicitly include the identifier of the software they describe.
F4. Metadata are FAIR, searchable and indexable
Accessible: Software, and its metadata, is retrievable via standardized protocols
A1. Software is retrievable by its identifier using a standardized communications protocol.
A1.1. The protocol is open, free, and universally implementable.
A1.2. The protocol allows for an authentication and authorization procedure, where necessary.
A2. Metadata are accessible, even when the software is no longer available
Interoperable: Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.
I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards.
I2. Software includes qualified references to other objects.
Reusable: Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software).
R1. Software is described with a plurality of accurate and relevant attributes.
R1.1. Software is given a clear and accessible license.
R1.2. Software is associated with detailed provenance.
R2. Software includes qualified references to other software.
R3. Software meets domain-relevant community standards.

Software Development Outputs for Management

The development of the GeoSocial service will follow the Agile Dynamic Systems Development Method[2] (DSDM) that uses an iterative approach to development that broadens the traditional focus on delivering a software product to consider project requirements as a whole. Key team members involved in this process include the Social Data Scientist and Software Developer at AURIN, with oversight from the Partner Lead (AURIN) and Work Package Manager ISSR). Specific outputs from development are described below in Table 2.

[2] The DSDM Agile Project Framework, https://www.agilebusiness.org/page/TheDSDMAgileProjectFramework

Table 2: Software development outputs and their assignments.

Output	Description	Responsible	Accountable	Consulted	Informed
Prioritised requirements list (PRL)	Prioritised list detailing all project requirements. Contents initially drawn from the ISSR-AURIN Workshop 16 August 2022. Items on this list are initially functional, drawn from data, user, environment levels, and will expand to non-functional through iterative development. Administration of this is critical for software development to evolve effectively.	Social Data Scientist (AURIN)	Partner Lead (AURIN)	Work Package Lead (ISSR)	Work Package Members (ISSR + AURIN)
Timeboxes	Structured time periods within which requirements are actioned by the Software Developer. These are expected to be fortnightly and planned, tracked and reviewed during work package meetings.	Software Developer (AURIN)	Partner Lead (AURIN)		Work Package Lead (ISSR) Work Package Members (ISSR + AURIN)
Examples and demonstrators	Developed demonstrating a requirement's functionality. Tracking of these is a responsibility of the Partner Manager with oversight from the Work Package Manager.	Social Data Scientist (AURIN)	Partner Lead (AURIN)	Work Package Lead (ISSR)	Work Package Members (ISSR + AURIN)

Testing	Testing of incremental solutions are key stages in development. Outputs are formal testing procedures for use by the developer to ensure suitable coverage and acceptance criteria are met. To be report on in regular fortnightly meetings	Social Data Scientist (AURIN) Software Developer (AURIN)	Partner Lead (AURIN)		Work Package Members (ISSR + AURIN)
Versioning	Development will follow version control methods to handle revisions and their tracking.	Social Data Scientist (AURIN) Software Developer (AURIN)	Partner Lead (AURIN)		
Supporting materials	Supporting materials are critical to ensure project quality remains at a certain level. Outputs are well-documented codebases, documentation and user guides.	Social Data Scientist (AURIN) Software Developer (AURIN)	Partner Lead (AURIN)		Work Package Lead (ISSR + AURIN)
Evolving solution	The current solution under development, together with work in progress within timeboxes. Outputs are combined in increments that contribute to the evolving IRISS solution.	Social Data Scientist (AURIN) Software Developer (AURIN)	Partner Lead (AURIN)	Work Package Lead (ISSR)	Work Package Members (ISSR + AURIN)
Project technical reports	Feedback, learnings and benefits from solutions are to be recorded and this continues until the project's successful delivery and closure.	Social Data Scientist (AURIN) Software Developer (AURIN)	Partner Lead (AURIN)	Work Package Lead (ISSR)	Work Package Members (ISSR + AURIN)

Key meetings within the Work Packet include the following:

Weekly technical meeting (AURIN)

Fortnightly progress meeting (ISSR + AURIN)

Monthly IRISS project meetings (Work Package Members)

Interdependencies with other work packages, such as Work Packages 2 and 4, are expected to be identified and progressed by the Work Package and Partner Leads through regular fortnightly meetings and cross-work package meetings coordinated by the IRISS Project Manager.

6. Next Steps and Future extensions

Following the preliminary solution design, the formal design and development stage is planned to commence with early demonstrators to explore examples and start to address the must haves from the PRL as well as connections to other work package outputs, such as VASSSAL (vocabulary service) and SPIRE (survey package). These early demonstrators are also expected to both inform and upskill team members in the problem domain. This is expected to further contribute to the initial design of the GeoSocial service solution architecture, which is a key activity to be led by AURIN and supported by ISSR and other project partners through to 31 December 2022. Fortnightly meetings are planned to be held between ISSR and AURIN to progress this work towards the delivery of demonstrator and operation pilot solutions through to 30 June 2023.

The workshops and consultations led to the identification of several issues that are out of the scope of the project's prototyping phase but are important for the long-term future of the service.

First, a new administrator needs to be chosen to run the service after June 2023. The future administrator will be responsible for maintaining and updating the service. For example, the scripts will need to be reviewed and possibly updated every time a new wave of a panel survey is published. In addition, changes in existing software, e.g., R libraries, must also be monitored.

Second, the service will need to expand in terms of available data. This means including more ADA survey data and more spatial data in the service, as well as offering new types of linkages, i.e., linking data that use different spatial identifiers or different versions of the same identifiers. Such linkages would require geographical dictionaries with concordances between various classifications, which are currently unavailable.

Third, the service will need to improve existing functionalities as well as add new ones. Some ideas for future additions have already been listed in the previous section. GUI and a data analytics and visualisation module are particularly important to include in order to be able to attract a broader user base, including those with lower technical skills.

6. References

ISO (2019) *ISO 941-210:2019, Ergonomics of human-system interaction – Part 210: Human-centred design of interactive systems*. DOI: ISO - ISO 9241-210:2019 - Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems.

Sharp, H., Preece, J. and Rogers, Y. (2019) *Interaction Design: Beyond Human-Computer Interaction*, John Wiley & Sons, Inc., 5 Edition, Indianapolis, IN, USA.

Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A-L, Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., Honeyman, T., et al. (2022). *FAIR Principles for Research Software version 1.0. (FAIR4RS Principles v1.0).* Research Data Alliance. DOI: https://doi.org/10.15497/RDA00068.

Appendix A – Iterative Software Development

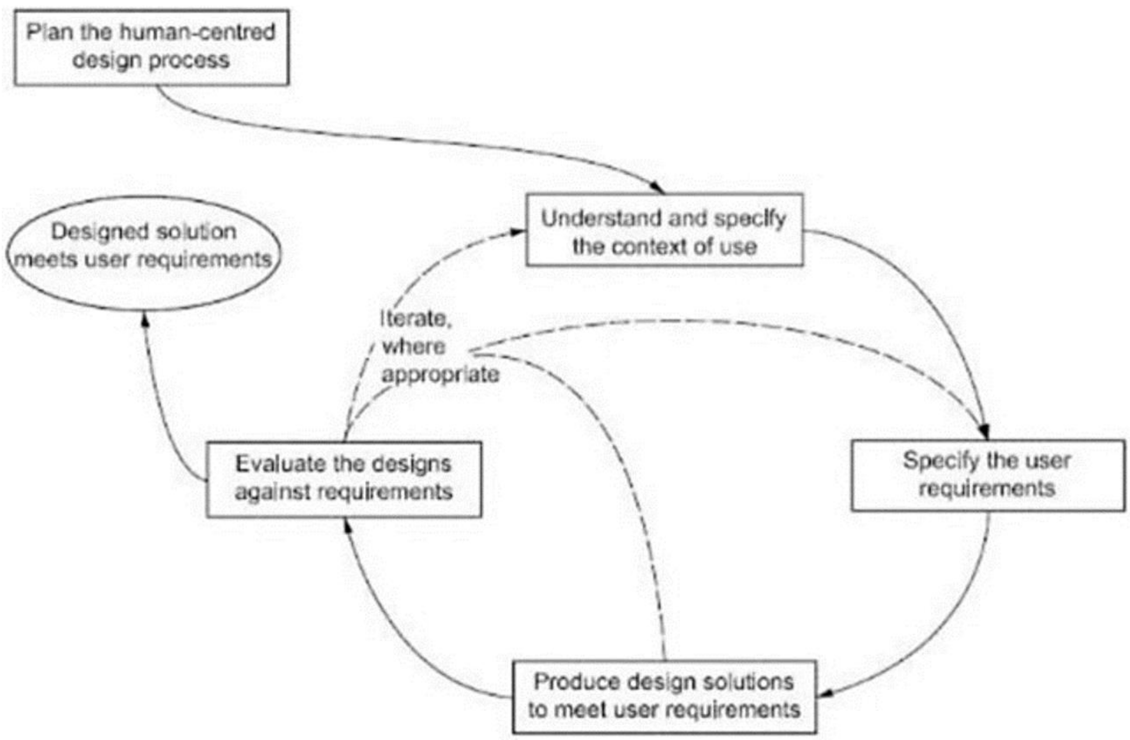

Figure 8: Iterative Software Development. Source: ISO 941-210:2019, Ergonomics of human-system interaction – Part 2016: Human-centred design of interactive systems (https://www.iso.org/standard/77520.html).

Appendix B – Prioritised Requirements List

Last update 31 August 2022

Must have	
UR1	No login required to browse/explore scripts so that users can explore what the service has to offer without investing too much time.
UR2	Example Script
UR3	Script selection (selection of datasets to be integrated)
UR4	Accessible coding style - widely used language (R), scripts easy to understand and modify
UR5	Provide a correspondence flag (warning) when a classification might have changed, e.g., when survey data include an older version of SA2 than geographical data that will be added.
UR6	Ability to read/input survey data (HILDA)
UR7	Ability to browse and select survey data (e.g., respondent/HH files)
UR8	Ability to browse and select spatial data (e.g., year, SA1/2 files)
UR9	Join survey data first to perform left-join
UR10	Integrate multiple waves of survey data
UR11	Add multiple versions (years) of geographical data
UR12	Read spatial data
UR13	Check input data (ids, variable names, e.g., duplicated variable names or whether the variables to be added are already present in the dataset)
UR14	Meaningful error statements if encountered
UR15	Ability to link one survey file with another geospatially (using geographical identifier such as SA2)
UR16	Ask user to define a file input/output location
UR17	Read Stata input files
UR18	Provide a link to correspondence information (ABS)
UR19	Provide a link to signup/access AURIN data
UR20	Provide high level information on what survey datasets can be linked and to what
UR21	Provide links to some survey metadata (HILDA, ABS)
UR22	Retain everything from the input files (e.g., data structure, variable values and labels) for both survey and spatial
UR23	Outputs integrated data (e.g., in Stata and R formats)
UR24	Outputs the script executed
UR25	Outputs a list of input data used and basic information (e.g., id used for integration)
UR26	Design and develop following FAIR principles for research software
UR27	Outputs a brief data linkage report (input data, output data, join ID).
Should have	
UR28	Graphic user interface (in addition to the example script)
UR29	Ability to select loading/save locations etc.

