SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

Special Focus

May 2013, Vol. 56 052112:1–052112:13
doi: 10.1007/s11432-013-4830-9

A Web-based visual analytics system for real estate data

SUN GuoDao[1], LIANG RongHua[1*], WU FuLi[1] & QU HuaMin[2]

[1]*College of Information Engineering, Zhejiang University of Technology, Hangzhou* 310023*, China;*
[2]*Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hongkong, China*

Received September 7, 2012; accepted February 3, 2013

Abstract Real estate is an important industry in most countries. However, the analysis of the real estate market is very challenging as the data are high dimensional and have complex spatial and temporal patterns. In this paper, we present a novel Web-based visual analytics system, which integrates state-of-the-art interactive visualizations to enable end users to create their own visualizations and gain insight into the real estate market. The system is implemented using the new features in HTML5, which are natively supported in current browsers. We adopt a coordinated view design in our system consisting of four major components: a map view to show the geographical information of houses, a stacked graph view to show the evolution of house sales over time, a pixel-bar view to visualize multiple attributes of houses, and a treemap view to present the hierarchical structure of the data. Novel clutter reduction methods and rich user interactions are further proposed to enhance the flexibility and analytical ability of the whole system. We have applied our system to real property market data and obtained some interesting findings. Moreover, feedback from the end users of our system is very positive.

Keywords visual analytics, spatio-temporal data, multiple coordinated views, real estate

Citation Sun G D, Liang R H, Wu F L, et al. A Web-based visual analytics system for real estate data. Sci China Inf Sci, 2013, 56: 052112(13), doi: 10.1007/s11432-013-4830-9

1 Introduction

The real estate market in China has become very active in recent years. Allied to the rapid growth of the Chinese economy, the real estate industry has also developed rapidly. After the housing system reform in 1998, development of the Chinese real estate market only lasted for a short period of ten years, laws and policies pertaining to the housing market had various problems, and major cities in China experienced more than a decade of skyrocketing house prices. In 2003, the Chinese State Council issued a“notification on promoting the sustained and healthy development of real estate market”. In this notification, the real estate industry is characterized as a pillar industry in the country’s economic development. Over the years, the country has frequently published macro-control policies, such as the 2010 “three most stringent real estate market regulations in history” and the 2011 “real estate tax”. However, house prices still remain high, and there is no evidence of a downward trend.

[*] Corresponding author (email: rhliang@zjut.edu.cn)

© Science China Press and Springer-Verlag Berlin Heidelberg 2013

Like most Chinese cities (e.g., Beijing or Shanghai), Hangzhou, which is the capital of Zhejiang Province, has also experienced a golden decade in the real estate market. The property market in Hangzhou has flourished from the outset and is still showing strong momentum signs of development without experiencing large fluctuations. Therefore, since the beginning of 21st century, the property market in Hangzhou has been described as the "phenomenon of Hangzhou". According to reports on Xinhua Net [1)], Hangzhou was ranked first in the 2010 China urban housing price list, followed by Beijing and Shanghai. Thus, an analysis of Hangzhou's property market has high practical significance and social value.

Real estate market data are a kind of classical spatio-temporal data, indicating geographical distribution of the houses, trends in house prices and sales volumes, and other unknown trends. Visualization helps present, analyze, and discover these hidden stories intuitively, efficiently, and interactively.

With the advancement of Web technology, visualization is no longer restricted to desktop environments. Native support in Web browsers has enabled developers to implement computation-intensive, interactive visualizations without the use of third party software. Users are able to explore and analyze data in visualization software embedded in Web browsers, and since there is no reliance on static files, the data are always up-to-date.

In this paper, we implement a Web-based system that visualizes and analyzes Hangzhou real estate market data. The system includes various visualization components that encode different attributes of the real estate market. We also develop a coordinated mechanism to connect the various visualization components. Through these techniques, our system shows the development of the market and the correlation among house prices, sales volumes, time, culture, and policy.

The major contributions of our work are: 1) We present different visualization techniques to analyze the attributes from different aspects according to the characteristics of the data. 2) We develop an interactive Web-based visualization system to study the real estate market and assist a variety of users to understand the data visually. 3) We demonstrate using case studies that our system can facilitate analysis of the development of the real estate market and discovery of correlations between different aspects of the market.

2 Related work

• **Geographical visualization.** It is estimated that 80% of all digital data generated nowadays include geospatial referencing [1]. Geovisualization communicates geospatial information through human understanding to realize data exploration and decision-making processes. This is a broad field that is closely related to other visualization fields, such as scientific visualization and information visualization. Takatsuka et al. [2] presented GeoVISTA Studio, an open-source, Java-based visual development environment designed for geospatial data. GeoVISTA Studio is a programming-free environment that allows users to build applications rapidly for geocomputation and geographic visualization. Based on GeoVISTA Studio, Hardisty et al. [3] proposed the GeoViz Toolkit, which uses a component-oriented coordination method to aid geovisualization application construction. Our system, which was inspired by this coordinated visualization mechanism, adopts a similar mechanism to visualize real estate data. Malik et al. [4] also developed a visual analytics system with multiple collaborative views for maritime resource allocation and risk assessment, containing views of geographic distribution and time series curves. Slingsby et al. [5] used a treemap to discover spatio-temporal patterns through several layouts including spatial ordered layout, squarified layout, and temporal ordering layout. Treemap is also used in our system to correlate the house location in the treemap and the original house position on the map.

