Privacy Preserving Geo-Linkage in the Big Urban Data Era

Richard O. Sinnott, Christopher Bayliss, Andrew Bromage, Gerson Galang,
Yikai Gong, Philip Greenwood, Glenn Jayaputera, Davis Marques,
Luca Morandini, Ghazal Nogoorani, Hossein Pursultani,
Muhammad Sarwar, William Voorsluys, Ivo Widjaja
Department of Computing and Information Systems,
University of Melbourne, Melbourne, Australia
Contact Author: rsinnott@unimelb.edu.au

ABSTRACT

Big data technologies and a range of Government open data initiatives provide the basis for discovering new insights into cities; how they are planned, how they managed and the day-to-day challenges they face in health, transport and changing population profiles. The Australian Urban Research Infrastructure Network (AURIN – www.aurin.org.au) project is one example of such a big data initiative that is currently running across Australia. AURIN provides a single gateway providing online (live) programmatic access to over 2000 data sets from over 70 major and typically definitive data-driven organizations across federal and State government, across industry and across academia. However whilst open (public) data is useful to bring data-driven intelligence to cities, more often than not, it is the data that is not-publicly accessible that is essential to understand city challenges and needs. Such sensitive (*unit-level*) data has unique requirements on access and usage to meet the privacy and confidentiality demands of the associated organizations. In this paper we highlight a novel geo-privacy supporting solution implemented as part of the AURIN project that provides seamless and secure access to individual (unit-level) data from the Department of Health in Victoria. We illustrate this solution across a range of typical city challenges in localized contexts around Melbourne. We show how unit level data can be combined with other data in a privacy-protecting manner. Unlike other secure data access and usage solutions that have been developed/deployed, the AURIN solution allows any researcher to access and use the data in a manner that meets all of the associated privacy and confidentiality concerns, without obliging them to obtain ethical approval or any other hurdles that are normally put in place on access to and use of *sensitive* data. This provides a paradigm shift in secure access to sensitive data with geospatial content.

Keywords: *Big data; Data privacy; Geo-spatial systems; Urban research.*

I. INTRODUCTION

The Australian Urban Research Infrastructure Network (AURIN) project (www.aurin.org.au) is a major federally funded project across Australia [1]. The project commenced in 2010 with a $24m investment by the Department of Industry and has since received $4m (2013) and a further $2m investment (2015) from the Department of Education as part of the Australian SuperScience and the Education Investment Fund initiatives. AURIN was tasked with providing urban and built environment researchers across Australia with a research environment offering seamless and secure access to a wide array of highly distributed and typically completely heterogeneous data sets from a multitude of autonomous data providers. The data providers were selected on their holding the definitive data sets across the cities and urban settlements of Australia. These data providers stem from federal government, State-based organizations, commercial/industrial organizations and research and academia.

AURIN has also been tasked with providing a rich and extensible portfolio of state of the art analysis and visualization tools reflecting best practice in geospatial data analytics. The AURIN platform allows deep understanding of key research issues surrounding Australia's past, current and future urban settlements [2]. AURIN is unique in Australia and indeed globally with the capabilities for large-scale integration of a rich range of heterogeneous data holdings that were hitherto largely independent silos of data, information and ultimately knowledge. At the heart of the AURIN platform has been providing online, i.e. *programmatic* access, to diverse and heterogeneous data sets through a range of targeted (web) services in a manner that supports the researchers and their associated research processes. Through the deployment of the infrastructure within the Australian Access Federation

(www.aaf.edu.au) researchers are able to authenticate once at their home sites using their own username/passwords, and are subsequently able to access the AURIN portal and all remote data sets without further authentication challenge/response demands – so called *single sign-on*. The fact that the data being requested is from a remote and autonomous agency is completely transparent to the end users. All of the data and tools within AURIN are accessible through a single gateway that is openly accessible to all academics, with features and capabilities to allow access to non-academics. Increasingly industry and especially local government users are accessing and using the AURIN resources, e.g. to compare their own local data and compare it with other national data sets.

The development and evolution of the AURIN gateway and associated infrastructures has not been without its challenges. One of the major challenges that AURIN faced at the project commencement was the lack of maturity of data providers in supporting online programmatic access to their data resources. Whilst many sites were willing to make their data available, e.g. through the websites directly or in the form of Excel spreadsheets that could be downloaded from those websites, AURIN did not wish to pursue a centralized data warehouse-based approach since these data sets would rapidly become obsolete and researchers typically want to know that they were accessing the official data rather than some copied version of the data. To address this, AURIN pursued a model of live (programmatic) access to the definitive and potentially evolving data resources reflecting Australia's urban settlements. In undertaking this, AURIN could not mandate any particular technical solution to the data providers. Instead AURIN was tasked with interworking with whatever systems and solutions these sites were willing to host and support in accessing their systems – this included the data access and use technologies as well as their underlying security systems. The web service solutions that have since been developed have covered a diverse portfolio of technical solutions including Simple Object Access Protocol (SOAP) services; Representational State Transfer (ReST)-style web services; Open Geospatial Consortium (OGC) enabled web services such as web feature services (WFS) and web mapping services (WMS) (www.opengeospatial.org), through to Statistical Data Markup Exchange (SDMX – www.sdmx.org) services.

AURIN has engaged with a huge array of organizations across Australia that hold data sets that are fundamental to urban research and enabling smart, data-driven management of cities. These include data sets at a national (federal) level from major governmental organizations such as the Australian Bureau of Statistics (ABS - www.abs.gov.au) and the Bureau of Infrastructure, Transport and Regional Economics (BITRE - www.bitre.gov.au) amongst many others. At a State-level, this includes government agencies and authorities such as the Department of Transport Victoria (VicRoads - www.vicroads.vic.gov.au) and Department of Human Service (DHS – www.dhs.vic.gov.au) amongst many others. Furthermore, numerous commercial organizations also hold data sets that need to be unlocked for urban researchers, e.g. data from commercial energy and water suppliers and geospatial data from organizations such as the Public Sector Mapping Agency (www.psma.com.au). Similarly, a wide array of research projects and research groups hold urban research data sets of national significance.

The AURIN project has successfully delivered an advanced Cloud-based infrastructure supporting seamless access to all of these data holdings from all of these data providers and many more. This has only been possible through close collaboration with the many distributed and importantly, autonomous government, industry and research agencies. The implementation of the AURIN platform leverages best practice in web-based service-oriented architectures utilizing Cloud-based technologies for federated data querying and integration using a range of data clients targeted to the demands of the remote data providers based on the data formats they possess, and the ways in which they are willing to make this data accessible. Specifically the AURIN platform has been developed on the openStack-based National eResearch Collaboration Tools and Resources (NeCTAR – www.nectar.org.au) Research Cloud. The actual production version of the AURIN platform is hosted on a VMware-based Cloud infrastructure hosted at the University of Melbourne. This includes three database servers and two compute servers with failover redundancy through VMware. This system has been extremely robust and reliable with less than three days outage over the last five years, due to a networking hardware failure.

The system as a whole supports a range of components driven by rich metadata provided by the data providers themselves. The data itself is retrieved from data providers and converted into GeoJSON format (www.geojson.org) and subsequently stored within a no-SQL database (CouchDB – www.couchdb.apache.org) for analysis and visualization. The core AURIN technological platform is described in more detail in [3]; the typical use cases that have driven the use of aggregate level data sets described in detail in [4]; the way in which the existing data as a whole is re-used is described in [5],

and the tools and workflows used for analysis of the various data sets accessible through the AURIN platform are described in [6].