UR30	Join multiple (more than two) input datasets in a single step, i.e., add multiple spatial datasets to the survey file at the same time.
UR31	Outputs detailed data linkage report (input data, output data, join ID, filters, (e.g., temporal, spatial, variables), linkage rate %, lists codes with no match).
Could have	
UR32	Check input data quality (e.g., Numeric/string, value range)
UR33	Vocabulary service integration
UR34	Ability to filter based on labels
UR35	Link to AURIN API (instead of pointing to spatial data in the local environment)
UR36	Create/publish R package and index in CRAN
UR37	Ability for a user to select/map between original vs modified input
Won't have	
UR38	Visualisation
UR39	Loc-I Demonstrator (concordances between areal identifiers)
UR40	Join with flexibility (e.g., right join)
UR41	Full search and discovery of data
UR42	Curated linked ready-to-use datasets
UR43	Service to aggregate individual data to areas
UR44	Maintainability (ongoing service agreement)
UR45	Easy calculation of derived variables
UR46	Data access service and authentication/authorisation (CADRE)
UR47	Secure environment to process sensitive data
UR48	Ability to automatically convert between different input data types/formats
UR49	Easy access to data documentation (vocabularies, classifications, data comparability, limitations etc) ranging from technical info for advanced users to broad info for less advanced users

Appendix 1: User requirements note

Purpose

The purpose of this document is to discuss the audience profile relevant for the WP4 Demonstrator 1, and the WP3 GeoSocial Service more broadly. It covers the technical skills that different users have, as well as different thematic/substantive interests. Recommendations for the target user profile for the purpose of WP4 Demonstrator 1 and WP3 are also outlined in the document.

Technical requirements

User requirements change with the user's skills levels. Most advanced users have the technical skills and tools to merge and transform data on their own. They are also more likely to be interested in using sophisticated statistical methods that require unit-level data. For such users, ease of access to the data as well as high-quality documentation are essential. Moreover, for such users, flexibility is important.

Less skilled users will require more advanced tools to be able to utilise available data. For example, users less comfortable with data processing need easy access to the data and documentation too, but would also require more support in data preparation. Ideally, they could access ready-to-use or easy-to-link datasets, with curated content such as derived variables. Such data enhancements (e.g. linked dataset, derived variables) might not be crucial for more advanced users but could be still appreciated as time-saving measures (if flexibility is sufficiently retained).

The least advanced users, who neither can process data themselves nor are interested in complex statistical modelling, would require significantly more features provided as part of the service. They would need not only ready-to-use data but also additional functionalities allowing data analysis and visualisation. Again, such tools might not be necessary for more advanced users but could prove useful, e.g. by allowing quick descriptive analysis.

The user base is likely inversely proportional to the level of technical skill. The tools with more features and aimed at low-skilled users might potentially have a wider user base, possibly stretching beyond the academic and research community – such as policy analysts working in government agencies. However, the service aimed at this type of user would need to be more developed and would require more curated data assets to operate. Developing this kind of service and data requires significant time and resources. Such more curated service and data are also much less flexible. Therefore, more time is required for researching the needs and potential use cases before commencing work on developing them.

Below we present user requirements for three discrete types of users. The actual users will quite often fall in between these categories, but the classification helps to illustrate what types of functionalities are required at various levels of skills.

Table A1.1: User requirements and benefits.

Type of user	User requirements	Benefits
High skills level - Advanced user, capable of performing advanced data transformations, merging datasets, deriving variables, and interested in sophisticated statistical methods. IRISS personas: Evan – data scientist/analyst Martin – researcher (data collector) Danielle – researcher (data analyst)	Easy access to the data frees the user from negotiating data access with data custodians and securing permissions for data linkage. Flexibility in data formats. Information on linkage keys for geosocial data (e.g., SA2 codes) or concordance tables between different geographies allowing individual-to-spatial data linkage (e.g., postcodes to geographical classifications). Data documentation that includes data vocabularies and classifications, data comparability (e.g., akin to IPUMS), description of data limitations (e.g., limitations to survey data aggregation by geography arising from sampling, notes on data quality/ reliability – especially in the case of administrative data, etc.). Instructions for data derivations (e.g. on how to derive income from ATO data). Confidentiality and privacy protection measures. Secure environment to process data if data require protection. Search & discovery – particularly important as the list of datasets grows.	Faster access to the data. Removes the need for data conversions and some processing. Faster and easier data merging. Certainty regarding data meanings. Reduced time needed for developing data transformations code. Less room for errors leading to data breaches. Improved discoverability of data. Easier access to the right data.
Medium skills level - A user advanced in applying statistical sophisticated statistical methods but less	In addition to the previous: Linked ready-to-use datasets (e.g. individual-level survey data appended to geographical/spatial	Wider access to the data.

comfortable with data manipulation. IRISS personas: Martin – researcher (data collector) Danielle – researcher (data analyst) Yosef – data analyst	data). A user could potentially select from a list of datasets and variables to be downloaded/analysed. Derived variables or a library of code that could be used for data derivations.	Even easier and faster access to the analysis-ready data. Standardised and comparable measures. Reduced data processing
Low skills level – A user who cannot manipulate data and who might be more interested in descriptive analysis than advanced statistical modelling. IRISS personas: Serena – policymaker Yosef – data analyst	Easy access to the data frees the user from negotiating data access with data custodians and securing permissions for data linkage (like for previous types of users). Data documentation explaining the variables and data limitations (focused more on the meaning of the data than the technical process of variable derivation than in the case of more advanced users). Safeguards preventing incorrect use of data (e.g. making sure that a dataset can be used to produce representative community profiles) Interface for data analysis and visualisation (e.g. Gapminder data animations, Shiny, Tableau). Built-in confidentiality and privacy protection measures ensuring that only safe outputs are available.	Easy access to the data. Certainty regarding data meanings. Less room for analytic errors. Increased data usability and utility to untrained users. Reduction of the risk of data breaches.

Thematic requirements

The work package aims to create a data product that integrates people, place, time, and space. There are numerous ways in which data on people can be linked with data on places. There are researchers working in a number of disciplines across the social sciences with interest in such data, including sociologists, education researchers, political scientists, economists, social planners, human geographers and others. Attempting to cater to too many types of users and to integrate too many data sources at once is flawed with risks related to feasibility and the robustness of the data infrastructure that is being developed. Instead, an initial stage of the project should focus on maximising the impact/ usefulness of the tool while keeping feasibility at the forefront. This can be achieved by focusing on a small number of datasets with relatively broad user base and topic coverage. Future expansions of the system should involve a bigger number of, and more diverse data sources.

Recommendations for WP3 and Demonstrator 1

The precise data selection criteria are still being developed (for more details see the 'Data requirements' document), however, the following core considerations taken from the project proposal will be applied to select the data[3]:

The Service developed as part of WP3 and Demonstrator 1 will integrate data on people and spaces, while also capturing the longitudinal dimension.

For the purpose of the Demonstrator, data on people will come from a 'high-profile' longitudinal survey available through ADA, with demonstrated wide user base (see 'Data requirements' document for more detail).

Spatial data, which will be at least in part based on the Census and supplied by AURIN, will need to use geographies that are identical to those used in the selected survey(s), or that can be easily converted to those used in the survey(s).

The datasets will be selected to produce useful and methodologically sound analysis.

It is recommended that the core pilot service developed under WP3, and illustrated with Demonstrator 1, is developed with a relatively high-skilled user in mind, to ensure the feasibility of delivering these within the project timeframe. To this end the following parameters are proposed for WP3 and Demonstrator 1:

The Demonstrator will consist of:

An integrated dataset, combining a longitudinal survey from the ADA with spatially-structured data derived from the Census and other relevant sources;

[1] Other future expansion could cover creating 'community profiles' out of other sources of administrative data, Land use data, Satellite data, Private sector data (e.g. CoreLogic, Uber) and more.

A set of scripts performing data integration, employing relevant data vocabularies, including those developed as part of WP2;

A technical report documenting the steps that need to be undertaken to perform data integration and outlining issues that will need to be resolved as part of WP3 service (including legal and technical issues);

Statistical analyses to demonstrate the utility of the integrated dataset – these can be shared with the research community in the form of an academic paper, and shared with the relevant government stakeholders to demonstrate potential policy applications and impact.

Although the ultimate goal is to offer an online system that researchers can use to access enhanced data, it might not be feasible within the project timeframe. Establishing governance frameworks and addressing related legal issues is a time-consuming process. Therefore, for the purpose of the Demonstrator, the data integration process will be carried out locally at the researcher/user end (UQ) and the integrated dataset will also be deposited locally.

The GeoSocial service prototype will consist of:

A set of scripts – such as Python scripts packaged as a Jupyter notebook – to carry out data integration, employing relevant data vocabularies, including those developed as part of WP2;

A technical report documenting the steps that need to be undertaken to perform data integration and outlining issues that will need to be resolved as part of further scaling up/industrialising the service (including legal and technical issues);

In WP3, consideration will be given first to the data integration process being carried out and the integrated dataset being deposited outside of the local user environment – for instance to make it available to the broader research community as part of the ADA Dataverse. However, each step of the development process will include exploring the options for transferring the system to an online environment and mapping out potential challenges.

While the core service and the Demonstrator will be initially designed with a relatively skilled academic/researcher user in mind, consideration will also be given to provide solutions for relatively lower-skilled policy user, which might include tools to facilitate spatial visualisations of data.

Appendix 2: Data requirements note

The purpose of this document

The purpose of this document is to outline the process for selecting the data for Demonstrator 1 and WP3 of the IRISS project. The focus here is mainly on the Demonstrator, but the key points raised in the document also apply more broadly to the GeoSocial data integration service developed as part of WP3. Once the process of data scoping has been finalised, recommendations for the datasets to be selected for the purpose of Demonstrator 1 will also be presented as part of this document.