• **Time series visualization.** While the most frequently used visualization of time series is still the line graph with a time axis and a price/sale volume axis, many techniques for visualizing time series have been proposed and developed during the past 20 years. Havre et al. [6] proposed ThemeRiver to visualize thematic variations over time across a collection of documents. To show multiple time series,

[1)] House price ranking in 2010. http://news.xinhuanet.com/fortune/2011-02/10/c_121059274.htm

Byron et al. [7] developed a stacked graph, which pays attention to the geometry and aesthetics of the design. Keim et al. [8] proposed pixel bar charts, which visualize very large multi-attribute data sets without aggregation. Based on the pixel bar charts, Ziegler et al. [9] integrated dense pixel-displays to visualize financial time series data. The pixel bar chart transforms a two-dimensional (2D) line graph into a one-dimensional bar. With respect to time series clustering, the time series data can be classified into two categories: equal length and unequal length data series. As the time series for price or number of sales in the real estate market are discrete, we focus on unequal length time series. According to Liao's survey [10] of the analysis of clustering of time series data, time series clustering methods can be grouped into three major categories, depending on whether they work directly with raw data, indirectly with features extracted from the raw data, or with models built from the raw data. Liao et al. [11] applied time series clustering in battle simulation, using a distance measure of DTW (dynamic time warping) and the K-method-based genetic clustering algorithm. Fu et al. [12] extracted perceptually important point (PIP) feature from the time series data, and applied this technique to analyze the Hong Kong stock market. Oates et al. [13] and Wang et al. [14] used model-based time series clustering algorithms, with a discrete hidden Markov model (HMM) as the model and log-likelihood as the distance measure.

• **Web technology.** Native technologies supported by Web browsers include scalable vector graphics (SVG), 2D raster graphics using HTML5's canvas element, and WebGL where the context of the canvas HTML element provides a 3-dimensional (3D) computer graphics API without the use of plug-ins. SVG is an XML specification file format for both static and dynamic 2D vector graphics. The HTML5 canvas is designed to support rich and interactive Web-based applications, with the contents rendered on canvas using JavaScript. WebGL is based on OpenGL ES 2.0 and is a royalty-free Web standard for 3D graphics rendering. It uses JavaScript as the programming language and the HTML5 canvas element to render the contents. In our system, we use the HTML5 canvas to create each visualization component, and the WebGL to buffer the visualization results to speed up the rendering and interaction process. Current online visualization systems include Spotfire[2)], ManyEyes[3)], and so on. Spotfire is a comprehensive commercial software platform with sophisticated visualization features which enable users to develop dynamic analytic applications. The ManyEyes site offers various stand-alone interactive visualization components using Java Applet, and users can upload data, construct visualizations, and leave comments on either data sets or visualization results.

Currently there are some similar online visualization systems for real estate data. Tableau[4)] offers online real estate analysis for users to understand market trends across metro areas, but it requires considerable expertise for users to create their own visualizations. Search.ch[5)] collects advertisements from various house sales website, and provides house sales information visually.

3 System and data

3.1 System overview

An overview of our system is shown in Figure 1. The system has the following major components: the geomap view, the stacked graph view, the clustered pixel bar view, and the treemap view. The geomap view shows the geographical distribution of the houses on the map and provides house clustering based on the house locations. The stacked graph view displays the number of house sales over time using different layouts, layer orderings, color encodings and time intervals such as month, week, and so on. The clustered pixel bar view visualizes the attributes of the house such as house price or sales volumes without aggregation, and offers time series clustering based on single or multiple attributes of the houses. The treemap view presents a comprehensive view of all the attributes of the houses and offers various layouts and orderings to enhance visual cognition. Since the component's interface is semi-transparent,

2) Spotfire. http://spotfire.tibco.com
3) Many Eyes. http://www-958.ibm.com
4) Tableau. http://www.tableausoftware.com/solutions/real-estate
5) Search.ch. http://immo.search.ch

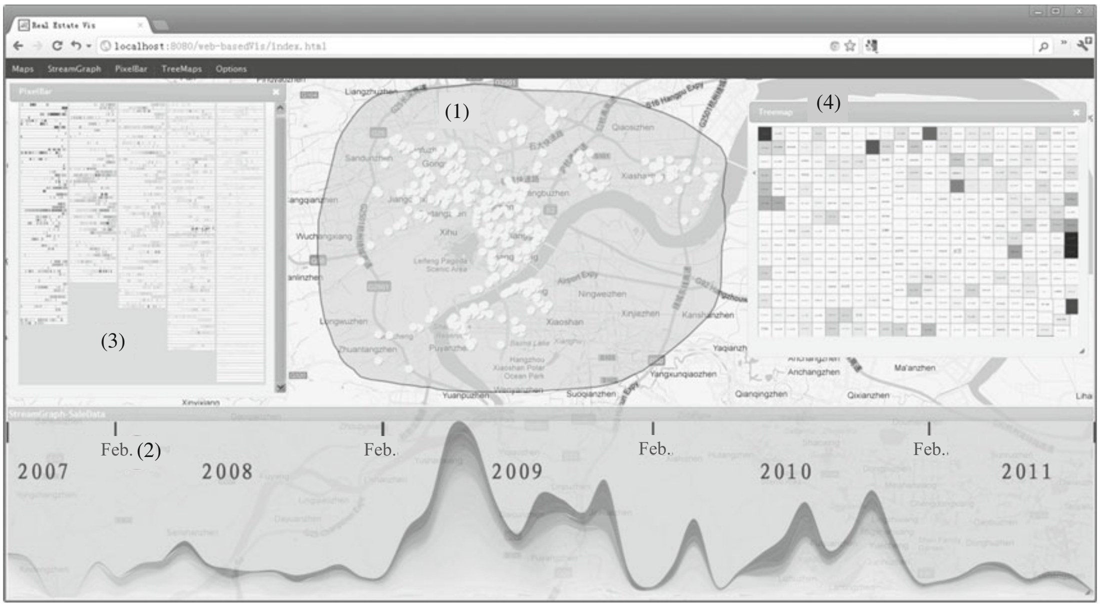


Figure 1 An overview of the system. The system consists of the following components: (1) the geomap view which shows the geographical distribution of the houses and house clusters on the map; (2) the stacked graph view which visualizes the number of house sale over time using various layouts and layer orderings; (3) the clustered pixel bar view which visualizes changes in price or number of sales over time; (4) the treemap view which presents the hierarchical structure of the data using various combinations of layouts and orderings.

the overlay of components dose not cause occlusion.

