University Library

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Sinnott, RO;Voorsluys, W

Title:
A scalable Cloud-based system for data-intensive spatial analysis

Date:
2016-11-01

Citation:
Sinnott, R. O. & Voorsluys, W. (2016). A scalable Cloud-based system for data-intensive spatial analysis. International Journal on Software Tools for Technology Transfer, 18 (6), pp.587-605. https://doi.org/10.1007/s10009-015-0398-6.

Persistent Link:
https://hdl.handle.net/11343/56969

A Scalable Cloud-based System for Data-intensive Spatial Analysis

R.O. Sinnott, W. Voorsluys
Department of Computing and Information Systems,
University of Melbourne, Melbourne, Australia

Advances in Cloud computing technology and the availability of affordable and easy to use Cloud services are enabling a multitude of scientific applications to use these resources as primary or secondary computing infrastructure. The urban and built environment research domain is one area that can benefit greatly from Cloud computing. The global population growth and increase in the size and population of cities, raises many challenges for governments, planners and researchers alike. The Australian Urban Research Infrastructure Network (AURIN – www.aurin.org.au) project has been tasked with developing an advanced platform (e-Infrastructure) across Australia to tackle these challenges. The platform leverages large-scale Cloud resources to provide federated data access to, at present over 1100 data sets from major and often definitive government and industry data-rich organisations, and for scalable data processing and visualisation. The original AURIN tools were developed using the Object Modelling System (OMS) and supported integrated workflows to define and enact/re-enact scientific processes. More recently the work has evolved to focus more on delivery of a workbench offering a rich range of tools delivered through an extensible workflow environment. In this paper, we provide the background to AURIN including the scientific drivers that are shaping the work and the realisation of the Cloud-based AURIN environment. We focus in particular on the workflow environment and show how it seamlessly utilizes the Cloud for urban research processes focused especially on data-intensive spatial analysis. We illustrate the utilisation of this workflow environment across a range of case studies reflecting urban research activities.

Keywords— *e-Infrastructure, urban research, Cloud computing, geospatial systems, spatial analysis.*

1. INTRODUCTION

A staggering amount of data across all walks of life is now being generated: social, scientific, by government, by industry and across all research endeavours. This is often referred to as the data deluge [1]. In order to process this ever-growing amount of data, large-scale e-Infrastructures are essential – often realised as data centres. Historically these have been based upon cluster-based technologies to support high performance computing at a given centralised location, however more recently Cloud computing has been the predominant technology for big data processing and analysis [2]. This model is based upon the premise that the data is brought to the location where it can be processed/analysed. However more often than not, big data is highly distributed, completely heterogeneous and evolving at a rapid pace. To overcome this, and to facilitate the sharing of heterogeneous data supporting real-time collaborative tasks between geographically distributed science teams, many organisations leverage a range of data access and harvesting solutions. Service-oriented architectures (SoA) comprised of distributed web services offer an ideal model for federated data access and delivery [3].

Cloud computing-enabled web service frameworks enable potentially thousands of computers to be utilised to share data, applications and computing capacity to achieve a desirable outcome in a manner transparent to the end-user who ideally only interacts with a single entity [4]. The implementation of such infrastructures must meet the growing demands of domain-specific needs: accessibility and interoperability as well as support a multitude of research endeavours more generally [5-7]. In a web service environment comprising distributed

(autonomous) organisations, a need exists for intelligent access and use of these distributed resources that allow simplifying the access to data and the scalable processing of this data. Scientific workflows accessing distributed services offered through a SoA represent one way in which such coordination of access to and use of data can be realised [8] although other models such as distributed query processing also exist.

The focus of this paper is to describe the technical solutions that have been established to support scalable use of Cloud resources in the urban research domain. Specifically this work has been undertaken to support the research needs of the AURIN project [9]. The $24m federally funded AURIN project supports nationwide urban and built environment research across Australia. The AURIN e-Infrastructure offers a flexible architecture that delivers a data-driven research environment offering secure access to distributed and highly heterogeneous data sets from numerous government, industry and academic organisations across Australia. This is delivered through a web-based research environment. Key to this environment is the extensive portfolio of tools for data analytics. We describe the domain drivers and challenges of AURIN and the technical solutions that have been developed to meet these challenges. We focus in particular on the workflow environment and how it has been developed to support scalable data intensive analyses.

The remainder of the paper is organized as follows. In section 2 we summarise the domain drivers that have shaped the AURIN efforts focusing in particular on urban research challenges, the need for workflows and the role of the Cloud. In section 3 we present the AURIN e-Infrastructure and describe key capabilities that it offers to deliver seamless, secure access to large-scale distributed and heterogeneous data sets with integrated data analysis capabilities completely delivered through an integrated workflow environment. Section 4 describes this workflow architecture and its implementation. Section 5 presents a range of case studies on the application of the workflow environment for large-scale data analytics problems representative of those facing urban researchers. Finally in section 6 we give concluding remarks and identify areas of future work.

2. URBAN RESEARCH CHALLENGES

It is common knowledge that the global population is increasing. It is perhaps less widely known that increasing numbers of people are moving to live in urban areas – this is especially the case in Australia. According to the Australian Bureau of Statistics (ABS – www.abs.gov.au), Australia's population at 30 June 2012 was 22.7 million. However it is projected to increase to between 36.8 million and 48.3 million in 2061, and reach between 42.4 million and 70.1 million in 2101. The vast majority of these individuals will reside in cities (http://www.abs.gov.au/ausstats/abs@.nsf/mf/3222.0). Given this, it is unsurprising that there are significant challenges for the sustainable development of cities. Cities that exist today are typically not being designed/planned to cope with issues on the horizon such as the increase in the population and the challenges that this gives rise to for transport, health, housing amongst many other considerations.

Contemporaneously there has been a significant rise in computing capabilities and the compilation of many data sets across all aspects of urban activities and life: social media, transport related data, health related data, population and employment related data etc. There are increased needs and demands for information technologies to help make sense of these data sets; ideally to improve the current urban quality of life as well as help to plan for the future. Whilst there have been a range of activities related to improving *aspects* of urban environments, e.g. focused on transport or on health, it is the combination of multiple factors and data sets that are key since cities are complex, evolving entities. The impact of a growing population will directly impact on transport and on health and on multiple social science areas of concern. There thus exists a need for systems that allow such big data to be accessed and processed in a manner that reflects urban research needs.

AURIN is one such initiative. It is tasked with providing a security-driven, online platform allowing access to an extremely heterogeneous variety of big data sets from definitive (authoritative) sources. In developing this platform it was identified [10] that it needs to facilitate the integration of multiple diverse types and sources of data, and offer real-time interrogation of data using a sophisticated suite of statistical, spatial, modelling and visualisation tools. The vision of the AURIN e-Infrastructure is to provide a unified research environment for urban and built environment researchers across Australia. Whilst it is quite possible to develop a heterogeneous collection of (individual) data services and resources targeted to subsets of the urban research landscape, AURIN was tasked with a grander vision: a unified and integrated environment that could be used for a multitude of urban research endeavours through a single one-stop-shop.

Many of the challenges facing AURIN and indeed urban research itself can be classified as data-driven. Across Australia a huge array of organisations exist that hold data that is fundamental to supporting urban research and understanding the future challenges facing urban settlements. Whilst many of these data providers provide access to data on the Internet, e.g. the ABS makes extensive data available for direct download as zipped packages, typically as formatted, multi-sheet Excel Spreadsheets, this model of data delivery places major challenges for researchers needing to access and subsequently analyse data. There are tens of thousands of Spreadsheets on the ABS website and analysing particular data in a given situational context requires the user to locate, access and download the relevant data. This situation is magnified when juxtaposed with other national and State-wide government organisations and industry bodies holding data that must be combined to analyse urban phenomena. A multitude of organisations exist in Australia, that possess significant and important data sets that are essential to underpin urban research, but at present these data sets are hard to access, use, interpret and place many hurdles on researchers wishing to understand the bigger picture of urban settlements. For example, often these data providers have strict terms and conditions on the access to and use of their data resources. This can include limitations on download, further distribution, as well as limits on records that can be made accessible and to whom.

Overcoming the barriers caused by this diversity is at the heart of the AURIN e-Infrastructure. Urban researchers should be able to access diverse data sets as simply as possible. Key to this is the notion of *single sign-on* whereby users authenticate once and are subsequently able to access distributed resources without further challenge/response demands. Following successful authentication, depending on their privileges they should be able to access diverse data sets and analyze them according to their research needs as if the data was available directly through the web site (portal) they are accessing. To deliver this requires that programmatic access to data is achieved, or more specifically federated access to the distributed databases and systems delivered through a targeted and heterogeneous SoA.

A major challenge in undertaking this effort is the variety and heterogeneity of the data itself. The era and hype associated with “big data” is a reality that AURIN must address to meet the demands of the urban research community. Whilst urban research data sets are not especially large when compared to certain disciplines, e.g. genomics or astrophysics, they do fall under the typical big data classification categories: *velocity* (they can/do change repeatedly, e.g. actual traffic data); *variety* (they come from many autonomous sources and are in the main, completely heterogeneous) and *veracity* (the original source of the information is often essential to establish). In this context, there is no common vocabulary or ontology that has been developed that is widely used across government and industry bodies of Australia. Indeed even within a given urban research domain such as transport, a multitude of challenges exists in understanding and relating data sets. As one representative example, household travel surveys are often used by government agencies to better understand commuting patterns of city populations. Typically these surveys are conducted over different time periods, at different levels of aggregation, have a range of different questions that are used to create a variety of different travel measures and indicators, and as a result are hard to compare.

One of the major obstacles to addressing this is the richness, complexity and domain knowledge associated with the data itself. Capturing detailed metadata is essential in this context, and essentially human readable metadata. Thus while a given survey may refer to variables used in a particular statistical analysis, the shorthand (abbreviated) notation often used to describe this information is typically cryptic and prohibits future data re-use. Systems that allow a richness of metadata to be captured and used in a manner that supports intelligent data access and usage are thus essential. It is noted that this situation is not solely related to survey-type data, but many other data sets and research disciplines.

Across Australian and indeed many countries, the geospatial landscape and land use is continually evolving. Cities are growing and their boundaries are being continually refined: local government authorities and suburbs are changing both in terms of their geographical footprints and their associated utilisation. A multitude of geospatial classification systems exist: those that have been standardized and widely accepted, e.g. the states of Australia (NSW, Victoria etc); others that are widespread and commonly used but continually evolve, whilst others still have been proposed by researchers themselves. The ability to discover and use data classified according to a particular geospatial classification in a given geographical context, and navigate across different geo-classification schemes is essential. Thus knowing what data sets exist for Melbourne might include a multitude of data geo-classifications in the same geographical context, e.g. the local government authorities of Melbourne, the statistical local authorities of Melbourne, the road network of Melbourne, the labour force regions of Melbourne, the polling booth catchment areas of Melbourne etc.