It is the case however that the vast majority of the data sets accessible through AURIN are aggregated, e.g. to specific geospatial areas (postcodes, statistical local areas, local government authorities etc). Such levels of aggregation are often used to overcome privacy issues. Whilst use of aggregate level data is important, it is often non-aggregated (disaggregated) data that are essential to understand the challenges in particular situational contexts. For example, understanding spatial patterns of human interaction and engagement in particular situational settings is often critical to understand challenges of cities, e.g. where should the next train station be built? Why is there so much crime here? Why do people have particular health challenges in a given locality?

It is noted that not all disaggregated data sets have sensitivities, e.g. the location of public transport stations is not in itself private information, however asking people about their commuting behaviour and the distance they live from a local train station is sensitive information and has privacy demands. Such individual-level data is often captured from attitudinal responses to targeted questionnaire-based surveys. There is thus a demand for solutions that provide online access to and use of unit-level data related to situational contexts. It is noted that these questions and responses can be on a variety of topics: health, crime, transport, housing etc. Importantly in AURIN, the typical hurdles in supporting access to and use of unit level data should be removed as far as possible. Thus many organizations provide solutions where researchers are obliged to obtain ethics approvals before data can be accessed and linked [7]. This can often be time consuming and dissuades researchers from accessing and using data for research purposes. Alternatively, some solutions exist that require researchers to physical visit the location where the secure data is housed and support a range of checks to ensure that their linkage/use of the data meets all necessary demands placed by the data holders [8]. These solutions may often be based on completely valid demands, but they place hurdles on researchers that make access to and use of data difficult and more challenging than they might otherwise be. Ideally any researcher should be able to access and use sensitive data, where the usage is largely unrestricted and where the technical solutions enforce privacy for all users. In the technical realization of such a system it is absolutely essential that the privacy and confidentiality rights of the individuals that might complete such surveys are fully supported, and all concerns of the data providers are fully recognized and explicitly supported.

This paper provides a case study in how AURIN supports individual-level geospatial privacy exploiting *privacy preserving* geospatial data linkage solutions utilizing unit-level geocoded data. The work is based on a major health survey conducted by the Department of Health in Victoria where over 25,000 Victorians were interviewed on aspects of their health and lifestyles based on their own individual situational contexts. The italicized term *privacy preserving* here indicates that the unit-level data, e.g. the address of the individual, should never be released directly. Rather analysis is undertaken on the unit-level through a black box-based approach and aggregate results returned - where appropriate, e.g. subject to privacy limitations.

The rest of this paper is structured as follows. Section II covers related privacy preserving work. Section III presents the architecture of the privacy preserving solution and its application in a range of case studies. Finally section IV draws conclusions and identifies areas of future work.

II. Related Work on Data Privacy

Data privacy is an area of increasing research importance, especially as more organizations move towards an outsourced model of IT provision and data sharing through the Cloud. A rich body of work has been undertaken in the area of information governance and protection of user privacy [9-11]. In the science gateway domain, much works has focused on authentication, authorization, auditing and accounting, e.g. [22,23] are prime examples of this. This security model does not tackle privacy however – just secure and monitored access to sensitive data. For many organisations, this model is simply not possible for legal and confidentiality reasons. It is the case that a balance needs to be struck between protection of individual privacy and balancing the positive benefits that can be derived from use and re-use of existing data. Thus access to private/sensitive information should only be supported if there are clear benefits to be obtained from data sharing. It is the case that data can include both privacy information and non-private information that can be used for research purposes. There is an increased concern that once data goes beyond the organizational firewall then there is a danger that it could *eventually* be used to re-identify an individual. Given this, ideally the data should stay within the organizational firewall, but secure forms of access to the data are supported without direct data release.

Within AURIN, data can have a range of aggregated and non-aggregated characteristics. The location of a train station is non-aggregated and not necessarily sensitive information, whilst the names and addresses of individuals that live within 100m of that train station in sensitive and thus has to be protected for privacy concerns. Aggregation of such information is one common model of preserving privacy, e.g. the actual number of individuals that live within 100m (or more!) of the station compared to the specific locations where those individuals live.

This is not without its challenges however. For example, de Montjoye et al [12] showed that over 90% of individuals could be identified using no more than four given geospatially coded data elements. Other work requires that organizations and society more generally consider all phases of the data lifecycle: from data creation, collection, use, sharing, subsequent uses, and disposal. A major issue is the future potential to use and misuse data if and when it is released into the research community. Thus what might be considered as anonymous now, e.g. through suitable encryption techniques, may ultimately be broken by brute force approaches or more likely through the encryption keys being compromised by human errors or errors in security process. Direct release of data is therefore a risk that many organizations are unwilling to contemplate with sensitive data.

With regards to the specific needs and challenges of geospatial data, a range of solutions has been explored for protecting user location privacy while still guaranteeing the utility of information. K-anonymity [13] and (k, δ)-anonymity [14] are two examples that protect privacy information based on partition-based location privacy models. K-anonymity protects privacy in location-based systems based on the hypothesis that it is impossible for attackers to differentiate an individual from k other different individuals. When it is used for location privacy preservation, the set of k points should collectively be indistinguishable from one another. There are many ways to implement this method, such as introduction of dummy locations [15] and by cloaking [16]. The former solution adds k-1 properly selected dummy points and uses both the real and dummy locations for analysis. Cloaking uses artificial cloaking areas that include k points sharing some property of interest for analysis. The drawback of K-anonymity is that it is built on assumptions about the quantity of a potential attacker's auxiliary knowledge, i.e. the approach fails if dummy locations can be distinguished from real locations. Although some improvements have been proposed such as considering ubiquity, congestion and uniformity when dummy points are generated, e.g. to make them look more similar to real locations or taking an individual's auxiliary information into consideration, there are also some defects that can be exploited. For example, assumptions are made regarding how much additional information an attacker might have.

In recent years, differential privacy [17] has been widely used for the protection of location-based data [18]. In this model, adding moderate degrees of noise can be used to obfuscate the location and thereby preserve location privacy. The advantage of differential location privacy is that it allows to protect individual location information whilst still allowing the data to be used for analysis and/or mining. The problems with such approaches are that when noise is added to data, it becomes increasingly difficult to utilize and has the potential for introducing inaccurate results [18]. A better solution is to use the *unit-level* data as is and hence avoid the computational overheads of adding noise or redundancy into the data.

From an urban research perspective, the names and addresses of any individual are typically inconsequential. It is the statistical patterns over geospatial regions that are important to understand and analyze. Hitherto most researchers have considered geospatial regions as centroids or polygons of various forms within which behavioural patterns occur, e.g. the commuters living with a circle of radius 1km from a particular train station, and undertake their analysis based upon direct (Euclidean) geometries associated with such geometries. However such approaches are naïve since they ignore the true topological information given by the local context, i.e. using the road/street networks for the distance measurements. A more accurate specification is given by those individuals that live within 1km walking distance of the train station, where the distance is the smallest summation of the individual street network segments, i.e. the shortest route from where they live to the train station. Thus the Euclidean (as the crow flies) distance might be 100m, but this may be across a river or where no associated roads exist connecting the two points. A further complicating factor in this regard is that the street network itself evolves: new roads are built; houses created/destroyed; the population itself also changes.