The system is designed to help expert users such as decision makers to find these patterns and trends, and gain insight from the complexity of the data. Furthermore, it helps ordinary users who have rigid requirements of the real estate market, understand the current situation and the evolution of the market.

3.2 Data

We collected the Hangzhou real estate data using a customized Web spider program from four Hangzhou official and authoritative websites, namely, the Hangzhou Real Estate Information website[6], Hangzhou Transparent Selling House website[7], Living in Hangzhou website[8], and Hangzhou Planning Bureau[9]. The Hangzhou Real Estate Information website is the official website of the Department of Hangzhou Houses Management, which carries out the functions of the Government. The Hangzhou Transparent Selling Houses website is a house information publishing website under the Department of Hangzhou Houses Management and is the most professional, most authori-tative real estate information site. Living in Hangzhou is currently the most influential real estate media website. The Hangzhou Planning Bureau offers geographically relevant information for Hangzhou such as planning units for the city.

The data used in our system includes 382 houses constructed and sold between September 2007 and October 2011, with geographically relevant information for over 40,000 sale entries. The attributes of a house consist of dimensions such as the house identifier, house name, house longitude and latitude, administrative unit, planning unit, property type, and property developer, as well as numerical dimensions such as the total number of houses in a building, the area of the houses, and the opening times. The geographically relevant information consists of the additional data related to houses such as urban roads, highways, bus carriers, hospitals and schools. The time relevant data refers to daily house prices, the book number and the number of transactions over the four years.

[6] Hangzhou Real Estate Information website. http://www.hzfc.gov.cn
[7] Hangzhou Transparent Selling House website. http://www.hzfc365.com
[8] Living in Hangzhou website. http://zzhz.zjol.com.cn
[9] Hangzhou Planning Bureau. http://www.hzplanning.gov.cn

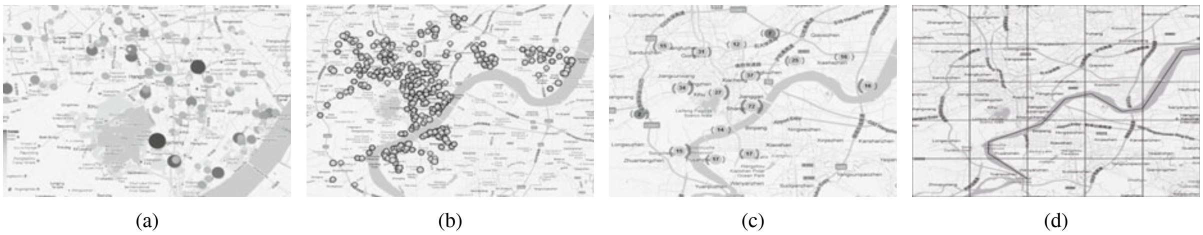


Figure 2 The geomap view of our system. (a) House markers on the map colored and sized by price; (b) sun, cloud, rain, and snow glyphs represent the different intervals in the value, which denotes the price from high to low in this figure; (c) house clusters surrounded by arcs represent the summary attributes (price and sales volume) in a cluster, while the number in the cluster icon represents the number of markers in this cluster; (d) segmentation polyline between the grids.

4 Geomap view

4.1 Design scheme

In our system, markers of houses are added to Google Maps using the Google Maps API[10] according to the house's longitude and latitude. Each house is represented by a dot on the map; here, the size or color of the dot can be encoded with different data dimensions (see Figure 2(a)), while glyphs can be used to represent different intervals of the data (see Figure 2(b)).

As hundreds of houses are crowded into one city and a large number of locations need to be displayed in a small area, directly plotting too many markers can easily lead to a degraded user experience, including overload in performance and slow interaction with the map.

To overcome this situation, the Google Maps API provides Google Fusion Tables[11] to store the data for rendering on Google servers as a layer on the map. This solves the bottleneck for performance, although the problem of overlapping house markers still exists.

Thus, we intend aggregating the markers that are close to one another on the map to improve visualization. According to the Google Geo APIs team[12], a grid-based clustering method for markers is provided. First, the map is divided into grids of fixed size, with the size of the grid varying according to the change in the map's zoom level. Then the markers are grouped within each grid. This method can be executed fairly quickly by the browsers, but there is an obvious problem in that two markers that are quite close to each other may be allocated into two different grids. Another method, distance-based clustering, is relatively traditional and similar to k-means clustering. Clusters are created based on the distance among the markers and a cluster centroid is built with the position of the newly added marker and the position of the old cluster centroid. This method iterates over each marker to determine how far it is from the center of each cluster centroid and adds the marker to the nearest cluster. This cluster algorithm solves the problem mentioned above, but it needs to iterate over the markers several times and calculate the Euclidean distance between two markers, causing degradation in performance as it runs in the user's browser.

Based on the Google Geo Developer's Marker Cluster, this paper presents a modified grid-based clustering algorithm: for each marker, if the marker lies within a cluster, it is added to that cluster, and the center of the grid representing that cluster is updated with the geometric center of the markers in the cluster. Otherwise, the marker is incorporated into a new cluster, and a new grid is also created, where the grid bound of the cluster is first specified by the users (see Figure 2(c)). As the modified grid-based clustering only needs to traverse all the house markers once, it has good performance and rapid response ability when users scroll through the map zooming in and out.

In addition, we summarize the attributes of each cluster, and propose a novel visualization that places arcs around the cluster icon. As shown in Figure 2(c), the arc to the left of the cluster icon represents the total sales in this cluster, while the length of the arc is encoded as the sale volume. The right arc is

[10] Google Maps API. https://developers.google.com/maps/

[11] Google Fusion Tables. http://www.google.com/fusiontables

[12] Mahe L, Broadfoot C. Too Many Markers! https://developers.google.com/maps/articles/toomanymarkers

used to denote for the price.

