CPSC 533 Analysis Project

Construction Change Order Visualization Analysis

Chiu, Chao-Ying

Department of Civil Engineering, University of British Columbia

1 INTRODUCTION - DESCRIPTION OF PROBLEM: TASK, DATA

1.1 Global Problems of Utilizing Data in Construction Industry

During the course of a construction project, voluminous data pertaining to the construction process is generated and recorded in different formats. Those formats of construction process data are: filled out preprinted forms, workbooks, and logs thereof; plain text documents like contracts, memorandum, e-mail, meeting minutes, etc; pictorial documents like drawings, pictures, and videos, etc. Except for some forms, workbooks, or logs are recorded as digitized abstract database, rests of other data are stored in paper format or digital format (digital video for example). This current practice of modeling and organizing construction data makes analyze them very difficult. For example, how can we associate a monthly cost account database with several digital videos and conduct some analysis? However, it is doable to transform physical data into categorical data so that they can be stored in databases for data analysis. So a collection of video clips can be categorized as video clip of first floor construction, of second floor construction, and so on. Or we can mine text documents and categorize the content. Therefore, a long description of change orders can be classified into whether the description is about "reason of change order" or "location of change order", etc. Except for very large sized contractor company, seldom are the aforementioned data modeled and organized in information systems. The reason that the construction industry is refrained from doing so is the effort of integrating them into an abstract database system is costly and needs know-how. Furthermore, the heterogeneous nature of the modeled data also makes data analysis extremely difficult. Although visualizing data for exploratory data analysis is recognized, lack of adequate medium for doing so still hampers the breakthrough of effectively and efficiently analyzing data thereby assisting the management function and enhancing project performance. Since the industry can not analyze data in support of enhancing performance, the driver of modeling and organizing data is lost in return.

Currently, we have successfully modeled the construction process data in Repcon, a research system developed in the University of British Columbia. However, due to the lack of good data analysis technology, its functionality of enhancing all management functions is limited. We believe data visualization could be one of the effective data analysis solutions. However, the current visualization techniques the industry or the academia of construction management field use are still limited to Excel built-in chart features or graphics tools. It turns out that those tools are ineffective when dataset is large in dimensions. The amount of construction process dataset is not large comparing to business transaction data, but the number of dimensionality of data is high. Russell and Udaipurwala's nine views [1] of a construction project represent the nine major aspects of a complete project though some of them may overlap. Those nine aspects all have their own data hierarchy ranging from four levels to six levels of detail. It is not hard to imagine the dimensionality of constriction project data. Currently, except the construction activity scheduling data, resource usage data, cost spent data, and physical product data, the only small portion of the entire construction data, have been stored and interpreted visually by scheduling software (Microsoft Project and Premivera P3) and CAD software (Autodesk Revit Building), rest of the other data is seldom analyzed in a regular manner for the purpose of supporting management functions. Excel's built-in charts might be able to visualize a few of them mostly for the rare presentation situations, but they are incapable of visualizing data when full spectrum of construction data is included. The dimensionality of data scales up exponentially and so is the number of images required for visually analyzing data. Therefore, this project is trying to search latest information visualization technology in support of utilizing multidimensional construction data, prove the technology is superior to current industry practice of using Excel, and finally establish some guidelines of interpreting data by using information visualization technology. The selected domain application is change order management and the targeted information visualization techniques are whatever specializes in visualizing multi-dimensional data. In the following subsection, we will briefly explain the domain, task, and data of the change order application. Then we will review the related work of both the data visualization of related domains and most relevant information visualization technology in section two. We will explain how we choose adequate tools and how to use the techniques more efficiently and effectively by running through task scenarios in section three and section four. Images of those task scenarios generated by selected information visualization tools will be presented in section five. In section six we will summarize some guidelines of interpreting data using information visualization technology and validate the success of this project by explaining the improvement. Lastly, we will give out some lessons learned from this project.