It is important to follow a systematic process when assessing the suitability of data for geo-social integration in the context of the present project. First, systematically documenting the issues that need to be considered an be used to demonstrate how and on what basis specific datasets have been selected for the purpose of the Demonstrator. Second, such systematic process can then be followed when expanding the service to cover more and more diverse datasets in the future.

Introduction

This document assumes that the data integration of IRISS Work Package 3 concerns (i) survey data (from the ADA) on (ii) area-based data (such as Census data or other data held by AURIN).

The two main scenarios of combining survey and non-survey area-based information are:

Non-survey area information appended to survey household/person unit record data (i.e. the resulting dataset is a person-level dataset)

Area information derived from survey appended to non-survey area data (i.e. the resulting dataset is an area-level dataset, based on some spatial unit)

There are limitations for pursuing either option due to the underlying sampling designs and sub-sample sizes associated with the larger national (including longitudinal) surveys. The key limitations are twofold:

The sample design of national surveys (which is often oriented on ABS survey designs, such as the Labour Force Survey), usually entails a first cluster sample step that selects areas, which results in poor coverage of small areas in Australia. For example, of about 38,200 Collection

Districts (CDs) in Australia in 2011 only 488 were selected for the first HILDA wave. All consecutive sampling steps were nested within those 488 CDs[4].

The sampling (and weighting) is not designed to allow building up reliable estimates from the survey to various levels of geography, such as suburbs/SA2s, SA3s or LGAs. It is usually designed to allow, at best, metropolitan vs non-metropolitan breakdowns within the different jurisdictions in Australia.

These issues constitute particularly severe limitations for pursuing option (b) above as they dramatically constrict the possibility of deriving unbiased and reliable estimates for smaller geographical units from survey data. While there have been occasions where small area estimates were derived with the help of survey data the resulting estimates are often presented as ‘experimental’ and come with many disclaimers. None of these attributes are attractive for the IRISS project, neither is the fact that ‘building’ such estimates from or with the help of survey data is a project in and of itself that falls outside the scope of the IRISS proposal.

There are also some implications from the above points for pursuing option (a) (for the definition of the Demonstrator and WP3 more generally):

A credible analysis from option (a) would need to retain the original survey units of observation – households or individuals – as units of analysis. Questions that could then be answered are of the type ‘Are persons living in areas with <merged area characteristics from external non-survey sources> more likely to <some outcome of interest>?’ Or: ‘Which <merged area characteristics> predict individual or household attributes/behaviours?’.

> They cannot be of the type ‘Which named areas (such as LGAs, RA2s etc) are most associated with…?’. In this sense, the analyses would (still) be person/household-based, not place-based. Note though that data analysts could, in technical terms, undertake a place-based analysis with the same data if they wanted. However, the robustness and credibility of such analysis would likely be limited.

The closer the external non-survey area information fits the original unit of the sample selection, the more credible/valid the data analysis of the above type can be. In the case of the HILDA survey, for example, that means that the non-survey area information would ideally be at the level of the CD (as well as relating to the time of the HILDA data collection(s)) because

[4] We can map the 38,200 CDs and highlight the 488 as part of our technical report to visualise the area non-coverage of HILDA (or other large surveys).

then the area characteristics most likely represent the survey respondents' environment at the time. In this case, it may even be possible to ask place-based questions at the level of the selected 488 CDs, assuming that survey households/respondents are representative of the CDs, which would also need to be assessed.

> Non-survey area information might not be publicly available at the level of geographical units used in the sampling designs of surveys, such as the CD level. Likewise, identifiers for the original sampling units (such as CDs) might not be included in the accessible national survey datasets. The relationship between the original area units that were sampled and the area units included in the dataset would need to be investigated.

> Technically, it is possible to create higher level geographies from the available geographic information using available concordances to align the geographical areas in the survey data with those available in non-survey data. For example, information on CDs could be converted to SLAs in survey data to allow merging area information from non-survey data at the level of SLAs.

> However, with higher level spatial aggregations it becomes less and less likely that the non-survey (average) area information at these levels reflect the environment of households and people in the survey as, in the above example, the CDs in the survey were not selected to be representative of SLAs . Additional limitations arise when the aggregation to higher level geographies converts information from non-ABS geographies (e.g. postcode) to ABS geographies (e.g. SA2) or vice versa as these geographies do not neatly concord as many-to-1 or 1-to-many, but rather come with many overlaps with one postcode area falling into multiple SA2s, for example.

Selection of data for integration

We propose that selection of data (both survey and non-survey) for integration follows a set of considerations falling into three categories:

Strategic considerations, aimed at maximising the potential of the data for showcasing the advantages and value-add of socio-spatial data integration for research, policy and practice.

Technical considerations, aimed at maximising the feasibility of data integration, and the robustness of the analyses based on the data.

Concordance considerations, aimed at maximising the alignment between the survey and non-survey data to be integrated, including geographical and temporal alignment.

We further propose that the strategic considerations are used to pre-select a list of datasets (both survey and non-survey data), while technical and temporal criteria are subsequently applied to the list to prioritise datasets for subsequent integration.

The following sections describe the proposed strategic and technical considerations for selection of both survey and non-survey data.

Strategic considerations

Strategic considerations are similar for both survey and non-survey data. They include the user base and the likely impacts on policy (which may be associated with the level of prior investments from the Government into running a particular data collection) as well as the geographical, temporal and topical coverage. Tables A2.1 and A2.2 capture the key considerations for both survey and non-survey data respectively.

Table A2.1: Strategic consideration for selecting survey data

What	Why	Notes
Prominence of survey (number of users/gvt investment/ public profile)	The more popular a dataset the more likely the results of work package 3 and demonstrator 1 are of interest to the research community	This will likely correlate with sample sizes and identifying the 20 or so most prominent surveys may be a good point of departure in the selection process. The indicator used in this process may be based on number of downloads
Topics covered	Relevant for defining the Demonstrator	The topic should be of interest to a good section of the research community
Target population	The wider the population scope the wider the interest (potentially). Some surveys target specific sub-populations, cohorts or groups of people	General population surveys to take preference over surveys targeting distinct sub-populations or cohorts. National coverage should be prioritised over surveys targeting more localised geographies (e.g. state-based surveys)
Previous linkage requests	May indicate user interest, but also the potential for the survey data to be enhanced by spatial data	Might be able to get this from ADA. This can also inform the selection of the non-survey area-based information.

Temporal dimension	Integrating data over time as well as people and places is a key feature of the project.	Longitudinal surveys to take preference over cross-sectional surveys. Longitudinal surveys with a longer time span to take preference over surveys with shorter coverage

Table A2.2: Strategic consideration for selecting non-survey data

What	**Why**	**Notes**
Prominence of the data	The broader user base/the stronger investment/interest from the Government, the more potential for the impact and uptake of the integrated data	Information derived from Census is likely to be of high interest. National coverage should be prioritised over data covering more localised geographies (e.g. state-based data.
Topics covered	Needs to be convenient and useful for defining demonstrator and work package 3. Needs to complement well (i.e. ad value to) the survey data that is being integrated with	Could include considering the popularity of a topic
Temporal dimension	Integrating data over time as well as people and places is a key feature of the project.	Spatial data available repeatedly over time (e.g. Census, AEDC data) should be considered a part of the project, subject to temporal concordance of spatial information over time

Technical considerations

Technical considerations for surveys include key issues around sample structure, size and coverage, as well as the availability and accessibility of area-level identifiers that could be used as linkage keys to be integrated with spatially-structured data, with additional considerations including the intersections between geographical units identifiable in the data and the sampling units, and weights provided with the data (Table A2.3).

For non-survey data, it is assumed here that this non-survey data is based on whole populations (e.g. data drawn from the ABS Census or from reliable administrative data), rather than from samples, so that we do not need to consider sampling, weighting or estimating techniques used to arrive at any area estimates. Therefore, the key points of interest include geographical level/units, coverage of spatial system (the extent/completeness of coverage), and the reliability of information provided in the data, as well as the accessibility of the datasets (Table A2.4).

Table A2.3: Technical considerations for selecting survey data

What	**Why**	**Notes**
Sample design	This matters for presenting unbiased results at particular levels of geography and for identifying the appropriate geographical levels at which external area-based information can be meaningfully merged to. Was there any sample design that would allow place-based analyses beyond capital city vs rest of state? What was the original unit selected?	The sample design will likely be a major limitation for all national level surveys when it comes to generating reliable smaller area estimates, which translates into a limited potential for analysis of geographical units based on survey data
Sample size	The larger the sample size, the more analysis will be possible (e.g. undertaking statistical tests) but also at lower levels of geography	Preference for surveys with larger samples
Type of geographical information available	Geographical information that can be used to merge external information to or to aggregate to higher levels of geography is essential	The smaller the better. In theory it is possible that street address details from the survey frame exist. This could be of interest for longer-term projects following from this demonstrator.
Access to geographical information	The ease with which data with geographical information will be available to the relevant project teams will influence what can be achieved during the project period. This could be impacted by having to address privacy legislation.	For this project, this is more important than the above. But this step also includes exploring processes of getting more detailed data than customarily included in survey data (e.g. at the level of CD)
Distribution of geographical units across the sample in the data	This would particularly matter for place-based analyses, but could also influence possible household/person-based analysis with external area information added, as it could influence the prevalence with which area characteristics are then represented in the data, and this can affect the questions we can ask or the extent to which we can answer them	This is a later step that requires access to the data in some way (probably already requires access to restricted datasets or customised data provided by ADA). This may have to be undertaken and documented for all waves of a longitudinal survey.

Weighting	Weighting is closely related to sample design and affects the analysis and influences what can be reliably reported (e.g. at which level of geography). In the context of geo-spatial integration it is important to explore and understand the properties and performance of weights (particularly design weights) provided with surveys as these might be needed for subsequent data analyses.	Survey weights reinforce the survey design need to be considered in the context of household/person-based analysis.

Note that exploring some of the above aspects will be (considerably) more labour intensive for longitudinal surveys with multiple waves and potentially multiple cohorts (sampling, weighting, sample sizes, distribution of cases, topics).