A further challenge facing urban researchers (and researchers more generally) is the socialisation of the skillsets associated with research and the repeatability of scientific endeavour. Urban researchers use a range of data sets and analytical tools and the common understanding and application of these is not ubiquitous. Different areas of expertise exist, but this is often difficult to share amongst different urban-related domains. AURIN is expected to address this phenomenon through delivering tools and data sets in a common research environment, and augmenting these with workflows that capture the research process and allow this to be repeated. A core element of this whole process is the visualisation and understanding of data [20]. Whilst some urban researchers have a background in and access to geo-visualisation tools and are familiar with the use of map data and tools for mapping/analysis geospatial data, the vast majority are not geospatial experts. Furthermore, many of the more advanced geo-visualisation tools are only commercially available. AURIN aims to provide an environment where the common spectrum of visualisation and analytical tools is made available through an open-source software environment. A key aspect here is that the systems need to scale. In the big data arena, the processing of data on a desktop machine as might be typical with a geographical information system (GIS) tool is often impossible. The volume, variety and velocity of data require larger-scale resources.

Big problems need solutions for optimal use and processing of big data. The Cloud offers many capabilities that make it suitable for such environments. Most importantly is that the Cloud should allow flexible and scalable utilisation of large-scale infrastructure resources to researchers on demand. Thus rather than researchers (or organisations) having to establish their own data/compute centres, the Cloud allows to dynamically leverage such resources on demand.

There have been many Cloud initiatives that have established a range of capabilities: Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Solution (S3) - http://aws.amazon.com, Microsoft Azure (http://azure.microsoft.com), Rackspace (www.rackspace.com) through to a range of widely adopted services such as Dropbox (www.dropbox.com) and iCloud (www.icloud.com). The earliest inception of Cloud computing focused on Infrastructure-as-a-Service (IaaS) allowing dynamic creation/release of virtual machines (VM), often through elastic-scaling. Other models now also exist including Platform-as-a-Service (PaaS) such as Azure, and Software-as-a-Service (SaaS). An example of SaaS is Google's gmail (www.gmail.com). In SaaS – which is closest to the work described here, there is a need to support a variety of software systems in an outsourced manner, i.e. they are not hosted locally on the user's desktop but elsewhere on the Internet. The Cloud paradigm of outsourcing has many potential advantages for urban research and AURIN, however it does raise a range of significant issues – most notably security and information governance related. At some point it has to be recognised that tackling the many challenges facing urban research requires access to considerable computational and data resources. Such capabilities rarely exist within a single organisational context.

Across Australia the federally funded National eResearch Collaboration Tools and Resources (NeCTAR – www.nectar.org.au) project has been established to realise a national Cloud facility for all researchers across Australia - irrespective of research discipline. NeCTAR utilises the OpenStack Cloud middleware and has multiple Cloud nodes (also known as availability zones) established across Australia. At present Cloud nodes / resources have been made available in Melbourne (3840 cores), Monash (2560 cores), Brisbane (3808 cores), Canberra (3200 cores), Adelaide (2992 cores) and Tasmania (320 cores). The primary focus of NeCTAR has been to establish an Australia-wide IaaS service. Researchers are able to dynamically create a range of VMs from small VMs offering single core 4Gb machines with 30Gb storage, through to extra large machines offering 16 cores with 64Gb memory and 480Gb storage. In addition to this, the federally funded Research Data Storage Infrastructure (RDSI – www.rdsi.uq.edu.au) project is rolling out major storage capabilities across Australia. This is expected to provide over 100 Petabyte data storage for Australian researchers. AURIN has been allocated 40 VMs and 35TB storage through the NeCTAR and RDSI projects respectively. AURIN also uses its own virtualised resources to complement these offerings. This includes three large-scale compute and two data servers that have been virtualised with VMware.

It is the case that the vast majority of non-computing-specific researchers require more than pure compute and data storage however. They require SaaS where their tools and applications are available in a manner that imposes minimal hurdles to their utilisation. A range of applications and services are utilised in the urban research domain: statistical tools such as R and STATA are widely adopted. GIS tools are also utilised however the expertise in targeted utilisation of such tools is not widespread, or more accurately the definitive algorithms for data processing reflecting best practice in urban data analytics are often not widely utilised. Within AURIN a major focus has been on taking best of breed applications and delivering them as workflow-based tools to the wider research community. AURIN is especially focused upon open source software systems development and delivery.

3. AURIN SYSTEM OVERVIEW

The solution put forward here for elastic compute and workflow-based data scaling dovetails into a greater and more complex architecture that realises the AURIN project. AURIN aims to develop an e-Infrastructure supporting research in the urban research disciplines [11]. This field is very broad and covers population and demographic futures and benchmarked social indicators; economic activity and urban labour markets; urban health, wellbeing and quality of life; urban housing; urban transport; energy and water supply and consumption, and innovative urban design. Each of these areas represents a significant urban research area in its own right. However a key challenge (and research opportunity) is that all of these areas are them-

selves inter-related. The AURIN platform is tasked with the integration of tools and data sets that allow such inter-related analyses to take place.

At the heart of the AURIN approach is providing programmatic access to *in situ* data sets in a manner that supports the researchers and their associated research processes whilst ensuring that the data access is consistent with policies from the data providers. To achieve this AURIN supports secure federated data access to many of the definitive urban and built environment data sets from multiple data providers from across Australia. These include the Australian Bureau of Statistics (ABS), Departments of Transport, Departments of Health (amongst many others) as well as a range of data sets from industry stakeholders such as the Public Sector Mapping Agency (PSMA) – the definitive geospatial provider for Australia, and Fairfax Australian Property Monitors (APM) amongst many others. The data driven framework realised by AURIN is described in more detail in [12]. The (high-level) architecture of AURIN is shown in Fig. 1.

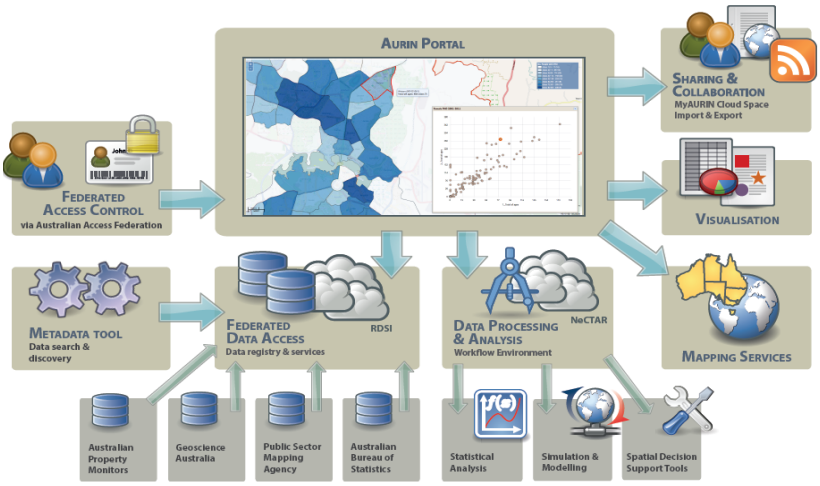


Fig. 1: Simplified AURIN Architecture (from [11]).

In this model, secure (federated) access is offered to all academics across Australia through the Internet2 Shibboleth Australian Access Federation (AAF – www.aaf.edu.au). Access for non-academics is supported through the AAF Virtual Home Organisation. In this model, authentication is achieved through username/passwords associated with the researchers own organisation. A major focus of AURIN is supporting access to (distributed) definitive data from major Australia-wide organisations. Wherever possible, hosting copies of external data sets (which are liable to change over time) is avoided. A range of tools is offered for data processing and visualisation within AURIN. Multiple expert committees associated with the project identified these tools as essential for urban research. These are discussed in more detail in section 4. A simplified overview of AURIN technical architecture is shown in Fig. 2.

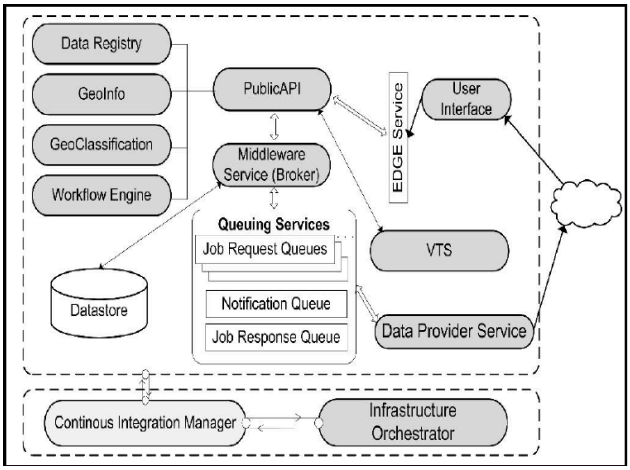


Fig. 2: The simplified technical realisation of the AURIN Architecture

The AURIN e-Infrastructure shown in Fig. 2 is composed of numerous components/modules:

- User Interface: is delivered through a targeted research portal offering seamless and secure access to all of the data sets and tools through a single point of access. The user interface makes heavy use of a range of Javascript libraries including Ext.js, Processing.js, Node.js and Ajax. All user access to the AURIN environment is through a web browser, i.e. there are no client side software requirements.
- Vector Tile Service (VTS): this service is a Map Server that provides vector tiling support for the infrastructure to deliver fast rendering of topographic maps in the user interface.
- EDGE service: this service provides and supports fine-grained authentication and authorisation – leveraging federated access control system information offered through the AAF. This service is based upon the Apache Shiro (http://shiro.apache.org) technology.
- Public API: provides a front end for the internal modules and components realised within the AURIN e-Infrastructure – this includes interfaces to the user interface and the middleware components.
- Middleware: is responsible for the control flow of information between the various components, i.e. the business logic. It offers a scalable messaging service that interacts as a broker between all components.
- Queuing services: provides support for asynchronous communications to help in scalability. The system utilises ActiveMQ for its message passing (http://activemq.apache.org). This supports scalability by introduction of asynchronous communications whereby multiple requests can be buffered in times of load.
- Data Registry: provides detailed information on the data sets and the metadata associated with the range of data providers offering resources to AURIN. At present AURIN makes available over 1100 national data sets with rich metadata on the data sets down to the individual attributes associated with any given data set;
- Geoinfo: enables the use of geographical boundaries across multiple resolutions. This supports providing dynamic data interrogation on larger data sets, e.g. disaggregated data across all of Australia.
- Geoclassification: supports a range of geospatial aggregation information across Australia. This includes graphs of spatial information that allow for example to establish all of the local government areas in a given State. There are many geo-classifications that have been put forward by different organisations and these have evolved over time (e.g. Melbourne in 2006 has a different spatial footprint to Melbourne in 2011 and the polygons reflecting the local authority boundaries across Victoria have also changed over time). This component supports intelligent navigation across the graph of spatial classifications, and especially how they may have changed over time.
- Datastore: provides a NoSQL store for the (heterogeneous) data sets returned from federated data providers. This has been realised using CouchDB.
- Data Provider Service: this module is tasked with providing the interface to an extensive range of heterogeneous data providers required by the AURIN infrastructure.
- Workflow Engine: provides the analytical toolbox that incorporates an extensive range of analytical tools realised as workflows that utilise the NeCTAR Research Cloud and AURIN VMs available through VMware. This component is the primary focus of this paper.