When users navigate from one place to another on a map full of marker clusters, they can click on a marker cluster to see the detail for that cluster. In a real map, a grid may include rivers, lakes, and so on, and thus clicking on a cluster may result in the presentation of details for surrounding terrain, for example, houses on both sides of the river. However, the users may only interested in a small area of the grid, e.g., one side of the river, and thus, segmentation of grids is necessary.

The proposed method is further developed as the following method: when a new grid is created followed by a new cluster and the grid contains certain geographical features like a river, we place a segmentation polyline presenting the river into the grid to form polygon grids, and the subsequent markers, which should have been added into the previous grid, are added into these polygon grids (see Figure 2(d)).

4.2 Interaction

Users can zoom in or out to view the details or overall distribution of the houses. While the house markers are clustered, fewer clusters appear when users zoom out, but the number of markers in each cluster increases, and vice versa.

Users can also change the visual encoding scheme, e.g., the color, size, or shape of the markers.

We also implemented a lasso selection tool with the Google Maps API to select the house markers with multiple lassos for future use.

5 Stacked graph view

After displaying the geographical distribution of the houses, we need to focus on the temporal attributes of the houses, especially the price and number of sales. As the number of sales is stackable whereas the price is not, i.e., the sales volume of different houses can be added together at any time point, we would like to display the number of sales of houses in a stacked graph view together with different layouts and orderings over the years.

The stacked graph view in Figure 1 shows the sales volume of all the houses which were conveniently selected using the lasso selection tool.

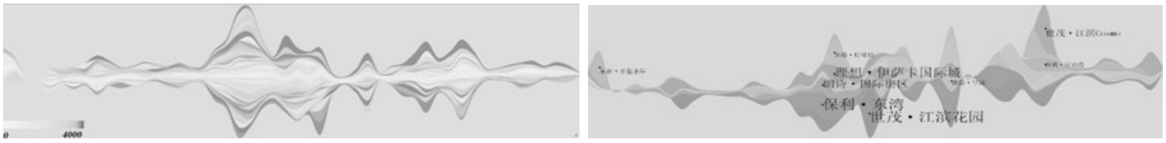
5.1 Layout and ordering

The stacked graph has various graph layouts such as the traditional stacked graph layout, ThemeRiver layout [7], Streamgraph layout [7], and minimized wiggle layout [7], as well as various layer orderings such as an ordering on each layer's onset [7], volatility [7] or size.

The traditional stacked graph layout simply uses one horizontal or vertical axis as the baseline and adds layers onto this. ThemeRiver uses a symmetric layout where the baseline is exactly one half of the total height of the graph at any time. Based on the ThemeRiver layout, the Streamgraph layout aims at minimizing wiggle of each layer and creating beautiful aesthetic forms of the stacked graph.

Ordering of the layer's volatility [7] places the most volatile layers along the outside of the graph to avoid distortion as far as possible, or on the inside to demonstrate how turbulent a stacked graph can be. Ordering on the layer's size simply places the layers from top to bottom or from bottom to top in the graph with the total sum of each layer's value at each point.

Combinations of different layouts and orderings can be applied to real estate data. The stacked graph view in Figure 1 displays the time series of all houses using the traditional stacked graph layout and descending size-based ordering, while the color and thickness of one strip are encoded as the total sum of the layer's value at each point, where the darker and thicker the strip is, the higher is the number of sale. Users can move the mouse over the stacked graph and the corresponding strip is highlighted. Users can see that nearly two-thirds of the houses have a relatively low sales volume, and that the sale of these houses only lasts for a short time. Similarly Figure 3(a) shows all stacked graph for the houses using the Streamgraph layout and volatility ordering, with the same color and strip shape encoding scheme. We find that most of the higher sales volume houses, which have a darker color, are placed outside the graph



(a) (b)

Figure 3 Stacked graph view of our system. (a) Stacked graph with Streamgraph layout and volatility ordering; (b) stacked graph with layer labeling.

as they experience bursts like a sudden surge or fall in sales volume. Consequently, we can deduce that the sales volume has some connection with the duration and volatility of the sale.

5.2 Labeling

Hopefully, with a better solution for labeling a stacked graph, the label can be placed in a reasonable position, and adjusted to the appropriate size, without overlapping other labels or layers. Thus, labeling hundreds of layers remains a tough challenge and seems inextricable only with a single static picture. Besides, labeling using Chinese is more difficult as the shapes of Chinese characters are more complicated than those in English.

In this paper, we would like to display labels representing higher sales volume houses directly on the graph with the remainder of the labels only visible when the user's mouse hovers on the graph, as shown in Figure 3(b). The font size S of the label is computed using the following formula:

$$S = L/T \times H,$$

where L is the number of sales of one house represented by the strip to be labeled, T is the total sum of the numbers of house sales dis-played in the graph, and H is the height of the display area. S should be smaller than or equal to a certain threshold, however, if S is too small, the display of the label needs to be disabled.

The streamgraph and minimized wiggle layouts contribute to the readability of the labeling because these place the baseline of the graph in the center of the display area, and thus the labeling can also be distributed around the baseline to reduce overlap.

5.3 Interaction

Users can resize the stacked graph window smoothly in the browser and the repainting of the graph is fast enough to satisfy real-time inter-action. Users can also highlight each layer to view the trend of a single house by hovering the mouse over the graph in the browser, and having detailed values pop up.