Table A2.4: Technical considerations for selecting non-survey data

What	**Why**	**Notes/Suggestions**
Geographical level/units	For linking meaningfully with survey data. The ‘closer’ to the geographical units in the survey data the better. Ideally, this would also link with the survey’s sampling design to increase the validity of the analyses and to potentially allow place-based analysis	
Geographical coverage in relation to survey sample	To maximise analysis options the geographical units should cover as much of the survey sample as possible (to minimise survey cases with missing information).	This step would come later and would have to be undertaken in conjunction with some pre-selected datasets
Accessibility	Need easy access to meet timelines for this project	Future possibilities beyond the scope of this project (including those requiring more difficult access) could be mapped out as part of the process
Reliability of information	Needs to consider where information came from and how it was aggregated, if applicable.	Would be less/no issue with Census data.

Concordance considerations

To undertake data integration between survey and non-survey (spatially structured) data it will be necessary to assess the concordance between the two data sources in terms of their geographical and temporal alignment.

Part of considerations around the geographical concordance includes assessing whether it is possible to identify the same geographical unit in both data sources, e.g. CD or SA2 identifier that can be used as a linkage key to connect survey and non-survey data. If no direct correspondence is identified (e.g. it is possible to identify CDs in the survey but relevant non-survey data is only provided at SLA level), it would be necessary to consider whether it is possible to aggregate lower level geographical information (e.g. CD, postcode) to higher levels of aggregate (SLA, SA levels, LGAs, electoral divisions…) This process should prioritise identifying aggregations/correspondences that maximise the 1:1 or m:1 relationships while minimising any splits of geographical units. The assumption here is that we would focus on population weighted correspondences.

Similarly, the temporal alignment of the survey and non-survey data should also be considered. For instance, for a longitudinal survey running over a number of years, it will be important to consider adding spatial data that correspond to some of the survey waves/rounds, rather than trying to append spatial data collected outside of the years covered by the survey. In principle, multiple rounds of spatial data (e.g. multiple Census rounds) could be integrated with a single long-running survey (such as HILDA). However, the changes to geographical classifications that might have taken place over the relevant time period might complicate this process to the point that arriving at temporally concordant spatial datasets is beyond the scope of the current project.

The concordance considerations constitute a vital component that will influence which non-survey area information can be meaningfully linked with which survey information. Specifically, if survey and non-survey datasets are independently selected based on strategic and technical considerations, the data integration project might still fail due to the lack of geographical or temporal concordance between the two. This third set of considerations should therefore be applied simultaneously and in conjunction with the other two (to minimise pursuing wrong leads with the survey and non-survey data).

To illustrate potential concordance issues, Figure A2.1 gives an example of how a non-ABS structure (postcode, which is available in some of the restricted ADA survey datasets) relates to SA2 (which is a prominent level of geography [often approaching the concept of a suburb] for which area data is compiled) purely at a spatial (not population) level.

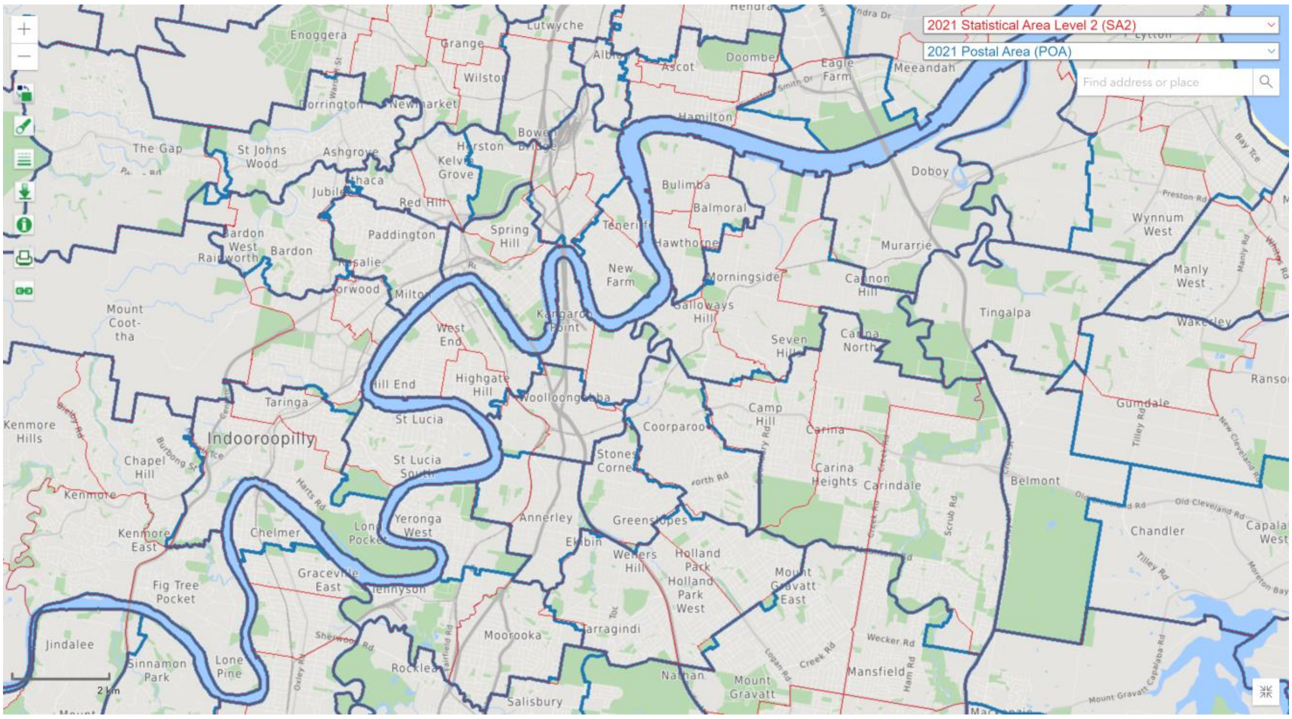

Figure A2.1: Sample of postcode and SA2 boundaries in parts of Brisbane

Appendix 3: Most downloaded ADA surveys

Table A3.1: The number of downloads for most downloaded ADA surveys

	Survey name	Dataverse ID	Download count
1	Household, Income and Labour Dynamics in Australia	354	33924
2	Australian Election Study - Voter Studies	96	14619
3	Longitudinal Study of Australian Children [both cohorts]	888	8864
4	ANU Poll	38	6693
5	Australian Survey of Social Attitudes	2	5892
6	National Drug Strategy Household Survey	284	3269
7	PIA Synthetic Data	431	2737
8	Australian Gallup Poll	1221	2103
9	Longitudinal Study of Indigenous Children	809	2080
10	Historical and Colonial Census Data Archive (HCCDA)	15305	1860
11	Australian Child and Adolescent Surveys of Mental Health and Wellbeing	177	1548
12	Longitudinal Surveys of Australian Youth [200x]	47	1513
13	Building a New Life in Australia	2128	1332
14	ADA General Collection	1847	1032
15	Australian Candidate Study	6501	1012
16	World Values Survey	17	914
17	Australian Historical Criminal Justice Data	15300	673
18	The Australian Longitudinal Study on Male Health	62	660
19	The Comparative Study of Electoral Systems (Australia)	15549	589
20	National Social Science Survey	553	573

Appendix 4: Previous work with HILDA involving spatial analysis

Introduction

This document presents work in progress to catalogue all published work that involved HILDA data and some spatial component to date. The main interest in compiling such work was in identifying the topics or main research interests of such work, and in which way spatial information was derived and used in such work. Documenting this can inform the IRISS project by understanding what has been done and why, and in which way spatial analysis components were implemented and associated issues identified, discussed and tackled.

Approach

This work was executed in three steps:

Identifying published work involving HILDA

Lists with HILDA publications were downloaded from the website of the Melbourne Institute. Three types of outputs were considered:

Books and book chapters (from https://melbourneinstitute.unimelb.edu.au/hilda/publications/books-and-book-chapters);

Journal articles (from https://melbourneinstitute.unimelb.edu.au/hilda/publications/journal-articles);

Other reports (from https://melbourneinstitute.unimelb.edu.au/hilda/publications/reports);

HILDA technical papers (from https://melbourneinstitute.unimelb.edu.au/hilda/publications/hilda-technical-paper-series); and

HILDA discussion papers (from https://melbourneinstitute.unimelb.edu.au/hilda/publications/hilda-discussion-paper-series).

There were 2,072 HILDA-related outputs as per Table A4.1.

Table A4.1: HILDA outputs

Type of output	Number
Books and book chapters	92
Journal articles	1,540
Other reports and output	391
HILDA technical papers	19
HILDA discussion papers	30
Total	**2,072**

Forthcoming work was excluded from these lists (n=59) as no further investigations of such work could be undertaken. The resulting lists were appended to one master list with 2,013 outputs and treated as the universe of existing HILDA publications.

Identifying HILDA work that involved some spatial component

Consecutive searches for title words were then performed on the master list using these search terms:

“spatial”, “geogra”, “region”, “remote”, “rural”, “metropolitan” and “area”.

Hits for any of those searches were copied over to a new list of HILDA publications. Hits did occur under the categories of books and book chapters, journals and other reports. No spatial work was identified under the categories of HILDA technical papers and HILDA technical discussion papers. The new list (n=70) was treated as the universe of HILDA publications involving a spatial component.

Identify topics of work and in which way spatial information played a role/was handled

Each publication on the new list was googled and downloaded where possible. None of the three books and book chapters were accessible online and neither were two journal articles and one report.

Each of the 64 downloaded publications was then scrutinised in relation to methodical information and the overall topic of the publication. Of particular interest in this process were what geographical information was used at what level of the geography, where it came from/how it was derived, and how it was used in the analysis. To this end four fields were used for each publication in the process of their documentation:

A field indicating whether and which spatial information was merged to HILDA data

A field for documenting at which geographical level the information was merged and/or used in the analysis. The latter also included works involving spatial components that did not entail merging external information to HILDA. In some cases, when external information was merged with HILDA data, the geographical level at which the analysis took place varied from the geographical level used in the merge.

A field for capturing the topic. This field documented the main relationships that were investigated in abbreviated format. For example, *area SES → wellbeing* would relate to work that was primarily interested in investigating the influence of area socio-economic characteristics on subjective wellbeing of people. Not all works involving spatial components were interested in relationships, some consisted of uncovering spatial distributions of some characteristics.

A field for documenting the analysis applied to the (merged) data. This initially concerned the main modelling undertaken to investigate the relationships of interest or the methods used to estimate distributions and was later extended to include some more detail on spatial aspects of an analysis.