1.2 Domain, Task, and Data of the Change Order Application

1.2.1 Domain Description

In a normal construction project, the facility owner hires architects and engineers to design the facility and monitor the construction process. Also there is a general contractor chosen to build the facility according to the design. Due to the usual omissions or mistakes of designs, unexpected environmental conditions, and owners' changing needs of the facility, many alterations of planned construction work are required during the construction phase. This change is normally initiated by contractors when they find discrepancy between actual conditions and design conditions assumed by clients. Usually the contractor will first submit the RFI (request of information), and then the architect or the engineers will reply by issuing SI (site instruction). If the

discrepancy is admitted by architects and engineers and reflected on the SI, the contractor will start to prepare a request of additional cost and time extension by submitting request of change orders. In addition to the discrepancy-caused change orders, sometimes the owner will make architect or engineers change original designs, hence another type of change orders will be generated due to the design changes. Whenever there are change orders, contractor-initiated or owner-initiated, impacts on project cost and schedule are always generated. Usually the owner would assume the additional cost and time by approving the submitted requests of change orders. However, sometimes the owners disagree with the amount of responsibility and reject the request, hence underlie the future claim.

The major issues centre on the so called "change management" are to minimize the cost impact resulting from changes, minimizing changes, reducing claims, and reducing the cost of responding to rejected change orders. For the aforementioned management functions, there is corresponding information required for executing them. Unfortunately, the information is hidden in large amount of unstructured data like RFIs, SIs or daily site reports. As a result, it is difficult to identify information so as to manage changes effectively. Therefore, the information visualization techniques appear to be a solution for efficiently and effectively extracting information.

1.2.2 Data Description

The dataset available are spreadsheet data that the construction staffs organizes from paper forms of change orders registry, site instruction, and request of information. This dataset comes from a real building project on campus of University of British Columbia, and were collected during March 2004 through April 2005. The dimensionality of this dataset is sixteen and there are 176 logs in the change order registry. The names and types of data items are shown in the Figure 1.

Change Order Registry							
Data Field Name	*Change Order Number*	*Issued Date*	*Projected Cost*	*Date Approved*	*Approved Cost*	*Reason of Change*	*Reference Number*
Data Abstration Type	Raw	Raw	Raw	Raw	Raw	Raw	Raw
Data Field Type	Dimension	Dimension (Time)	Measure	Dimension (Time)	Measure	Dimension	Dimension

Affected Sub-Trade				
Data Field Name	*Change Order Number*	*Trade*	*Trade Revision Number*	*Trade Change Order Amount*
Data Abstration Type	Raw	Raw	Raw	Raw
Data Field Type	Dimension	Dimension	Dimension	Measure

Initiated Document		
Data Field Name	*Change Order Number*	*Initiated Document*
Data Abstration Type	Raw	Raw
Data Field Type	Dimension	Dimension

Affected Physical Location			
Data Field Name	*Change Order Number*	*Location*	*Sub-location*
Data Abstration Type	Raw	Raw	Raw
Data Field Type	Dimension	Dimension	Dimension

Affected Physical Component				
Data Field Name	*Change Order Number*	*Major Group Elements*	*Group Element*	*Individual Element*
Data Abstration Type	Raw	Raw	Raw	Raw
Data Field Type	Dimension	Dimension	Dimension	Dimension

Figure 1: Data structure of change order data

1.2.3 Task Description

With change order data at hands, we want to identify suitable visualization tools, conduct exploratory data analysis by using the selected tools, rationalize the choices of visualization techniques, and conclude some principles of visualizing construction data. Since the nature of construction data is multidimensional abstract data, we deem automating optimum visual encodings of selected data, easiness and effectiveness of querying data of different dimension and value ranges, and addressing time dependent property of data as the most important criteria of choosing visualization tools. The available techniques fulfilling these criteria are discussed in the section two.