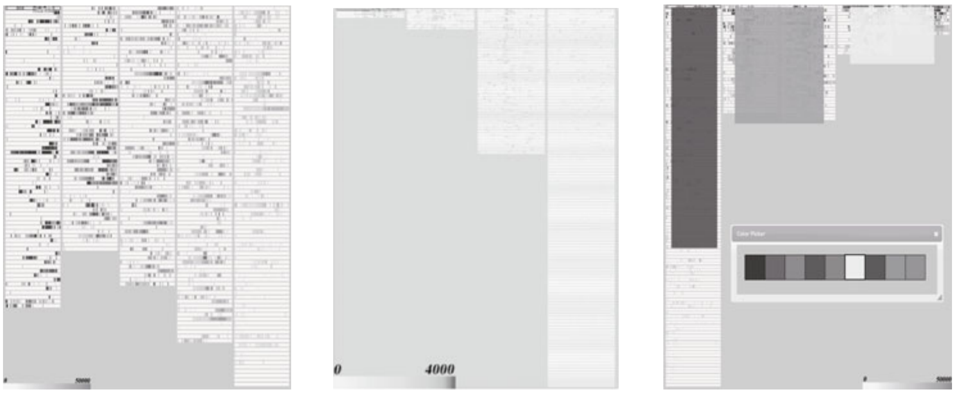
Users can also alter the time granularity, e.g., weekly or monthly, to view overall or detailed trends of the sale.

6 Clustered pixel bar view

6.1 Design scheme

In the previous section the sale volume was presented by aggregating all the data into one graph. However, users cannot look at the individual visualization of a single entry without any interaction. In this section, we present the clustered pixel bar view to visualize each house's data with one pixel bar and segment the price or number of sales into different ranges.

Each pixel bar in the pixel bar view in Figure 1 represents a house, where the horizontal axis denotes the time dimension which was set to four years based on our data, each vertical line of pixels represents the summary value (the sum of sales or the average price) in one week or month, and color is used to encode this value. The price ranges widely from ten thousand to forty thousand, but since the number



(a) (b) (c)

Figure 4 Clustered pixel bar view of all the houses. (a) Clustered pixel bar view based on house price; (b) clustered pixel bar view based on house sales; (c) five clusters representing different intervals of sales volume and price, with the color picker shown in the pixel bar view.

of sales ranges from single digits to thousands, we apply the logarithm operation to the number of sales to avoid color monotony. With the color distribution in the pixel bar, we can find the duration of the house sale and general trends in the real estate market.

As each pixel bar must be the same height as the display area, whereas the height of the line graph is determined by the actual data, pre-senting many entries using this technique helps save display space and a direct comparison among these tidy pixel bars can be perceived more easily than traditional line graphs through color.

Directly plotting the pixel bar on the graph may lead to confusion, making it difficult for users to discover existing and hidden patterns in the data. Therefore, according to the specificity of real estate data, automatic computation for the visualization is necessary. We would like to per-form clustering on the pixel bar based on the price or the number of sales.

Currently there are various clustering algorithms in the data mining field. As the clustering is carried out in the user's browser, we selected the k-means algorithm [15] since it is fast, has the ability to set the number of clusters, and is easy to implement.

First, users select the houses they want to visualize using lasso, and the desired number of clusters (denoted as n) in the option interface. Next, the system randomly selects n entries as the original cluster centroids. Then the algorithm iterates over all the entries to find the nearest cluster to each entry, where the distance is simply defined as the difference in the feature attributes of two entries such as the average price or total sales, and adds the entry to the nearest cluster. Thereafter, the cluster centroid is updated based on the average value of the entries in the cluster. The iteration terminates when there is no further change in cluster centroids; we found that generally the algorithm terminates within five iterations.

Each cluster is placed in one column in the pixel bar view's windows, and each pixel bar in a cluster is sorted in descending order, as shown in Figure 4. From the number of lines in each column in the graph, we can see the distribution of the different intervals of price or number of sales of the selected houses.

The method mentioned above can only reveal a single attribute, such as the house price or number of sales. We propose a cluster method on the pixel bar which combines both house price and number of sales. As we can see, different houses have different time intervals for sales records and the sale record of each house generally occurs at a different time, i.e., the time series is either continuous or discrete.

Thus, it is a challenging problem to cluster such time series. We propose dividing the houses into several segments where the houses in one segment are close in terms of number of sales, and sorting the houses in one segment based on the house price. First, we calculate the sum of sales for each house and sort them in ascending order, as shown in Figure 5. We use the method known as the identification of perceptually important points (PIPs) [12] and extract points that can represent the main characteristics of the time series graph. The PIP identification process is demonstrated in Figure 5, where the starting and end points are the first two PIPs, while the third PIP is the point with the maximum distance to the

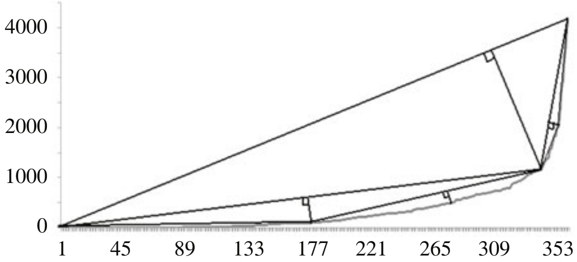


Figure 5 PIPs identification process. The x- and y-axis represent the house ID and the number of sales, respectively.

straight line connecting the first two PIPs. The distance is calculated as the vertical distance between the next PIP and the line as follows:

$$d = \frac{k \times x_3 - y_3 + b}{\sqrt{k^2+1}}, \quad k = \frac{y_2 - y_1}{x_2 - x_1}, \quad b = y_1 - \frac{y_2 - y_1}{x_2 - x_1} \times x_1,$$

where $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ are the starting and end points, respectively, and $P_3(x_3, y_3)$ is the point with the maximum distance to the line connecting P_1 and P_2.

The following PIPs will then be the points with the maximum distance to the previous adjacent PIPs. After obtaining the PIPs of the time series graphs, we generate several house sales volume segments. For each segment, we place the houses in one column from top to bottom in descending order based on average price (see Figure 4(c)).

