Journal of Visual Languages and Computing 45 (2018) 1–16

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

HomeSeeker: A visual analytics system of real estate data

Mingzhao Li[a,*], Zhifeng Bao[a], Timos Sellis[b], Shi Yan[a], Rui Zhang[c]

[a] *RMIT University, Melbourne, Australia*
[b] *Swinburne University of Technology, Hawthorn, Australia*
[c] *The University of Melbourne, Parkville, Australia*

ARTICLE INFO

Article history:
Received 20 June 2017
Revised 3 January 2018
Accepted 2 February 2018
Available online 5 February 2018

Keywords:
Real estate data
Visual analytics
Data integration
Multidimensional visualization
Geographic visualization

ABSTRACT

In this paper, we present HomeSeeker, an interactive visual analytics system to serve users with different backgrounds of the local real estate market and meet different degrees of user requirements. As a result, HomeSeeker augments existing commercial systems to help users discover hidden patterns, link various location-centered data to the price, as well as explore, filter and compare the properties, in order to easily find their preferred properties. In particular, we make the following contributions: (1) We present a problem abstraction for designing visualizations that help home buyers analyse the real estate data. Specifically, our data abstraction integrates heterogeneous data from different channels into a location-centred integrated real estate dataset. (2) We propose an interactive visual analytic procedure to help less informed users gradually learn about the local real estate market, upon which users exploit this learned knowledge to specify their individual requirements in property seeking. (3) We propose a series of designs to visualize properties/suburbs in different dimensions and in different granularities. We have collected, integrated and cleaned last 10 year's real estate sold records in Australia as well as their location-related education, facility and transportation profiles, to generate a real multi-dimensional data repository, and implemented a system prototype for public access (http://115.146.89.158). At last, we present case studies based on real-world datasets and real scenario to demonstrate the usefulness and effectiveness of our system.

1. Introduction

Finding a suitable property to purchase or rent is critical to people's life. People normally find candidate properties from online or by talking with real estate agents, and then inspect those preferred properties in person. The process of finding candidate properties can be both complicated and time-consuming since users could have many factors to consider, such as budgets, facilities like supermarkets and nearby transportation, as well as public schools associated with children's education. Moreover, most home buyers (esp. first home buyers) also prefer a property that will grow in value over time so they can take a step up the property ladder in a few years [1,2].

There are several commercial systems [3–8] in different countries that provide access for people to browse or search for real estates. We have noticed three common problems in using those existing systems. **Problem 1**: data in those systems are only about the description of properties themselves such as the number of bedrooms and bathrooms. However, a lot of useful geographically related information which is critical to users' choices is not captured, such as the regional, educational and transportational profiles (see Section 3.1), not to mention presenting them to users. Without such information, it is also very difficult to evaluate the robustness and effectiveness of the current visualization design of those commercial systems in visualizing such geo-related multidimensional information. **Problem 2**: existing commercial systems fail to provide an efficient way for users to learn about the local real estate market and understand their own needs based on their understanding of the market. Since *"users do not know what they want until they see it"* [9], many users, especially first home buyers, do not know much about the local real estate market; and they might continually change their preferences while they browse properties. For example, a user has a budget of one million dollars to buy a 3-bedroom house in Camberwell and he wants the property within a 10-minute walk to the nearest train station. However, after studying the market, he understands that such kind of a house might be not easy to find in Camberwell, so he will either find a property that is further to the station or in other suburbs. **Problem 3**: although some of the systems (such as Real Estate Australia [10]) provide map-based presentation, the design can

[*] Corresponding author.
E-mail address: mingzhao.li@rmit.edu.au (M. Li).

https://doi.org/10.1016/j.jvlc.2018.02.001
1045-926X/© 2018 Elsevier Ltd. All rights reserved.

only display the locations on the map, which is not sufficient to support comparison of properties in multiple aspects.

In a nutshell, existing systems fail to provide location-aware services, not to mention offering those services at different levels of details, due to the lack of a systematic approach to model and collect a comprehensive location-centered real estate data. To address those problems, we introduce HomeSeeker (http://115.146.89.158), an interactive real estate visualization system to augment current commercial systems. In particular, HomeSeeker is able to assist users in understanding the local areas and local real estate market, exploring and finding candidate properties based on their individual requirements, and visually comparing properties/suburbs in multiple aspects as given by the user. In particular, we make the following contributions from the perspective of visualization design.

1. We present a systematic problem abstraction that guides the visualization design of real estate data for home buyers. Particularly, we have collected data from different channels and integrated a comprehensive location-centred real estate dataset (**Section 3**).
2. We propose an interactive visual analytic procedure to help zero-knowledge users gradually understand the real estate market, and ultimately find appropriate properties based on their individual preferences (**Section 4**).
3. We design and implement a system with existing visualization design choices and novel visualization designs, and multiple coordinated views to visualize real estate data, which can also be used to solve other geographically related multidimensional analysis problems (**Section 5**).

The rest of the paper is organized as follows. Section 2 discusses related works with regard to the real estate domain and visualization. Section 3 presents the problem abstraction, followed by a system overview in Section 4. In Section 5, we present our visualization design choices in detail together with design justifications. Section 6 discusses the system implementation. We present experiments based on real-world datasets in Section 7, followed by conclusion and future works in Section 8.

2. Related work

In this section, we first present related work in the domain of this paper (i.e., visualization of real estate data), and then discuss the literature in a broader sense (i.e., geo-related multidimensional visualization).

2.1. Visual analysis of real estate data

Visual exploration of real estate data has been considered as an interesting application and attracted attention from both the academic and industry communities.

Many online platforms in different countries allow people to search for real estate properties, such as Real Estate [3] and Domain [4] in Australia, Right Move [5] in UK, Zillow [6] and Trulia [7] in the USA and Soufang [8] in China. Most of those systems provide functions such as searching among or filtering numerous properties available in the real estate market, and then display the result in the form of a list or a map. Some systems such as Zillow [6] also allow users to virtually look into properties of interest using Matterport 3D view. Although these commercial systems have provided users a convenient access to available real estates, they still have some drawbacks. First, the data that most of these commercial systems [3,4,8] have used is mainly about the property itself, while a lot of useful information that is critical to users' choices is neither captured nor presented. Although some systems [5,7] have provided information about the neighbourhood, they have not linked those information with individual properties. For example, Trulia [7] has provided school locations on top of the map, but it does not have the school zone information, and users cannot link individual properties with schools. Secondly, the visual design of these platforms cannot support presentation and comparison of properties in multiple aspects. In this paper, we try to augment existing commercial systems by providing a systematic data abstraction related to real estate and a visual analytics procedure that help users understand the local real estate market and explore the properties by themselves. Table 1 represents the comparison between our data profiles and existing systems. The detail will be illustrated in Section 3.

Many researchers from the data mining community have proposed methods and systems [11–14] related to real estate data. For example, Fu and Xiong [15] designed a discovery system for users to find high-value homes. Shahbazi et al. [16] estimated the investability of real estate properties through text analysis. Zhu et al. [17] proposed a prediction model to measure liquidity in real estate markets based on DOM (Days on Markets). Different with those works, we using interactive visualization to help users first understand the local real estate market, then find properties based on their own preferences. Research on visualization of real estate data dates back to Williamson and Shneiderman's classic Home Finder project [18]. They used dynamic query interfaces to help users explore a real estate database and find houses that meet specific search criteria. They also use sliders to allow limits to be placed on different dimensions such as house price and time to reach oneâs workplace; only those houses satisfying all criteria are represented by red dots on a map. Later, Tweedie et al. proposed Attribute Explorer [19], which added graphical feedback to the sliders themselves. Both works have focused on the efficient query processing part but failed to provide a way for users to compare the result houses/homes. With real estate data as examples, Spence presented different kinds of visualization, including scatter plots, parallel coordinates and iconic representations to compare the attributes of houses in the visualization textbook [20]. Nevertheless, there is still a lack of framework/procedure of how to combine those visualization designs and help users explore real estate data. Closest to our work is a web-based system [21] to visually analyse Hangzhou real estate market data. Comparing to it, our system is built upon a more comprehensive dataset that has more location-based attributes which are critical to users' purchasing decisions and we provide a way for users to explore individual properties.

2.2. Geo-related multidimensional visualization design

Since the regional, transportational and educational profiles centered around the properties/suburbs (detailed in Section 3.1) are all highly related to geography, the visualization of real estate data can be considered as a type of geographically related multidimensional data visualization. The design challenge for visualizing this kind of data is to provide space-efficient presentations that neither clutter visually nor confuse cognitively [22,23].

Multidimensional visualization methods. Multidimensional visualization methods have been proposed to help users discover hidden patterns and understand the complex relationships with respect to multiple dimensions. Those methods are designed based on different graph presentations [24], such as points (scatterplots [25]), lines (parallel coordinates [26]), regions (Heatmap [27]), and combined elements (multivariate glyphs [28]). Researchers have also proposed some combined methods, such as SPPC [29], P-SPLOM [30] to solve real-world problems. Although it is possible to use those existing multidimensional visualization methods to visualize the multiple attributes in our real estate data, they hide important information on how information (such as the housing price) varies across locations. In this paper, we have chosen care-

Table 1
Summary of our data profiles, and the comparison with existing commercial systems.

	Data Examples	Realestate AU [6]		Rightmove UK [4]		Trulia US [9]		Our System		Our Data Sources	Query Examples
		Data*	Index*	Data	Index	Data	Index	Data	Index		
Basic Information	price, address, number of bedrooms	✓	✓	✓	✓	✓	✓	✓	✓	[6]	apartments with 3 bedrooms
	geo-information, address, suburb	✓	✓	✓	✓	✓	✓	✓	✓	[6][47]	properties within an area
	airconditiong, heating, water views	✓	✗	⍻	✗	⍻	✗	✓	✓	[6]	houses with air conditioning
	market Information	✓	✗	✓	✗	✓	✗	✓	✓	[6][48]	houses in a region with an annual growth rate>10%
	images, floorplan	✓	✗	✓	✗	✓	✗	✓	✗	[6]	-
Educational Profile	nearby schools	✓	✓	✓	✓	✗	✗	✓	✓	[49][50]	houses near Camberwell High School
	school zones	✗	✗	✗	✗	✗	✗	✓	✓	[52]	houses inside Camberwell High School
	school ratings/rankings	✗	✗	✓	✗	✓	✗	✓	✓	[51]	houses with school zone of a top 10% high school
Transportational Profile	nearby public transport	✗	✗	✓	✗	⍻	✗	✓	✓	[49][53]	houses near Camberwell train station
	travel time to work	✗	✗	✗	✗	⍻	✗	✓	✓	[53]	houses within 10 mins walk to the nearest train station & 20 mins train to the users' working place
Facility Profile	supermarkets, hospitals, clinics	✗	✗	✗	✗	⍻	✗	✓	✓	[54-56]	where those facilities locate
	distance to the nearest hospital	✗	✗	✗	✗	✗	✗	✓	✓	[49]	houses within 1km to the nearest shopping center
Regional Profile	cultural diversity, median income	✓	✗	✗	✗	⍻	✗	✓	✓	[57]	a region with more people speaking Chinese
Enviromental Information	street view	✓	✗	✓	✗	✓	✗	✓	✗	[49]	houses within an area with good street views
	tree-lined street	✗	✗	✗	✗	✗	✗	✓	✗	[6]	houses within an area with tree-lined streets
	quiet environment	✗	✗	✗	✗	✗	✗	✓	✗	[6]	houses within a quiet area

* Data: whether the system has the data; Index: whether the system has the data linked to each property (as index) and support location-aware queries over the indexed data.

fully among existing visualization methods based on our data/task abstraction to build up our system. Section 5 will illustrate our visualization design choices one by one together with justification in detail.

Geographical visualization. Scientific visualization methods applied to portray geographical information from the 1990s [31] have focused on providing users imagining spatial relationships and was applied to a number of visualization types such as animations, spatial modelling and collaboration, using GIS and virtual reality [32]. In recent years, more information visualization methods have been applied to solve the geographical visualization problem. The design of geoViz (geographical visualization) is not limited to mapping elements based on their exact geo-locations. One of the most notable examples is the design of modern train maps [33,34]. The design challenge behind is to reduce connection overlap while places are still easy to find. With more and more urban data (such as the census data) becoming available, many online visualization tools have been developed in recent years to help citizens understand urban life [35–37]. However, those methods are limited to very few dimensions since position- and size-based visual variables are already used by maps. Therefore, they cannot support visualizing the multiple attributes of housing properties.