Privacy preserving algorithms should ideally utilize the most accurate road network information. The most accurate street network data for Australia is from the commercially available Public Sector

Mapping Agency (PSMA – www.psma.com.au). This is one of the data sets accessible and available through the AURIN platform.

III. AURIN URBAN DATA PRIVACY REQUIREMENTS AND SCENARIOS

Dynamic aggregation of *unit-level* data collected through surveys into sub-regions of a city to identify spatial variability has not been given much attention due in part to the issues and challenges in releasing the unit level data directly to perform such analyses. Instead it is predominantly the case that unit-level survey results are statically aggregated by agencies and made available to give a broader population perspective by agencies. As a typical example of this, in 2011 the Department of Health in Victoria (VicHealth) undertook a survey of 25,000+ Victorians to assess their overall health, wellbeing amongst a range of other factors. The questions covered health and wellbeing, community and social issues, work-life balance, discrimination and prevention of violence against women, drinking and smoking habits and basic demographics. Typical interview questions included:

- *On an average day, how long does it take you to commute TO and FROM <work>/<your mutual obligation/work for the dole placement>?*
- *How often does a lack of time keep you from spending the time you would like with family and friends?*
- *In the last 7 days, how much money did you spend on alcoholic drinks in bottles, cans or other containers, to take-away?*
- *In the last 7 days, how much money did you spend on alcoholic drinks, which you purchased and drank in a restaurant, pub or club?*

As seen, some of these questions have responses given as categorical variables, e.g. yes or no; others are given as numeric values, e.g. time (hours/minutes) or particular monetary ($) values, whilst other responses to survey questions result in a Likert scale, e.g. *Always*, *Often*, *Sometimes*, *Rarely*, *Never*. VicHealth aggregates the results of these questions to particular geospatial regions – typically local government areas (LGAs) or statistical local areas (SLAs) as shown in the spreadsheet in Figure 1. It is important to note that the survey also includes information on the names, addresses and other information of the respondents, however for obvious privacy reasons this is not included in the aggregated results that are released for research purposes or as indicators of the health and wellbeing of the regions of Victorians.

	B	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1		Alcohol															Work, life and time				
2		Purchased alcohol in the last 7 days					7 day $ spend on packaged liquor (of those purchasing)					7 day $ spend at a licensed premises (of those purchasing)					Long commute (≥2 hours per day)				
3		ww.vichealth.vic.gov.au/~/media/81AA1D74C78440F7B920A0A92A27E					w.vichealth.vic.gov.au/~/media/81AA1D74C78440F7B920A0A92A27					vichealth.vic.gov.au/~/media/81AA1D74C78440F7B920A0A92A2						vichealth.vic.gov.au/~/media/902312DE9AC9490EA8EC48F07B6				
4		numeric					numeric					numeric					numeric				
5	LGA_Name		CI low	CI high	significance	Victorian average		CI low	CI high	significance	Victorian average		CI low	CI high	significance	Victorian average		CI low	CI high	significance	Victorian ave
6	Alpine	43.3	36.2	50.6	none	36.3	35	29	42	better	45	52	12	91	caution	45	8.3	3.9	16.7	caution	11.6
7	Ararat	33.3	25.9	41.6	none	36.3	35	25	44	none	45	27	14	41	caution	45		1.6	12.5	don't report	11.6
8	Ballarat	40.0	37.0	43.2	none	36.3	36	33	40	better	45	35	26	43	none	45	4.7	3.3	6.8	better	11.6
9	Banyule	43.1	36.2	50.2	none	36.3	48	37	59	none	45	40	29	50	none	45	20.6	14.2	28.8	worse	11.6
10	Bass Coast	33.6	27.1	40.9	none	36.3	35	28	42	better	45	25	8	43	caution	45	6.8	3.1	14.2	caution	11.6
11	Baw Baw	34.3	27.4	41.9	none	36.3	37	28	47	none	45	24	17	31	better	45	10.8	6.0	18.6	caution	11.6
12	Bayside	46.5	39.1	53.9	worse	36.3	50	35	65	none	45	47	25	69	none	45	14.1	8.3	22.9	caution	11.6
13	Benalla	31.7	25.8	38.2	none	36.3	36	29	43	none	45	25	15	36	better	45		0.6	5.5	don't report	11.6
14	Boroondara	44.7	37.3	52.3	none	36.3	53	41	65	none	45	43	32	53	none	45	5.5	2.5	11.6	caution	11.6
15	Brimbank	24.3	19.9	29.3	better	36.3	37	29	45	none	45	52	23	82	caution	45	16.7	11.2	24.2	none	11.6
16	Buloke	29.6	22.6	37.7	none	36.3	59	26	92	caution	45	54	15	94	caution	45		0.1	5.4	don't report	11.6
17	Campaspe	34.6	28.2	41.6	none	36.3	46	33	59	none	45	34	19	48	none	45	2.5	1.0	6.1	caution	11.6
18	Cardinia	32.9	25.9	40.9	none	36.3	40	31	50	none	45	31	18	44	none	45	16.7	9.9	26.8	caution	11.6
19	Casey	29.7	25.1	34.8	better	36.3	43	27	60	none	45	91	35	147	caution	45	16.6	12.1	22.3	none	11.6
20	Central Goldfields	32.2	26.1	38.9	none	36.3	34	27	42	better	45	36	11	62	caution	45	9.3	4.1	20.1	caution	11.6
21	Colac-Otway	48.7	41.3	56.1	worse	36.3	38	29	47	none	45	36	21	52	none	45		1.1	8.6	don't report	11.6
22	Corangamite	31.8	25.0	39.6	none	36.3	41	29	53	none	45	41	22	60	none	45	4.2	1.6	10.6	caution	11.6
23	Darebin	35.1	28.1	42.7	none	36.3	48	33	64	none	45	33	21	46	none	45	14.8	8.9	23.6	caution	11.6
24	East Gippsland	38.6	31.7	46.0	none	36.3	48	27	68	none	45	37	23	51	none	45	7.5	3.0	17.9	caution	11.6
25	Frankston	40.8	33.5	48.6	none	36.3	49	28	70	none	45	45	8	81	caution	45	22.0	14.1	32.6	worse	11.6
26	Gannawarra	37.9	31.0	45.2	none	36.3	36	28	43	none	45	24	17	30	better	45	0.0	.	.	better	11.6
27	Glen Eira	38.4	31.4	45.8	none	36.3	50	37	64	none	45	43	16	71	caution	45	9.5	5.0	17.4	caution	11.6
28	Glenelg	38.8	31.8	46.4	none	36.3	37	29	44	none	45	23	17	28	better	45		0.2	3.1	don't report	11.6
29	Golden Plains	42.0	35.7	48.5	none	36.3	35	29	41	better	45	20	14	26	better	45	6.3	3.1	12.3	caution	11.6
30	Greater Bendigo	35.9	29.6	42.7	none	36.3	36	30	41	better	45	27	18	36	better	45		0.6	4.5	don't report	11.6
31	Greater Dandenong	28.3	21.5	36.1	none	36.3	39	30	48	none	45	36	9	62	caution	45	10.3	5.7	17.7	caution	11.6
32	Greater Geelong	41.3	36.0	46.8	none	36.3	45	37	53	none	45	39	27	52	none	45	9.8	6.2	15.0	none	11.6
33	Greater Shepparton	28.3	22.2	35.3	better	36.3	37	30	44	none	45	31	13	48	caution	45	2.4	1.0	5.8	caution	11.6
34	Hepburn	35.1	28.6	42.2	none	36.3	48	22	75	caution	45	31	20	43	none	45	8.5	4.2	16.2	caution	11.6
35	Hindmarsh	27.2	20.6	34.9	better	36.3	24	18	30	better	45	39	7	72	caution	45	8.0	3.0	19.7	caution	11.6
36	Hobsons Bay	38.2	31.2	45.6	none	36.3	43	33	54	none	45	40	18	63	caution	45	14.0	8.4	22.6	caution	11.6
37	Horsham	38.7	31.8	46.1	none	36.3	37	31	43	none	45	16	12	20	better	45		0.7	8.1	don't report	11.6
38	Hume	29.1	23.1	35.9	none	36.3	47	34	60	none	45	51	30	72	none	45	8.4	4.7	14.6	caution	11.6
39	Indigo	39.3	33.1	45.9	none	36.3	44	33	55	none	45	29	19	39	better	45		0.2	3.9	don't report	11.6
40	Kingston	36.4	29.7	43.6	none	36.3	37	27	48	none	45	27	16	38	better	45	14.1	9.0	21.2	none	11.6
41	Knox	32.0	25.1	39.7	none	36.3	40	29	52	none	45	28	19	36	better	45	14.2	8.9	21.9	none	11.6
42	Latrobe	37.2	30.3	44.6	none	36.3	48	32	65	none	45	75	22	128	caution	45		0.6	4.6	don't report	11.6
43	Loddon	36.3	27.6	46.0	none	36.3	38	20	56	none	45	34	11	56	caution	45		0.7	7.6	don't report	11.6
44	Macedon Ranges	35.9	29.6	42.7	none	36.3	42	31	53	none	45	30	20	39	better	45	21.0	15.1	28.5	worse	11.6
45	Manningham	32.0	25.5	39.4	none	36.3	52	30	73	none	45	58	35	82	none	45	18.8	12.0	28.1	none	11.6
46	Mansfield	44.7	38.0	51.7	worse	36.3	37	31	42	better	45	26	20	31	better	45	5.7	2.6	12.1	caution	11.6
47	Maribyrnong	40.3	33.1	47.8	none	36.3	38	32	45	none	45	36	23	49	none	45	11.8	6.5	20.5	caution	11.6
48	Maroondah	38.2	31.3	45.6	none	36.3	36	28	44	none	45	19	14	24	better	45	10.5	5.9	17.8	caution	11.6
49	Melbourne	45.0	36.8	53.5	none	36.3	53	34	73	none	45	42	32	51	none	45	4.8	1.9	11.7	caution	11.6
50	Melton	29.1	22.9	36.2	none	36.3	28	23	33	better	45	52	11	93	caution	45	21.9	15.1	30.7	worse	11.6
51	Mildura	37.6	32.3	43.1	none	36.3	50	38	62	none	45	46	28	65	none	45		0.5	4.3	don't report	11.6