It is noted that the AURIN e-Infrastructure itself has been defined and ultimately constructed through integration of a large range of sub-projects (over 65 separate subprojects) involving software tools and data services offered from multiple providers/research organisations from across Australia. The Melbourne eResearch Group within the University of Melbourne has been responsible for the overarching integration of all of these subprojects into the AURIN platform.

4. AURIN WORKFLOW ENVIRONMENT

The original AURIN workflow environment has been iteratively established deriving inspiration in technologies such as flow-based programming, data flows, and scientific workflows. The AURIN workflow environment is essentially a component-oriented workflow orchestration framework into which various data transformation components have been made available and linked together to perform analytical routines defined by expert AURIN users. Through the simple integration interface, developers of workflow components need not to know or be involved with the inner workings of the workflow engine. In fact, most of the existing components were produced by external groups of researchers contributing new or existing software libraries into the wider AURIN ecosystem. The workflow management system that powers the analytical toolbox is composed of three subsystems, each of which attends to very specific use-cases. Fig. 3 illustrates such use cases, presenting the actors and their interaction with the AURIN system.

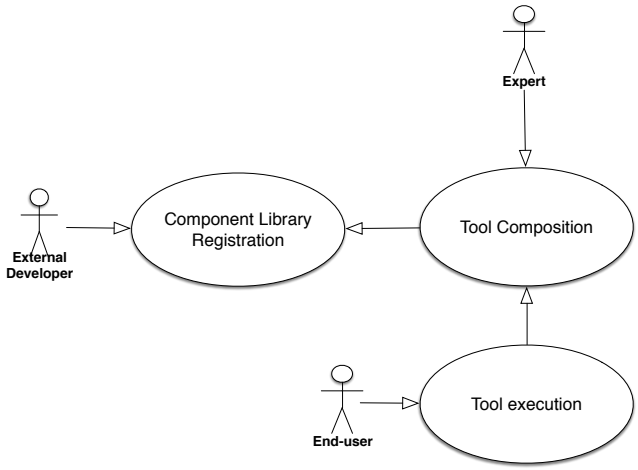


Fig. 3: Use-case diagram with actors and their interaction with the analytic toolbox

We recognise that there are multiple other workflow environments that are available: Galaxy, Taverna amongst many others [13-16]. Several of these are compared in [17]. AURIN explored many of these and identified that they did not tackle the specific scenarios facing many urban researchers. Firstly it was identified that there should be no requirement for installation and/or configuration of any software on the client-side, e.g. the users desktop. At the time of the AURIN project commencement many of the workflows environments, e.g. Taverna, were based on desktop applications. The solution was also expected to scale across highly heterogeneous systems including applications originally written in R, Java, C and allow to invoke/enact remote and independently developed web services of various flavours including geospatial web services, statistical web services amongst others. Essentially a solution was required that allowed simplifying the refactoring of these tools and applications. Supporting multiple different service and component types was also demanded and importantly this was expected to scale across heterogeneous Cloud resources and address the key need for secure access to distributed data and services. The workflow environment was also required to support highly secure scenarios where individual (unit level) data was to be accessed and used.

4.1 Workflow-based Component Library

Researchers and developers were commissioned by the AURIN project to contribute software libraries in their areas of expertise. Many of these libraries are existing pieces of software written in languages such as C, Java, R, or sometimes exposed as Web Services. The workflow analytical toolbox's component library is a repository of libraries that expose relevant metadata about the interface of each component. Contributors add metadata to their code via specially designed annotations and wrappers. At its most basic form, a component is a Java class with annotated fields and methods. The code annotation system was inspired by the Ob-

ject Modelling System (OMS), which provides a set a Java annotations used to tag fields and methods with control and informational metadata. For code not originally written in Java, a wrapper class is created and annotated, and the original code is called via technology specific bridged, such as Java Native Interface (JNI). A minimum set of metadata is required to allow the registration of a component into the toolbox's library, namely:

- Component name, description, category label, and author;
- Input and output named ports, including type, name, and description;
- Initialisation, execution, and finalisation methods identification;

Control annotations are instructions to the AURIN workflow framework on how to execute a component and how to chain it together with other components, while informational annotations are used in user interface constructs. A typical component will have a set of fields and methods that carry a particular role. Input fields carry parameters defined by a user or produced by other components. An execution method will run the component process. Output fields will contain a process' results to be displayed to users or used as input by other component. Table 1 lists some of the used metadata annotations and their purpose in describing a component's interface. In addition, Fig. 4 presents an example of an annotated Java class that makes use of most of the available annotations.

Table 1: Component metadata annotations

Annotation	Description	Applies to	Mandatory	Example
@Author	Component author	Class	No	@Author(name = "John Smith", contact = "http://www.aurin.org.au")
@Name	Component or input/output port name	Class, Field	Yes	@Name("lregression")
@Description	Component or input/output port description	Class, Field	Yes	@Description("Linear regression for a given vector.")
@Label	Component category	Class	No	@Label(AURINWFConstants.STATISTICS)
@In	Input port	Field	Yes (if there are input ports)	@In
@Out	Output port	Field	Yes (if there are output ports)	@Out
@Execute	Process execution	Method	Yes	@Execute
@Initialize	Process initialisation, input validation	Method	No	@Initialize
@Finalize	Process finalisation, clean up	Method	No	@Finalize

```java
@Name("AttributeUniqueValues")
@Description("Computes the unique values of a dataset attribute")
@Author(name="John Smith", org="AURIN")
@Label(AURINWFConstants.ANALYTICAL)
public class AttributeUniqueValues {

  @In
  @Name("Dataset")
  @Description("Dataset containing an attribute of type String")
  public SimpleFeatureSource dataset;

  @In
  @Name("Attribute")
  @Description("Attribute of type String for which unique values will be comput-
ed")
  public String attributeName;

  @Out
  @Name("Results")
  @Description("Resulting dataset containing a single attribute representing
unique values of the selected attribute")
  public SimpleFeatureSource uniqueValuesDataset;

  @Initialize
  public void init() throws Exception {
    // input validation goes here
  }

  @Execute
  public void exec() throws Exception {
    // execution goes here
  }
```

Fig. 4: Annotated Java class representing a component interface with added metadata

4.2 Workflow-based Tool Composition

Domain experts in fields such as statistics and spatial analysis access the toolbox via a workflow composition system. A meaningful workflow of linked components is called a tool. Tools are composed using a Domain Specific Language (DSL), which is processed by a composition engine that checks for link compatibility and automatically inserts data conversion components where necessary. At composition time, experts also describe the inputs of the tool, which are often data sets, data set attributes, numbers, Boolean values, and text fields; as well as outputs, which can are usually data sets, images, or text. Tool inputs are linked to input ports of components, while tool outputs must be fed from a components output port.

Fig. 5 depicts the composition steps, expressed in the DSL, of a tool aimed at computing the unique values of a data set attribute. The DSL is a subset of Gherkin, used by the Cucumber behaviour-driven development framework (http://cukes.info). It's adoption as a workflow composition language has allowed AURIN experts to express workflows in a verbose, yet friendly language, and at the same time produce definitions that can be directly deployed into the toolbox composition and testing engine.

```
Scenario: Compose workflow "Attribute Unique Values"
  When I create new workflow with ID "unique-values", name "Attribute Unique Val-
ues", description "Computes unique values of a dataset attribute", group "Data Ma-
nipulation"
  When I add component "aurin_geojson_reader" as "reader"
    And I add component "AttributeUniqueValues" as "uniqueVal"
    And I add component "aurin_geojson_writer" as "writer"
  When I link port "source" of "reader" to port "dataset" of "uniqueVal"
    And I link port "uniqueValuesDataset" of "uniqueVal" to port "features" of
"writer"
  When I create input variable "input_dataset" of type "String", visible "true",
link to port "url" of "reader"
    And tag input variable as "dataset"
  When I create input variable "attribute_name" of type " AttributeSelector", visi-
ble "true", link to port "attributeName" of "uniqueVal"
  When I create output variable "uniqueValuesResult" from port "result" of "writer"
```

Fig. 5: Workflow composition using a Domain Specific Language

4.3 Workflow-based Tool Execution

Once composed and tested, workflow-based tool definitions are stored in AURIN's analytical toolbox and made available to end-users via the AURIN web-based graphical user interface (GUI). User interface constructs are automatically generated and rendered, based on the tool definition and component metadata, resulting in simple and generic web forms. These GUIs are purposely kept simple through limiting the types of widgets to the most commonly needed and adopting a single column stacked layout.

Fig. 6 shows the UI form generated for the "*Attribute Unique Values*" workflow described above.

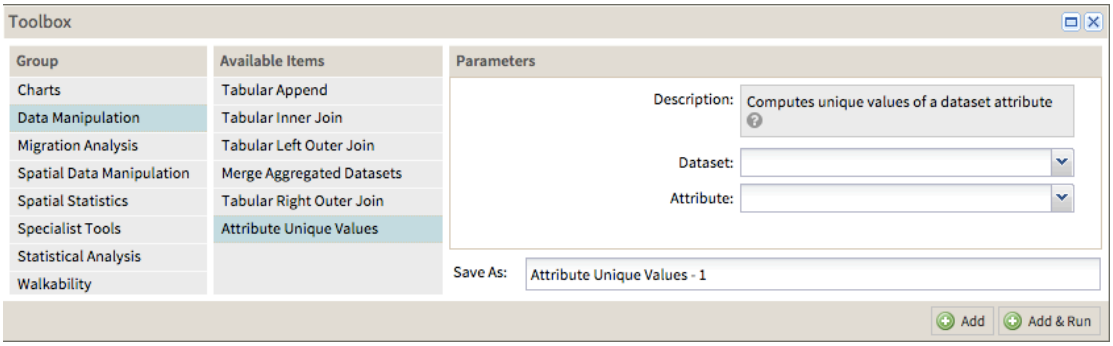


Fig. 6: Automatically generated UI based on the workflow definition and component metadata

4.4 Cloud-based Execution (Enactment) System

The architecture of the Cloud-based execution system is shown in Fig. 7. Tool execution (workflow enactment) is initiated by end-users through submission of a targeted user interface offered through the AURIN portal. The AURIN portal communicates with the back-end services via the AURIN public API, which itself is realised as a web service that directs (orchestrates) execution requests to the workflow engine's own REST-based interface. An execution request contains information such the tool to be invoked and values for all its inputs. If an input is a data set, a reference URL to that data set within the Datastore service is provided.