Limitations

The approach outlined above relies on in the Melbourne Institute having accurately compiled all HILDA-related publications (step 1). The search methodology to identify relevant work that involves some spatial component relied on such work being reflected in the title of the publications (step 2). Some publications could not be downloaded to date and have not been scrutinised as a consequence (step 3). The scrutinization of publications that were downloaded only considered particular sections of publications (most often Methods and Data sections and Abstracts) to identify relevant information, in the process of which such information may have been missed in other sections of the publications (step 3). Authors of publications also displayed different levels of engagement when documenting their methods, which may be related to the journal they published in, but also to the authors' capabilities in understanding methodological issues, their attitudes to transparency/disclosure of methodical information (step 3).

The character of the 'scrutinising' of existing work, at this point, relied more on scanning than on detailed reading to get through all available publications. In this process, not all details of methodological steps were documented (step 3). Publications were not always sufficiently explicit to decipher spatial and methodological detail of interest. For example, it was not always clear at what level of geography SEIFA index scores were derived and/or applied (step 3).

All of these matters constitute limitations for the work presented in this document. As this is a work in progress some limitations could still be minimised in the future, for example, by expanding search techniques (including the utilisation of data bases) in step 2 and/or by gaining access to publications not accessible to date and/or by revisiting individual publications to explore more detail than was apparent when scanning the publications.

Despite the limitations, this document should fulfill its main purpose of informing work on the IRISS project by identifying in which ways spatial information has been considered in analyses of HILDA data. A summary of insights from the documentation is provided next.

Summary of insights

Types of data analyses involving spatial data and HILDA

There are four general ways in which spatial data have been used in conjunction with HILDA data.

Work where spatial areas are selected as an area of interest on the basis of which some analysis is performed.

This is reflected in selecting a sample in the HILDA data by some geographic criterion/criteria. In the works investigated this involved selecting households or people who live in metropolitan areas in Australia, non-metropolitan areas or in individual cities like Melbourne or Sydney.

Work where spatial areas or their characteristics are controlled for in the analysis.

In the works investigated here, this most prominently involved using categories of remoteness and/or derived categories from SEIFA scores (e.g. deciles, quintiles) as control variables in models.

Work where spatial areas, types of areas or their characteristics are considered as possibly influencing some 'outcome'.

In the works considered here popular 'outcomes' in such research are in the areas of subjective wellbeing, physical activity and labour market statuses.

Work where spatial characteristics are considered as 'outcomes' influenced by other things.

Ways of creating spatial information used in analysis of HILDA data

In the publications considered so far, work under (b), (c) and (d) could involve:

creating spatial characteristics within HILDA data; and/or

adding spatial characteristics from other sources to HILDA data.

Usually, spatial characteristics are created within HILDA when they are of interest and when they are not more reliably available in other datasets. Spatial characteristics are added from external sources when they are either not available in the HILDA data (e.g. information about green spaces or area infrastructure) or when the information is more reliably available from other data sources. The latter often involves adding demographic type of data from the ABS Census as the Census captures such data more reliably at the level of small areas.

On few occasions work under (d) included estimating small area characteristics combining HILDA and other (usually Census) data. In these cases, the information estimated was included in HILDA (e.g. life satisfaction) but not at the level of small areas. One way of creating small area estimates was to model relationships between some predictors and the information of interest, such as life satisfaction, within HILDA at geographical levels reliably available, for example at national or state level. The modelling identified the statistical predictors/correlates for life satisfaction whereby the predictors/correlates included in such models were already selected on the basis of their availability at small area levels in other data sets. Based on the prevalence of these predictors/correlates in small areas in other datasets, the outcome of interest, such as life satisfaction was then estimated for small areas using the relationships between predictors/correlates and outcome identified at higher levels of geography in HILDA. This work relied on various assumptions (in addition to those that usually apply when modelling relationships) and comes with limitations. The most notably ones are:

that possible predictors are selected on the basis of their availability in reliable data sets; and

that relationships between predictors and the outcome at high levels of geography (in HILDA) are assumed to hold at lower levels of geography.

The former is a limitation as there are few data sources that hold reliable data at the level of small areas that one could consider from which to identify potential predictors for some outcome. In the considered cases here, they all involved the ABS Census, which covers various topics, however, does so at relatively generic levels of operationalisations. Statistically, this limitation would be reflected in poor model fits when modelling the outcome in the LSAY data due to unobserved factors that influence the outcome.

The plausibility of the assumption that relationships hold across different levels of geography will depend on the characteristic of interest. As people in Australia are free to move anywhere they like and as geographical mobility is relatively high it is possible that people in different regions may be influenced in different ways and that their location reflects their preferences in relation to spatial characteristics. Staying with the example of life satisfaction, it is then possible that relationships between spatial characteristics and life satisfaction vary across small areas as people in different areas value different things.

Topics investigated

Outcomes influenced by spatial matters[5]

Outcomes have been investigated in the context of potential spatial influences in a variety of domains.

Wellbeing and physical activity

Perhaps the most prominent outcomes domain investigated in the context of spatial features with HILDA data has been wellbeing including life satisfaction, mental health/stress, and obesity. At times, closely related to wellbeing and sometimes considered simultaneously has been physical activity as an outcome of interest in spatial contexts. Publications concerned with social participation of people with disability could also be seen as falling under this broad domain.

Labour market

Another prominent outcomes theme in relation to spatial characteristics lies in the domain of labour market outcomes: unemployment (e.g. in regional areas), unemployment duration, entry into self-employment, wages, employment underutilisation, job mobility, perceived job

[5] At times, the interest was purely on investigating the spatial distribution of the prevalence of something. These somethings are also considered as outcomes here.

insecurity, parents' labour force participation (in relation to spatial childcare supply) have all been at the centre of publications involving HILDA data.

Migration/mobility

Another research interest in some HILDA publications was concerned with relationships between spatial context on the one hand, and types of migration, mobility propensities or migration patterns on the other.

Social disadvantage

A further outcomes domain of interest lies around social stratification/disadvantage. Issues investigated under this domain in the context of spatial dimensions were unhealthy housing, persistent disadvantage, poverty transitions, higher education disadvantage, and housing and social inclusion. Publications concerned with social capital and financial pressures may also be included under this domain.

Other outcomes

A number of outcomes in other domains were also considered. Among those outcomes were fertility, ageing preferences (e.g. ageing in place vs other), trust, sibling interactions, appreciation of homes, neighbourhood satisfaction, concentration of ITC infrastructure and commuting patterns.

Spatial data of interest (as a filter or as influencing outcomes)

There were five domains which were particularly prominent in spatial data: the built and natural environment, socio-economic disadvantage, remoteness, labour markets, and the demographics domain.

Built and natural environment

Examples of spatial characteristics of interest under this domain were housing density, green space, vegetation, road coverage, land-use diversity, housing diversity, households with no cars, commuting by public transport, and distance to CBD.

Labour market

While labour-market statuses and behaviours were a prominent outcomes theme, regional labour markets were also considered as influencing other things, such as migration and labour market behaviours. Dimensions within the labour market domain of interest in such context were industry structure, employment growth, population growth, human capital (level of education), unemployment rates, employment rates, self-containment rates and the prevalence of part-time work.

Demographic

Treating area demographic characteristics as potential influencers of other matters was also relatively common. This included regional proportions of persons born overseas, measures of cultural diversity or ethic concentrations, as well as regional housing tenure, dwelling and income structures. Such information was most often sourced from ABS Census data.

Socio-economic disadvantage

SEIFA scores were used in a number of publications, perhaps sometimes because it was convenient to do so (as already included in the HILDA data). The most often used SEIFA indices were the Index of Relative Socio-Economic Disadvantage (IRSD) and the Index of Relative Socio-Economic Advantage and Disadvantage (IRSAD). Beyond SEIFA scores spatial structures of disadvantage were considered in the form of crime rates and unemployment rates, the latter overlapping with the labour market theme. Possibly related to area disadvantage, spatial patterns of housing affordability were also assessed as influencers on outcomes.

Remoteness/urbanisation

Categories indicating the remote or urban character of a region were also commonly investigated as potential influencers on outcomes. These were most often sourced from the remoteness (ARIA[+]) and Section of State (SOS) and, at times, the Section of State Range (SOSR) classifications within the broader ABS ASGS or earlier ASGC classifications.

Other

Other spatial characteristics that have been used in HILDA analysis as independent variables are:

the share of the ‘no’ vote in the Australian Marriage Law Postal Survey 2017 as indicating structural stigmatism as potentially affecting the wellbeing of those stigmatised;

home care packages ratios (from administrative data) as potentially influencing mental health; and

perceptions on childcare availability, quality and costs (from within HILDA) as potentially affecting parents’ labour market behaviours.

Levels of geography used in analysis

This section outlines the levels of geography previously used in analysis of HILDA data. This is regardless of whether spatial data was merged to HILDA data from external sources or whether spatial information was aggregated from within HILDA.

Individually identified areas

Researchers used various spatial geographies and levels, which mostly related to the ASGC or ASGS for analysing and reporting. This included Census Collection Districts, Statistical Local Areas from the ASGC, Statistical Areas 1, 2, 3 and 4, and Greater Capital Statistical Areas from the ASGS. All these levels/classifications identify individual, exclusive regions.

Non-ABS structures that identify individual regions that have been used in such analyses were postcodes, Local Government Areas and Electoral Divisions.

While these geographical areas are mentioned here as individually identifiable areas, the analysis and reporting was usually not concerned with individually identified areas: when individual areas were identified, this was in the context of selecting an area (as a filter) on the basis of which some analysis was performed (e.g. Sydney). On most other occasions, characteristics such as aspects of infrastructure, greenspace or demographics were merged to individually identifiable areas (e.g. at level SA2) and the analysis was then interested in how area characteristics influence some outcomes. In this sense, the individually identified areas were only relevant for merging the right information to them, but became irrelevant in the actual analysis. (This reflects our earlier methodological conclusion that HILDA data can be more legitimately used to investigate in which way area characteristics affect some outcome that is captured in HILDA than to investigate differences between identified small areas, such as between the LGAs Newcastle and Gold Coast.)

Types of areas

Some researchers were not concerned with individual areas or characteristics of individual areas but with a typology of areas. These were also most often sourced from the ABS structures with Sections of State (major urban, other urban, bounded locality and rural balance), the more detailed Section of State Range and the ARIA(+) remoteness categories being particularly prominent.