Geo-related multidimensional visualization. Some researchers have been working on visualizing geographically related multidimensional data [38–42]. One of the most popular and effective solutions to visualize the multiple attributes in geographic data is to use multiple coordinated views [43]. For example, TripVista [44] is designed with multiple views to analyse the traffic at a road intersection. Attribute Signatures [45] dynamically generates summaries in another view after users select elements on a map view. Zhang et al. [46] proposed a visual analytics framework to explore and compare the impact of geographical variations for multivariate clustering. In our paper, we adopt the design of multiple coordinated views as well. Specifically, besides connecting different views with mouseover/click+highlighting, we directly link parallel coordinates with a geo-coded scatter plot and a colored boolean table, which is able to efficiently visualize multiple attributes (of different types) in a single view.

3. Problem abstraction

In this section, we first describe real estate domain situation, then present the design of our task and data abstraction.

3.1. Domain situation

Our system is mainly targeted at home buyers who are interested in finding properties to live in, as well as real estate agents who help them to do so.

We use Melbourne and Sydney, two of the largest cities in Australia, as examples to understand the domain situation and to collect the data. To understand the information (i) that most home buyers are concerned about, and/or (ii) that they are suggested to pay attention to from domain experts' viewpoint, we first reviewed related articles (from related blogs[1] such as [1,2]) written by real estate domain experts, and got a list of information which is critical to users' property purchase choices; then we discussed with two real estate agents to further refine the data list.

Table 1 presents our conclusion of the information that is critical to home buyers' choices, as well as whether our system has the data and support data index comparing to existing popular real estate websites in Australia [3], the UK [5] and the USA [6].

As shown in Table 1, we have defined five profiles: (1) the basic profile describes basic information of each property itself; (2) the transportational profile is associated with distance and travel time to other places; (3) the facility profile defines nearby facilities, such as supermarkets, shopping centers, etc.; (4) the educational profile contains the information of schools; (5) the regional profile is related to the census data and describes the statistical information of the neighbourhood. For environmental information: it is critical to users' choice as well, but the data is hard to obtain and verify. Therefore, in our system, we extract some of the environmental information from the text description and include them in the basic profile, such as whether the property has mountain views.

Table 1 shows some example data in each profile. Although including everything in our system seems impossible, we claim that

[1] Source: http://www.realestate.com.au/advice/buying/ , https://www.domain.com.au/advice/

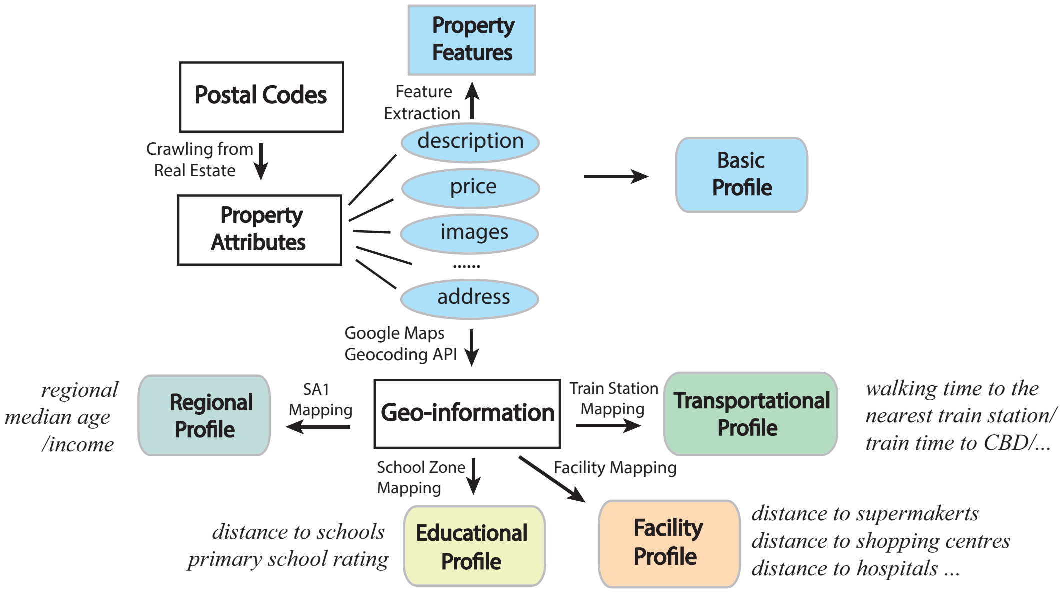


Fig. 1. Data collecting and processing.

the data that is critical to users' choice of properties should at least fit in one of our profiles. For example, nearby restaurants (which our system does not include) should fit in facility profile. The safety of a region can be described by the regional profile. Price per m^2 is in the basic profile.

3.2. Data collection and integration

Fig. 1 shows the process of how we collect data from different sources [47–57] and integrate them to form a location-centered real estate dataset.

We first crawl the basic information of sold properties from real estate websites [3] (**basic profile**). Based on the address, we get the geo-location (latitude and longitude) of each property using Google Maps API [49]. Based on the geo-location, we describe the other four profiles for each property in the following way:

- **Transportational profile**: we calculate (i) the walking time from each property to its nearest train station using Google Maps API [49], and (ii) the travel time between each pair of train stations based on the GTFS (General Transit Feed Specification) data [53]. We are then able to compute efficiently how long it takes from each property to a specific place (such as the userâs workplace) at runtime.
- **Regional Profile**: we get Australian census data from [57] which describes the census information based on different levels of statistical areas in Australia. We then map each property to an SA1 (Statistical Area Level 1, the smallest unit for the processing and release of Australian census data [57]), and then have the information such as median age and median income of the SA1 to describe the region around each property.
- **Educational profile**: we find either the exact school zone [52] that each property belongs to, or map the property with its nearest primary/secondary school. We get the corresponding school rating from [51] which is also linked with individual properties.
- **Facility profile**: we first crawl the information (including geo-locations) of different facilities such as supermarkets [54,55] and hospitals [56]. Then, for each facility, we map each property to the nearest one and calculate the distance between them using Google maps API [49].

Table 2
Statistics of the crawled data.

	Melbourne (VIC)	Sydney (NSW)
# of properties	680,327	757,352
# of train stations	218	373
# of supermarkets	447	507
# of shopping centres	125	194
# of clinics	2673	3536
# of hospitals	469	622
# of public primary schools	1149	1607
# of public secondary schools	282	401
# of private primary schools	465	567
# of private secondary schools	167	200
# of regions	13,339	17,895
# of total dimensions	72	

A summary of all data sources can be found in Table 1, and some data statistics are shown in Table 2.

3.3. Data and task abstraction

3.3.1. Data characteristic analysis

Based on different characteristics, we divide all attributes associated with each property into the following categories:

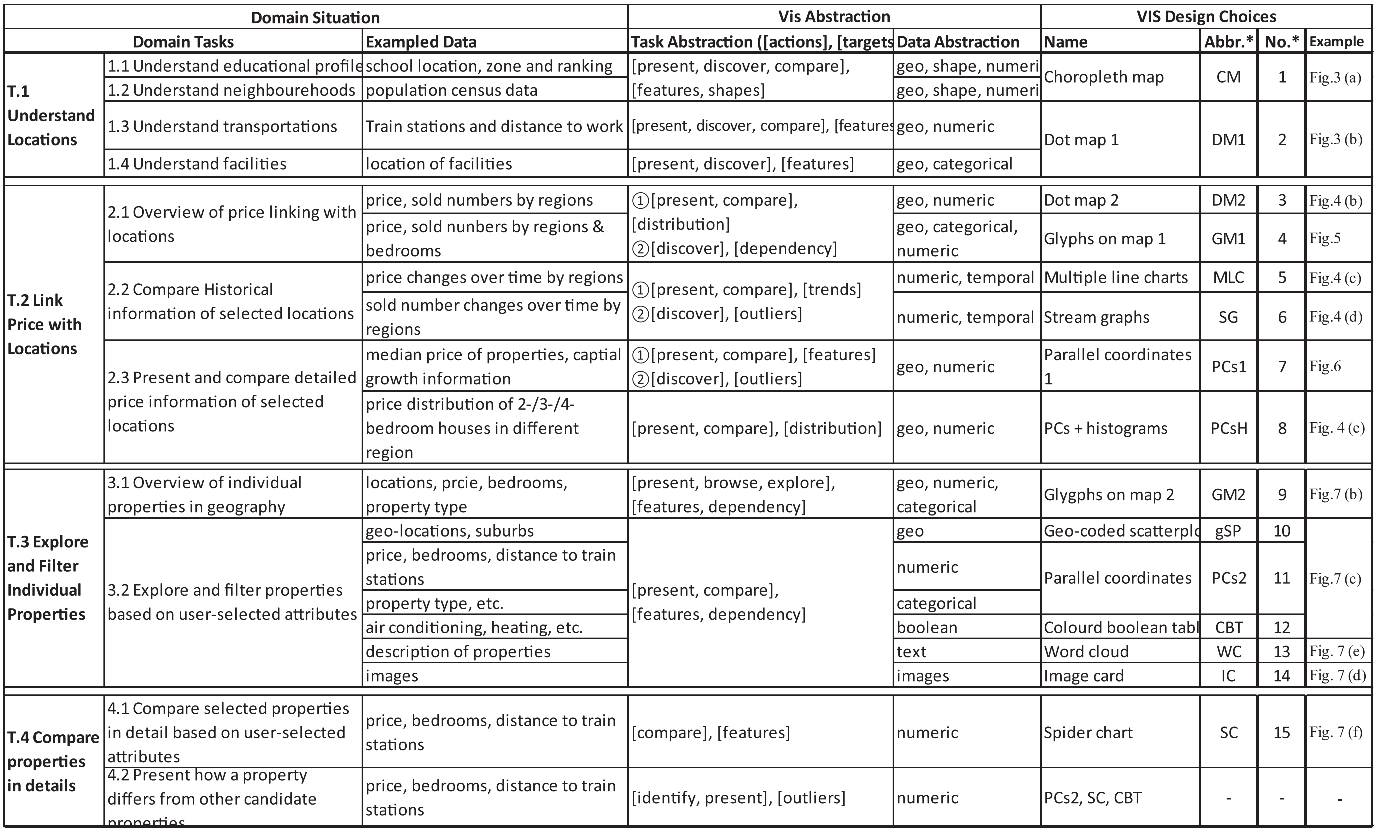
- **Geographical Attributes**, i.e., the geo-information;
- **Numeric Attributes**, such as price, number of bedrooms, the distance to the nearest train station, related secondary school rankings, etc.;
- **Categorical Attributes**, such as property type;
- **Boolean Attributes**, i.e., 34 features generated from the text description, such as whether a property has air conditioning.
- **Other Types of Attributes**, such as text and images.

3.3.2. Task abstraction

Based on [1,2], and as confirmed by domain experts, most home buyers want a property which suits their lifestyle and will grow in value over time so they can take a step up the property ladder in a few years time.

There are two main challenges to define specific tasks. On the one hand, users have various levels of knowledge on the local areas and the local real estate market. On the other hand, users have different individual requirements and their requirements could

Table 3
Domain situation, visualization abstraction and visual encoding idioms.