Figure 1: Aggregated Data from VicHealth Survey Data (2011)

These aggregated results can be used to show the big picture of health and wellbeing across the various regions (LGAs/SLAs) of Victoria through the AURIN platform as typified by Figure 2. This shows the VicHealth variables related to Lack of Time with Family (indicated by the colour-coded choropleth map) and Long Commute (>2 hours) (indicated by the size of the centroids). Combinations and exploration of the relationships between the data sets is supported through the AURIN platform. An example of this is shown in Figure 2 through the scatterplot. This shows the correlation between these two variables for the given regions (LGAs/SLAs) of Victoria. As might be expected, regions with increased numbers of individuals with an increased commuting time correlates with regions with individuals that have decreased time to spend with their family, i.e. the higher commuting time, the less time for friends and family. Each point in the scatterplot of Figure 2 represents a region of Victoria. Figure 2 also shows the actual (aggregated) data from the survey data itself. This includes the actual data and confidence levels associated with the data. A rich range of metadata is also available that describes the context of the data as was collected by VicHealth and the associated license under which this data was released. All of the data sets can be exported and saved outside of the AURIN portal, e.g. a CSV files, JSON or where geospatial data is included, as Shapefiles.

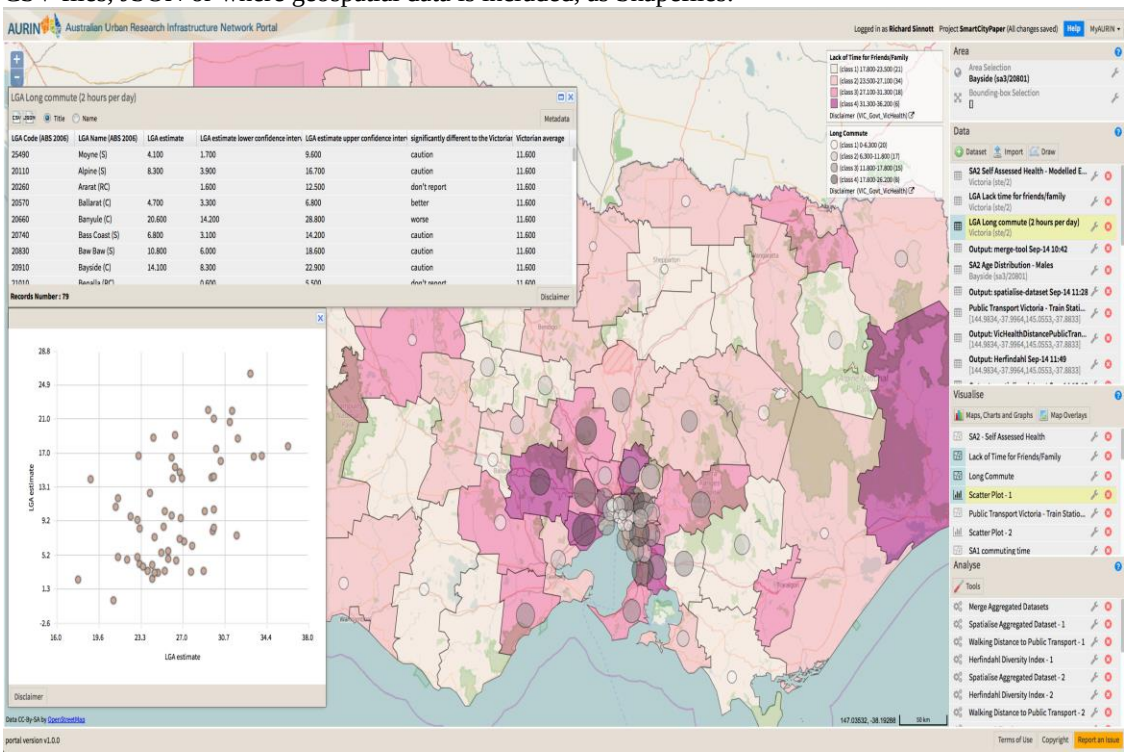

Figure 2: Lack of Time for Friends/Family and Long Commute Aggregated Data

Such *statically* aggregated indicators on health and wellbeing are typically used to understand the broader population health, however it is often the localized behavior that is the most important to understand. How does commuting times relate with the distance of the commuters in the survey from their nearest public transport possibility? How does engagement in physical activity relate to access to local parks or sports facilities? How does consumption of alcohol correlate with the distance to the nearest premise licensed to sell alcohol? How does feeling of safety at night correlate with availability of street lighting? Answering such questions raises challenges regarding privacy and information governance that must be tackled, since the precise location (address) of respondents in the survey is highly confidential. However the location of other information, e.g. train stations, parks, streetlights and bars/alcohol selling premises is not restricted information and is data that is directly accessible from data providers within the AURIN platform. To tackle this, novel algorithms were developed that allowed access to and use of unit level VicHealth data and their use in combination with other data to explain the patterns in a given localized context. Importantly, the algorithms have to support *dynamic* scenarios where pre-established answers cannot be given, but have to be calculated *on–the-fly* using a rich variety of other non-privacy demanding data sets. The answers to such questions must also be able to work at a variety of aggregation levels of interest to the researchers themselves.