2 RELATED WORK

2.1 Related Work of Relevant Domains

2.1.1 Construction Management Domain

Computerized visualization of construction management related data is used intensively in the application of monitoring the construction activity progress, product completed, resources consumed, and cost spent. The often used Gantt chart linked with 3D CAD (Figure 2.1) and charts of resources consumed and cost spent are already adopted by construction industry and several commercial software are available. In addition to the activity sequencing data, physical product data, cost account data, and resource usage data, much more data collected during the construction process that could provide the insight of construction project performance begins to draw researchers' attention and the need of visualizing more aspects of construction data increases. In [2] computer graphics software that offers the functionality of zooming/panning images in addition to 2D and 3D charts generation are used to visualize resources consumption data, location data, and productivity data. Simply by juxtaposing 2D bar charts or 3D bar charts of combination of aforementioned data, the insight of resources usage distributed along time and space, and productivity distributed along different locations are obtained (Figure 2.2). In the case of visualizing resources usage distributed along time and space, it helps quickly understand the site congestion problem if the clustering in the images are seen. Researchers also began to seek out non-traditional visualization images for interpreting the multidimensional construction process related data. In [3], Treemap[4] is used to present monthly cost report data. The original cost report data is eight dimensional, and the Treemap utilize texts, sizes of rectangles, two color saturations, and two levels of detail to visualize five dimensions (Figure 2.3). In the application of analyzing change order management domain data, use creativity and extensively utilize Excel's built-in chart features to visualize them. By drawing two 3D bar charts, six dimensions of the change order data are visualized (Figure 2.4). The distribution of change order along location, time, and sub-trade are easily observed.

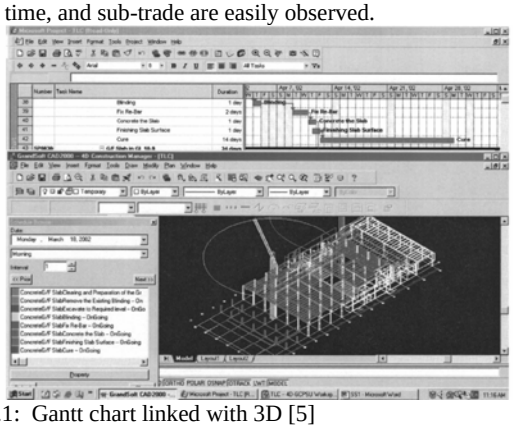


Figure 2.1: Gantt chart linked with 3D [5]

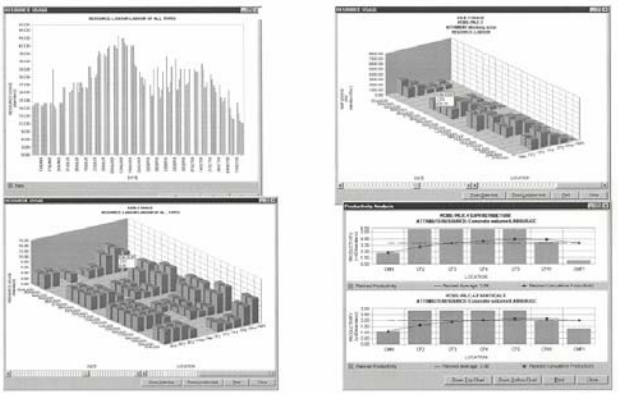


Figure 2.2: Juxtaposition of 2D and 3D bar charts to visualize resources consumption data, location data, and productivity data

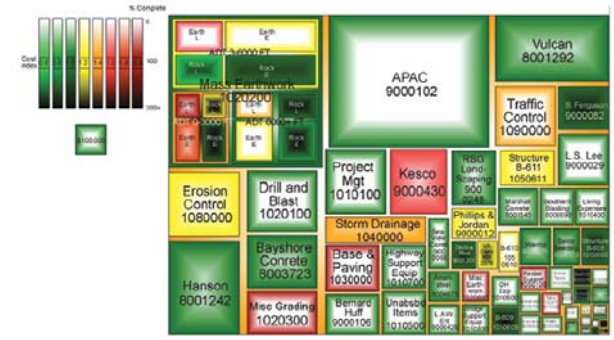


Figure 2.3: Treemap for visualizing monthly cost report data

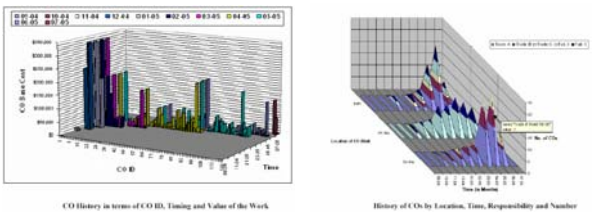


Figure 2.4: Excel generated 3D bar charts for visualizing 6 dimensional change order data

2.1.2 Business and Manufacture Domain

The advance of computer technology is partly driven by business activity needs. Therefore, the utilization of information system in business industry is more mature than construction industry. And so does the computerized data analysis techniques. The business industry is more willing and capable to adopt new technology, and it has much higher demand of atheistic aspects of the visual effect, even they are costly and redundant. In [6], researchers survey the state-of-the-art information visualization techniques used by business industry. Coincided with findings in other reports, virtual reality metaphors for visualizing abstract data are used intensively. 3D landscapes integrating aesthetic maps and charts hung on the "walls or floors" (Figure 2.5) are usually observed in the visualization tools specific to business application.