Domain Situation			Vis Abstraction		VIS Design Choices			
Domain Tasks		Exampled Data	Task Abstraction ([actions], [targets	Data Abstraction	Name	Abbr.*	No.*	Example
T.1 Understand Locations	1.1 Understand educational profile	school location, zone and ranking	[present, discover, compare], [features, shapes]	geo, shape, numeri	Choropleth map	CM	1	Fig.3 (a)
	1.2 Understand neighbourehoods	population census data		geo, shape, numeri				
	1.3 Understand transportations	Train stations and distance to work	[present, discover, compare], [feature	geo, numeric	Dot map 1	DM1	2	Fig.3 (b)
	1.4 Understand facilities	location of facilities	[present, discover], [features]	geo, categorical				
T.2 Link Price with Locations	2.1 Overview of price linking with locations	price, sold numbers by regions	①[present, compare], [distribution] ②[discover], [dependency]	geo, numeric	Dot map 2	DM2	3	Fig.4 (b)
		price, sold nunbers by regions & bedrooms		geo, categorical, numeric	Glyphs on map 1	GM1	4	Fig.5
	2.2 Compare Historical information of selected locations	price changes over time by regions	①[present, compare], [trends] ②[discover], [outliers]	numeric, temporal	Multiple line charts	MLC	5	Fig.4 (c)
		sold number changes over time by regions		numeric, temporal	Stream graphs	SG	6	Fig.4 (d)
	2.3 Present and compare detailed price information of selected locations	median price of properties, captial growth information	①[present, compare], [features] ②[discover], [outliers]	geo, numeric	Parallel coordinates 1	PCs1	7	Fig.6
		price distribution of 2-/3-/4-bedroom houses in different region	[present, compare], [distribution]	geo, numeric	PCs + histograms	PCsH	8	Fig. 4 (e)
T.3 Explore and Filter Individual Properties	3.1 Overview of individual properties in geography	locations, prcie, bedrooms, property type	[present, browse, explore], [features, dependency]	geo, numeric, categorical	Glygphs on map 2	GM2	9	Fig.7 (b)
	3.2 Explore and filter properties based on user-selected attributes	geo-locations, suburbs	[present, compare], [features, dependency]	geo	Geo-coded scatterplo	gSP	10	Fig.7 (c)
		price, bedrooms, distance to train stations		numeric	Parallel coordinates	PCs2	11	
		property type, etc.		categorical				
		air conditioning, heating, etc.		boolean	Colourd boolean tabl	CBT	12	
		description of properties		text	Word cloud	WC	13	Fig. 7 (e)
		images		images	Image card	IC	14	Fig. 7 (d)
T.4 Compare properties in details	4.1 Compare selected properties in detail based on user-selected attributes	price, bedrooms, distance to train stations	[compare], [features]	numeric	Spider chart	SC	15	Fig. 7 (f)
	4.2 Present how a property differs from other candidate properties	price, bedrooms, distance to train stations	[identify, present], [outliers]	numeric	PCs2, SC, CBT	-	-	-

* Abbr. (Abbrations) and No. are used in Fig. 4 and Table 5; No. is corresponding to the VIS Design Number in Section 5.

change during the process of discovering and knowing better about the local real estate market.

Based on the above observation, we define four levels of basic tasks, as shown in Table 3. Users with different background of the local real estate market and different requirements can use different kinds of combinations of the four basic tasks to discover their preferred properties. For example, a user who does not know much of the local real estate market and/or who is not sure of his requirement might go through all the four tasks to find candidate properties; while an advanced home buyer might skip some of the steps and find the properties based on his requirements (Examples will be illustrated as case studies in Section 7.2, with a summary of different cases in Table 5).

Each basic task includes several sub-tasks, as shown in Table 3. The table also presents the corresponding data to support the tasks, as well as the process of how we analyse the tasks and the corresponding data based on Munzner's *what-why-how* model [58].

4. System overview

In this section, we first define some design maxims, then design system procedure.

4.1. Design maxims

Based on our problem abstraction and the characteristics of our targeted users, we define three design maxims (DMs) to help us design the system and choose proper visualization idioms.

DM1 **Incremental learning**. Our targeted users are home-buyers, who might have a very limited knowledge of the local real estate market. This also greatly affects their ability to describe their requirements. For example, since a person does not know what kind of a property he can buy with one million dollars in Melbourne, he will also easily feel confused in defining the properties that he prefers. Therefore, our system should allow such less informed users to learn the about local real estate market step by step.

DM2 **Intuitive and simple designs**. Our users, as home-buyers, might not know much about data analytics/visualization. While being informative, our visualization design choices should be also intuitive and simple for end users to easily understand. We should be careful with complex visualization designs; and if we do use them, we should provide detailed guidelines.

DM3 **Consistent visualization designs**. Since we may involve several different visualization design choices, we should try to use a consistent set of visual encodings in order not to confuse users. For example, colours should have the same meaning in different visualization design choices.

4.2. System pipeline

Based on the task abstraction and the DMs, we design an interactive procedure (Fig. 2) to help different levels of home-buyers incrementally learn about the local real estate market and find their preferred properties.

First, models in our system refer to different descriptive statistics. We have defined different abstracted models based on our basic tasks, as shown in Fig. 2. For example, we have defined different granularities [23] of summarized information at the suburb level. At the high granularity, we calculate the median price of properties; at the medium granularity, we compute the median price of, for example, 3-bedroom houses or 2-bedroom units; at

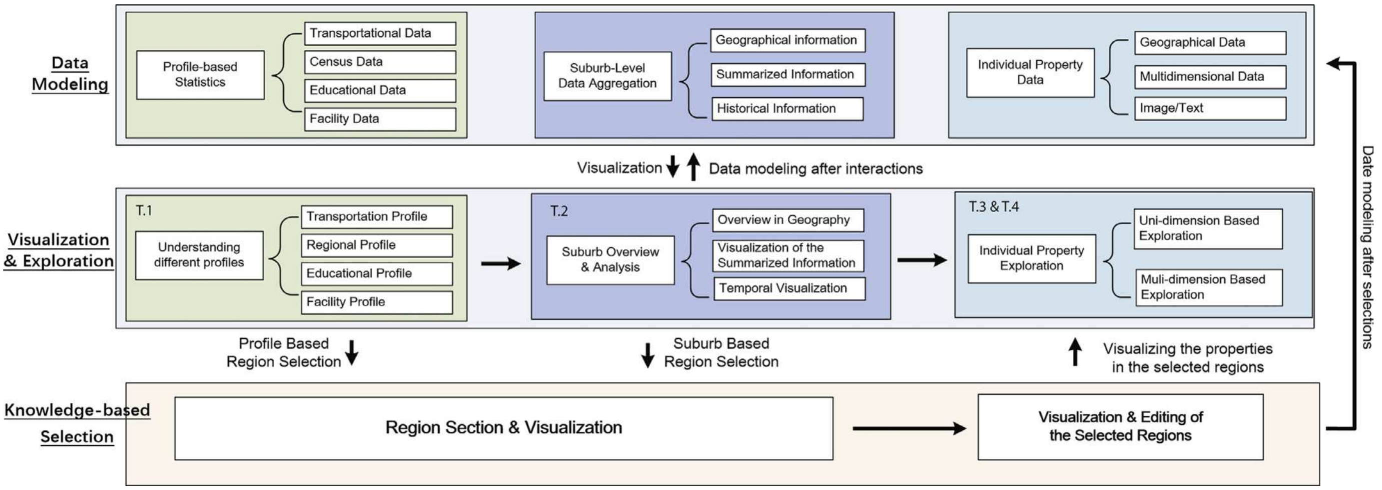

Data Modeling
Profile-based Statistics
Transportational Data
Census Data
Educational Data
Facility Data
Suburb-Level Data Aggregation
Geographical information
Summarized Information
Historical Information
Individual Property Data
Geographical Data
Multidimensional Data
Image/Text
Visualization
Data modeling after interactions
Date modeling after selections
Visualization & Exploration
T.1
Understanding different profiles
Transportation Profile
Regional Profile
Educational Profile
Facility Profile
T.2
Suburb Overview & Analysis
Overview in Geography
Visualization of the Summarized Information
Temporal Visualization
T.3 & T.4
Individual Property Exploration
Uni-dimension Based Exploration
Muli-dimension Based Exploration
Profile Based Region Selection
Suburb Based Region Selection
Visualizing the properties in the selected regions
Knowledge-based Selection
Region Section & Visualization
Visualization & Editing of the Selected Regions

Fig. 2. The proposed interactive procedure to visually analyse the real estate data.

the low granularity, we summarize the detailed price distribution of all 2-bedroom units, all 3-bedroom houses, etc.

Secondly, we design three levels of visualization, as profile-based region visualization (T.1, Section 5.1), suburb-based visualization (T.2, Section 5.2) and property-level visualization (T.3 & T.4, Section 5.3) to answer the four basic domain tasks. All the visualization designs are interconnected with each other. On the one hand, we use multiple coordinated views to visualize different information in the same screen, and designs are coordinated with each other with interactions like mouseover/click+ highlighting. On the other hand, designs cross different levels link with each other based on consistent visual encoding idioms and through user interactions such as region selections. For example, a user can select some school zones in profile-based region visualization, or select several suburbs in the suburb-based visualization, and then move to property-level visualization to explore the properties in those school zones and/or those suburbs; also, the user can search for all the properties in the same school zone in the property-level visualization.

5. Visualization design and justification

In this section, we first introduce our visualization design choices based on our task and data abstraction. Then, we describe how we have involved users in our iterative design process to make the system fit for general users.

5.1. Profile-based region visualization (T.1)

To help home buyers understand different lifestyles in different areas (T.1), we present each of our four profiles using choropleth maps or dot maps.

VIS Design 1: Choropleth map. We use Choropleth map to visualize the *educational and regional profiles*. For the educational profile, we visualize the school zone on top of the Google Maps. Schools with a better rating will be mapped with a darker colour (Fig. 3(a)). Similarly, regional exploration presents census information on top of the map. Users can select different information such as median age, median income, the percentage of people coming from China, India, etc.

VIS Design 2: Dot map 1. Since *transportational profile* does not contain shape information, we draw circles on top of the map to represent the location, with colour illustrating the travel time to the city or to users' working place (Fig. 3(b)). Similarly, we use different colours to represent different types of facilities (supermarkets, or shopping centers) for *facility profile*.

Justification. We choose the Choropleth map which is perhaps the most commonly used and effective method for summarized information with geo-shapes; since the data value (e.g. school ranking) is numeric, we choose colour luminance to represent the value. We choose colour luminance, instead of the area size of the circle to represent the travel time, because otherwise, it might confuse users whether a larger size means a longer time or not. An alternative for visualizing different types of facilities is to use different shapes, but we choose colour hue considering the effectiveness ranking of different visual channels [58].

5.2. Suburb-level visualization (T.2)

At the suburb level, we have three main coordinated views: the Google Maps view, the historical and the multidimensional view (Fig. 4).

5.2.1. The Google Maps view (T.2.1)

We design two levels of data abstractions to provide users an overview of how property prices in different suburbs vary across locations.

VIS Design 3: Dot map 2. To visualize the median price and number of properties in each suburb, we draw circles on top of Google Maps (Fig. 4(b)). The position of each circle corresponds to the centre location of a suburb in geography. The size and colour of the circle represent the number of sold properties in this suburb, and the median price of those properties, respectively.

VIS Design 4: Glyphs on map 1. Since prices could be easily affected by the number of bedrooms (or building size in some countries), we further divide the properties in a suburb into different groups based on the number of bedrooms. Therefore, we depart the circle into different sections to display one more dimension. As shown in Fig. 5(a), different sections represent different numbers of bedrooms, and the number of properties that have a particular number of bedrooms are presented is size of the section. The color is based on the median price of the properties in each section. Fig. 5(b) shows the result when we only select 3-bedroom properties.

Justification. The two visualization designs are consistent by sharing the same visual channels, i.e., colour and size. We design a color scheme to present the median price by using different color saturation and two different colour hues (blue and red). Such colour design is widely used in other applications such as altitude

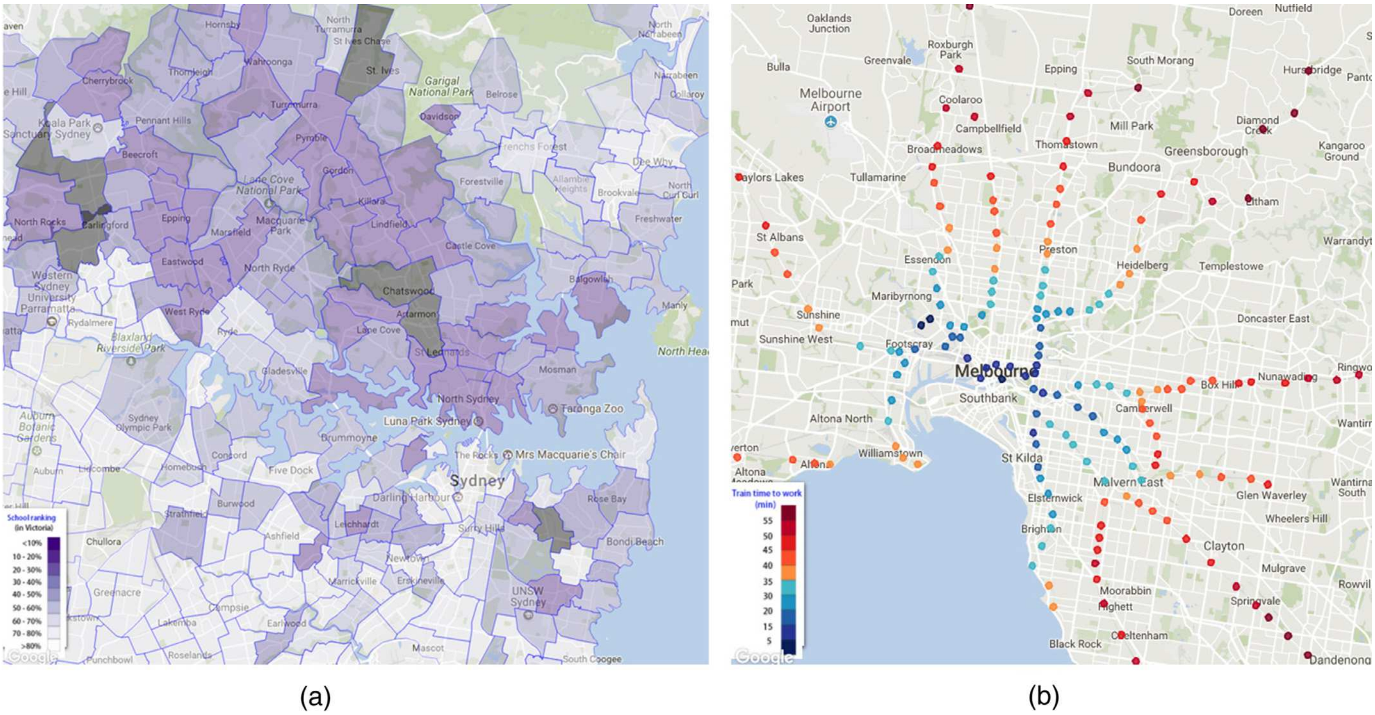


Fig. 3. Design of profile-based region visualization (a) VIS design 1: Choropleth map of Sydney's educational profile - primary school exploration; (b) VIS design 2: dot map 1 based on Melbourne's transportational profile - train time exploration.