IV. Architectural Considerations for Data Privacy

Addressing such access and use scenarios also poses several challenges to data providers such as VicHealth that ultimately must be addressed by any proposed technical solution. Firstly, such organizations are extremely wary of exposing their data/systems to external software systems, e.g. by opening their ports and the subsequent risks of direct incoming connections/queries from the Internet. For many organizations, their firewall is their security system and incoming connections from the Internet are of great concern. Secondly the dangers of disclosure risk control and information governance must be considered and reflect the organizations needs and demands, i.e. the potential to identify individuals from the survey is a major concern that must be addressed.

The AURIN system has included capabilities to avoid the possibility of compromises caused by exposing VicHealth data directly to the Internet as well an empowering them with regards to data disclosure risk control as shown in Figure 3. Figure 3 also shows the many other data clients allow direct that are available that allow access to remote services/data sets. In these cases again, security can be defined and enforced at varying levels. For example, data sets from commercial organizations such as PSMA are only accessible to academic collaborators, whilst industry/government-based users are prohibited access to these data sets. Data such as the Australian Tax Office Australian Business Registry have even further data access restrictions where access to the company data is granted on an individual-level only. The AURIN platform allows group and individual based permissions (blacklists and whitelists). In principle these could be used to provide direct access to unit level data, however the risk to VicHealth and the demands for privacy/confidentiality prohibit this model of security, since individuals outside of VicHealth are not allowed direct access to the confidential data.

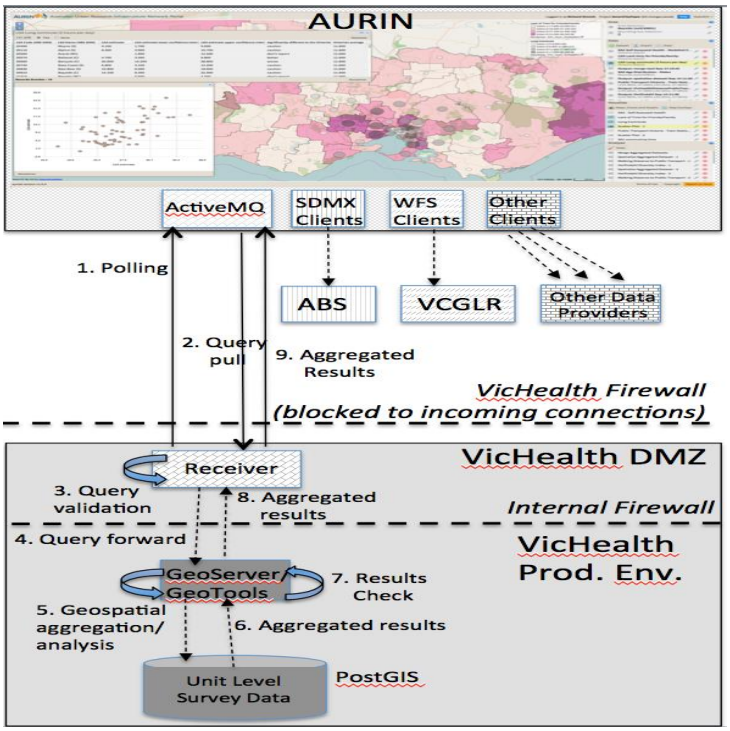

Figure 3: Architecture of AURIN VicHealth Data Linkage Solution

In the AURIN VicHealth model, the *unit-level data* from VicHealth is not directly accessible through an Internet-facing service – rather pull-based polling is used (step 1). In this model, digitally signed, queries are asynchronously pulled from secure, targeted AURIN service endpoints (supporting asynchronous interactions through ActiveMQ (http://activemq.apache.org) through services hosted in a virtualized demilitarized zone (DMZ) of VicHealth (step 2). These queries include the region of interest such as a particular situational context along with relevant data related to that area, e.g. bottleshop locations (accessible from other organisations within the AURIN platform). The service end-point receiver in the VicHealth DMZ checks the validity of requests (step 3) and if valid (step 4), forwards them on to services hosted on VicHealth services for analysis (step 5). The analytical routines run using

the unit level data and aggregated results are returned (step 6). These are validated to ensure that they meet any further limitations and/or constraints imposed by VicHealth (step 7), e.g. the number of individuals in that localized area is above a given pre-agreed data disclosure risk value. Finally, the aggregated results using the unit-level location information are released to the receiver (step 8) and subsequently on to the service end-point (step 9). At no point is the identifying data of the individuals released - only aggregated statistics are released. Importantly however, these aggregated statistics use the location of the VicHealth survey respondents with other public data to support analyses that would have been impossible otherwise.

In realizing this, a range of algorithms and services has been implemented that utilize the actual location (address) of survey respondents. These include standard clustering measures, basic statistics and a range of targeted distances measures, e.g. distance from the location (address) of the individual to another set of point-based locations. The pseudo-algorithm to traverse the street network and identify the actual distance from the location of interest (such as the train station) and the address of the individual represent the geo-privacy preserving solution are shown in Algorithm 1. As seen the algorithm calculates a network neighbourhood through traversing the road network. In this model the PSMA road network data is used as the street network. In this model, if there are insufficient numbers of respondents in a given sub-region (currently agreed with VicHealth to be three), then zero results are returned.

```
READ origins[], destinations[], lines_in_roadNetwork[]
map_graph=generateMap(origins, destinations, lines_in_roadNetwork)
FOR origin IN origins
   shortest_distance = maximum
   FOR destination IN destinations
      distance = length(map_graph, origin, destination)
      IF distance < shortest_distance
         shortest_distance = distance
      END IF
   END FOR
   result[i] = shortest_distance
END FOR
IF result.size > 3  WRITE avg = average(result)
ELSE WRITE avg = 0
END IF
```

Algorithm 1: Average Distance Calculation from a Set of Points to another Set of Points for a Set of Regions

A representative example showing the resultant street network is shown in Figure 4, where blue circles represent an arbitrary (representative) location and red dots the destinations of interest. The algorithm traverses the road network (PSMA) data and calculates the average distance between these the set of points (the VicHealth respondent locations) and a set of known locations, e.g. train stations or places to purchase alcohol in the local vicinity.

Figure 4: Network Traversal and Distance Measure Calculation Example

These algorithms can be invoked through a range of client side tools that are available within the AURIN portal. These include walkability measures, Herfindahl indexes as well as a range of basic statistics. The client side user interface to these tools is shown in Figure 5, where the walking distance from places to purchase alcohol is shown with a spatialised region data set as input (SA2).