The workflow engine retrieves the tool definition from the definitions database, computes the component execution order based on a directed acyclic graph, creates an execution instance, and subsequently forwards it to the execution manager. The execution manager is responsible for identifying the resources required for executing each step of the tool's workflow, based on the definition of each component. The component type is obtained from the component definitions database.

As discussed, scalable execution of components on distribute resources is coordinated through the use of the ActiveMQ message broker. This provides high performance, reliable, and scalable transport of information across all involved systems. The message broker keeps multiple control queues and one data exchange queue. Possible component types, e.g. Java, R, or Graphics workers, utilise separate control queues. The workflow execution manager produces messages on control queues, which are consumed by workers. Workers are specialised software running on Cloud-based VMs. These workers are only capable of executing components of a certain type and thus only need to consume messages from one queue. The worker provisioning system dynamically creates and destroys worker VMs based on information about queue length and service time. Several workers of each type are available.

A data manager service runs in a separate VM, and is responsible for coordinating data transfers through the message broker's data queue. It obtains control messages from the execution manager, specifying a data reference identifier and information about the worker that requires the data. The data manager initiates a data transfer from the Datastore service to the data queue. The worker requiring the data set can then initiate a data transfer from the message queue into its local storage. Data streaming is supported by the message broker, thereby aiding the performance of the infrastructure and allowing it to cope when large amounts of data need to be transferred. However, it is noted that in the case of some data-intensive opera-

tions such as geospatial tabular joins, data is not moved to workers, rather, these operations are delegated to the Datastore service.

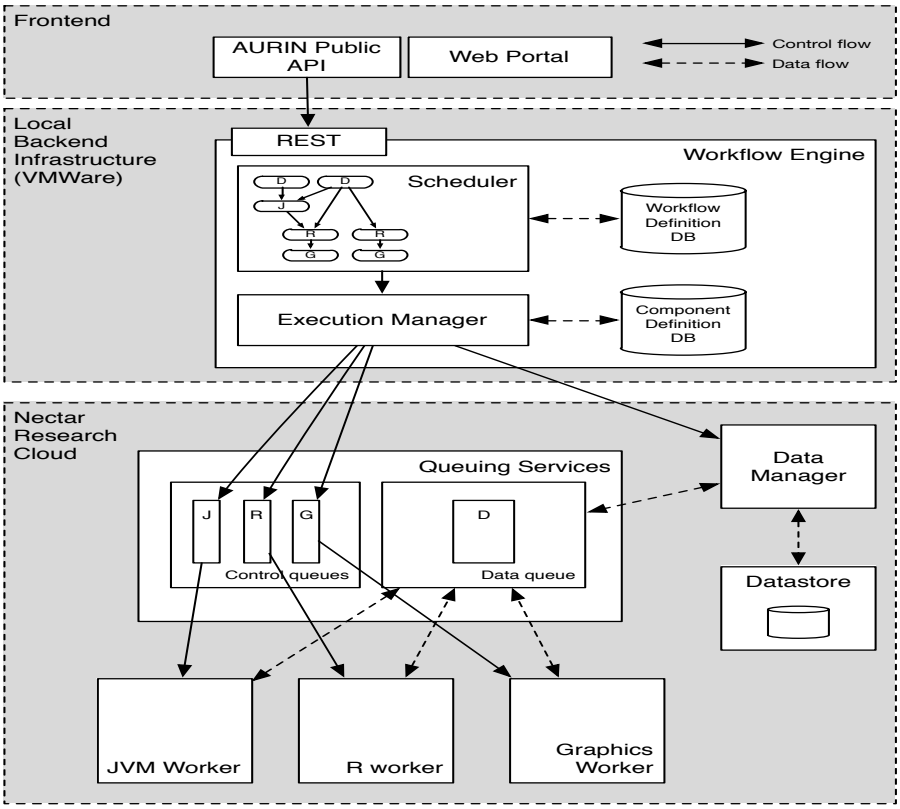


Fig. 7: Cloud-based workflow execution architecture

A hybrid-cloud approach is used in the deployment scheme of the various modules of this architecture. The workflow engine and its associated REST-based web service are deployed in a virtualised cluster alongside the AURIN frontend. The message broker, the workers, and the data manager run in a Cloud-based infrastructure offered by the NeCTAR Research Cloud.

5. CASE STUDIES

To illustrate the way in which the workflow environment supports the analysis of complex urban research phenomena, we explore two case studies. The first explores the voting patterns within the State of Victoria based on the 2010 federal election results. The second explores the establishment of a walkability index around a range of schools in / around Melbourne.

5.1 Victorian Voting Analysis

Politics and politicians have a huge influence on the planning and policies that ultimately shape urban settlements. As such, it is highly relevant to understand the way in which trends in voting patterns between and across political parties arises. Are there situational contexts of key importance to the voting results for cities and States? We focus here on the analysis of the federal election results of Victoria in 2010. To commence this process, the selection of the region of interest is made (Victoria). As shown in Fig. 8, the state of Victoria has numerous geospatial classifications (the map shows the 2011 Suburbs for Victoria).

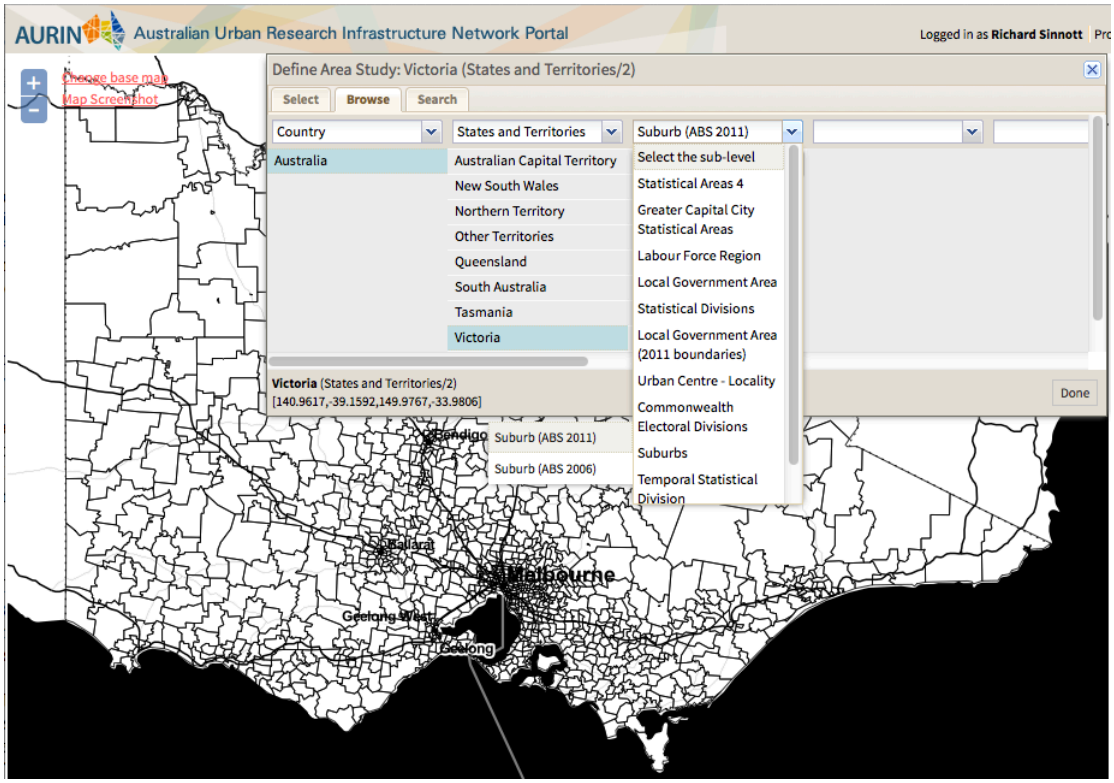


Fig. 8: Selection of Geospatial Region of Interest

Once the region of interest is selected, the data that is of interest can be obtained. The AURIN platform provides a range of mechanisms to search for and obtain data. Key to this is the geo-spatial filtering that takes place. Thus when Victoria is selected as the region of interest then data sets specific to NSW are excluded as one trivial example. A range data search capabilities are supported and shown in Fig. 9. Here any data that is related to the search term "voting" is retrieved. This search results are based upon the metadata provided by the data providers using targeted tools supported through the AURIN platform. The complete set of AURIN data is listed in http://data.aurin.org.au. As shown in Fig. 10, the data and importantly the variables related to voting data are accessible. In this case location quotients[1] related to voting patterns for the Australian Labour Party, the Liberal-Nationals Coalition, the Liberal Party and the Greens are selected. A subset of the data itself is shown in Fig. 10, where as seen (bottom left of Fig. 10) there are 1719 polling booth catchment areas across Victoria.

[1] A location quotient is a way of quantifying how concentrated a particular industry, cluster, occupation, or demographic group is in a region as compared to the nation. It can reveal what makes a particular region "unique" in comparison to the national average.