6.2 Interaction

We implemented a color picker tool to help users select an appropriate color to interact with the pixel bar view, where the color category is a qualitative color group in the ColorBrewer [16]. The qualitative color scheme does not imply order; instead it is the differences in kind, which are used to differentiate clusters in our system. After selecting the color, users can select the pixel bar that needs to be highlighted in other visualization components with a rectangular marquee using that color (see Figure 4(c)). A time slider was also implemented to select a sales interval to help analyze the sales trend and geographic distribution at different times.

The window of the clustered pixel bar view can also be resized or hidden to fit the browser in real-time.

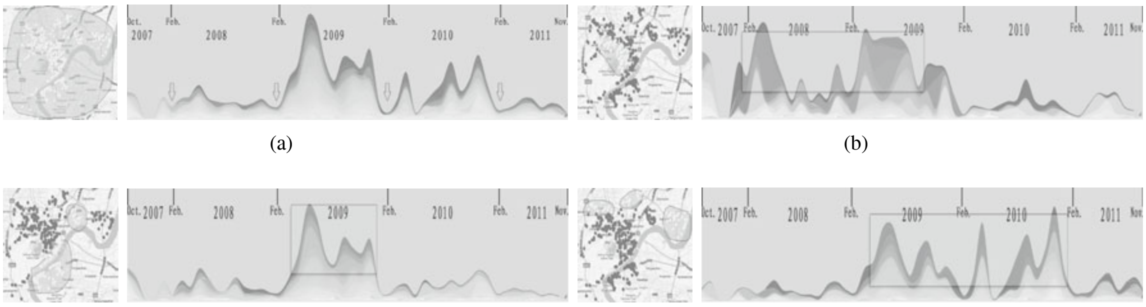
7 Experiments

Our system was implemented using J2EE at the server side and HTML5 through the Processing.js[13)] library at the client side. Processing.js is a port of the Processing Visualization language. The system can work in any HTML5 compatible browser without any plug-ins, including current versions of Chrome, Firefox, Safari, Opera, and Internet Explorer. With WebGL supported in HTML5, the rendering of complex pictures is much faster using personal computers. Our system was tested on an Intel® Pentium® Processor E5300 (2.66 GHz) desktop with 3 GB RAM and an NVIDIA Geforce G100 GPU with 512 MB RAM.

The data grabbed from the sites mentioned in Section 3 were preprocessed to generate corresponding data structures to adapt to different visualization components, with outliers such as zero values removed. We were able to obtain valid sale data for more than 40,000 sales entries, and our system supported interactive real-time visual display and user interaction.

In the following, we demonstrate the usability of our system through several case studies.

13) Processing JS. http://processingjs.org



(a)

(b)

(c)

(d)

Figure 6 Stacked graphs of houses in different areas. (a) Stacked graph of all the houses; (b) stacked graph of houses in the CBD and western part of the city; (c) stacked graph for BinJiang, JiuBao; (d) stacked graph of houses in XiaSha and the northern part of the city.

7.1 Usability for decision-makers

First, users can use our system to explore the patterns and trends in the numbers of sales using the stacked graph view over the years to obtain some macroscopic patterns and help decision-makers to gain a better understanding of the large dataset.

To begin with, users select all the houses with the lasso tool to create a stacked graph view with the normal stacked layout as shown in Figure 6. After scrutinizing the image carefully, we find that a relatively low ebb occurs in January or February every year, because the Chinese Lunar New Year occurs at this time followed by the Spring Festival holidays. This situation is mainly caused by the closing of part of the property market and transport during the Spring Festival, which leads to the phenomenon that millions of people are at home.

As a result of the Subprime Mortgage Crisis in 2008, which was directly related to the real estate market bubble, the market went from bad to worse; visualization helps present the effect. As shown in Figure 6, the number of sales in 2008 was lower than that in any of the other years.

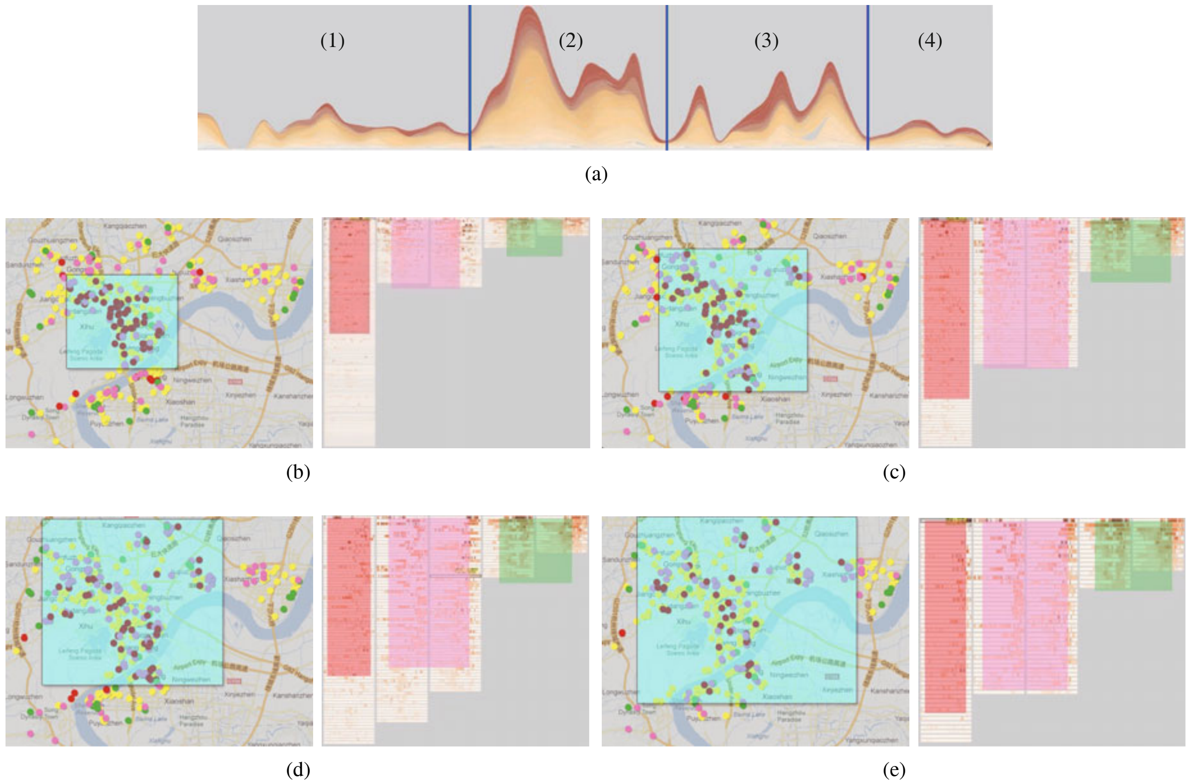
Although the global economic crisis still continued, house sales in the following year, i.e., 2009, behaved quite differently from that in 2008. The sales in that year grew phenomenally and it were almost three times greater than in 2008. After observing this anomaly, we found out that towards the end of 2008, the state and local government in China proposed a series of bailout policies to stabilize the real estate market, including loosening restrictions on mortgages for buying a second home, reduction and exemption of building tax, and the buying homes residence policy. Thus, real estate had great appeal for buyers with rigid demands, which had been accumulating for nearly half a year in 2008. The trading peaks in 2009 were caused by the housing trade fair in the spring and autumn, which also affected the trading in other years.