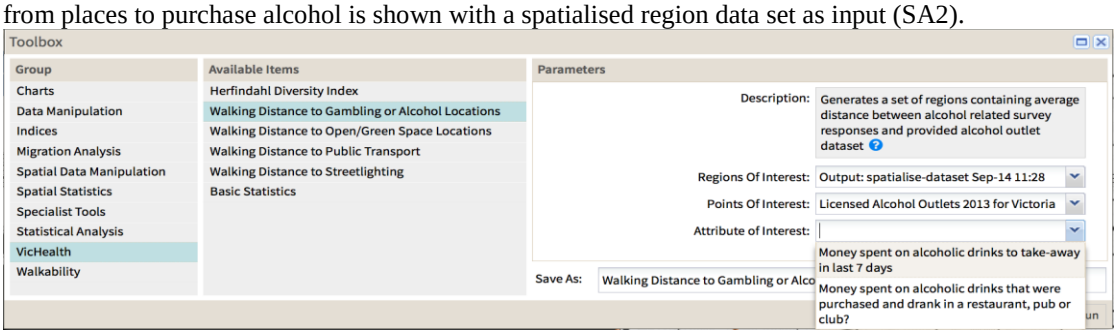

Figure 5: Security-oriented Portal Tools for VicHealth Unit-Level Data Usage

V. Realisation of Privacy-oriented Urban Case Studies

To understand the capabilities offered through these tools with VicHealth unit level data we present a range of case studies.

A. Distribution of VicHealth Respondents

In undertaking localized statistics, it is often necessary to understand the number and distribution of individuals that have completed the survey within a given region. Figure 5 shows the distribution of the number of individuals within a given Statistical Area (SA3 - Bayside) that completed the VicHealth survey – this is one of many SA3s across Victoria. In this case it is possible to consider less aggregated regions (SA2/SA1) as the sub-regions of interest. In this case SA2s are demonstrated.

As seen, each sub-region (SA2) within the Bayside SA3 has a varying number of responders to the survey (between 35-66). The variability is shown in Figure 6 as a choropleth map with darker shades representing regions with higher number of respondents. As noted previously in algorithm 1, if there were fewer than three respondents in a given area, e.g. due to a smaller region of interest or it was a more sparsely inhabited area, then the results would be suppressed completely. All results are shown here since there are at least 35 respondents in each area. On average, the VicHealth survey included around 300 individuals for each SLA across Victoria

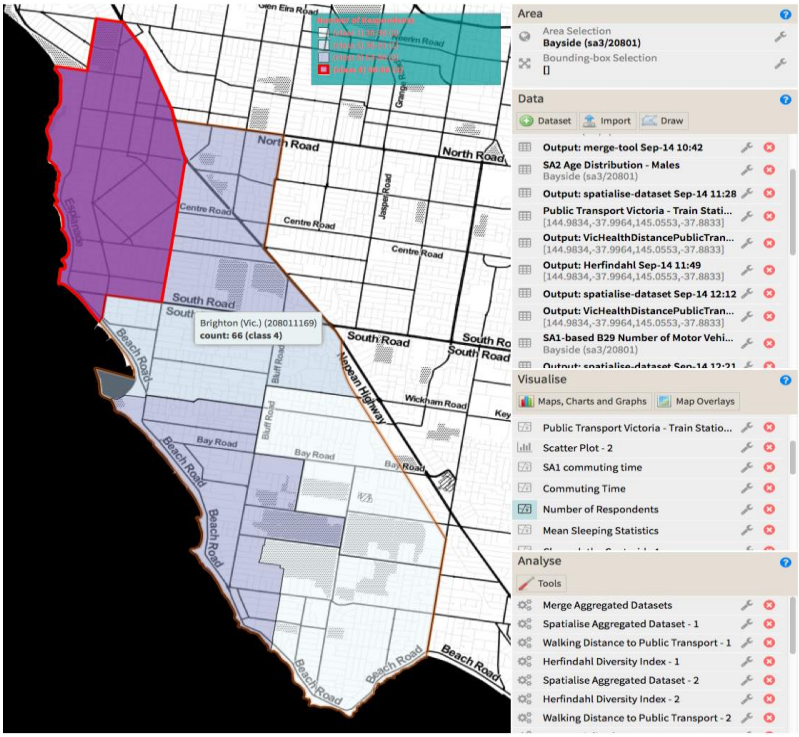

Figure 6: VicHealth Distribution of Respondents for SA3 Bayside

In addition to the distribution of respondents to the survey, localized decomposition of behaviour can be explained by the responses of individuals to the survey. To highlight this we consider several scenarios that reflect the typical use of geospatial privacy supported through AURIN.

B. Commuting Patterns of Baysiders

Firstly we consider the impact of commuting patterns on location of train stations within the Bayside area of Melbourne. To support this scenario, the location of train stations from Public Transport Victoria and shown as solid circles in Figure 7 and compared (measured) with the commuting times for survey respondents living within the SA2.

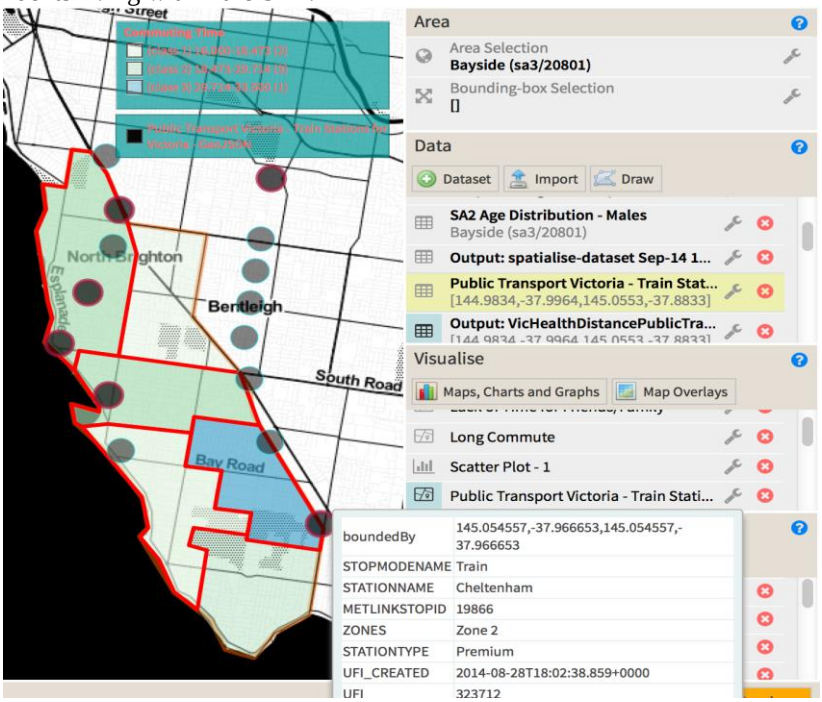

Figure 7: Commuting Patterns and Relation to Train Stations

The results of this analysis indicate that the commuting times for the VicHealth respondents vary across the SA2s from between 18mins-33.5mins (as shown in Figure 8). This variance is indicated in Figure 7 through the colour coding of the choropleth map with darker colours indicated longer commuting times. This localized analysis can be used to determine the potential future train stations and/or bus stops that might supplement areas where commuting times are excessive.

Output: VicHealthDistancePublicTransport Sep-14 11:31

CSV JSON SHP Title Name Metadata

geometry	null	count	feature_id	distance	c6mins	correlation
		35	upload-13543...	1645.971	29.714	0.637
		66	upload-13543...	621.788	29.712	0.657
		55	upload-13543...	936.800	18.473	0.543
		36	upload-13543...	753.861	33.500	0.660
		53	upload-13543...	817.283	27.981	0.431

Records Number : 6

Figure 8: Commuting Times for Bayside VicHealth Correspondents

It is important to note that the times are calculated using the actual locations of the correspondents, but without ever revealing their location. This model of access to and use of the unit-level data does not require any specific needs and requirements (ethics or permissions).