In the manufacturing setting, statistics graphing charts are used often in support of the task of controlling product quality and problems identification. James uses multiple views of data, interaction with those views, and links amongst views to interactively analyze data of semiconductors and optical fibers manufacturing [7]. Figure 2.6 shows a data browser consists of views of map of machines, metrics legend, line charts corresponding to the metric and machine observed, and the box-plots.

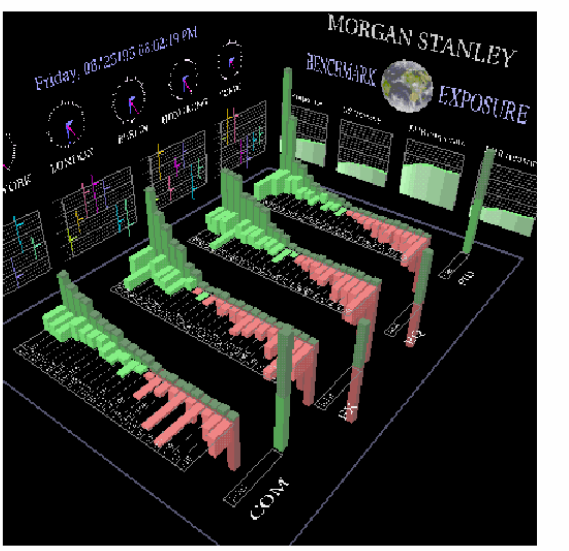


Figure 2.5: "Floor and wall" metaphor for visualizing business data (from [6])

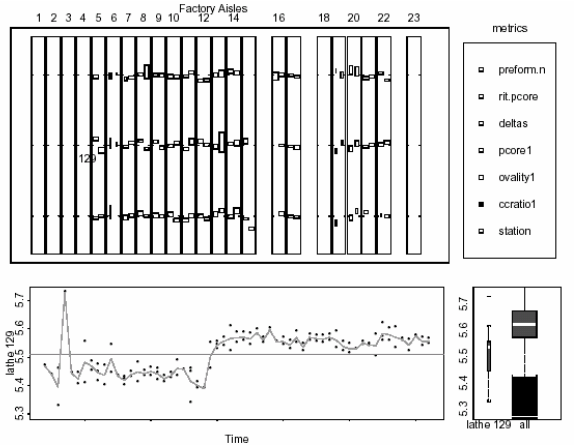


Figure 2.6: Linked views of data of optical fibre manufacture domain

2.2 Related Work of Visualization Technique

2.2.1 Visual Encoding Formalism

Bertin is the first one trying to formalize the generation of still images of data [8]. The formalism of mapping data of different measurement nature (categorical, ordinal, quantitative), visual marks (point, line, area) and retinal variables (e.g. size, shape, color), and organizations thereof were established. Cleveland and Robert further provide that there is an order of visual encodings in terms of the effectiveness of conveying measurement scale of quantitative data [9]. Position of visual encoding gets the highest rank. Mackinlay continues to expand the visual encoding

effectiveness ranking to ordinal data and categorical data [10] (Figure 2.7). He also develops a primitive graphical language that automates the design of graphical presentation of relational data. Stolte, et al. develop a visual specification and system that can automatically generate optimum images within a single table according to measurement scales of data (visual encoding selection) and combinations of data (layout selection) [11]. In Eick's guidelines [12] of engineering perceptually effective visualization of abstract data, it recommend better choices of positioning visual encoding (the term of glyph is used in the paper) and data encodings.