International Labor Days in May 2009 and other years also lead to a trading peak every year because of the long holiday. However, followed by the housing boom in March and April in 2010, the government introduced several policies in late April, for example, suspending mortgages for third houses. With the market more heavily regulated, the number of sales decreased.

In 2011, restrictions on the purchase of houses, which in fact, constituted the most severe regulation in history, were proposed by the state and local government, resulting in much lower house sales than in previous years.

After observing the stacked graph of all the houses, we focus a comparison among different districts and in an attempt to unveil urban development throughout these years.

Figure 6(b) displays the stacked graph for the central business district (CBD) and the western part of the city, while Figure 6(c) depicts that for BinJiang district in the southeast and JiuBao district in the northeast of Hangzhou. The stacked graph for the XiaSha district in the northeast and the northern part of the city is shown in Figure 6(d). We can see that house sales in Figure 6(b) mostly occurred during 2008 and 2009, while the houses outside the CBD and along the Qiantang River were mainly sold in 2009



(a)

(b)

(c)

(d)

(e)

Figure 7 Time segmentation and pixel bar for different time periods, together with house distributions of different attributes. (a) Time segmentation of all the years based on the number of sales; (b) graph from Oct. 2007 to Apr. 2009; (c) graph from Mar. 2009 to Jan. 2010; (d) graph from Feb. 2010 to Mar. 2011; (e) graph from Apr. 2011 to Oct. 2011.

(see Figure 6(c)). The majority of houses further from the CBD, which are shown in Figure 6(d), were sold in 2009 and 2010. Thus, these stacked graphs are useful for discovering and verifying that there was an obvious trend of outward expansion in Hangzhou during these years. We can see that the city expanded mainly to the east and north, and most of the newly built and sold houses are distributed along the Qiantang River.

This example clearly confirms the ability of our system to find patterns and trends. Moreover, decision-makers from different departments can learn about the complexity of the data, and gain valuable insight, which is conducive to macroeconomic regulation and control.

7.2 Usability for ordinary users

Our next case study shows that the clustered pixel bar view in our system helps reveal the relevance among the expansion of the houses' geographic locations, the rise in house prices, and the distribution of house sales. Convenient access to our system helps users who have rigid demands and wish to buy houses or users who want to invest in the real estate market to ascertain both the current situation and the evolution of the market.

First, we apply a time filter to the pixel bar, so that pixel bars with different time periods can be compared. The time segmentation of these years is based on Figure 7(a). According to the characteristics shown in the stacked graph view in Figure 7(a), we divide the sales period into 4 time intervals, i.e., Oct. 2007 to Apr. 2009, Mar. 2009 to Jan. 2010, Feb. 2010 to Mar. 2011, and Apr. 2011 to Oct. 2011. For each period, house sales are either booming or quiet. We would like to visualize the houses sold during these intervals using the pixel bar view and to cluster the pixel bars using the algorithm mentioned in Section 5.

We start by filtering out the houses sold between Oct. 2007 and Mar. 2009 when the number of sales was relatively low, and then we cluster the filtered houses into 5 clusters, arranged from left to right in the pixel bar view (see Figure 7(b)). The numbers of houses sold in the right clusters are greater than those in the left clusters. For each cluster, the houses are arranged in descending order from top to bottom based on the house's average price, and the color in each pixel bar is used to encode the house price. In this way, classes with different intervals of price and sales volume can be generated; e.g., houses with higher prices and lower sales volumes appear in the upper-left corner of the pixel bar view, while houses with the highest sales volume appear on the right side of the view where the price is relatively low as confirmed by the color.

Similarly, we create other clustered pixel bar views for the remaining time periods (see Figure 7 (b), (c) and (d)).

Thereafter, we use a ColorBrewer-based color picker tool to select the houses in the pixel bar view that we want to highlight on the map using rectangle marquees.

In Figure 7(b), we can see that most of the houses with higher prices and lower sales volumes are located in the center of the city and are colored red, while highest sales volume houses (colored green) are mainly located in the suburbs of the city and their price is relatively low. The higher sales volume houses (colored pink) are mostly located between the center and suburbs of the city where the price is at an inter-mediate level. The yellow markers denote the houses that are not highlighted.