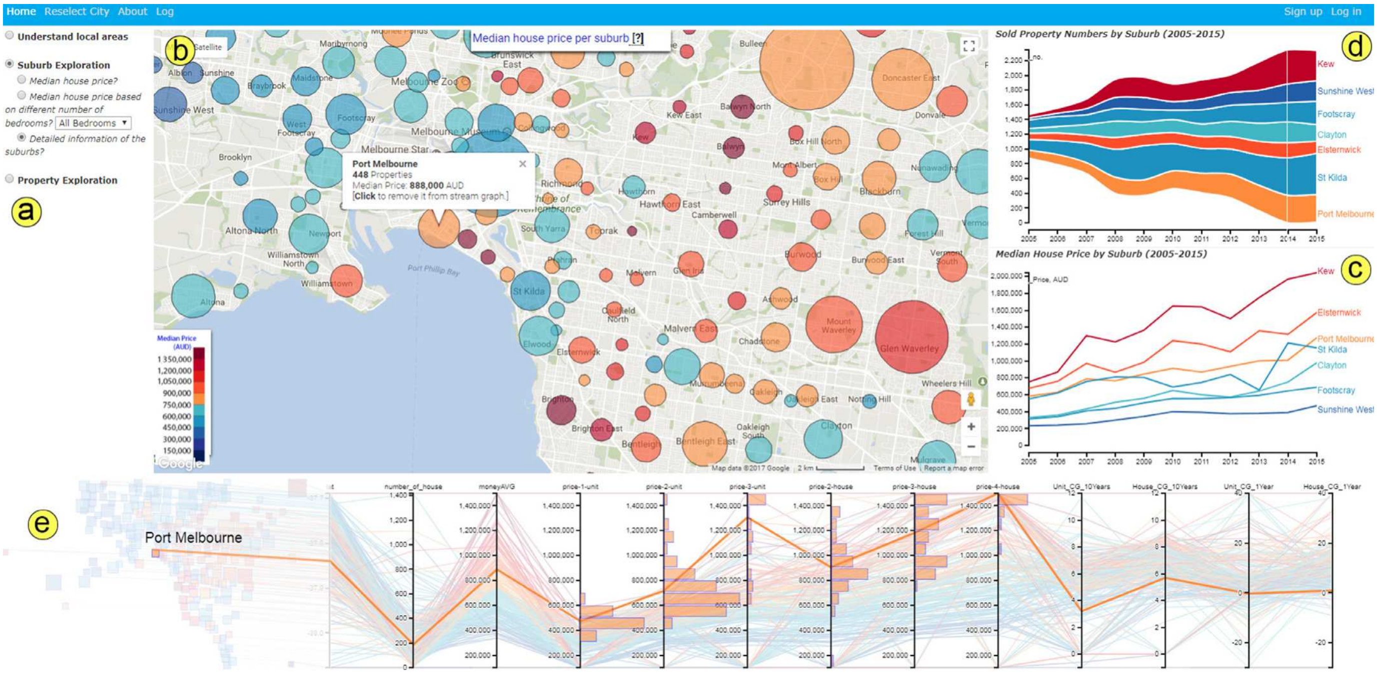


Fig. 4. Design of the suburb-level visualization (a) navigation bar; (b) map view (a dot map); (c) historical view - multiple line graphs; (d) historical view - stream graphs; (e) multidimensional view.

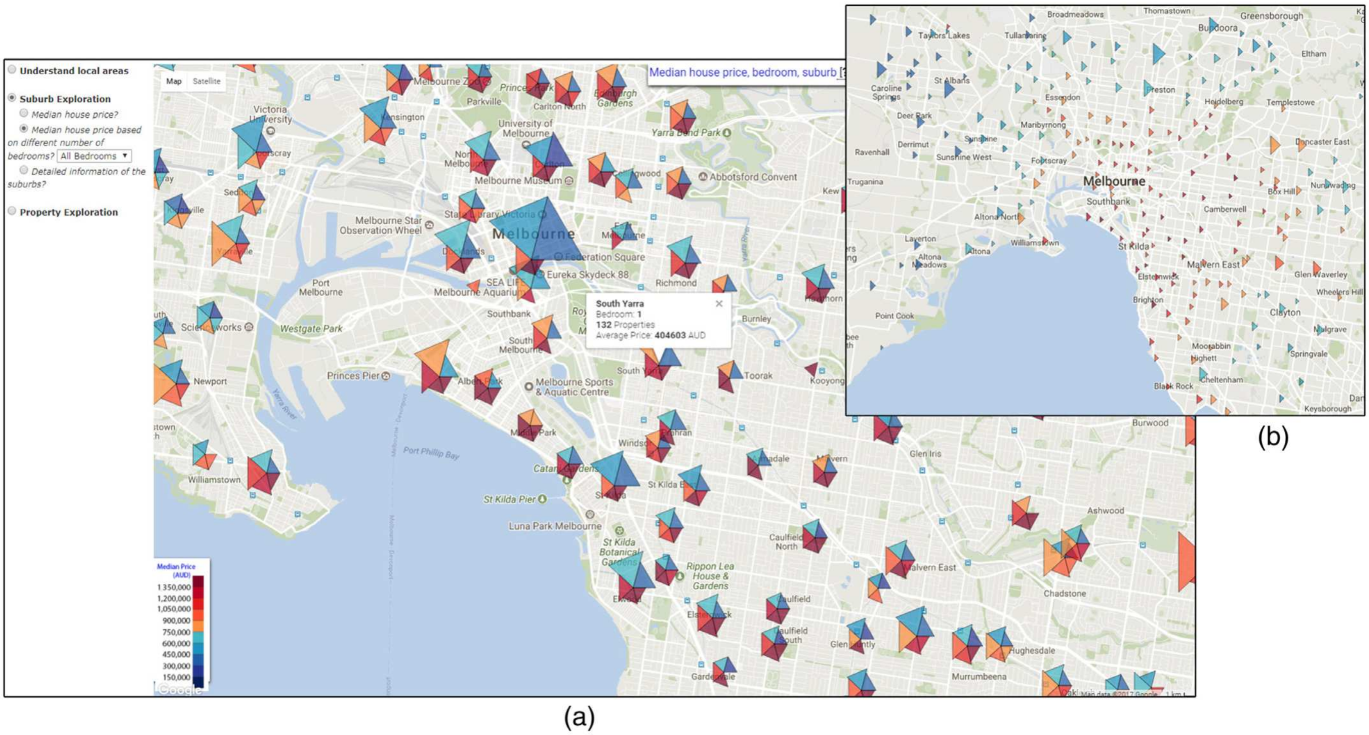
maps [59]. The color design considers the recommendation from [60]. Instead of using choropleth maps which show boundaries of suburbs, we choose circles and glyphs on maps, so that we could use more retinal variables, i.e., size and shapes.

5.2.2. The historical view (T.2.2)

This view presents how median prices (multiple line charts) and sold numbers (stream graphs) of houses/units in each suburb change over time. Those suburbs can be selected by clicking on the circles (i.e. suburbs) in the Google Maps view (Section 5.2.1), or by filtering the attributes in the multidimensional view (Section 5.2.3).

VIS Design 5: Multiple line graphs. As shown in Fig. 4(c), the horizontal axis represents time, i.e., from 2005 to 2015, and the vertical axis represents median price. Each line represents the price changes of one suburb.

VIS Design 6: Stream graphs. As shown in Fig. 4(d), each suburb is represented as a stream. The horizontal axis and colour have



○ Understand local areas
● Suburb Exploration
○ Median house price?
● Median house price based on different number of bedrooms? All Bedrooms ▾
○ Detailed information of the suburbs?
○ Property Exploration

Median house price, bedroom, suburb

South Yarra
Bedroom: 1
132 Properties
Average Price: 404603 AUD

Fig. 5. Illustration of VIS Design 4: glyphs on maps 1: (a) Median house price based on different number of bedrooms; (b) only 3-bedroom properties are selected.

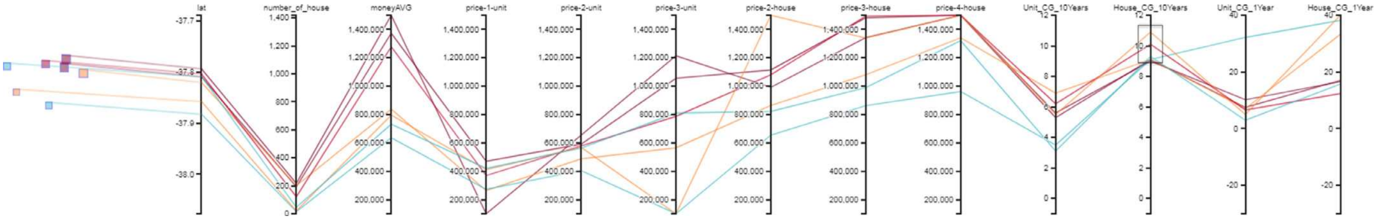


Fig. 6. Illustration of VIS Design 7: PCs.

the same meaning with that in multiple line graphs. The width of the stream indicates the number of sold house/unit in that particular year.

Justification. Two of the design choices are consistent as they share the same meaning of horizontal axis and colours. The colour, which means median price, is consistent with all the other designs in the system as well. We choose stream graphs to visualize the sold number, since the sum of the sold number of a group of suburbs make sense, and this is somehow consistent with the design in the Google Maps view (Section 5.2.1) since the width of a stream and the area (which can be both considered as size-related visual channels) both mean the numbers. On the other hand, if we use stream graphs to visualize the median price change, the height of stream graphs will easily confuse people, since the sum of median prices does not make sense.

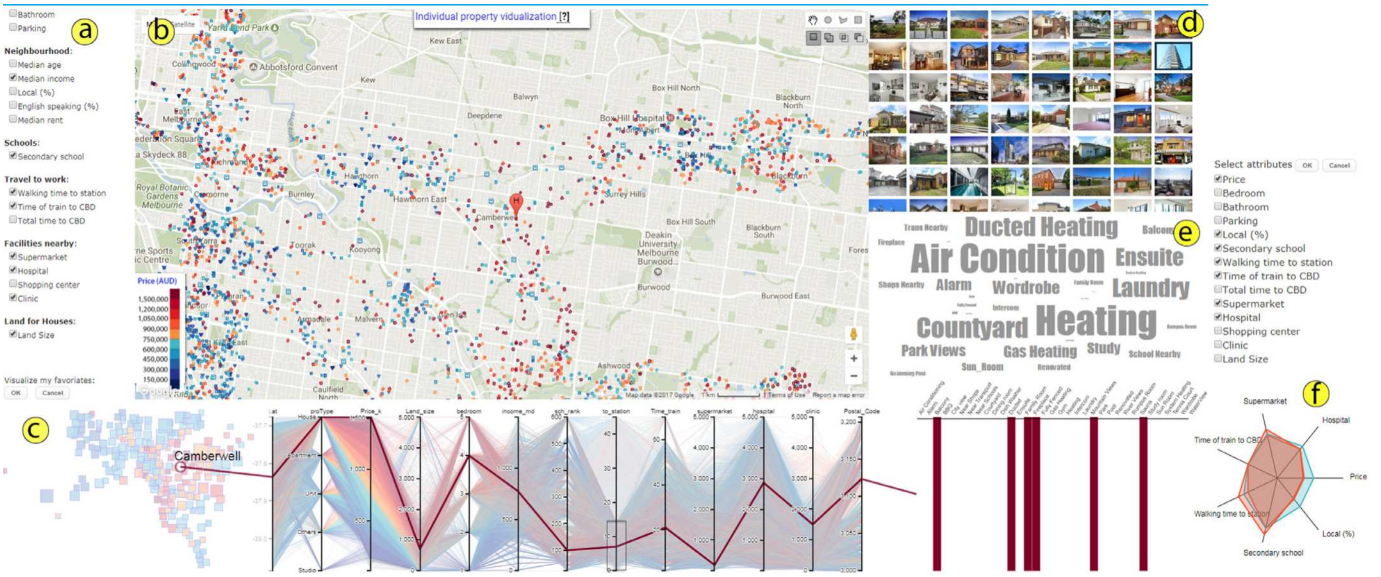
5.2.3. The multidimensional view (T.2.3)

The multidimensional view presents detailed statistical information of each suburb.