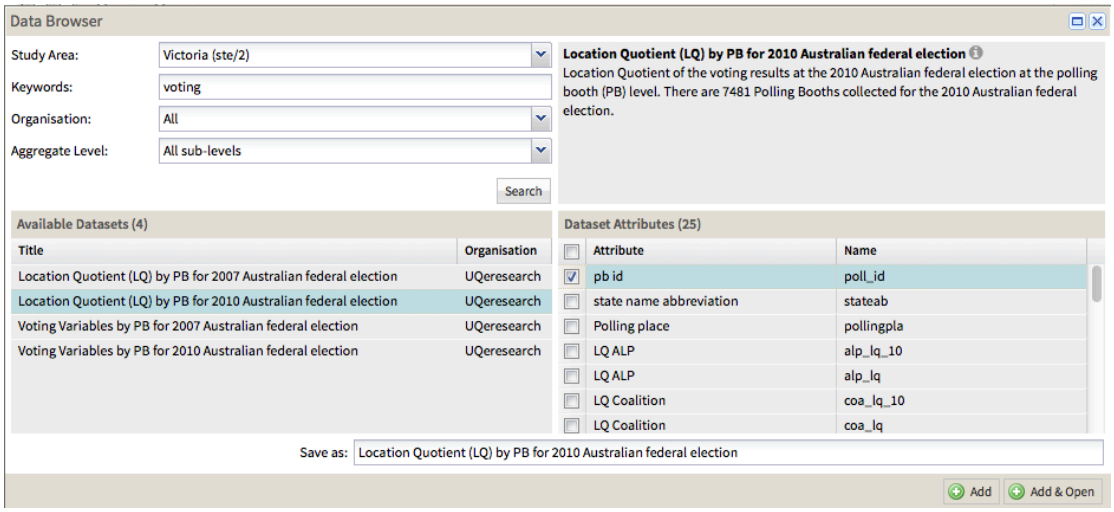
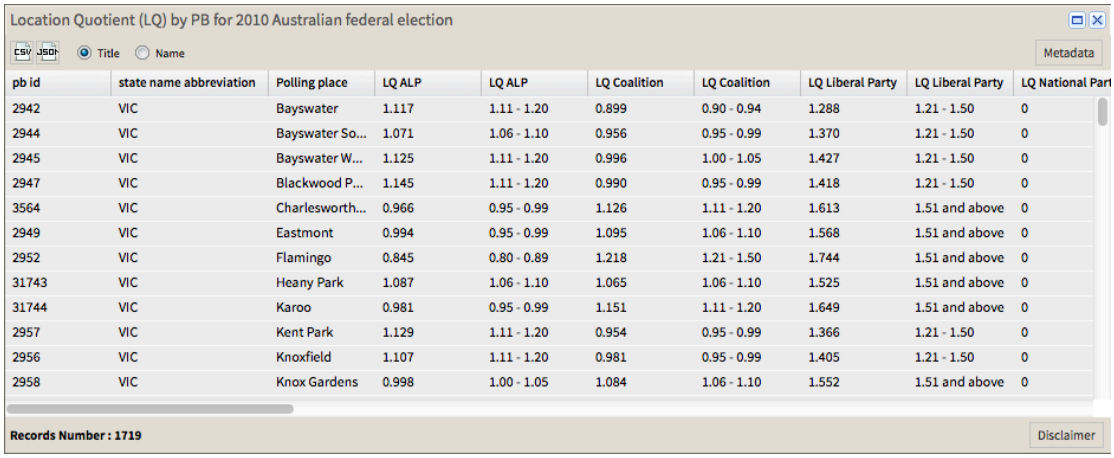


Fig. 9: Selection of Voting Data for Victoria



pb id	state name abbreviation	Polling place	LQ ALP	LQ ALP	LQ Coalition	LQ Coalition	LQ Liberal Party	LQ Liberal Party	LQ National Part
2942	VIC	Bayswater	1.117	1.11 - 1.20	0.899	0.90 - 0.94	1.288	1.21 - 1.50	0
2944	VIC	Bayswater So...	1.071	1.06 - 1.10	0.956	0.95 - 0.99	1.370	1.21 - 1.50	0
2945	VIC	Bayswater W...	1.125	1.11 - 1.20	0.996	1.00 - 1.05	1.427	1.21 - 1.50	0
2947	VIC	Blackwood P...	1.145	1.11 - 1.20	0.990	0.95 - 0.99	1.418	1.21 - 1.50	0
3564	VIC	Charlesworth...	0.966	0.95 - 0.99	1.126	1.11 - 1.20	1.613	1.51 and above	0
2949	VIC	Eastmont	0.994	0.95 - 0.99	1.095	1.06 - 1.10	1.568	1.51 and above	0
2952	VIC	Flamingo	0.845	0.80 - 0.89	1.218	1.21 - 1.50	1.744	1.51 and above	0
31743	VIC	Heany Park	1.087	1.06 - 1.10	1.065	1.06 - 1.10	1.525	1.51 and above	0
31744	VIC	Karoo	0.981	0.95 - 0.99	1.151	1.11 - 1.20	1.649	1.51 and above	0
2957	VIC	Kent Park	1.129	1.11 - 1.20	0.954	0.95 - 0.99	1.366	1.21 - 1.50	0
2956	VIC	Knoxfield	1.107	1.11 - 1.20	0.981	0.95 - 0.99	1.405	1.21 - 1.50	0
2958	VIC	Knox Gardens	0.998	1.00 - 1.05	1.084	1.06 - 1.10	1.552	1.51 and above	0

Fig. 10: Voting Data for Victoria

Based on this data, the workflow environment can be used to perform a range of deeper analyses. The AURIN toolbox contains an extensive range (over 100) of tools realised as workflows as shown in Fig. 11 which shows the user interface for one of these tools (creating a box plot of the data shown in Fig. 10). These tools are realised as workflows as described above and cover key capabilities including basic charting, e.g. bar charts/pie charts; data manipulation, e.g. joining of data sets; migration analysis related tools; spatial data manipulation; spatial statistics; specialist tools such as employment clustering; statistical analysis and walkability tools. It is important to emphasise that the tools were originally implemented in a heterogeneous and non-interoperable manner. However it is also key to note that the implementation of these tools reflects the state of the art and best practice in urban research data analysis. The original tools were identified by urban research experts involved in committees that advised and subsequently shaped the AURIN project including the tools and data that were ultimately offered through the AURIN platform.

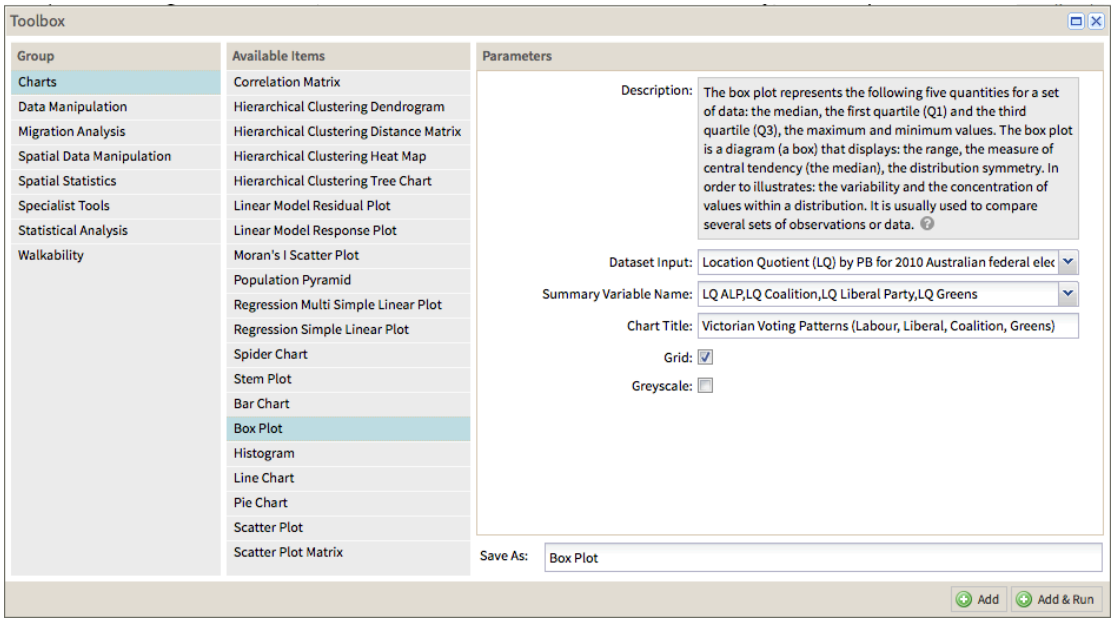


Fig. 11: AURIN Workflow-based Tool Environment

A typical initial scenario to analyse the data can include creation of a box plot representing key quantities for a set of numeric data such as the median, the first quartile (Q1) and the third quartile (Q3), the maximum and minimum values for voting patterns. The output of such a simple workflow (with a single tool) is shown in Fig. 12.

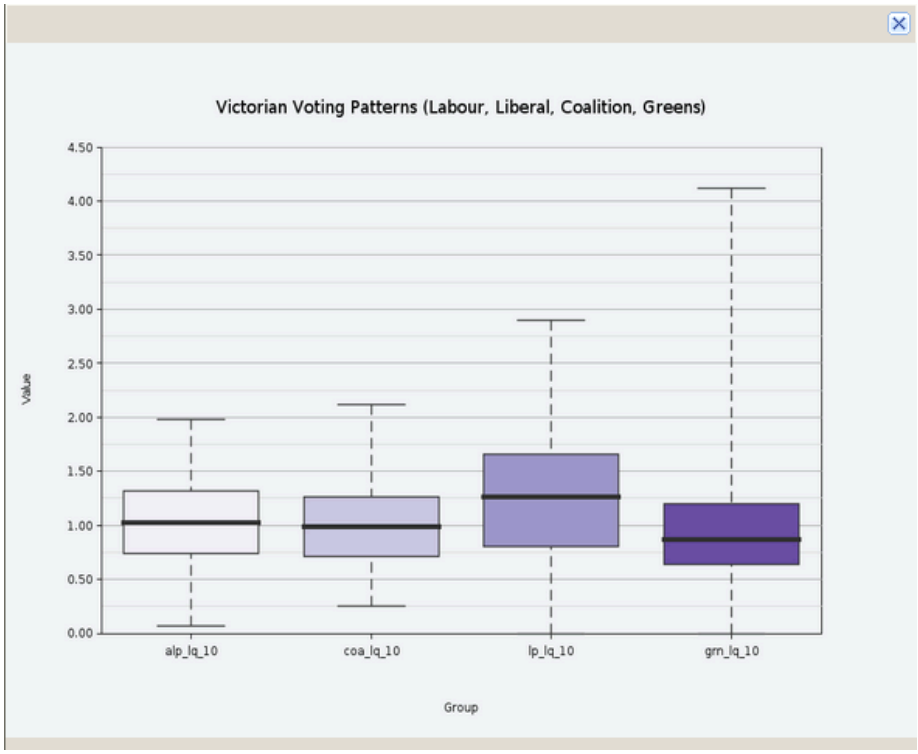


Fig. 12: Box Plot showing Voting Patterns for Labour, Coalition, Liberal and Green Parties of Victoria

Whilst offering a broad-brush analysis of data, it is often necessary to provide deeper data analytics of data to explore potential patterns in data sets. This can involve complex work-flows involving multiple tools. One example of such a workflow that is supported through the AURIN environment for analysis of complex data sets such as the voting data given above is

shown in Fig. 13. This workflow uses a range of advanced spatial statistics and visualisation capabilities.