In the same way as before, the areas are selected using rectangle marquees and the same color schemes, and then the house markers are also highlighted (see Figure 7 (b), (c) and (d)). By scrutinizing these images carefully, we find that most houses with higher sales volumes are distributed outside the city (colored green). Meanwhile, the red house markers, which represent higher price houses have expanded outwards over the years. Some red outlier markers can be found in the southwest of the city, which is quite far from the city in these figures. Having obtained detailed information for these houses, we know that they are villas in the south of Hangzhou.

7.3 User feedback and expert interview

We asked ten ordinary users who wanted to buy houses, five users who wanted to invest in the real estate market, and a research group consisting of seven people who had domain expertise in real estate market research, for feedback after using and exploring our system.

Both the ordinary users and expert users were impressed with this interactive visual analytics system, which provides different coordinated visualization components. The feature of direct access of the Web-based system was highly appreciated by the users. Compared with traditional visualization graphs on current real estate websites, where only limited functions such as line graphs and pie charts are supported, they found that our system, which integrates collaborative visualization components, helps to make the exploration process more visual and convenient. They also stated that our system assists in ascertaining exactly and quickly where, what, or when to buy or invest. Users who wanted to invest in the real estate market remarked that the stacked graph view gives a clear visual outward expansion of the city, which helps in making business decisions. Furthermore, those wishing to buy houses commented that geomap and treemap views provided an intuitive or adjusted distribution of house prices or sales volumes.

In addition to commenting on our system's functionality, the expert users also made several suggestions for improvement, which were quite valuable and constructive. For example, they talked about introducing more real estate models such as the Hedonic Price Model. Meanwhile, the price of land should also be considered in our system to visualize the interaction between different factors of the market. Evolution of urban areas and the process of population migration represented by the market in which they are interested should also be included in our system.

8 Conclusion

In this paper, we addressed the problem of the real estate market using information visualization and presented a Web-based system to help ordinary users and experts perceive the data quickly and gain appropriate insights. The system consists of various visualization components, which visualize different attributes of the data and work collaboratively as a whole. Some well established interaction techniques such as lasso selection, highlighting, and zooming are also integrated into our system. Our evaluation, which included user studies and feedback from ordinary users and domain experts, demonstrated the flexibility and effectiveness of our system. In addition, HTML5 plays as a powerful role for data visualization in our system, and we believe that HTML5 can be applied to other visualization applications that also have a variety of users.

There are several avenues for future work. First, we intend to add other cities' real estate data to our server and integrate an online database, containing financial and country policies and news, into our system, so that the analysis of real estate data can be based on reliable background information. We plan to include more interactive visualization techniques into our system to satisfy additional analysis needs and complement each of the components. Finally, we want to take advantage of this open Web platform and develop this system into a social ecosystem site that allows users to engage their collective intelligence, such as communication on the visualization results.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. 61070114), Zhejiang Provincial Natural Science Foundation (Grant No. Z1090630) and Program for New Century Excellent Talents in University (Grant No. NCET-12-1087).

References

1 MacEachren A, Kraak M. Research challenges in geovisualization. Cartogr Geogr Inf Sci, 2001, 28: 3–12

2 Takatsuka M, Gahegan M. GeoVISTA Studio: a codeless visual programming environment for geoscientific data analysis and visualization. Comput Geosci, 2002, 28: 1131–1144

3 Hardisty F, Robinson A. The GeoViz Toolkit: Using component-oriented coordination methods for geographic visualization and analysis. Int J Geogr Inf Sci, 2011, 25: 191–210

4 Malik A, Maciejewski R, Maule B, et al. A visual analytics process for maritime resource allocation and risk assessment. In: IEEE Conference on Visual Analytics Science and Technology, Providence, 2011. 221–230

5 Slingsby A, Dykes J, Wood J. Configuring hierarchical layouts to address research questions. IEEE Trans Vis Comput Graph, 2009, 15: 977–984

6 Havre S, Hetzler B, Nowell L. ThemeRiver: visualizing theme changes over time. In: IEEE Symposium on Information Visualization, Salt Lake, 2000. 115–123

7 Byron L, Wattenberg M. Stacked Graphs—Geometry & Aesthetics. IEEE Trans Vis Comput Graph, 2008, 14: 1245–1252

8 Keim D, Hao M C, Ladisch J, et al. Pixel bar charts: a new technique for visualizing large multi-attribute data sets without aggregation. In: IEEE Symposium on Information Visualization, San Diego, 2001. 113–120

9 Ziegler H, Jenny M, Gruse T, et al. Visual market sector analysis for financial time series data. In: IEEE Symposium on Visual Analytics Science and Technology, Salt Lake, 2010. 83–90

10 Liao T W. Clustering of time series data—a survey. Pattern Recognit, 2005, 38: 1857–1874

11 Liao T W, Bolt B, Forester J, et al. Understanding and projecting the battle state. In: Army Science Conference, Orlando, 2002. 2–5

12 Fu T, Chung F, Luk R, et al. Financial time series indexing based on low resolution clustering. In: 4th IEEE International Conference on Data Mining, Brighton, 2004. 5–14

13 Oates T, Firoiu L, Cohen P R. Clustering time series with hidden Markov models and dynamic time warping. In: Proceedings of the IJCAI-99 Workshop on Neural, Symbolic, and Reinforcement Learning Methods for Sequence Learning, Stockholm, 1999. 17–21

14 Wang L, Mehrabi M G, Kannatey-Asibu E. Hidden Markov model-based tool wear monitoring in turning. J Manuf Sci Eng-Trans ASME, 2002, 124: 651–658

15 MacQueen J. Some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, 2007. 281–297

16 Harrower M, Brewer C A. ColorBrewer.org: An online tool for selecting colour schemes for maps. Theor Mapp Pract Cartogr Represent, 2003, 40: 27–37