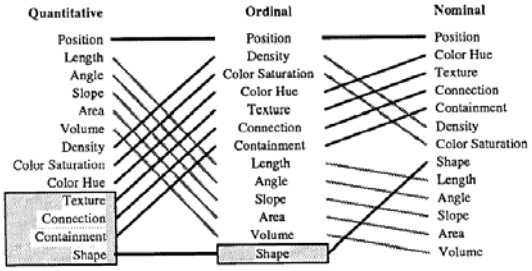
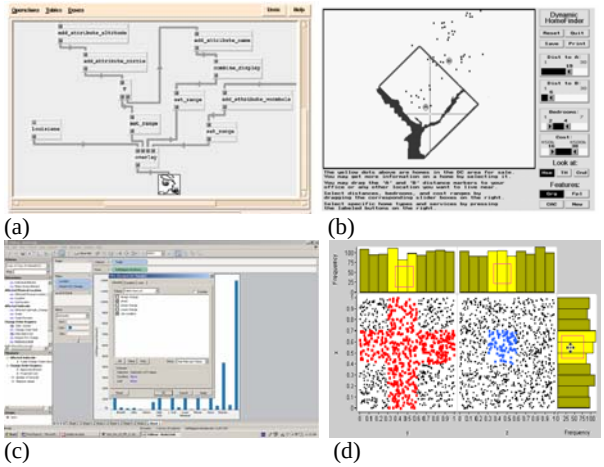


Figure 2.7: Visual encoding effectiveness ranking by Mackinlay

2.2.2 Visual Query

There are two types of visual query techniques. One is that users still have to instruct the query using SQL language, but in a more user friendly graphical interface. The graphical interface widgets for querying data include sliders [13], filter check boxes and data shelves [11], and visual primitives [14]. Then the queried data are retrieved and presented visually so that users grasp the content of data more quickly and can refine next query more easily. The other type is users can directly manipulate the visual objects of data by brushing thereby the selected data will be highlighted whereas filtered out data are dimmed. The interfaces of the mentioned techniques are shown in Figure 2.8.

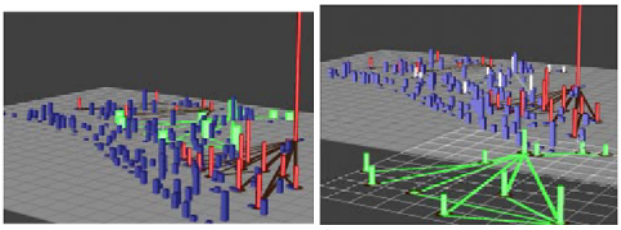


(a) (b) (c) (d)

Figure 2.8: (a) Visual primitives (b) Dynamic query slide (c) Filter check boxes and data shelves (d) Query by Multiple brush (from [15])

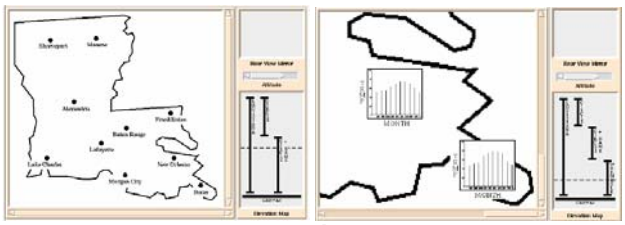
2.2.3 Interaction

The interaction with visualization includes zooming, panning, rotating, moving, and brushing images or visual encodings of the images. In [16], visual encodings of data are treated as objects so that users can rotate, lengthen, move, highlight or dim them through direct mouse manipulation on them (Figure 2.9). This kind of interaction helps solve the occlusion problems of 3D images. In [14], users can zoom in a map and see another related pool of data or level of detail of data, not just to magnify the visual encodings (Figure 2.10). This kind of interaction is called semantic zooming and it help user first gain overview of one data, and then proceed to see interested data in detail or see related data. Another useful interaction technique is coordinated multiple views (or called linked data views). Related data or same data are presented in different aligned views, horizontally or vertically. Any effect of interaction behaviors in one view is synchronized in other views. This coordinated synchronization include select in one view also select in other views; select in one view prompt navigate in other views; navigate in one view also prompt navigation in other views. Christopher did a comprehensive investigation on relevant techniques of coordinated multiple view [17]. This kind of interaction facilitates the use of linked views when conditional distributions are of interests [18]. Brushing is useful for interactively selecting subset of data directly on visualization of data using point devices. Cleveland uses two-dimensional/rectangular brushes to highlight, label, and delete data points in matrices of linked scatter plots [19]. Willis uses different types of brushing to replace, toggle, add, subtract, and intersect subset of data [20] (Figure 2.11).