VIS Design 7: Parallel coordinates (PCs) 1) We use parallel coordinates (Fig. 6) to visualize suburb-based statistics, such as the median price of 2-bedroom houses, the median age in this suburb, etc. For users to easily compare the price differences of different property types and bedroom numbers, we have all the maximal and minimal number of price-related axes consistent, such as 1.5M and 0.2M. As shown in Fig. 6, we connect the left of parallel coordinates with a geo-coded scatter plot, to allow users to easily link each suburb represented as lines in parallel coordinates with the one in the Google Maps view.

VIS Design 8: PCs + Histograms. We have a novel design of drawing histograms on top of parallel coordinates to show the detailed statistics of a group of properties (e.g. 2-bedroom houses in the suburb) besides the original median number (Fig. 4(e)). For example, we first show the median price of 2-bedroom houses for each suburb in parallel coordinates. When a user selects one suburb to highlight (from any views), we draw a histogram on top of parallel coordinates to show the detailed price distribution of all 2-bedroom houses in this suburb. By highlighting each suburb one by one, users could easily get the differences of property characteristics and the price distribution in each suburb.

Justification. One alternative method to visualize the multidimensional information in a limited screen size is to use Heatmap [27]. We choose parallel coordinates since it is easy to have filtering on its own axes, and it is easy to be combined with other methods, as what we have designed, with histograms. Such design of drawing histograms on top of parallel coordinates is novel and effective to show the detailed information of the selected suburb. It is different with Hansen's work [61] combining parallel coordinates and histograms. In particular, Hansen has provided a histogram at each parallel coordinate axis to indicate the instances of data associated with line values, i.e., the histogram value is a summary of what you can see from the axis; while histogram in our design



□Bathroom
□Parking

Neighbourhood:
□Median age
☑Median income
□Local (%)
□English speaking (%)
□Median rent

Schools:
☑Secondary school

Travel to work:
☑Walking time to station
☑Time of train to CBD
□Total time to CBD

Facilities nearby:
☑Supermarket
☑Hospital
□Shopping center
☑Clinic

Land for Houses:
☑Land Size

Visualize my favoriates:
OK Cancel

Select attributes OK Cancel
☑Price
□Bedroom
□Bathroom
□Parking
☑Local (%)
☑Secondary school
☑Walking time to station
☑Time of train to CBD
□Total time to CBD
☑Supermarket
☑Hospital
□Shopping center
□Clinic
□Land Size

Fig. 7. Design of property-level Visualization. (a) navigation bar, where users can select attributes to explore in the multidimensional view; (b) google maps view; (c) multidimensional view, filtered as properties within 15-min walk to the nearest train station; (d) image card; (e) word cloud; (f) spider chart.

is an expansion of the existing median price and shows more detailed distribution information related to the axis.

5.3. Property-level visualization (T.3&T.4)

Our property-level visualization (Fig. 7) includes the Google Maps view, the multidimensional view, the Image Card view and the Word Cloud view. By default, we only show the Google Maps view. Users can select factors that they care about to visualize them in the multidimensional view, which will also enable the other views.

5.3.1. The google maps view (T.3.1)

The Google Maps view provides users how properties distribute geographically and how attributes of properties vary across locations.

VIS Design 9: Glyphs on map 2. Each property is mapped as a regular polygon (Fig. 7(b)). The color of the polygon is determined by the price of a property. The shape of a polygon is defined by bedroom numbers, i.e., a triangle and a square represent a property that has one bedroom and two bedrooms, respectively.

We define a set of **selection tools** to allows users directly select regions that they are interested in from the Google Maps view (Fig. 7(b) - right top corner). First, we allow users draw polygons, circles and self-defined shapes (polygons) to select regions in the Google maps view. Secondly, inspired by Adobe Photoshop CC,[2] we define four selection operators as 1) *new*, 2) *add to*, 3) *intersect with* and 4) *subtract from*, so a user can select regions freely from the Google Maps view.

Justification. In this map view, we try to provide a multi-scale visual encoding scheme [23] to visualize different levels of details with regard to different levels of user interactions. We choose the colour to represent the most important information since it is still recognizable even when the map is zoomed out. Using colour to represent price also makes sure the design is consistent with the suburb level (Section 5.2). We choose polygons instead of circles to represent each property, so we could present another dimension with the shape of polygons; also, drawing polygons is more efficient than drawing circles. We use shapes rather than the size of properties, since visual clutter is less likely to happen.

5.3.2. The multidimensional view (T.3.2)

The multidimensional view presents and compares all the numerical, categorical and boolean data linked to each property.

VIS Design 10: Parallel coordinates 2. As shown in Fig. 7(c), parallel coordinates is used to display numeric and categorical attributes. It has two main functions in our system. First, it provides the comparison of different properties in multiple dimensions; by selecting and highlighting, users can also easily understand how one property is different with the others. Secondly, it provides an interactive way for users to do filtering of properties of their own preferences based on the attributes that they are concerned about (which can be selected from the navigation bar) .

VIS Design 11: Geo-coded scatterplot. We map each property as a point in a 2D coordinate system based on its geographic location (Fig. 7(c), left), similar to a dot map [62]. This geo-coded scatter plot connects the Google Maps view and the multidimensional view. To avoid visual confusion, we set the first axis in parallel coordinates as *latitude*, which greatly reduces the chance of intersection of connected lines.

VIS Design 12: Coloured boolean table. As shown in the right side of Fig. 7(c), it is a variation of Heatmap [27] to visualize boolean attributes (such as whether the property has air conditioning or not). Each column represents a feature, and each row is corresponding to a property. The feature of a property that has a value of *yes* will be filled with the colour that is the same as that in parallel coordinates and the geo-coded scatterplot. This design is directly connected with parallel coordinates' right end.

Justification. We have directly linked three different designs to visualize multiple attributes associated with each property. First, we choose parallel coordinates out of a bunch of multidimensional visualization methods since it allows users to filter out properties based on different attributes in a straightforward way, which also helps our filtering function outperform existing commercial systems. Secondly, we use a geo-coded scatter plot directly connecting parallel coordinates. We believe that it makes users more easily link the multidimensional view with the Google Maps view. We have gained feedback from some users that, the geo-coded scatter plot makes them easier to understand the meaning of a particular

[2] http://www.adobe.com/

polyline in parallel coordinate which represents different attributes of a property. Thirdly, with regard to visualizing boolean data, one alternation is using parallel sets [63]. We choose the boolean table since users are more concerned with whether one property has one particular feature other than the relations among the features; and for the information of how many properties share the same features, we will show it in the word cloud view.

5.3.3. The image card view (T.3.3) & word cloud view (T.3.4)

We provide the Image Card view and the Word Cloud view to display the additional image and textual information.

VIS Design 13: Image card. HomeSeeker by default displays 200 pictures of 200 properties (Fig. 7(d)). The actual number of pictures depends on the screen resolution. Also, a vertical scroll bar is provided. After users select a property from the Image Card view or from other views, we display more pictures of the property.

VIS Design 14: Word cloud. We visualize the most common features of the properties with word cloud (Fig. 7(e)) that the user has selected in the Google Maps view or in the multidimensional view. When users click on one feature (in the Word Cloud view), the properties that have the feature are highlighted on the map and in parallel coordinates, while the corresponding column in the coloured Boolean table is also highlighted. The Word Cloud view presents an overview of the general features of the properties that users are looking at, and help them understand a group of properties in a particular region (selected from the Google Maps view) or share the similar features (selected from the multidimensional view).

Justification. We have two additional views to visualize additional image and text information. For the images, we provide a basic function to display one image associated with each property. We have not provided any sorting function since the main purpose of this view is to provide users some basic idea of how each property looks like. When the number of properties decreases, the image can be enlarged to increase space utilization rate. To keep consistency with other views and avoid confusion, the colour of word cloud is set into the black.

5.3.4. The spider chart view (T.4)

The spider chart view provides users a detailed comparison among a limited number of candidate properties.

VIS Design 15: Spider chart. We use a spider chart [64] (Fig. 7(f)) to compare the properties in the user's favourite list based on the attributes that the user is concerned about. First, we have reversed the value in some axes so that the property that is placed more outside always means it is better in the corresponding dimension. Secondly, we have pre-defined the maximum and minimum value of each axe in the spider chart as the maximum and minimum value of each dimension in the whole data table. Such design allows users to understand whether the differences between two properties are huge or not.

Justification. The main purpose of the spider chart view is to compare the properties in the user's favourite list. The first controversial part of the design is the duplicate information as presented in parallel coordinates. We argue that the spider chart augments the parallel coordinates to show a more clear comparison of properties in the aspects selected by users. Another problem of the spider chart is the limited number of properties it supports since we display very limited number of properties in the spider chart. If there are more properties in the user's favourite list, he/she can still compare those properties in parallel coordinates. The third problem is the limited number of attributes that the spider chart supports. Normally, the spider chart can display at most 12–16 properties, which is able to meet most users' requirements.

5.4. Iterative redesign with users involved

Since HomeSeeker is targeted at general home buyers, one of our concerns is that some of the visualization designs might be difficult to understand by users that have no visualization background. To overcome the problem, we have involved users in the process of our visualization design. After a preliminary implementation of the visualization design, we asked eight users (all potential home buyers) to comment on our visualization choices. As shown in Table 4, we had two questions (Q0-A&B) to ask about reviewers' expertise on visualization and the local real estate market. Then we asked users to try our system. For each visualization design, we asked users to answer Q1 and Q2 by themselves. After that, we answered their questions and gave them some guidelines (that would be easier given by videos later), and then asked them to answer the remaining three questions (Q3–5).

The result of user feedback (Fig. 8) allows us to understand which part of the visualization design is difficult for non-experts (on visualization) to understand, which helps us to revise our design and also to design user guidelines in the future. For example, stream graphs is a controversial design. Most users are not familiar with it and are confused with the vertical axis, since they are more familiar with line charts and do not understand the meaning of the upper and lower bound of each stream. As a result, though we have argued (in Section 5.2.2) that stream graph is more appropriate to visualize the historical information of property sold numbers, we use multiple line charts to visualize it by default and allow users to change it to stream graphs as an advanced function. Another design choice that users find hard to understand is parallel coordinates. However, all involved users find it useful after they understood it (through guidelines), with a user particularly commented that "*I think parallel coordinates is useful since it allows me to filter out the properties and I can see the result refreshed immediately after my interaction. Also, since I can see that there are more properties that have a distance within 10-minute walk to the nearest train station, I do not need to worry about this attribute when I am doing filtering on other attributes.*" Based on the feedback from users, we by default disable the multidimensional view for both suburb-level and property-level exploration, and we will give detailed instruction to guide users interact with the multidimensional view if they are interested.

6. System implementation

We developed HomeSeeker[3] [65] in HTML5 and Javascript, with the library of d3.js and Google Maps JavaScript API. The database that we used is MySQL 5.6.17, and we used PHP to connect the database.

Particularly, in the Google Maps view, we sample properties if there are too many properties (>5000), and our index system based on the address enables that the sampled properties are in different areas. To guarantee the efficiency, we use progressive rendering techniques [66] in parallel coordinates.

7. Experiments and discussion

In this section we present four case studies subsequent to the description of the experimental dataset. This is followed by domain expert interviews and a general discussion.