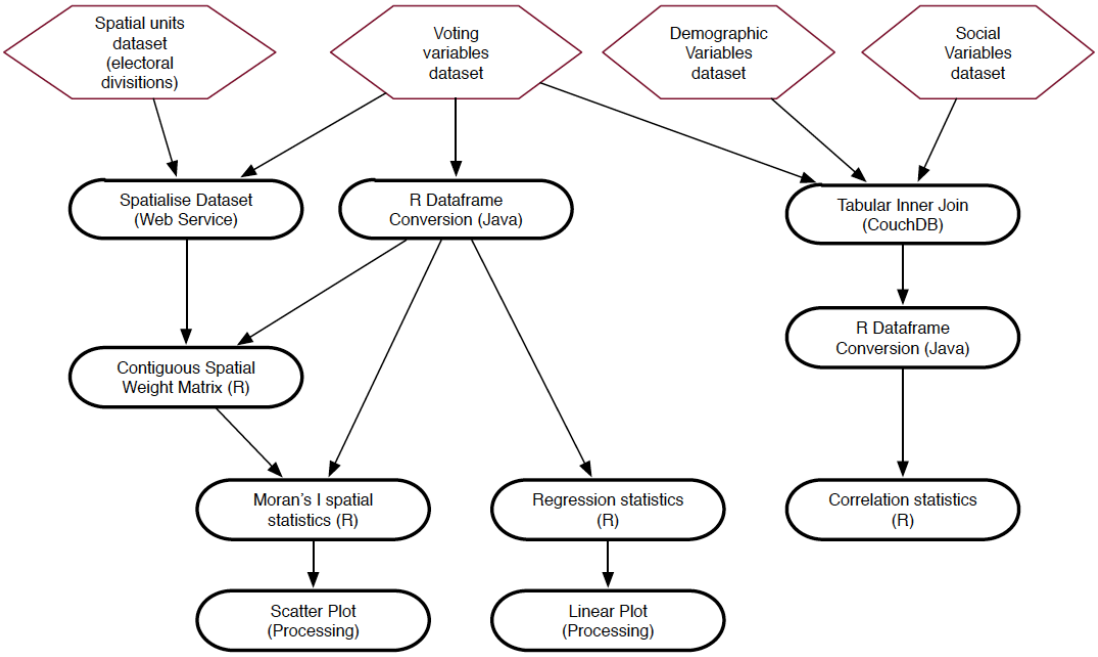


Fig. 13: Workflow to support data-intensive analytics associated with Voting Data

To realise this, the following components were incorporated into the workflow:

- Spatialise Data Set – a web service that joins a data set, aggregated by a geographical level, with the corresponding spatial geometries;
- Contiguous Spatial Weight Matrix – computes a matrix of contiguity weights, which counts locations as "near" if they share a boundary with another such location;
- R Dataframe Conversion – transforms a data set into an R Dataframe so that it can be used as input in R scripts;
- Moran's I – computes Moran's I global measure of spatial autocorrelation statistic and related information, i.e. the degree to which a set of spatial features and their associated data values tend to be clustered together;
- Regression – computes a linear regression statistic;
- Tabular Inner Join – joins columns of multiple data sets based on a specified key;
- Correlation – performs a statistical correlation;
- Scatter Plot – provides a graphical plot suitable for visualising the output of spatial statistics;
- Linear Plot – graphical plot for visualising the result of a linear regression.

Based on this workflow a range of deeper analytics are possible. A typical requirement is to see how votes are geospatially distributed spatially across Victoria. Choropleth maps are one of the key approaches used by urban researchers for achieving this. In this scenario the variables *Percentage Swing* to or away from the Coalition is shown in Fig. 14 for the regions (polling booth catchment areas) around Melbourne. Here reds indicate negative swings, i.e. swings to other parties, and blues indicate positive swings to the Coalition parties.

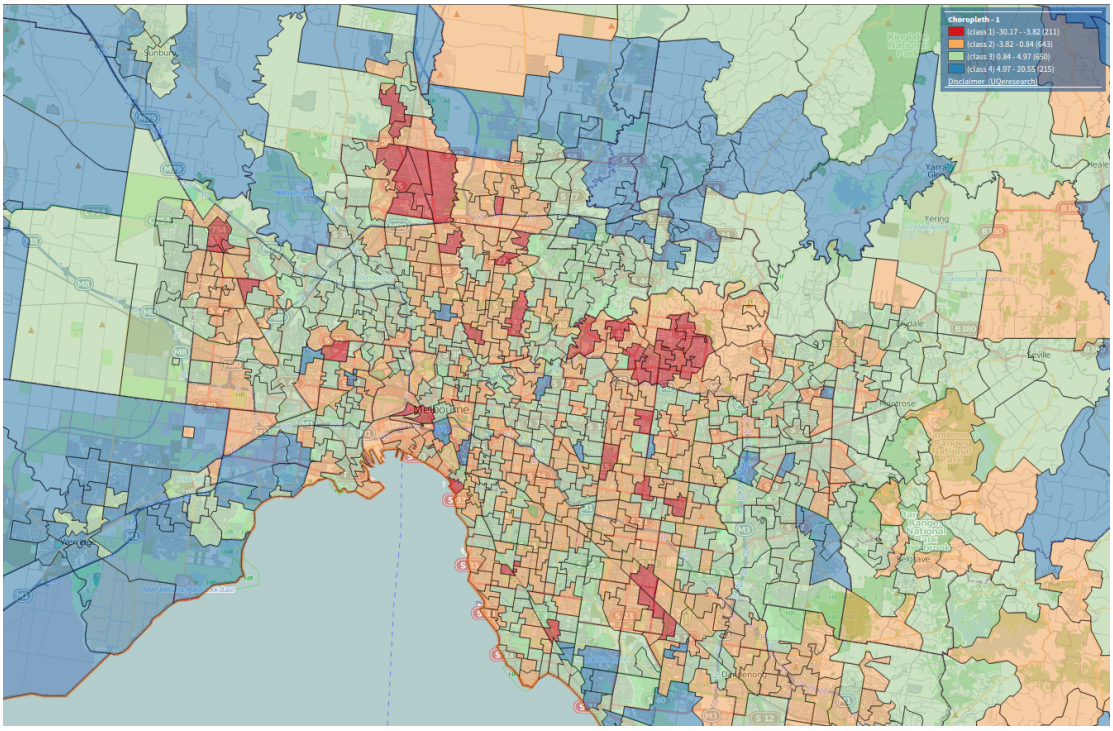


Fig 14: Choropleth Map Showing the Voting Swing to (green)/ from (red) the Coalition

The above mapping gives the impression that the swings are not randomly distributed: polling booths that are close to each other tend to have similar patterns of swings, e.g. reds and greens tend to cluster together. This workflow allows checking whether this is truly the case. Thus rather than by chance/intuition alone the workflow uses targeted Spatial Statistics tool. Specifically the workflow includes the Moran's I statistical approach, which compares similarity in location with similarity in attribute values. The output of this for the Victorian Voting Swing data is shown in Fig. 15 indicating that there is indeed a correlation.

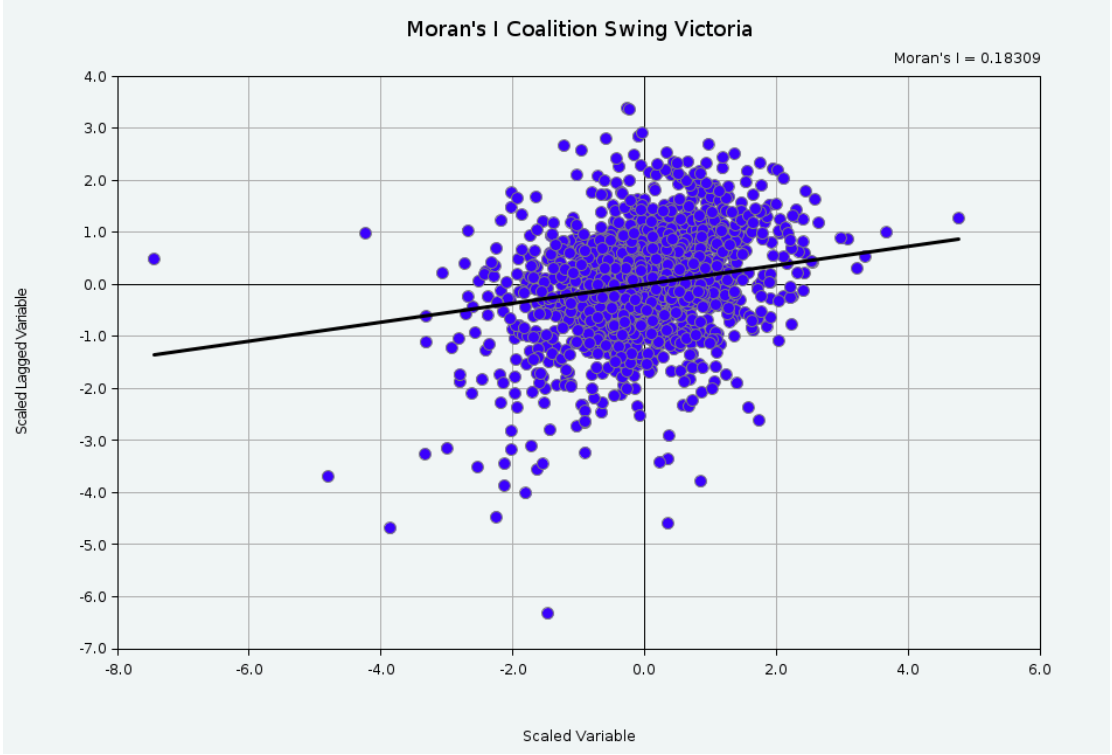


Fig 15: Moran's I Measure of Spatial Correlation for Voting Swing

To determine whether other factors are associated with the percentage of Coalition Swing in polling booths around Victoria, it is possible to superimpose (overlay) the percentage of Green voters over the percentage of the swing to the Coalition (Fig 16).