C. Alcohol Consumption Correlation with Availability of Bottleshops

Finer-grained analysis is also possible, e.g. below a given SA2 level of aggregation. Figure 9 shows the official places licensed to sell alcohol from the Victorian Commission for Gambling and Liquor Regulation (VCGLR – www.vcglr.vic.gov.au) for Bayside (shown as yellow centroids). Correlating these with VicHealth respondents and the amount of money they spend on alcohol is a typical urban research question, i.e. does availability of places to purchase alcohol encourage individuals to spend more on alcohol. In this scenario however the intention is support a much finer-grained aggregation level (down to the SA1-level which is typically of the order of a block of houses). In this case, many (most) of the SA1s that were identified in that particular SA3 for Bayside do not contain sufficient numbers of respondents from the VicHealth survey (at least 3), hence the results are set to null (as indicated by the darker gray polygons). The coloured polygons indicate those SA1 regions where the amount of money spent on alcohol increases, i.e. darker polygons indicate an increased amount of money that is spent on alcohol within that SA1.

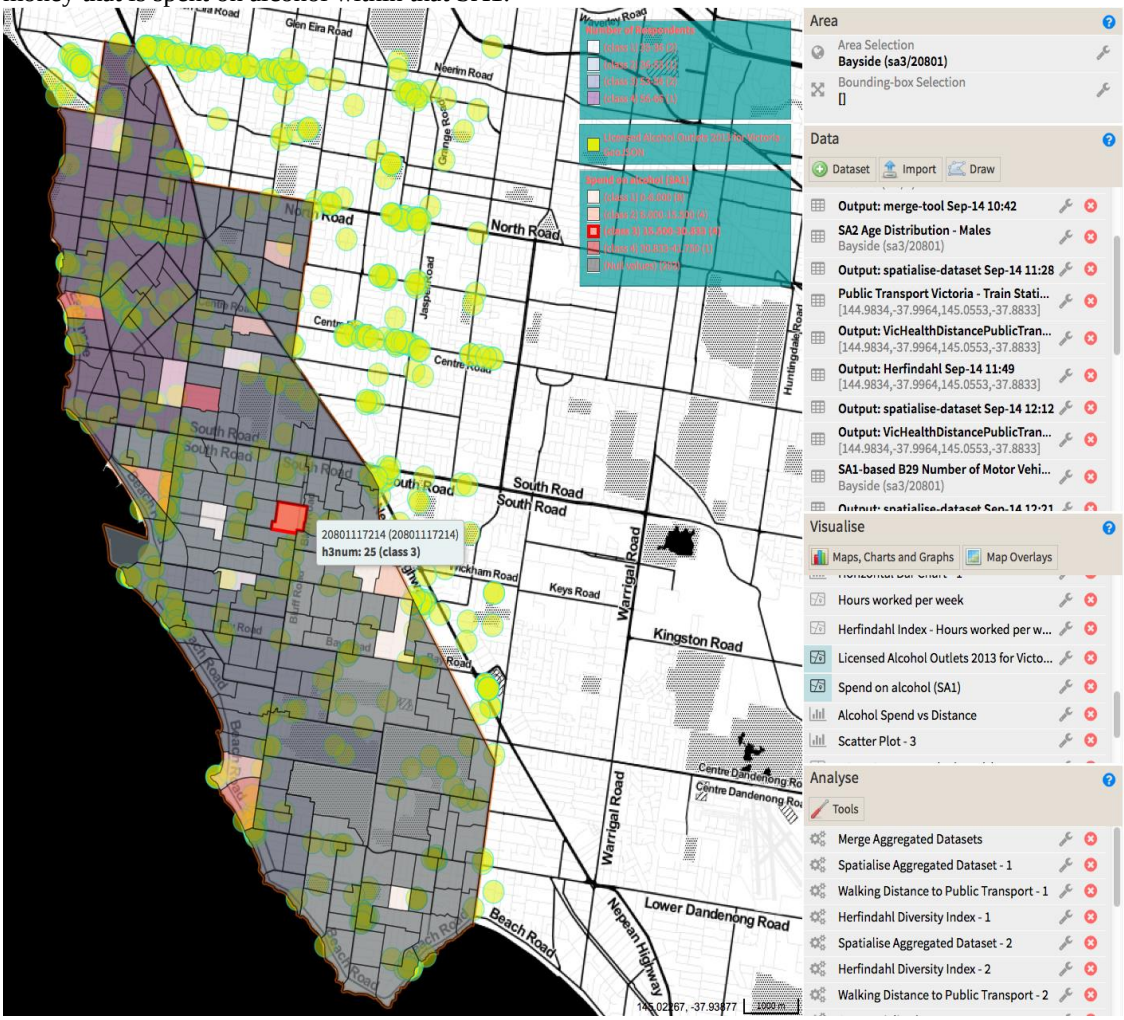

Figure 9: SA1-based Alcohol Consumption Patterns for Bayside

This information can be used to determine a range of alcohol related issues that may occur at a local level, e.g. correlating this kind of data with crime or socio-economic challenges is possible through the AURIN platform. One key factor to consider here is whether there is in fact a correlation between alcohol consumption and average distance to places to buy alcohol for those individuals that live within the SA1s in Figure 9. This is shown in Figure 10. In this case 220 SA1s exist within the SA3 Bayside, of which 13 have sufficient numbers of respondents in the survey where values and patterns can be observed. As seen one outlier exists within this correlation that shows most money is spent on alcohol ($41.75) even though the average distance is furthest away (558m). The weekly alcohol spends vary between $2.50-$41.75. This localized analysis augments the big picture statistics that is typically

released by VicHealth across all of Victoria at the SLA/LGA levels. Essentially, the scenarios are fully automated and results returned directly to the AURIN end users.

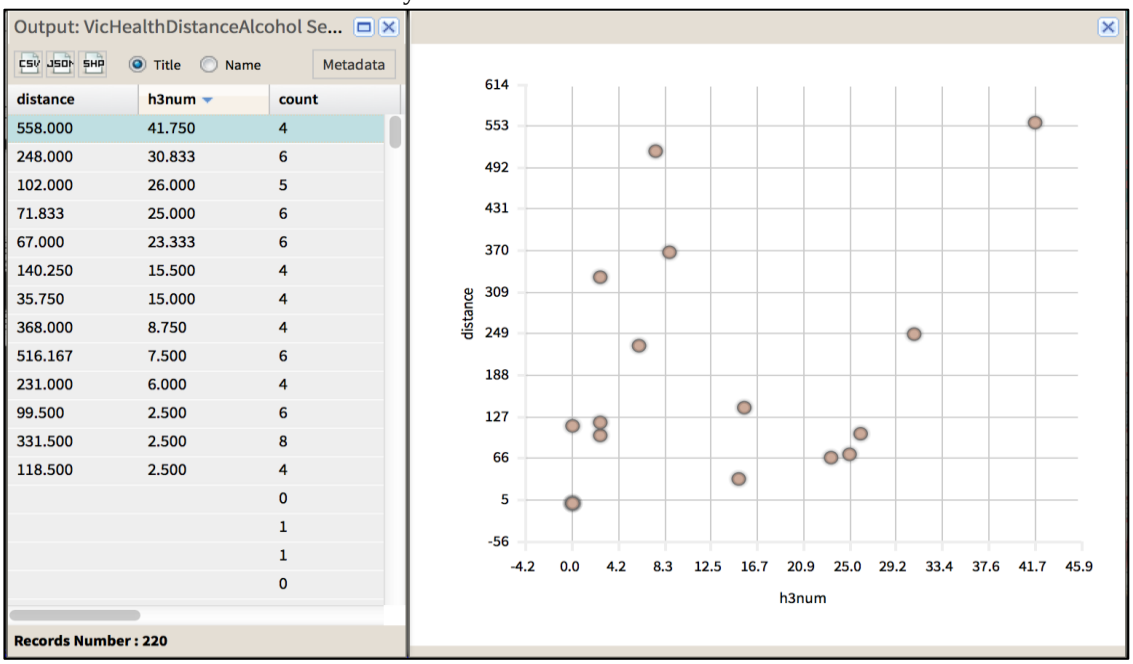

Figure 10: Correlation between Alcohol Consumption and Places to Purchase Alcohol in Bayside