7.1. Dataset description

As discussed in Section 3, we have crawled the real estate data from different channels and integrated them to form a comprehen-

[3] The demo is available at http://115.146.89.158/

Table 4
Likert scale question design.

General Questions:
Q0-A. Are you familiar with visualization? (1–5, 1: not familiar at all, 5: very familiar)
Q0-B. Are you familiar with the Melbourne real estate market? (1–5, 1: not familiar at all, 5: very familiar)

For each visualization design choice:
Q1: Are you familiar with this kind of presentation? (1–5, 1: not familiar at all, 5: very familiar)
Q2: Is it easy for you to guess the meaning of it? (1–5, 1: very hard, 5: very easy)
Q3: Do you think it is easy to understand/learn? (1–5, 1: very hard, 5: very easy)
Q4: Do you think it is interesting? (1–5, 1: not interesting at all, 5: very interesting)
Q5: Do you think it is useful? (1–5, 1: not useful at all, 5: very useful)

Table 5
User requirements of the four cases and corresponding tasks and visualization design choices.

User	Buget	Requirement	Task	T.1		T.2						T.3 & T.4						
			VIS Design	VD1	VD2	VD3	VD4	VD5	VD6	VD7	VD8	VD9	VD10	VD11	VD12	VD13	VD14	VD15
			Sequence	CM	DM1	DM2	GM1	MLC	SG	PCs1	PCsH	GM2	gSP	PCs2	cBT	WC	IC	SC
A	1,200,000	a 3-bedroom property within a good public secondary school zone and close to facilties	T.1 -> T.3 -> T.4	✔	✔							✔	✔	✔	✔	✔	✔	✔
B	900,000	a property in a suburb with more Chinese people, and might increase in value	T.2 -> T.1 -> T.3 -> T.4	✔		✔		✔		✔	✔	✔	✔	✔	✔	✔	✔	✔
C	800,000	invest a property that makes more profits	T.2 -> T.3 -> T.4			✔	✔	✔	✔	✔	✔	✔	✔	✔	✔		✔	✔
D	-	new to the city, want to explore the local real estate market	T.1 -> T.2 -> T.3	✔	✔	✔	✔	✔	✔	✔	✔	✔		✔		✔		

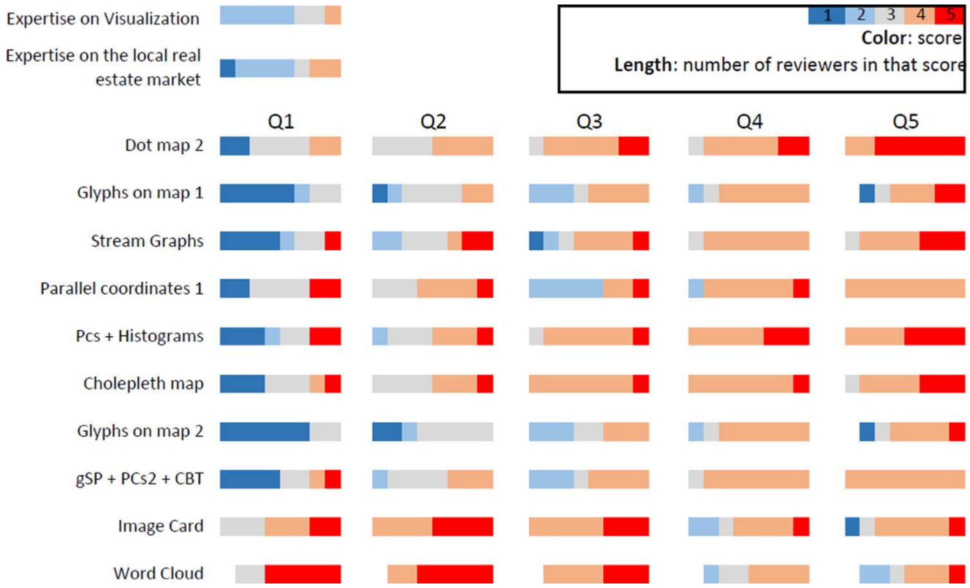


Fig. 8. User feedback based on some of the VIS design choices.

sive location-aware dataset. We use Melbourne dataset to present case studies in this section.

7.2. Case studies

We present four typical cases of real home buyers/investors using HomeSeeker to find properties. Table 5 describes the budgets and requirements from the four users, and summarizes how users can use our basic tasks to finish different tasks and the involved visual designs.

7.2.1. Case I. From school zone to discover properties

User A had a budget of 0.8 million dollars to buy a 3-bedroom property in the Eastern Melbourne suburbs where most of her friends live. As a parent, she preferred the property within a good public secondary school and close to facilities like supermarkets, shopping centres and clinics.

She first went to profile-based region visualization (**T.1**) and selected secondary public school exploration. After discovering and comparing the school ratings, she selected those schools having a good academic performance (within the top 20% in Melbourne), including Mckinnon Secondary College, Balwyn High School, etc. (Fig. 9(a)).

She then clicked on property exploration (**T.3**) and found that there were too many available properties. After selecting those attributes that she was interested, she enabled the multidimensional view. She then filtered out the properties based on her original requirements (houses, below 0.8 million, 3-bedroom, within 500 m to the nearest supermarket, 2 km to the near shopping center and 1 km to the nearest clinic (Fig. 9(b)-[1–6]). After comparing the attributes of the remaining properties, she realized that she also wanted to live in a region with more local people (Fig. 9(b)-7). Since her son might need to go to the city by train at weekends, the desired property needed to be near the train station as well

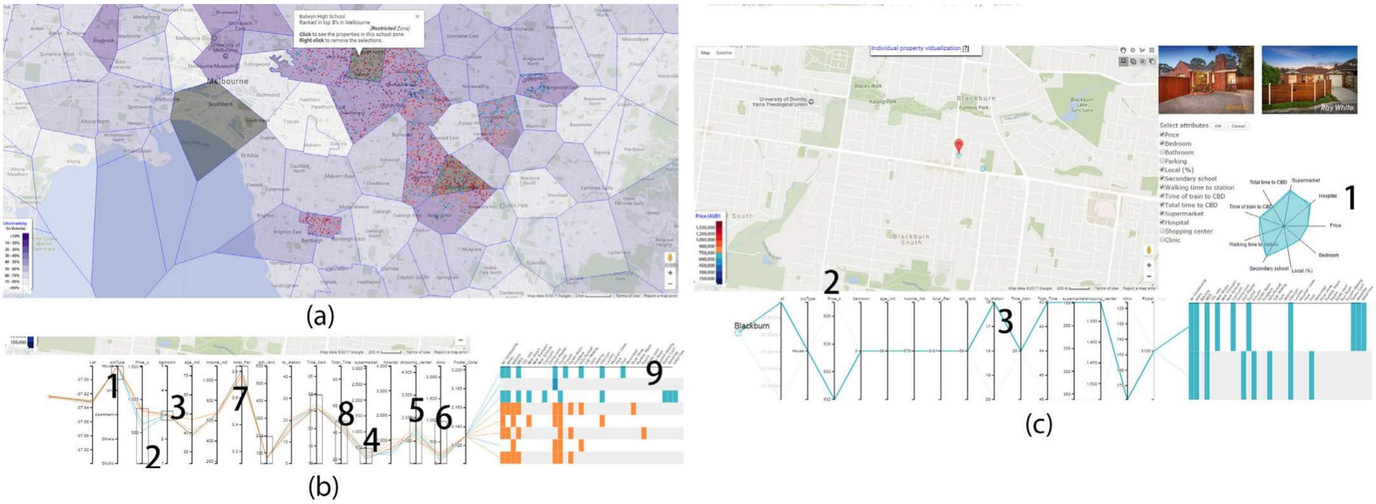


Fig. 9. Illustration of Case I: from school zone to discover properties. (a) selecting properties within the zone of good public secondary schools; (b) selecting properties based on the user's requirements in the property-level exploration; (c) comparing candidate properties in detail.

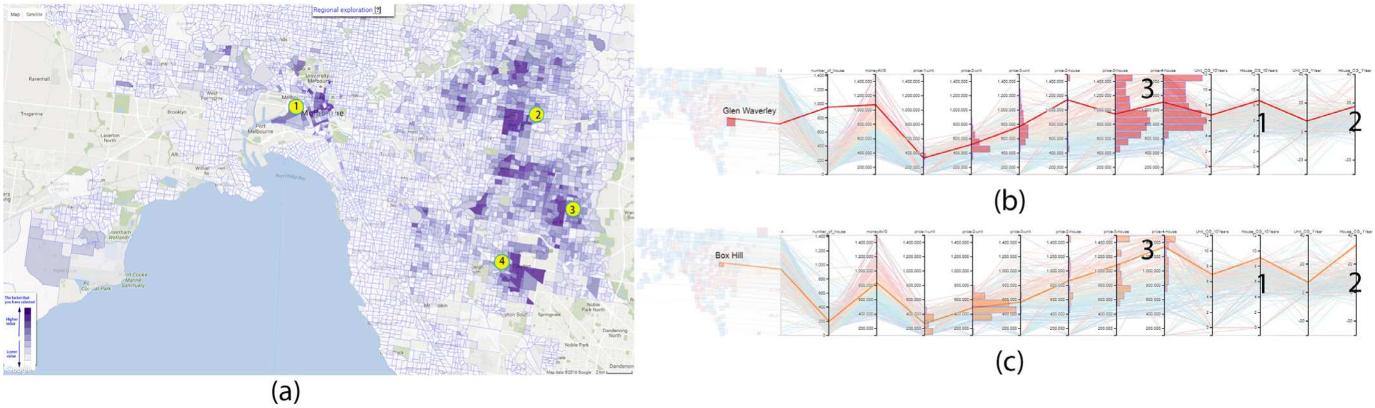


Fig. 10. Illustration of Case II: from understanding regions to discover properties. (a) regional exploration (regions with more Chinese residents are discovered); (b & c) detailed information of two selected suburbs in Melbourne: Box Hill (b) and Glen Waverley (c).

(Fig. 9(b)-8). After the filtering, there left ten properties all in the suburb of Blackburn. She picked two of the houses in lower prices (Fig. 9(b)-2) and with most of the facilities nearby (Fig. 9(b)-9).

After saving the two properties into her favourites, she compared those two houses in detail (**T.4**) (Fig. 9(c)). She found that the two houses were very similar to each other (Fig. 9(c)-1), except that House 1 had a slightly shorter walking time to the nearest train station from. Then, she further discovered from the multidimensional view that House 1 was 6k more expensive than House 2 (Fig. 9(c)-2), but was closer to most of the facilities (Fig. 9(c)-3). At last, she decided to inspect those two houses in person and had House 1 (Fig. 9(c)-4) as a priority.