On occasions, researchers aggregated existing categories or derived them via combining categories of remoteness with section of state, for example, to use the derived categories in the analysis. Of further interest in this context could be the work of the Bureau of Transport and Regional Economics (2005), which presents a process for generating 69 regions as the basis for creating reliable estimates for social capital indicators from HILDA data.

On other occasions, researchers merged information to HILDA at a small level of geography (e.g. postcode) and then aggregated the information to a higher level (e.g. Statistical Division) using some concordance before the data analysis.

Years for the data merged to HILDA

HILDA has a more than 20-year history and this is reflected in the variety and versions of spatial and other data that has been merged to it to undertake analysis that involves some spatial component. ABS Census data used in such processes dates as far back as to the 1991 census and reaches up to the 2016 Census. Census data was sometimes used to create regional growth measures, which partially accounts for going back in time so far. For example, in one journal article authors used Census data for 1991 to 2011 to derive regional population growth rates they merged to HILDA data. However, purely cross-sectional data from the 2006 and 2011 Census were also merged to HILDA in the works scrutinised here, which attests to such work involving some data integration having been undertaken for some time.

Treatment and discussion of (spatial) data integration issues

Collectively, across the publications considered here, the level of discussion of data integration issues was low. Many researchers did not discuss any limitations or implications whether that related to the aggregation of information within HILDA to some spatial level or the analysis at some spatial level.

A few researchers presented some rationale for selecting particular geographical levels and these rationales consisted of theoretical (e.g. larger areas suitable for studying labour market behaviours, smaller areas suitable for studying physical activities) and/or some practical (usually sample size) considerations although there did not appear to be an agreed or commonly used minimum threshold for an area sample in HILDA data across the different publications.

Very few researchers explicitly considered the HILDA sampling design and issues of representativeness in the selection of spatial areas for some analysis. Most of those ended up selecting levels consistent with the sampling design – Census Collection District (CCD, one of

HILDA's sampling units) and Greater Capital City Statistical Areas (GCCSA), the areas for which HILDA was designed to deliver reliable estimates for.

One publication (Kubiszewski, I., Zakariyya, N., Jarvis, D, 2019) modelled outcomes at different levels of geography (SA1, 2, 3 and 4) and discussed differences in the model output in the context of differences in spatial scales.

A few researchers encountered issues of spatial inconsistencies over time that they tackled with some concordance using available correspondence tables (e.g., Dekker, K., Brouwer, W., and Colic-Peisker, V., 2019). This included issues of inconsistent vocabularies for Statistical Local Areas between years (Parkinson, S, Ong, R., Cigdem, M. and Taylor, M , 2014).

Summary

There is a long and visible tradition of undertaking some type of spatial analysis using HILDA data. Of the 64 downloaded publications scanned, at least[6] 24 involved merging some spatial information to HILDA data[7]. It was also relatively common to use spatial information already supplied in the HILDA data sets (particularly SEIFA deciles and remoteness categories). On a few occasions HILDA data was aggregated at some spatial levels, and on very few occasions HILDA data was used in conjunction with other data to generate small area estimates (for areas that did not contain such information in HILDA).

The most prominent outcomes in research involving HILDA and some spatial component appear to lie in the domain of wellbeing, especially subjective wellbeing. Prominent spatial influencers on outcomes were seen in the domain of the built and natural environment, but also in the labour market and demographic domains as well as in types of remoteness/urbanisation.

Thematic domains used as independent and dependent concepts in analyses involving spatial features can also be the same: aspects of labour markets, migration and social disadvantage were, at times, treated as potential influencers and at other times as potential outcomes.

Most of the 64 downloaded publications did not, or did very little elaborate on data integration issues and implications. Few publications made explicit reference to representativeness in the context of the HILDA sample design. Those that did, tended to use information at the level of

[6] It was not always clearly documented whether something, and if so, what was merged to HILDA data.
[7] Note that sometimes the same data merge process was used to inform multiple publications, so the number of times that external data was merged to HILDA was lower than 24.

the area sample unit (CCD) or the level the HILDA survey was designed to be representative for (GCCSA).

Another aspect that influenced the selection of a spatial level in analyses lay in conceptualising spatial relationships in the context of particular themes. For example, labour market behaviours were theorised to be influenced by circumstances in larger areas (e.g. labour market structures at levels of SA4 or higher) while physical activity was seen to be influenced by circumstances closer to one's residence (e.g. green space at the level of SA2 in metropolitan regions[8]).

The scan did not identify many treatments of longitudinal spatial data integration issues and no treatment of longitudinal non-spatial data integration issues[9].

[8] This reasoning may apply less to SA2s outside metropolitan areas as these can cover much larger areas.
[9] It is possible/likely that such treatments of non-spatial longitudinal inconsistencies are included in some of the publications included in the master list of HILDA publications that were not scrutinised here.

Documentation of publications involving some spatial component

The publications listed below were identified using the process outlined in the Approach section. This section especially is a work in progress. There is scope for revisiting the publications and extending on the documentation that was started here, for example, to document in some detail how longitudinal spatial inconsistencies were dealt with.

Table A4.2: Books and Book Chapters

Title	Information merged to HILDA	Spatial level of analysis	Topic	Analysis method
Clark, W.A.V., and William Lisowski, W., Unpacking the Nature of Long-Term Residential Stability in Rachel Franklin (ed) *Population, Place and Spatial Interaction: Essays in Honor of David Plane*, Springer. 2019	?	?	Internal migration	*No access*
Ghasri M., Rashidi T.H., 'Investigating the internal compromise between wife and husband's commute time changes in residential relocation', in Briassoulis H., Kavroudakis D., Soulakellis N. (eds), *The Practice of Spatial Analysis*, Springer, Cham, pp. 325-339. 2019	?	?	Internal migration	*No access*
Crown, D., Corcoran, J. and Faggian, A., 'Migration and human capital: The role of education in interregional migration: The Australian case', in K. Kourtit, B. Newbold, P. Nijkamp and Mark Partridge (eds), The Economic Geography of Cross-Border Migration, Springer, Cham, pp. 247-267. 2021	?	?	Education → Internal migration	*No access*

Table A4.3: Journal articles

Title	Information merged to HILDA	Spatial level of analysis	Topic	Analysis method
Wang, S., Liu, Y., Lam, J. and Kwan, M.P., 'The effects of the built environment on the general health, physical activity and obesity of adults in Queensland, Australia', *Spatial and Spatio-temporal Epidemiology*, vol. 39, article 100456. 2021	SEIFA (IRSAD), housing density, green space coverage, road coverage, land-use diversity, housing diversity, commuting by public transport, households with no cars, distance to CBD	SA2	Built environment → Health, physical activity and obesity	Multilevel mixed effects models after PCA for reducing built environmental variables to factors
Kubiszewski, I., Jarvis, D. and Zakariyya, N., 'Spatial variations in contributors to life satisfaction: An Australian case study', *Ecological Economics*, vol. 164, article 106345. 2019	No merge involved?	Various ASGS	Spatial variation in predictors of life satisfaction	Geographically weighted regressions
Kubiszewski, I., Zakariyya, N., Jarvis, D., 'Subjective wellbeing at different spatial scales for individuals satisfied and dissatisfied with life', *PeerJ*, vol. 7, article 6502. 2019	Spatial vegetation data (NDVI)	SA1, SA2, SA3, SA4	Distribution of subjective wellbeing at different geographies; Natural, social, human and built capital → life satisfaction	Fixed effects models at SA 1 to 4
Gray, E. and Evans, A., 'Geographic variation in parity progression in Australia', *Population, Space and Place*, vol. 24, no. 2, pp. 1-11. 2018	No merge involved	purpose built regional classificati on with 5 categories	Geography (5 categories) → fertility	Multilevel logistic modelling (women aged 16 to 44)

Menigoz, K., Nathan, A., Heesch, K.C. and Turrell, G., 'Neighbourhood disadvantage, geographic remoteness and body mass index among immigrants to Australia: A national cohort study 2006-2014', *PLoS ONE*, vol. 13, no. 1, article e0191729. 2018	No merge involved?	SA1	Spatial disadvantage (IRSD) and remoteness (at SA1 level) → Obesity	Multi-level random effects linear regression models
Perales, F. and Todd, A., 'Structural stigma and the health and wellbeing of Australian LGB populations: Exploiting geographic variation in the results of the 2017 same-sex marriage plebiscite', *Social Science & Medicine*, vol. 208, pp. 190-199. 2018	Share of 'no' vote to plebiscite on same sex marriage	ED	Spatial stigma → Health and wellbeing of stigmatised (LGB)	Multilevel regressions
Han, J.H. and Kim, J., 'Variations in ageing in home and ageing in neighbourhood', *Australian Geographer*, vol. 48, no. 2, pp. 255-272. 2017	No merge involved	Selection of state	Influences on ageing preferences	Included a dummy for major city in models
Han, J.H., Kim, J.Y. and Kim, K., 'Dynamics of housing mobility in Australian metropolitan areas, 2001-2010: A Longitudinal Study', *Urban Policy and Research*, vol. 35, no. 2, pp. 122-136. 2017	No merge involved	Five capital cities	Internal migration in metropolitan areas	descriptive
Clark, W. and Maas, R., 'Spatial mobility and opportunity in Australia: residential selection and neighbourhood connections', *Urban Studies*, vol. 53, no. 6, pp. 1317-1331. 2016	No merge involved	CCD level	Predictors of internal migration flows (up and down as per SEIFA decile)	Regressions

Hermes, K. and Poulsen, M., 'The intraurban geography of generalised trust in Sydney', *Environment and Planning A*, vol. 45, no. 2, pp. 276-294. 2013	No merge involved	CCD level	Distribution of trust in Sydney	used HLDA data and Census 2006 data to create synthetic spatial microdata – small area estimates of generalised trust for Sydney starting from CD level) via combinatorial optimisation; also used GSS data
Keramat, S.A., Alam, K., Al-Hanawi, M.K., Gow, J., Biddle, S.J. and Hashmi, R., 'Trends in the prevalence of adult overweight and obesity in Australia, and its association with geographic remoteness', *Scientific Reports*, vol. 11, article 11320. 2021	No merge involved	Custom built remoteness variable using available info in HILDA	Remoteness → Obesity	Random effects logit models
Perales, F. and Plage, S., 'Sexual orientation, geographic proximity and contact frequency between adult siblings', *Journal of Marriage and Family*, vol. 82, no. 5, pp. 1444-1460. 2020	No merge involved	Something from within HILDA	Sexual orientation → Sibling interactions and sibling geographical proximity	Random effects multilevel models
Watson, N., 'Measuring geographic mobility: Comparison of estimates from longitudinal and cross-sectional data', *Survey Research Methods*, vol. 14, no. 1, pp. 1-18. 2020	No merge involved	none	Compared mobility estimates from HILDA with Census and GSS	Descriptive analysis and modelling