VI. CONCLUSIONS

The integration of unit record data collected through survey research with spatial information is essential to understand localized behaviours and in turn to tackle the challenges facing cities. AURIN has developed novel privacy-preserving approaches that allow direct and automated use of sensitive data whilst preserving the privacy of the location of the respondents. Importantly this data can be used with other data sets that exist external to the likes of VicHealth and thus provides a big data environment with extensive analytical capabilities. These systems have been developed through close collaboration with VicHealth and their technical staff. The system itself depends on degrees of trust. One major aspect of trust is the ability to deploy systems within the demilitarized zone (DMZ) of VicHealth. These systems provide indirect access to the unit level data, i.e. the unit level data is accessible through services hosted within the DMZ.

There are many other sources of unit-level based data. One prime resource is social media and resources such as Twitter, where individual tweets can include explicit locations (latitude/longitude) of tweeters when location based services are activated on the mobile devices used for tweeting. Exploring the real-time behaviour of the road transport system using social media is explored in [19] and associated issues of the privacy of the trajectories of tweeters explored in [20]. It is expected that twitter will become a key data resource supporting a range of real-time behaviour scenarios facing the urban transport system of Australia and indeed for real-time health monitoring activities as might be required in pandemic situations [21].

We note that the AURIN e-Infrastructure is very much a supporting activity. That is, the work in the e-Infrastructure development is not targeted at delivering novel IT solutions per se, but on supporting the urban research community in *their* research needs. Several thousand researchers are actively using the portal and associated e-Infrastructure. Since the project commenced, the AURIN platform has been accessed over 50,000 times by researchers across Australia (data provided by the AAF). The fastest growing community of users is from non-academic sources with Government and industry increasingly interested in the capabilities for big data analytics and security offered through the AURIN platform.

This geospatial privacy preserving solution will be explored in the future with VicHealth where a new survey is currently underway. Furthermore new technologies are being developed to make more use of the data. Mobile applications allowing analysis of the data are one mechanism and AURIN has developed an openAPI that allows access to data outside of the AAF protected portal/web-based environment. This openAPI utilizes geospatial services for geospatial data access as well as web processing services that support invocation of tools and services.

ACKNOWLEDGMENTS

The AURIN project is funded by the Department of Education. We gratefully acknowledge their support. We also thank the AURIN technical team for the inputs and support throughout the project. We are hugely grateful to VicHealth for their ongoing support and trust with special thanks to Felix Acker and the technical staff at VicHealth.

REFERENCES

[1] AURIN Final Project Plan, http://aurin.org.au/resources/final-project-plan

[2] R. Stimson, et al, *The Australian Urban Research Infrastructure Network (AURIN) Initiative*, State of Australian Cities, Melbourne, Australia, November 2011.

[3] R.O. Sinnott, et al, *A Data-driven Urban Research Environment for Australia*, IEEE e-Science Conference, Chicago USA, October 2012.

[4] R.O. Sinnott, et al, *The Australian Urban Research Gateway*, Journal of Concurrency and Computation: Practice and Experience, April 2014, doi: 10.1002/cpe.3282.

[5] R.O. Sinnott, et al, *The Urban Data Re-use and Integration Platform for Australia: Design, Realisation and Case Studies*, IEEE International Conference on Information Re-use and Integration, San Francisco, USA, August 2015.

[6] R.O. Sinnott, W. Voorsluys, *A Scalable Cloud-based System for Data-intensive Spatial Analysis*, Journal of Software Tools for Technology Transfer, Springer Verlag, May 2015.

[7] S.M. Randall, et al. *Privacy-preserving record linkage on large real world datasets*. Journal of biomedical informatics 50 (2014): 205-212.

[8] F. Ritchie, *Secure access to confidential microdata: four years of the Virtual Microdata Laboratory*, Economic and Labour Market Review 2, no. 5 (2008): 29-34.

[9] M. Smith et al. *Big Data Privacy Issues In Public Social Media*. 2012 6th IEEE International Conference on Digital Ecosystems and Technologies.

[10] M. Mendoza, et al. *Twitter Under Crisis: Can we trust what we RT*?. In Proceedings of the first workshop on social media analytics, pp. 71-79. ACM, 2010.

[11] W. Itani, et al. *Privacy as a service: Privacy-aware data storage and processing in cloud computing architectures*. In Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on, pp. 711-716. IEEE, 2009.

[12] Y.A. de Montjoye, et al. *Unique in the Crowd: The privacy bounds of human mobility*. Scientific reports 3 (2013).

[13] L. Sweeney, *K-anonymity: A model for protecting privacy*, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[14] H. Hu, et al, *Privacy-aware location data publishing*, ACM Transactions on Database Systems, vol. 35, no. 3, p.17, 2010.

[15] M. Xue, et al, *Location diversity: Enhanced privacy protection in location based services*, in LoCA, ser. LNCS, vol. 5561. Springer, 2009, pp. 70–87.

[16] Y. Xiao, et al. *Differentially private data release through multidimensional partitioning*". In Secure Data Management, pages 150-168, 2010.

[17] C. Dwork, et al, *Calibrating noise to sensitivity in private data analysis*, in Proceedings of 3rd Theory of Cryptography Conference, New York, USA, Mar. 2006, pp. 265-284.

[18] M. E. Andrés, et al, *Geo-indistinguishability: Differential Privacy for Location-based Systems*, in Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security. ACM, 2013, pp. 901-914.

[19] Y. Gong, et al, *Identification of (near) Real-time Traffic Congestion in the Cities of Australia through Twitter*, Understanding the City with Urban Informatics, CIKM 2015, Melbourne, Australia, October 2015.

[20] S. Wang, et al, *Follow-Me-Not: Protecting the Trajectory Privacy of Social Media Users*, submitted to Journal of Social Network Analysis and Mining, November 2015

[21] J.P. Zaldumbide, et al, *Identification and Verification of Real-Time Health Events through Social Media*, International Conference on Data Science and Data Intensive Systems, Sydney, Australia, December 2015.

[22] V. Welch, J. Barlow, J. Basney, D. Marcusiu, N. Wilkins-Diehr, N., 2007. A AAAA model to support science gateways with community accounts. Concurrency and Computation: Practice and Experience, 19(6), pp.893-904.

[23] T. Scavo, V. Welch, 2007. A grid authorization model for science gateways. In International Workshop on Grid Computing Environments (No. 3).