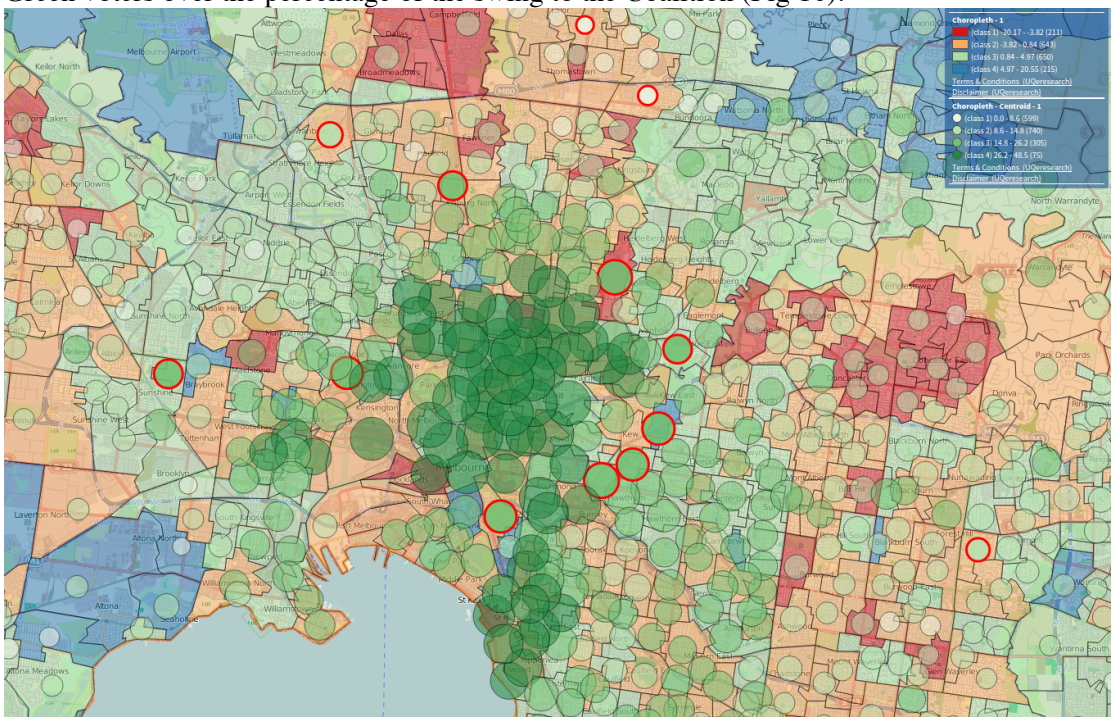


Fig 16: Choropleth Map of Green Voters (green centroids) and percentage of Coalition Swing (red choropleth) around Victoria

To check the statistical strength of this pattern we can use the result of the Linear Regression Plot (Fig. 17) to visualise and estimate the statistics regarding the strength of the correlation and hence the swing itself.

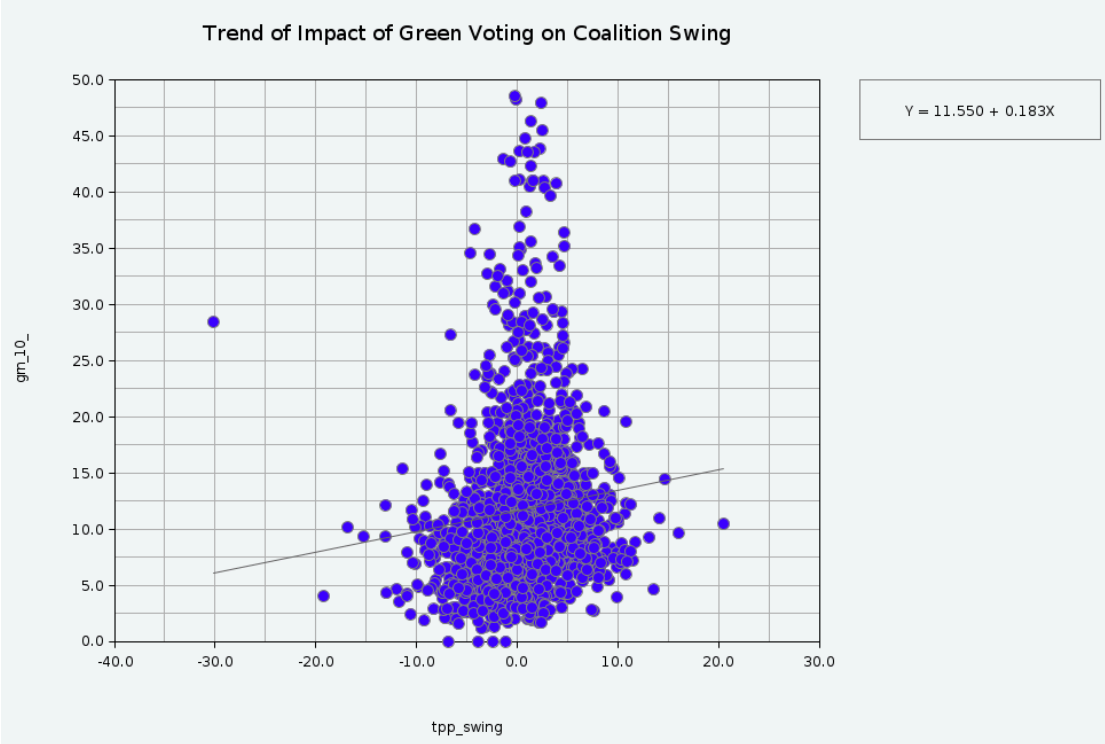


Fig 17: Linear Regression showing Impact of Green Voting on Coalition Swing

It is important to note that such detailed analytics allows far deeper understanding of a given data set than would otherwise be possible for most researchers. Hitherto such analysis would require a range of algorithms written in languages such as R, SPSS or STATA, but the deeper geospatial analytics would be beyond most researchers who are typically not GIS experts. This is just one workflow reflecting the analysis of voting data. The same workflow can be used for a range of other geospatial data explorations, e.g. relation between fast food restaurant locations and the average body mass index (BMI) for children in different regions of Australia (as one of many examples).

5.2 Walkability Index Workflow

A walkability index is a key measure that can be used to help answer many public health research questions related to health and wellbeing. A walkability index is used to assess how walkable a given neighbourhood is and what are the factors that determine this? Such information can be useful to better understand growing challenges of obesity, increasing dependence on cars as well as quality of life more generally.

The AURIN walkability workflow library is a set of automated components that generate walkability indices at user-specified scales. Typically these are at smaller community-based aggregation levels such as suburb, a Census Collection District and/or localised user-specified neighbourhoods. The realisation of the walkability workflow tools builds on a range of specialised spatial and statistical functions that have been contributed by urban researchers working in the population health domain. The original walkability tool was a desktop application used by a small cohort of population health scientists; its components were re-coded in Java especially for the AURIN project. The AURIN-enabled workflow version of this tool provides a readily and broadly accessible, spatial analytic tool available to assess the walkability of Australian urban areas across a multitude of situational contexts. These components were further combined with data pre-processing routines to create tools, made available together as the walkability toolkit.

Here, we describe a scenario utilising AURIN's workflow composition mechanism to create one such tool, aimed at providing walkability indices for a set of Melbourne school addresses, and producing a score based on three key measures: road connectivity, gross dwelling density and the land use mix. The following components were used in the workflow composition.

- Geocoder: transforms a set of addresses into geographic coordinates (latitude/longitude points). AURIN provides programmatic access to the (commercial) PSMA geocoder for this purpose. This is the most up to data geocoding system for Australia.
- Projection washer: it is assumed that data inputs come in a range of projections, thus requiring re-projection to an appropriate (preserving both area and distance) projection before analysis takes place. This component transforms a data set's projection by converting its coordinate reference system into linear coordinates.
- Neighbourhood generator: a routine to create polygons (neighbourhoods) by traversing a road network in all directions, based on user-supplied parameters of distance and radius. The network itself utilises the definitive road/street network for Australia offered through PSMA.
- Connectivity measure: a count of three way (or more) street intersections over the area of the neighbourhood. The street network used to calculate the number of intersections act as a proxy to show the potential mobility of pedestrians within the neighbourhood.
- Density measure: a ratio consisting of the count of the number of dwellings over a given region.
- Land use mix measure: examines the heterogeneity/homogeneity of land uses (of interest) within a neighbourhood. This might be commercial, industrial, residential, parkland etc.
- Z-score: computes the Z-score over all measures.

The components used for this workflow and their inter-connections are depicted in Fig. 18.

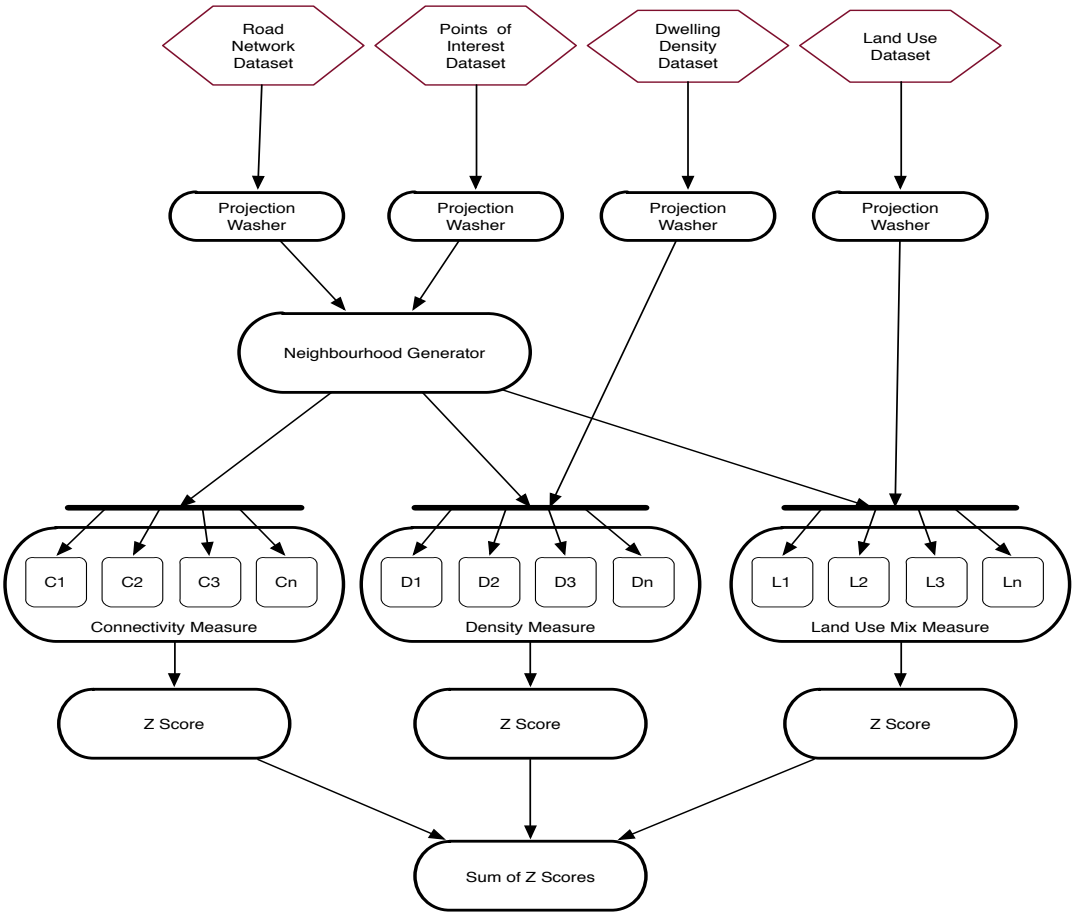


Fig. 18: Walkability workflow with geocoding of addresses

In this scenario a set of addresses of inner Melbourne schools is uploaded to the AURIN portal as a CSV file containing one address per line. The Geocoder service offered by PSMA produces a set of points, which are subsequently used as input for the neighbourhood generator component. For each neighbourhood, walkability measures are computed. Specifically users are able to specify the distance over which walkability is to be measured, e.g. from a given point, walk 1km along the road network and determine the overall walkability score along the routes that are taken.