Baker, E., Lester, L., Beer, A. and Bentley, R., 'An Australian geography of unhealthy housing', *Geographical Research*, vol. 57, no. 1, pp. 40-51. 2019	Coefficients/weights from modelling Australian survey of Housing and Wellbeing data (but these were not merged to spatial levels)	Remoteness (as part of analysis, not merge)	Distribution of unhealthy housing	Applied weights from modelling relationships in ASHW data to cases in HILDA based on their socio-demographic characteristics; then proceeded with cross-tabs/ graphs
Baker, E., Bentley, R., Lester, L. and Beer, A., 'Housing affordability and residential mobility as drivers of locational inequality', *Applied Geography*, vol. 72, pp. 65-75. 2016	?	?	Housing affordability + residential mobility→ Locational inequality	*No access to article*
Baum, S., Bill, A. and Mitchell, W., 'Unemployment in non-metropolitan Australia: Integrating geography, social and individual contexts', *Australian Geographer*, vol. 39, no. 2, pp. 193-210. 2008	Area characteristics from Census 1991/2001 (employment [incl growth over 91-01], population, industry mix, education level)	Non-metropolitan LGAs	Area and individual characteristics → unemployment (in regional areas)	Multi-level modelling
Butterworth, P., Rodgers, B. and Jorm, A. F., 'Examining geographic and household variation in mental health in Australia', *Australian and New Zealand Journal of Psychiatry*, vol. 40, no. 5, pp. 491-497. 2006	(Probably) no merge involved	CCD level	Area (SEIFA decile and remoteness [4 categories]), household and individual characteristics → mental health	Multi-level fixed effects modelling

Tran, M.M. and Gannon, B., 'The regional effect of the consumer directed care model for older people in Australia', *Social Science & Medicine*, vol. 280, article 114017. 2021	Use of home care package ratios from AIHW	SA2	Regional HCP variations and individual characteristics → Mental health	Difference in difference models (pre and post introduction of CDC model)
Crown, B.D., Gheasi, M. and Faggian, A., 'Interregional mobility and the personality traits of migrants', *Papers in Regional Science*, vol. 99, no. 4, pp. 899-914. 2020	No merge involved	GCCSA	Personality traits and other characteristics → internal migration prob (using 16 GCCSA for determining migration)	Probit fixed effects regressions
Nikolaev, B.N. and Wood, M.S., 'Cascading ripples: Contagion effects of entrepreneurial activity on self-employment attitudes and choices in regional cohorts', *Strategic Entrepreneurship Journal*, vol. 12, no. 4, pp. 455-481. 2018	No merge involved	GCCSA by age and gender	Regional cohort self-employment and individual characteristics → individual entry into self-employment	Regional self-employment proportions calculated within HILDA sample; Multi-level logit models
Perales, F., 'Dynamics of job satisfaction around internal migrations: A panel analysis of young people in Britain and Australia', *The Annals of Regional Science*, vol. 59, no. 3, pp. 577-601. 2017	No merge involved	none	Internal migration (distance and motivation) → job satisfaction	Linear fixed effects
Elias, A. and Paradies, Y., 'The regional impact of cultural diversity on wages: Some evidence from Australia', *IZA Journal of Migration*, vol. 5, article 12. 2016	COB from 2001 and 2011 Census	Merged as postcode	Changes in regional cultural diversity → Wages	PC level data transformed to LGA level (using 2016 concordance)– then calculation of regional

				characteristics; Shift-share models OLS and FE
Cobb-Clark, D.A. and Sinning M.G., 'Neighborhood diversity and the appreciation of native- and immigrant-owned homes', *Regional Science and Urban Economics*, vol. 41, no. 3, pp. 214-236. 2011	Proportion born overseas and SEIFA from Census 2001 and 2006	postcode	Neighbourhood diversity and other characteristics → appreciation of homes (value, as estimated by home owner respondents)	Perceptions on neighbourhood calculated at postcode level within HILDA; quantile regressions
McPhedran, S., 'Disability and community life: Does regional living enhance social participation?', *Journal of Disability Policy Studies*, vol. 22, no. 6, pp. 40-54. 2011	No merge involved	Regional vs major city (ARIA)	Remoteness → social participation of people with disability	Area characteristics (SEIFA, housing tenure, area attachment) aggregated from within HILDA; multiple linear regressions
McPhedran, S. Regional living and community participation: are people with disability at a disadvantage? 2010 https://research-repository.griffith.edu.au/bitstream/handle/10072/61514/95748_1.pdf?sequence=1	No merge involved		Area and other characteristics → social connectedness / life satisfaction	Area characteristics aggregated from within HILDA; multiple regressions

Kettlewell, N., 'The impact of rural to urban migration on well-being in Australia', *Australasian Journal of Regional Studies*, vol. 16, no. 3, pp. 187-213. 2010	No merge involved	Section of state (major urban, other urban, bounded locality, rural balance)	Rural to urban migration → life satisfaction	Dynamic fixed effects model
Baum, S., Bill, A. and Mitchell, W., 'Employment outcomes in non metropolitan labour markets: Individual and regional labour market factors', *Australasian Journal of Regional Studies*, vol. 14, no. 1, pp. 5-25. 2008	Employment growth (industry mix, region-specific), manufacturing share, services industry share, human capital from 1991 and 2001 Census	LGA	Regional and individual factors → employment under-utilisation	Multivariate logit models with clustering
Awaworyi Churchill, S. and Smyth, R., 'Locus of control and the mental health effects of local area crime', *Social Science & Medicine*, vol. 301, article 114910. 2022	Crime rates from police statistics (violent, property and total)	postcode	Local area crime rates → mental health	Fixed effects models (also including area characteristics from within HILDA)
Baffour, B., Chandra, H. and Martinez, A., 'Localised estimates of dynamics of multi-dimensional disadvantage: An application of the small area estimation technique using Australian survey and Census data', *International Statistical Review*, vol. 87, no. 1, pp. 1-23. 2019	No merge involved		Small area (at level SA 4) estimation of persistent disadvantage (using Census SEIFA scores in the process)	Area-level version of generalised linear mixed model with logit link function

Forbes, M. and Barker, A., 'Local labour markets and unemployment duration', *Economic Record*, vol. 93, no. 301, pp. 238-254. 2017	Local unemployment rates from Small Area Labour Markets (SALM) data (after concording this data to SLA)	SLA	Local labour markets → unemployment duration	Semi-parametric risk models and piecewise constant baseline models; Robustness checking of spatial aggregation via repeating analysis at SA4 level (involved some concordances)
Perales, F. and Plage, S., 'Losing ground, losing sleep: Local economic conditions, economic vulnerability, and sleep', *Social Science Research*, vol. 62, pp. 189-203. 2017	Local unemployment rates	SA4	Local labour markets and indiv economic vulnerability → sleep	Random intercept multilevel models
Tomaszewski, T., 'Living environment, social participation and wellbeing in older age: The relevance of housing and local area disadvantage', *Journal of Population Ageing*, vol. 56, no. 1/2, pp. 119-156. 2013	?	?	Local area disadvantage → social participation and wellbeing	Random effects models; *No access to publication*
Breunig, R.V., Weiss, A., Yamauchi, C., Gong, X. and Mercante, J., 'Child care availability, quality and affordability: Are local problems related to maternal labour supply?', *Economic Record*, vol. 87, no. 276, pp. 109-124. 2011	No merge involved	SD (generated from postcode)	Area perceptions on childcare availability, quality and costs → parents' labour force participation	Linear maximum likelihood models; Robustness testing with other geographies (SLA, LFR, MSR x SOS)

Baum, S., Arthurson, K. and Rickson., K., 'Happy people in mixed-up places: The association between the degree and type of local socioeconomic mix and expressions of neighbourhood satisfaction', *Urban Studies*, vol. 47, no. 3, pp. 467-485. 2010	Area information on housing tenure, income and ethnic backgrounds from Census 2011	CD	Local SES mix → neighbourhood satisfaction (in metropolitan areas)	Logit regressions with clustering
Baum, S., Bill, A. and Mitchell, W., 'Labour underutilisation in metropolitan labour markets in Australia: Individual characteristics, personal circumstances and local labour markets', *Urban Studies*, vol. 45, no. 5-6, pp. 1193-1216. 2008	Area employment rates, self-containment rates, perct part-time	LFR (n=36)	Area labour markets and individual characteristics → Labour market outcomes	multivariate logit models with clustering
Vidyattama, Y., 'Assessing the association between trust and concentration area of migrant ethnic minority in Sydney', *Australian Economic Review*, vol. 50, no. 4, pp. 412-426. 2017	Data from ABS 2006 and 2011 (2011 data was transformed from SA1 to CD level)	CD	Concentration of migrant minorities and other area (incl SEIFA at SLA) and individual characteristics → trust	Local Moran I statistics to identify spatial concentration of migrants; Multivariate regressions of trust
Dockery, A.M., Seymour, R. and Koshy, P., 'Promoting low socio-economic participation in higher education: A comparison of area-based and individual measures', *Studies in Higher Education*, vol. 41, no. 9, pp. 1692-1714. 2016	No merge involved	PC	Area (and individual) disadvantage → HE participation	Logit regressions to construct individual measures of SES; Crosstabulations of the categories of the new measure with SEIFA quartiles

Milner, A., Kavanagh, A., Krnjacki, L., Bentley, R. and LaMontagne, A.D., 'Area-level unemployment and perceived job insecurity: Evidence from a longitudinal survey conducted in the Australian working-age population', *Annals of Occupational Hygiene*, vol. 58, no. 2, pp. 171-181. 2014	No merge involved? (area unemployment rates calculated from within HILDA?)	Major statistical regions (n=13 – 2 per larger state)	Area level unemployment → perceived job insecurity	Mixed effects multi-level regressions
Ali, M.A., Alam, K. and Taylor, B., 'Measuring the concentration of information and communication technology infrastructure in Australia: Do affordability and remoteness matter?', *Socio-Economic Planning Sciences*, vol. 70, article 100737. 2020	No merge involved	16 GCCSAs	Remoteness + SEIFA → concentration of ITC infrastructure	various
Butterworth, P., Kelly, B.J., Handley, T.E. and Inder, K.J., 'Does living in remote Australia lessen the impact of hardship on psychological distress?', *Epidemiology and Psychiatric Sciences*, vol. 27, no. 5, pp. 500-509. 2018	?	?	Remoteness, financial hardship etc → psychological distress (in rural and remote regions)	Multi-level logistic regressions; HILDA only used in sensitivity testing
Venn, D. and Hunter, B., 'Poverty transitions in non-remote Indigenous households: The role of labour market and household dynamics', *Australian Journal of Labour Economics*, vol. 21, no. 1, pp. 21-44. 2018	No merge involved	None (HILDA is assumed to represent non-remote areas)	Household dynamics (trigger events) → poverty transitions	Cross-tabulations of transition events with significance tests