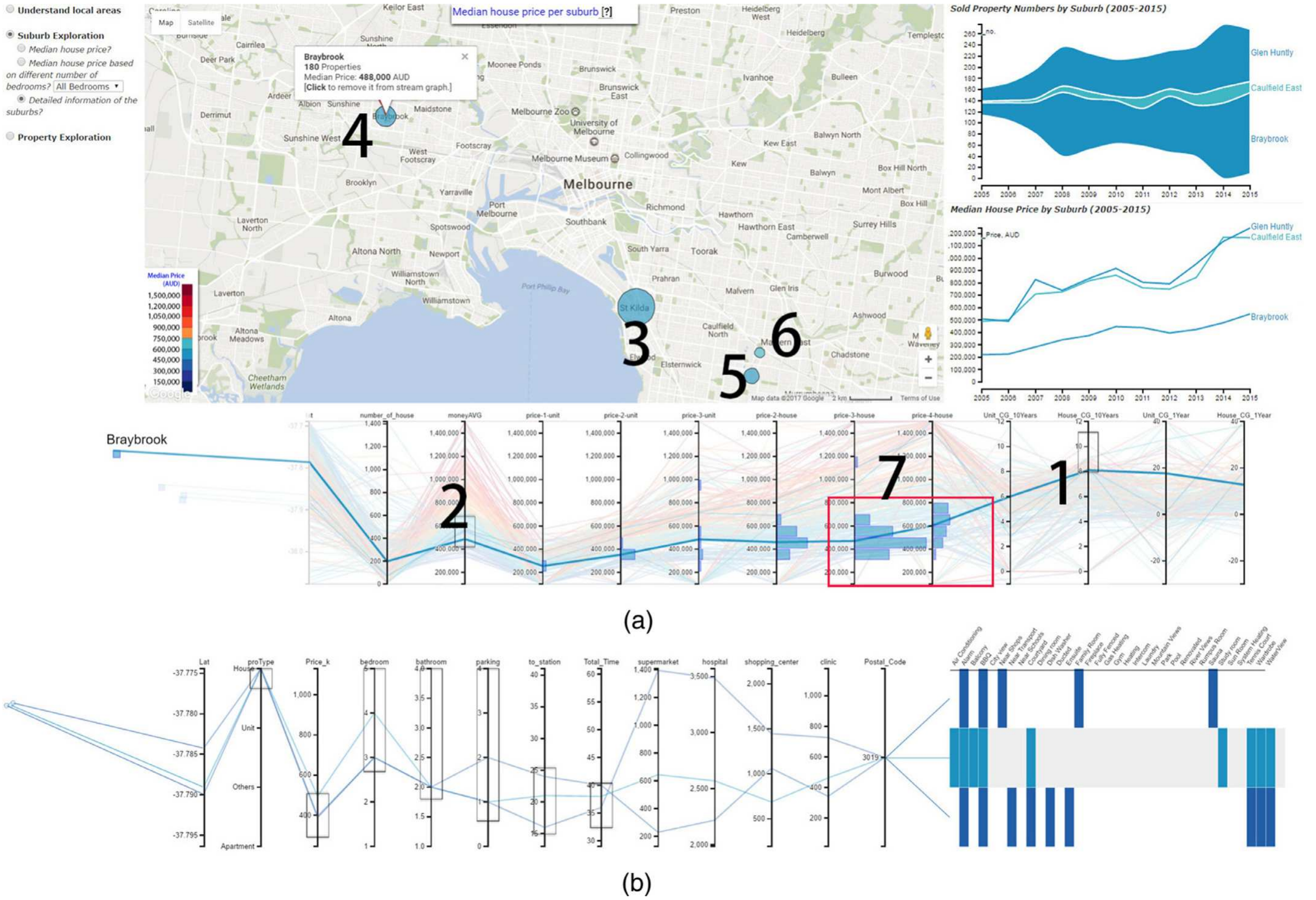
7.2.2. Case II. From understanding regions to discover properties

User B wanted to buy a house in a region where there were more Chinese residences so that his Chinese parents would have an easy life in the region. Since he might live there only for 3 or 4 years, he also preferred a region where the house value could increase in value over years.

He first went to regional exploration (**T.1**) and selected *Chinese_%* (percentage of Chinese residents) from the drop-down box (Fig. 10(a)). He identified four regions where there were more Chinese residences, as Melbourne CBD (1), Box Hill (2), Glen Waverley (3), and Clayton (4). He then selected the median age at the regional exploration and found that Melbourne CBD and Clayton had more young people living there. Therefore, considering his parents' age, he chose Box Hill and Glen Waverley as two candidate suburbs to buy a house.

Then he went to suburb-level exploration (**T.2**), and selected Box Hill and Glen Waverley from the Google Maps view to add them to the historical view. By comparing price-related information in the two suburbs (Fig. 10(b & c)), he noticed that 1) house price in both the two suburbs had an average capital growth of over 8% in the past ten years (Fig. 10(b & c)-1); 2) house price in Box Hill (Fig. 10(b)-2) increased over 30% in the past year, while the increasing rate for Glen Waverley (Fig. 10(c)-2) was around 20%; 3) there were much more 3-/4-bedroom houses within his budget (1 million) in Glen Waverley (Fig. 10(b)-3) comparing to Box Hill (Fig. 10(c)-3). Considering the available properties and the possibility of price bubbles in Box Hill at the moment, he decided to first consider properties in Glen Waverley.

He double clicked the suburb of Glen Waverley and enabled to explore the individual properties in Glen Waverley. Since he did not care about the suburb boundaries, he drew rectangles with our region selection tools to add more candidate regions near Glen Waverley Community centre, and then filtered the properties and compared the properties based on his own requirement, similar to how User A discovered the individual properties and compared them (**T.3** &**T.4**), User B found three candidate properties and had a ranking among them.


○ Understand local areas
● Suburb Exploration
○ Median house price?
○ Median house price based on different number of bedrooms? All Bedrooms ▾
● Detailed information of the suburbs?
○ Property Exploration

Braybrook
180 Properties
Median Price: 488,000 AUD
[Click to remove it from stream graph.]

(a)

(b)

Fig. 11. Illustration of Case III: investment on properties: (a) suburb exploration - selecting suburbs with higher house price growth; (b) comparing 3 candidate houses at the property-level exploration.

7.2.3. Case III. Investment on properties

User C had 0.6 million dollars and wanted to invest a property from which he could make the most profits.

He went directly to suburb exploration (**T.2**), and selected those suburbs whose average capital growth of houses was above 8% in the past ten years (Fig. 11(a)-1). Considering his budget, he further filtered the suburbs with limiting the average price below 0.7 million (Fig. 11(a)-2). There were only four suburbs left, as St Kilda, Braybrook, Glen Huntly and Caufield East (Fig. 11(a)-[3–6]). After comparing the detailed information of those suburbs, he found that, there were more 3-/4-bedroom houses in Braybrook (Fig. 11(a)-7) that were below the average price of this suburb (0.49 million, also certainly within his budget), which means that he could buy one of the properties, did some decoration if needed and then sold the property at a fair price. Therefore, he chose Braybrook as the target suburb, and clicked into the suburb to explore the individual houses.

At the individual property level (**T.3** &**T.4**), he chose the attributes that he thought critical to resell the property. Then he filtered properties (Fig. 11(b)) and had (3+)-bedroom houses which were below the average price in Braybrook, had at least two bathrooms and one packing, and within 40 minutes to the city by train. He had three houses left. After comparing the properties in detail, he chose House 1 as his primary target and the other two houses as candidates properties.

7.2.4. Case IV. Exploring the properties as beginners

User D was new to Melbourne, and wanted to explore the local real estate market, got to know different lifestyles in different regions, so that he could have a basic idea of what kind of properties he would buy in the future.

He first selected suburb exploration (**T.2**) (Fig. 4), and found that, in general, the prices of properties in the eastern suburbs and the southeastern suburbs were higher than that in the northern suburbs, and the prices of that in the western suburbs were the lowest; also, the price had a negative correlation with the distance from the Melbourne CBD (central business district) area. He was surprised to find that property prices in the CBD were not high, since he believed that renting a room in the CBD was quite expensive. He found the reason of this as he moved to another level to see how the prices of properties were different based on different numbers of bedrooms in different suburbs (Fig. 12(a)). There were more 1-bedroom and 2-bedroom apartments in the CBD area (Fig. 12(a)-1), which highly influenced the price; however, the regions around the CBD, such as Albert Park (Fig. 12(a)-2), Fitzroy (Fig. 12(a)-3) and Fitzroy North (Fig. 12(a)-4) had more 3-bedroom houses, which made the price very high. The patterns showing how the prices of Melbourne properties were different along with different directions and distances to the CBD were more clear to him when only the 3-bedroom properties were selected. He also explored the individual properties, and checked the differences between the (individual) properties of the CBD and the surrounding areas.

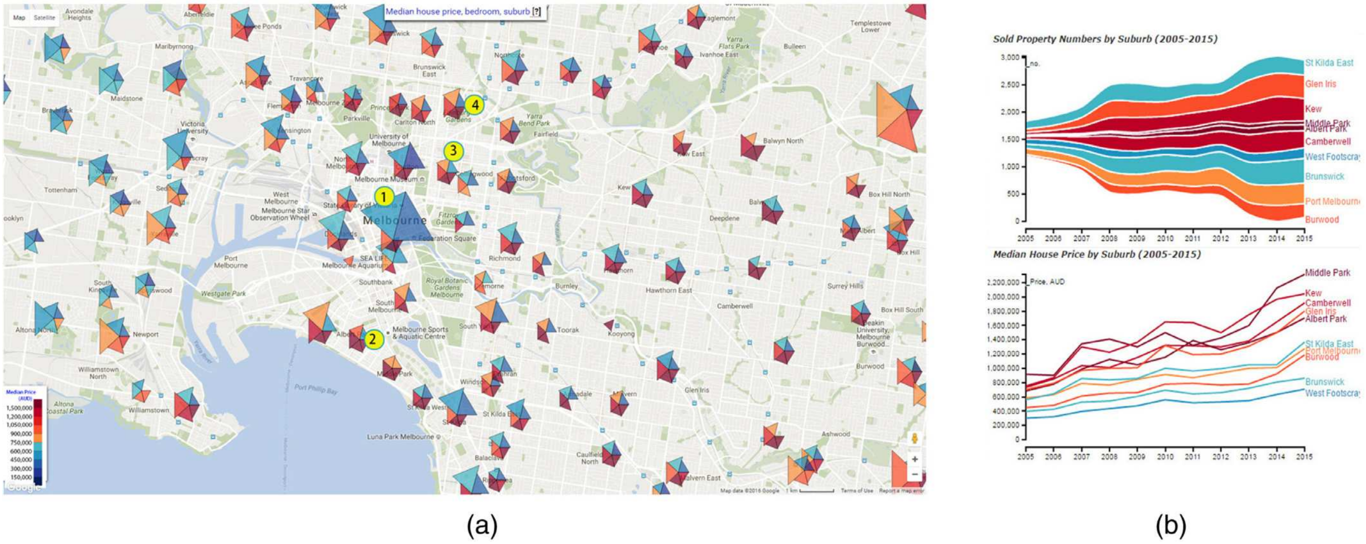


Fig. 12. Illustration of Case IV: exploring the properties as beginners. (a) visualizing property price based on different number of bedrooms; (b) changes of houses price and sold house property number over 10 years at different suburbs of Melbourne.

He also noticed that median house prices were much higher than unit prices. From the Google Maps view, he selected several suburbs which locate in the western, northern, eastern and south-eastern of Melbourne, and also had various distances to the CBD (Fig. 12(b)). Overall, the house price in 2015 had at least doubled comparing to ten years ago; while the unit price had increased too, but not significantly, especially in the latest five years. Further looking into the changes of house prices, he noticed that the price had been increased from 2005 and peaked at 2010/2011, and then decreased a little bit in the following two years and then increased at a much higher rate in the latest two years.

He then discovered the difference among regions in other aspects, such as people, public schools, facilities, etc. in the profile-based region exploration (**T.1**). At last, he clicked on property-level exploration (Fig. 7) and defined several kinds of filtering settings to see what kinds of properties he could afford in different regions (**T.3**).

7.3. Domain expert interview

To attract users, we demonstrated our demo system at the 27th Australian Database Conference[4] (for half a day) and the 2017 conference of Beyond Research - Pathway to Impact[5] (for two days). We also advertised our system on social media with an introduction video and demo links. Besides general users and researchers from different fields, we particularly got the attention from two real estate agents, two researchers in Business IT (related to real estate) and a professor in urban science, which are all considered as domain experts.

One of the real estate agents commented on our system (based on an early version of implementation), saying that *[The System is] Incredible, what a great translation of property buyer needs. It's refreshing to see it built from authenticity, instead of commercialism.* The agent contacted us from social media and met us in person. The same with how we interviewed with other domain experts, we introduced the system, asked them to try the system by themselves, and then sought feedback from them.

Most of the domain experts are fascinated about how we collected the data from different sources and integrated them together. One of the real estate agents commented that our tool could be very useful for him to recommend properties to home buyers, and help them illustrate why the property is worth to buy. He also suggested us to include the annual growth information which investors might be more interested (for example, they might search for the suburbs that have an annual growth larger than 10% in the last five years). Though our system mainly focuses on home buyers, we have included such information in our later implementation to suit for more user types.

7.4. Discussion

Based on the case studies, we have demonstrated that our system can be beneficial to both home buyers and investors; and users with different levels of knowledge (on the local real-estate market) can jump into a certain level of tasks to explore the local properties. Although users still need to inspect the properties before they make a final decision, our system provides rich sources for them to find candidate properties and understand the properties before the inspection. We have augmented existing commercial systems such as [3–5,8] in the following aspects. First, based on a unique location-centred dataset that we provide, all the data (related to different profiles) and services (exploration in the suburb level and individual property level) in our system are interconnected. Secondly, we have provided a way to help users understand the local real estate market, which also helps them further understand what kinds of property that they prefer and which they can afford. Thirdly, we allow users to explore the properties, find properties based on their requirements in a more interesting and effective way, since our system has provided instant feedback after they interact with the system. Last, we allow users to visually compare suburbs and properties, which the current commercial systems fail to do.