The end result of this workflow is a new data set, with results of each measure as well as Z-scores attached to each neighbourhood. Fig. 19 shows a map representation of the results, with neighbourhoods (distances walked along the road networks) coloured according to their Z-scores. In this representation, green neighbourhoods indicate more walkable areas, whilst red scores represent lower scores, i.e. less walkable neighbourhoods.

Correlating such information with other sources of data can explain a broader range of phenomenon. Do children not walk to school, as they have to cross many busy roads or the walk itself is through industrialised areas? Is there a correlation between the walkability of an area and ownership of cars? Are richer neighbourhoods more likely to be more walkable neighbourhoods? Such questions can only be answered through access to the definitive sources of data and the use of tools that can be used to merge and analyse such data.

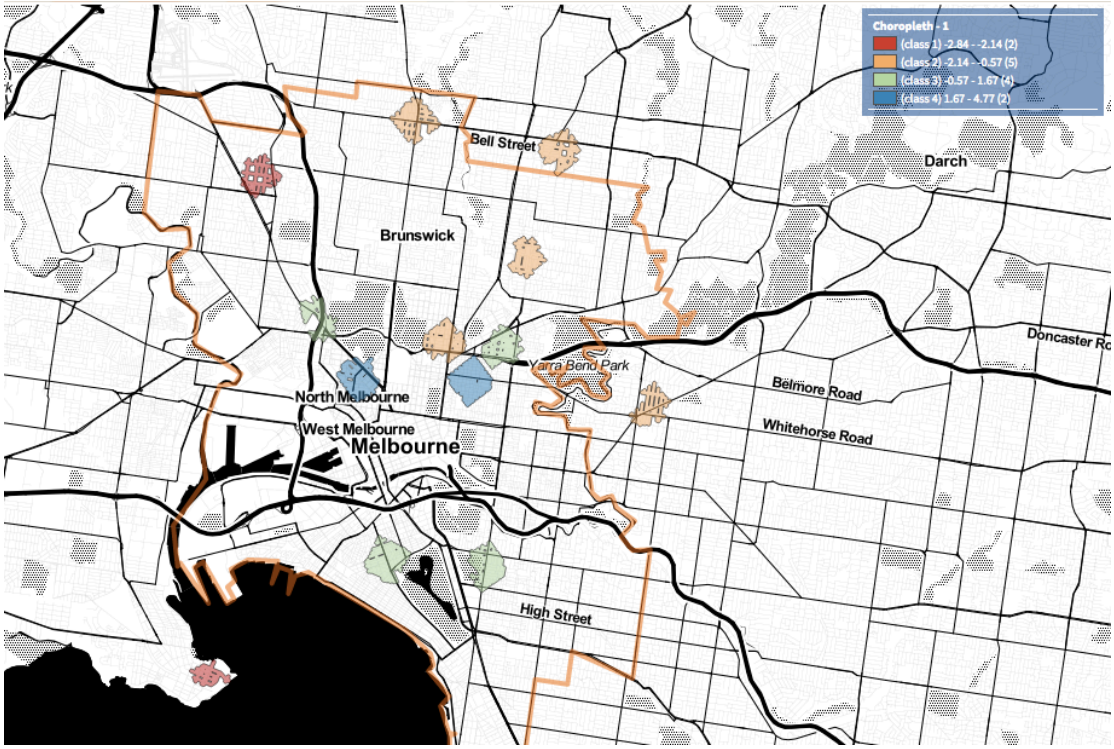


Fig. 19: Results of walkability workflow with walkability scores computed for a set of schools around Melbourne

6. CONCLUSIONS

The AURIN project has developed a data rich research environment that allows access to and use of urban research data at a scale that has hitherto been impossible. The project has made available an extensive range of tools that exploit major Cloud infrastructures. All of these tools have leveraged a common workflow-based approach that allows seamless interconnection and support. This includes support for security-oriented scenarios where pull-based queries and associated geospatial data analysis is required without revealing the identity of the individuals themselves [18].

The ultimate litmus test of the AURIN platform is not related to the implementation of the systems and services however. Rather it is through the uptake and adoption of the systems themselves. Basic statistics on the number of users and the organisations that have been accessing AURIN are obtained both within the AURIN environment and independently through the AAF. Currently (June 2015) there have been over 35,000 user sessions of the AURIN portal (i.e. log-ins) from *all* organisations involved in the AAF. It is interesting to observe that the fasting growing community of users of the AURIN platform are from non-academic organisations. Thus there is a clear trend in adoption at both the government and industry level, i.e. users accessing the platform through the AAF Virtual Home Organisation (VHO). The AURIN team is active in outreach and engagement activities, such as master classes, industry focused training and conference participation, with new users adopting the platform on a daily basis. Furthermore new data sets and services are continually being added to the platform to keep the community engaged.

The AURIN platform has demonstrated what can be achieved for large-scale federated data access and analysis. This solution is now providing the basis for a range of follow-on grants related such as the national platform for Clean Air and Urban Landscapes funded by the federal government Department of the Environment from 2015-2021 (http://www.environment.gov.au/science/nesp). The future funding for AURIN itself is currently under discussion at the federal level along with many other national initiatives such as NeCTAR and RDSI.

Acknowledgments
The AURIN project is funded through the Australian Education Investment Fund SuperScience initiative. The project began in July 2010 and is due to run to mid-2015. We gratefully acknowledge their support and the wider research community on their feedback. We are grateful to the rest of the AURIN team including those within the Melbourne eResearch Group (Dr. Glenn Jayaputera, Philip Greenwood, Luca Morandini, Dr. Ivo Widjaja, Sulman Sarwar, Dr. Marcos Nino-Ruiz, Dr. Hossein Pursultani, Christopher Bayliss, Ghazal Nogoorani, Gerson Galang, Andrew Bromage, Daghan Acay, Davis Marques, Rosana Rabanal), those within the AURIN Office (Jack Barton, Xavier Goldie, Chis Pettit, Bob Stimson, Stewart Wallace) and the wider network of AURIN collaborators. Map tiles in figures by Stamen Design (stamen.com) under CC BY 3.0 (http://creativecommons.org/licenses/by/3.0).

7. REFERENCES

[1] A. J. Hey and A. E. Trefethen, "The Data Deluge: An e-Science Perspective," 2003.

[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica, "A view of cloud computing," Communications of the ACM, vol. 53, pp. 50-58, 2010.

[3] R. Hirschheim, R. Welke, A. Schwarz, (2010). Service-oriented architecture: Myths, realities, and a maturity model. MIS Quarterly Executive, 9(1), 37-48.

[4] A.J. Hey and A. E. Trefethen, "Cyberinfrastructure for e-Science," Science, vol. 308, pp. 817-821, 2005.

[5] P. Papajorgji, H. W. Beck, and J. L. Braga, "An architecture for developing service-oriented and component-based environmental models," Ecological Modelling, vol. 179, pp. 61-76, 2004.

[6] J. Shiers, (2007). The worldwide LHC computing grid (worldwide LCG). Computer physics communications, 177(1), 219-223.

[7] V. Marx, (2013). Biology: The big challenges of big data. Nature, 498(7453), 255-260.

[8] C. Hoffa, G. Mehta, T. Freeman, E. Deelman, K. Keahey, B. Berriman, J. Good, (2008, December). On the use of cloud computing for scientific workflows. In eScience, 2008. eScience'08. IEEE Fourth International Conference on (pp. 640-645). IEEE.

[9] R.O. Sinnott, G. Galang, M. Tomko, R. Stimson, Towards an e-Infrastructure for Urban Research Across Australia, IEEE e-Science Conference, Stockholm, Sweden, December 2011.

[10] R. Stimson, M. Tomko, R.O. Sinnott, The Australian Urban Research Infrastructure Network (AURIN) Initiative: A Platform Offering Data and Tools for Urban and Built Environment Researchers across Australia, State of Australian Cities, Melbourne, Australia, November 2011.

[11] C. Pettit, R. Stimson, J. Barton, X. Goldie, R.O. Sinnott, T. Kvan, The Australian Urban Intelligence Network supporting Smart Cities, to appear in CUPUM 2015 Conference book on Smart Cities and Planning Support Systems, eds: S. Geertman, J. Stillwell, J. Ferreira, R. Goodspeed, February 2015.

[12] R.O. Sinnott, C. Bayliss, G. Galang, P. Greenwood, G. Koetsier, D. Mannix, L. Morandini, M. Nino-Ruiz, C. Pettit, M. Tomko, M. Sarwar, R. Stimson, W. Voorsluys, I. Widjaja, A Data-driven Urban Research Environment for Australia, IEEE e-Science Conference, Chicago USA, October 2012.

[13] J. Goecks, A. Nekrutenko, J. Taylor, (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology, 11(8), R86.

[14] B. Javadi, M. Tomko, R.O. Sinnott, Decentralised Orchestration of Data-centric Workflows in Cloud Environments, Future Generation Computing Systems, 2013, http://dx.doi.org/10.1016/j.future.2013.01.008

[15] E. Deelman, G. Singh, M. Su, J. Blythe, Y. Gil, C. Kesselman, D. Katz, (2005). Pegasus: A framework for mapping complex scientific workflows onto distributed systems. Scientific Programming, 13(3), 219-237.

[16] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, C. Goble (2013). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic acids research, gkt328.

[17] A. Barker, J. Van Hemert, (2008). Scientific workflow: a survey and research directions. In Parallel Processing and Applied Mathematics (pp. 746-753). Springer Berlin Heidelberg.

[18] R.O. Sinnott, P. Chhetri, Y. Gong, A. Macaulay, W. Voorsluys, Privacy-preserving Data Linkage through Blind Geo-spatial Data Aggregation, The IEEE International Symposium on Big Data Security on Cloud (BigDataSecurity 2015), New York, USA, August 2015.

[19] R.O. Sinnott, C. Bayliss, A. Bromage, G. Galang, G. Grazioli, P. Greenwood, G. Macauley, L. Morandini, M. Nino-Ruiz, G. Nogoorani, C. Pettit, M. Tomko, M. Sarwar, R. Stimson, W. Voorsluys, I. Widjaja, The Australian Urban Research Gateway, Journal of Concurrency and Computation: Practice and Experience, April 2014, doi: 10.1002/cpe.3282.

[20] I. Widjaja, P. Russo, C. Pettit, R.O. Sinnott, M. Tomko, Modeling Coordinated Multiple Views of Heterogeneous Data Cubes for Urban Visual Analytics, International Journal of Digital Earth, October 2013.