Dockery, A.M. and Lovell, J., 'Far removed: An insight into the labour markets of remote communities in Central Australia', *Australian Journal of Labour Economics*, vol. 19, no. 3, pp. 145-174. 2016	No merge involved	None (as far as HILDA is concerned)	Labour markets in remote central Australia	HILDA only used at national level to compare study results using other survey data
Inder, K.J., Berry, H. and Kelly, B., 'Using cohort studies to investigate rural and remote mental health', *Australian Journal of Rural Health*, vol. 19, no. 4, pp. 171-178. 2011	This is a paper that introduces various studies in the mental health space in rural and remote Australia that, at the time, were yet to be undertaken. This includes a project involving HILDA data.			
Sharifi, F., Nygaard, A. and Stone, W.M., 'Heterogeneity in the subjective well-being impact of access to urban green space', *Sustainable Cities and Society*, vol. 74, article 103244. 2019	Population density, distance to public transport from Census; Urban green space accessibility index (self-calculated from DELWP data)	SA1a in metropolitan Melbourne	Access to urban green space → subjective well-being	Various regressions involving location, time and individual-fixed and random effects
Ambrey, C.L., 'An investigation into the synergistic wellbeing benefits of greenspace and physical activity: Moving beyond the mean', *Urban Forestry and Urban Greening*, vol. 19, pp. 7-12.	Greenspace data from PMSMA Australia Limited Transport and Topography mapping data (using GIS)	CD (within major capital cities)	Green space, physical activity → subjective wellbeing	Conditional logistic regressions
Ambrey, C., 'Greenspace, physical activity and wellbeing in Australian capital cities: How does population size moderate the relationship?', *Journal of Public Health*, vol. 133, pp. 38-44.	Greenspace data and Population data from Census	CD (within major capital cities)	Green space, physical activity and population size → subjective wellbeing	Cluster-specific fixed effects models

Ambrey, C., 'Urban greenspace, physical activity and wellbeing: The moderating role of perceptions of neighbourhood affability and incivility', *Land Use Policy*, vol. 57, pp. 368-644.	Greenspace data from PMSMA Australia Limited Transport and Topography mapping data (using GIS)	CDs (within major capital cities)	Green space, physical activity and perceptions on affability of neighbourhood → subjective wellbeing	As above with clustering at LGA level
Rashidi, T., 'Dynamic housing search model incorporating income changes, housing prices and Life-cycle Events', *Journal of Urban Planning and Development*, vol. 141, no. 4, pp. 04014041.	none (Regional unemployment rates appear to be calculated within HILDA)	Major Statistical Region (unemployment)	Various → relocating	Various econometric models
Ambrey, C.L. and Fleming, C.M., 'Public greenspace and life satisfaction in urban Australia', *Urban Studies*, vol. 51, no. 6, pp. 1290-1321.	Not clear what was merged	CD (greenspace, SEIFA, population density) in Capital cities	Greenspace → life satisfaction	Various econometric including ordered logit models
Ambrey, C.L. and Fleming, C.M., 'Valuing ecosystem diversity in South East Queensland: A life satisfaction approach', *Social Indicators Research*, vol. 115, no. 1, pp. 45-65.	Simpson's diversity index	CD (within the South East Qld Bioregion	Ecosystem diversity → life satisfaction (in SE QLD)	Ordered probit and OLS regressions
Phelps, C., Harris, M.N., Ong, R., Rowley, S. and Wood, G.A., 'Within-city dwelling price growth and convergence: Trends from Australia's large cities', *International Journal of Housing Policy*, vol. 21, no. 1, pp. 103-126. 2021	HILDA data was only used for a descriptive table with non-spatial information			

Baker, E., Pham, N.T.A., Daniel, L. and Bentley, R., 'New evidence on mental health and housing affordability in cities: A quantile regression approach', *Cities*, vol. 96, article 102455. 2020	No merge involved	Cities and major regional towns	Housing affordability → mental health (cities)	Panel regressions with year and individual level foxed effects; (controls for states and major cities)
Biddle, N. and Montaigne, M., 'Income inequality in Australia – Decomposing by city and suburb', *Economic Papers*, vol. 36, no. 4, pp. 367-379. 2017	HILDA was only used to estimate the mean values of household income ranges from the census data. These were then used to look at the distribution of income inequality based on Census data.			
Black, D., O'Loughlin, K., Kendig, H. and Wilson, L., 'Cities, environmental stressors, ageing and chronic disease', *Australasian Journal on Ageing*, vol. 31, no. 3, pp. 147-151. 2012	No merge involved	SOS	Long-term residence by SOS + SEIFA + indiv characteristics → ageing and disease	Logistic regression and survival modelling
Bill, A., Mitchell, B. and Welters, R., 'Job mobility and segmentation in Australian city labour markets', *International Journal of Environment, Workplace and Employment*, vol. 3, no. 3-4, pp. 212-229. 2007	No merge involved	Metropolitan (ADL, BNE, Perth, SYD, MEL collectively) vs non-metropolitan	City labour markets and job mobility	Clustered (person) logit models
Terrill, M., Batrouney, H., Ha, J., and Hourani, D. (2018). *Remarkably adaptive: Australian cities in a time of growth*, Grattan Institute. (report)	HILDA was only used for sourcing commuting times.			

Table A4.4: Reports

Title	Information merged to HILDA	Spatial level of analysis	Topic	Analysis method
Dekker, K., Brouwer, W., and Colic-Peisker, V. (2019). *Suburb with a higher concentration of Muslim residents in Sydney and Melbourne: Spatial concentration, community, liveability and satisfaction* (RMIT Draft Report), Part of 'The impact of ethnic diversity, socioeconomic disadvantage and sense of belonging on Islamophobia and social cohesion locally and nationally: a mixed-method, longitudinal analysis'. RMIT University.	Census (age, gender, cob, education, emp status, occupation, tenure, household income, family size, IRSAD)	SA2 (used 2006 boundaries for merging 2006, 2011 and 2016 Census data)	Ethnic concentrations in suburbs over time → satisfaction and liveability (Syd, Mel)	Multiple regressions
Productivity Commission, *Geographic labour mobility*, Research Report, Canberra, April. 2014	HILDA data was used for investigating reasons for moving in relation to labour force and employment status (no explicit spatial component in this analysis).			
Johnson, L., Hossain, A., Thomson, K., and Jones, W. (2016). *Cities: Lengthy commutes in Australia*. Department of Infrastructure and Regional Development, Bureau of Infrastructure, Transport and Regional Economics, Research Report 144.	No merge involved	Remoteness, SOS, GCCSA	Distribution of commuting patterns	descriptive
Bureau of Infrastructure, Transport and Regional Economics, ***Population Growth, Jobs Growth and Commuting Flows in Melbourne***, Research Report No. 125, Canberra. 2011	HILDA was only referenced as part of referring to previous work.			
Bureau of Infrastructure, Transport and Regional Economics, ***About Australia's Regions – Jun 2008***, BTRE, Canberra. (various years)	No merge involved	Remoteness categories	Remoteness and social capital, financial pressures (as	Descriptive tables

			self-reported in HILDA)	
Bureau of Transport and Regional Economics, ***Focus on Regions No. 4: Social Capital***, Information Paper No. 55, BTRE, Canberra. 2005		State remoteness categories and urban centre size categories (from SOS – 9 [sub] categories used); also developed 69 regional categories from ASGC SD and SSDs	Regions and social capital	Process of deriving 69 regions to generate reliable estimates for social capital indicators from HILDA data; Considered risk of unrepresentativeness at these regional levels by using particular sample errors in sig testing Descriptive tabulating of social capital indicators Clustering of regions by type of social capital profile
Stone, W., Reynolds, M. and Hulse, K., ***Housing and Social Inclusion: A Household and Local Area Analysis***, Final Report No. 207, Australian Housing and Urban Research Institute, May. 2013	No merge involved	Remoteness (3 categories) and SEIFA IRSD (3 categories) (both CD based)	Housing and social inclusion and local areas Interest in area types rather than specific	Descriptive analysis and linear regression of social exclusion measure (0-7)

		Derived measure combining the two (3x3) with 9 categories	small areas (e.g. pc 4103) also because HILDA is not suitable for analysing the latter	
Stone, W. and Reynolds, M., ***Social Inclusion and Housing: Towards a Household and Local Area Analysis***, Positioning Paper No. 146, Australian Housing and Urban Research Institute, March. 2012	This paper preceded the above study and defined parameters of the methodological approach.			
Tanton, R., Vidyattama, Y., and Miranti, R. (2016). ***Small area indicators of wellbeing for older Australians (IWOA)***. Prepared for the Benevolent Society by NATSEM (National Centre for Social and Economic Modelling) for the Institute for Governance and Policy Analysis, University of Canberra. 2014	na	SA2	Small area estimates of Wellbeing indicators for older Australians	Area estimates for indicators derived from techniques using various survey (including HILDA) and Census data
Parkinson, S, Ong, R., Cigdem, M. and Taylor, M., *Well-being outcomes of lower income renters: A multi-level analysis of area effects*, Final Report No. 226, Australian Housing and Urban Research Institute, August. 2014	Median household income, tenure, landlord type by dwelling structure from 2001, 06 and 11 Census	SLA (CD also considered and rejected) (matching at SLA level was not straight forward though as differences in SLA	Areas → wellbeing of lower income renters	Imputation of area information for between Census years using linear interpolation Multilevel models also controlling for regional unemployment at level of Metropolitan Statistical Region

		vocabularies)		
Berry, H.L., Bode, A. and Capon, A., *Mental Health in Australia's Million-Plus Cities: Social Environments, Built Environments and Psychological Dynamics*, Department of Families, Housing, Community Services and Indigenous Affairs Report. 2010	*Not accessible*			