One of the limitations is related to the scalability. We have sampled the properties on top of the Google Maps view when there are too many (i.e., more than 5,000) properties at the selected region; and we have also used progressive rendering on parallel coordinates. Users have commented that the progressive rendering in a way makes them feel confused, and the massive lines in the parallel coordinates fail to provide much information. In our future work, we would like to discover more about the aggregation based on both the map view and the multidimensional view.

[4] https://adc2016.cse.unsw.edu.au/
[5] http://beyondresearchconference.com.au/

Another consideration is whether our framework works well in other cities/countries besides Melbourne/Australia. So far, we have also collected data from other major cities in Australia, such as Sydney, Brisbane, Perth and Adelaide. Since the data profiles that we have defined (Section 3.2) are general enough to cover all location information related to a real estate market, and new information can always be treated as one instance of one type of the profiles. Our framework works well with other cities, except that only some details in the process of data collecting need to be modified. For example, in Melbourne, most commuters travel by train or tram; while in Sydney, it could be by train or ferry. Certainly, there are different considerations across countries, we believe that our framework can inspire researchers from other countries to work on their real estate data.

8. Conclusion and future work

We have designed and developed HomeSeeker which augments existing commercial systems to help users understand the local real estate market, find preferred properties based on their personal preferences, and compare properties from the aspects that they are concerned about. We have provided a location-centred real estate dataset, which is exclusive in this paper. We have presented a systematic visualization design study following with justifications. Case studies based on real-world datasets have demonstrated the usefulness of our system, and users with different levels of knowledge on the market and different requirements (either general home buyers or investors) can benefit from the system.

In future work, first, we will include data from more cities and continually update real estates that are on sale from the real estate market. Secondly, we plan to address the scalability problem with clustering visualization, other than our current methods (sampling, filtering, progressive rendering, etc.)

Acknowledgements

This work was partially supported by ARC DP170102726, DP170102231, DP180102050, and National Natural Science Foundation of China (NSFC) 61728204, 91646204. Zhifeng Bao is supported by a Google Faculty Award.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jvlc.2018.02.001

References

[1] T. Jacqui, What makes a great suburb?, accessed March 2017, (http://www.domain.com.au/advice/makes-great-suburb/).
[2] P. Thornhill, Real estate buying guide - Melbourne: a first home buyers guide, 11 Mar 2015, (http://www.realestate.com.au/advice/melbourne-first-home-buyers-guide/).
[3] Real Estate Australia, accessed March 2017, (http://www.realestate.com.au/).
[4] Domain Australia, accessed March 2017, (http://www.domain.com.au/).
[5] Right Move UK, accessed March 2017, (http://www.rightmove.co.uk/).
[6] Zillow, accessed March 2017, (https://www.zillow.com/).
[7] Trulia, accessed March 2017, (https://www.trulia.com/).
[8] Soufang China, accessed March 2017, (http://www.fang.com/).
[9] R.S. Pressman, Software Engineering: a Practitioner's Approach, Palgrave Macmillan, 2005.
[10] Real Estate Australia: the map view, accessed March 2017, (http://www.realestate.com.au/buy/in-melbourne).
[11] Y. Fu, G. Liu, S. Papadimitriou, H. Xiong, Y. Ge, H. Zhu, C. Zhu, Real estate ranking via mixed land-use latent models, in: the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 299–308.
[12] Y. Fu, Y. Ge, Y. Zheng, Z. Yao, Y. Liu, H. Xiong, J. Yuan, Sparse real estate ranking with online user reviews and offline moving behaviors, in: IEEE International Conference on Data Mining, 2015, pp. 120–129.
[13] Y. Fu, B. Liu, The impact of community safety on house ranking, in: the 2016 SIAM International Conference on Data Mining, 2016, pp. 459–467.
[14] F. Tan, C. Cheng, Z. Wei, Modeling real estate for school district identification, in: IEEE International Conference on Data Mining, 2017, pp. 1227–1232.
[15] Y. Fu, H. Xiong, A discovery system for finding high-value homes, in: the 15th IEEE International Conference on Data Mining Workshop, 2016, pp. 1612–1615.
[16] M. Shahbazi, J.R. Barr, V. Hristidis, N.N. Srinivasan, Estimation of the investability of real estate properties through text analysis, in: the 10th International Conference on Semantic Computing, 2016, pp. 301–306.
[17] H. Zhu, H. Xiong, F. Tang, Q. Liu, Y. Ge, E. Chen, Y. Fu, Days on market: measuring liquidity in real estate markets, in: the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 393–402.
[18] C. Williamson, B. Shneiderman, The dynamic HomeFinder: evaluating dynamic queries in a real-estate information exploration system, in: ACM SIGIR, 1992, pp. 338–346.
[19] L. Tweedie, B. Spence, D. Williams, R. Bhogal, The attribute explorer, in: Conference companion on Human factors in computing systems (CHI'94), ACM, 1994, pp. 435–436.
[20] R. Spence, Representation, in: Information visualization: an introduction, 2014, pp. 41–110.
[21] G. Sun, R. Liang, F. Wu, H. Qu, A web-based visual analytics system for real estate data, Sci. China Inf. Sci. 56 (5) (2013) 1–13.
[22] H. Lam, T. Munzner, R. Kincaid, Overview use in multiple visual information resolution interfaces, IEEE Trans. Vis. Comput. Graph. 13 (6) (2007) 1278–1285.
[23] R. Beecham, C. Rooney, S. Meier, J. Dykes, A. Slingsby, C. Turkay, J. Wood, B. Wong, Faceted views of varying emphasis (FaVVEs): a framework for visualising multi-perspective small multiples, Comput. Graph. Forum 35 (3) (2016) 241–249.
[24] C. Ware, Information Visualization: Perception for Design, Elsevier, 2012.
[25] D.B. Carr, R.J. Littlefield, W. Nicholson, J. Littlefield, Scatterplot matrix techniques for large N, J. Am. Stat. Assoc. 82 (398) (1987) 424–436.
[26] H. Siirtola, K.-J. Räihä, Interacting with parallel coordinates, Interact. Comput. 18 (6) (2006) 1278–1309.
[27] L. Wilkinson, M. Friendly, The history of the cluster heat map, Am. Stat. (2012).
[28] M.O. Ward, G. Grinstein, D. Keim, Interactive Data Visualization: Foundations, Techniques, and Applications, CRC Press, 2010.
[29] X. Yuan, P. Guo, H. Xiao, H. Zhou, H. Qu, Scattering points in parallel coordinates, IEEE Trans. Vis. Comput. Graph. 15 (6) (2009) 1001–1008.
[30] C. Viau, M.J. McGuffin, Y. Chiricota, I. Jurisica, The flowvizmenu and parallel scatterplot matrix: hybrid multidimensional visualizations for network exploration, IEEE Trans. Vis. Comput. Graph. 16 (6) (2010) 1100–1108.
[31] D.R.F. Taylor, Perspectives on visualization and modern cartography, in: Visualization in modern cartography, 1994, pp. 333–341.
[32] A.M. MacEachren, D.R.F. Taylor, Visualization in modern cartography, Elsevier, 2013.
[33] J. Vertesi, Mind the gap the london underground map and users' representations of urban space, Soc. Stud. Sci. 38 (1) (2008) 7–33.
[34] M. Ovenden, Transit maps of the world, Penguin, 2015.
[35] D. Liu, D. Weng, Y. Li, J. Bao, Y. Zheng, H. Qu, Y. Wu, Smartadp: visual analytics of large-scale taxi trajectories for selecting billboard locations, IEEE Trans. Vis. Comput. Graph. 23 (1) (2017) 1–10.
[36] N. Ferreira, M. Lage, H. Doraiswamy, H. Vo, L. Wilson, H. Werner, M. Park, C. Silva, Urbane: a 3d framework to support data driven decision making in urban development, in: IEEE Conference on Visual Analytics Science and Technology (VAST), 2015, pp. 97–104.
[37] R. Chang, G. Wessel, R. Kosara, E. Sauda, W. Ribarsky, Legible cities: focus-dependent multi-resolution visualization of urban relationships, IEEE Trans. Vis. Comput. Graph. 13 (6) (2007) 1169–1175.
[38] S. Goodwin, J. Dykes, A. Slingsby, Visualizing the effects of scale and geography in multivariate comparison, in: IEEE Conference on Visual Analytics Science and Technology (VAST), 2014, pp. 251–252.
[39] S. Goodwin, J. Dykes, A. Slingsby, C. Turkay, Visualizing multiple variables across scale and geography, IEEE Trans. Vis. Comput. Graph. 22 (1) (2016) 599–608.
[40] P. Harris, C. Brunsdon, M. Charlton, S. Juggins, A. Clarke, Multivariate spatial outlier detection using robust geographically weighted methods, Math. Geosci. 46 (1) (2014) 1–31.
[41] A.D. Singleton, P.A. Longley, Geodemographics, visualisation, and social networks in applied geography, Appl. Geogr. 29 (3) (2009) 289–298.
[42] A. Slingsby, J. Dykes, J. Wood, Exploring uncertainty in geodemographics with interactive graphics, IEEE Trans. Vis. Comput. Graph. 17 (12) (2011) 2545–2554.
[43] W. Javed, N. Elmqvist, Exploring the design space of composite visualization, in: IEEE Pacific Visualization Symposium, 2012, pp. 1–8.
[44] H. Guo, Z. Wang, Tripvista: triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection, in: IEEE Pacific Visualization Symposium, 2010, pp. 163–170.
[45] C. Turkay, A. Slingsby, H. Hauser, J. Wood, J. Dykes, Attribute signatures: dynamic visual summaries for analyzing multivariate geographical data, IEEE Trans. Vis. Comput. Graph. 20 (12) (2014) 2033–2042.
[46] Y. Zhang, W. Luo, E. Mack, R. Maciejewski, Visualizing the impact of geographical variations on multivariate clustering, Comput. Graph. Forum 35 (3) (2016) 101–110.
[47] Department of Enviorment, Land, Water and Planning, Victoria State Government, accessed March 2017, (http://www.dtpli.vic.gov.au/).
[48] Land Channel, Department of Environment, Land, Water and Planning, Victoria State Goverment, accessed March 2017, (http://www.land.vic.gov.au/).
[49] Google Maps API, accessed March 2017, (https://developers.google.com/maps/).

[50] Australia Curriculum, Assessment and Reporting Authority, accessed March 2017, (http://www.acara.edu.au/).

[51] Better Eduation Australia, accessed March 2017, (http://bettereducation.com.au/).

[52] Melbourne School Zones, accessed March 2017, (http://melbourneschoolzones.com/).

[53] GTFS (General Transit Feed Specification) Victoria, accessed March 2017, (https://www.data.vic.gov.au/data/dataset/).

[54] Coles Supermarket Store Locations, accessed March 2017, (https://www.coles.com.au/store-locator).

[55] Woolworth Supermarket Store Locations accessed March 2017, (https://www.woolworths.com.au/Shop/StoreLocator).

[56] Health Engine Australia, accessed March 2017, (https://healthengine.com.au/).

[57] Australian Bureau of Statistics - Census, accessed March 2017, (http://www.abs.gov.au/).

[58] T. Munzner, Visualization Analysis and Design, CRC Press, 2014.

[59] J. Cohen, C. Small, Hypsographic demography: the global distribution of human population with altitude, accessed March 2017, (https://www.ldeo.columbia.edu/~small/Pop/HypsoDemo/HypsoDemoMain.html).

[60] C. Brewer, M. Harrower, Colorbrewer 2.0: color advice for cartography, accessed March 2017, (http://colorbrewer2.org/).

[61] M.D. Hansen, Combining parallel coordinates and histograms, Jun 2013, US Patent App. 13/931,785.

[62] J.R. Mackay, Dotting the dot map: an analysis of dot size, number, and visual tone density, Surv. Mapping 9 (1) (1949) 3–10.

[63] R. Kosara, F. Bendix, H. Hauser, Parallel sets: interactive exploration and visual analysis of categorical data, IEEE Trans. Vis. Comput. Graph. 12 (4) (2006) 558–568.

[64] M. Friendly, Statistical graphics for multivariate data, in: SAS SUGI 16, 1991, pp. 1157–1162.

[65] M. Li, Z. Bao, T. Sellis, S. Yan, Visualization-aided exploration of the real estate data, in: Australian Database Conference, Springer, 2016, pp. 435–439.

[66] C. Turkay, E. Kaya, S. Balcisoy, H. Hauser, Designing progressive and interactive analytics processes for high-dimensional data analysis, IEEE Trans. Vis. Comput. Graph. 23 (1) (2017) 131–140.