(a) (b)

Figure 2.9: (a) 3D bar chart (b) Green bars are moved for closer examination



(a) (b)

Figure 2.10: (a) A map of weather stations (before zooming) (b) Weather data of 2 stations appears after zooming in

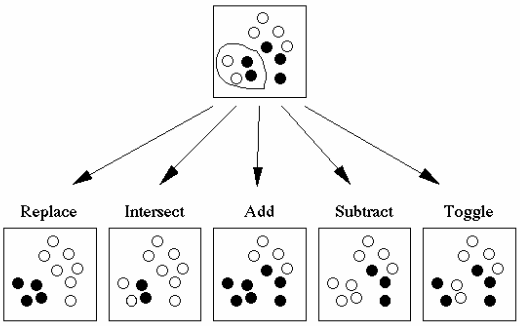


Figure 2.11: Five types of brushing and selecting

2.2.4 Visualization of time dependent data

Frank categorizes the time axis as four types. The four types are discrete time points as opposed to interval points; linear time as opposed to cyclic time; ordinal time versus continuous time; ordered time versus branching time [21]. MacEachren also identifies several crucial questions that we may pose on the time dependent data: Does data exist at a specific time? When does data exist? How long does data exist? How does data change along time? What is the sequence of data? How often does data occur? Which data coexist in a certain time point or period? [22]. Havre uses the river metaphor (Figure 2.12) to visualize liner time dependent, multi-dimensional data [23]. Carlis uses the spiral metaphor (Figure 2.13) to visualize cyclic time dependent, multi-dimensional data [24]. Ankerst [25] and Wijk [26] both use calendar metaphor (Figure 2.14) to visualize daily data. The former one focuses on observing trend of number of air planes maintenance events along the time in the unit of day (how does data change along time? which data coexist in a certain time point or period?) while the latter one is for grouping patterns of electricity consumption for different days (when does data exist?). Plaisant develops visualization similar to the Gantt chart [27] (Figure 2.15) to see the history of patients, which can answer the questions of: what is the sequence of data; which data coexist in a certain time point or period; how long does data exist.

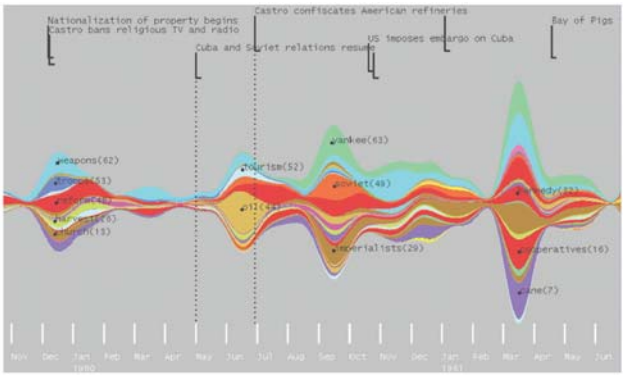


Figure 2.12: River metaphors for visualizing liner time dependent, multi-dimensional data

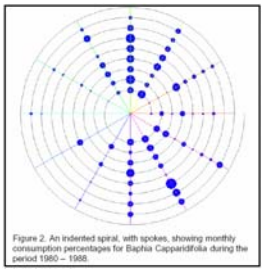


Figure 2.13: Spiral metaphor to visualize cyclic time dependent, multi-dimensional data

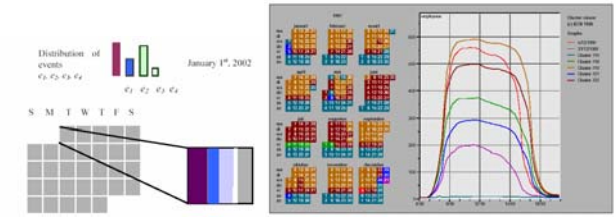


Figure 2.14: Calendar metaphor visualization of time dependent data

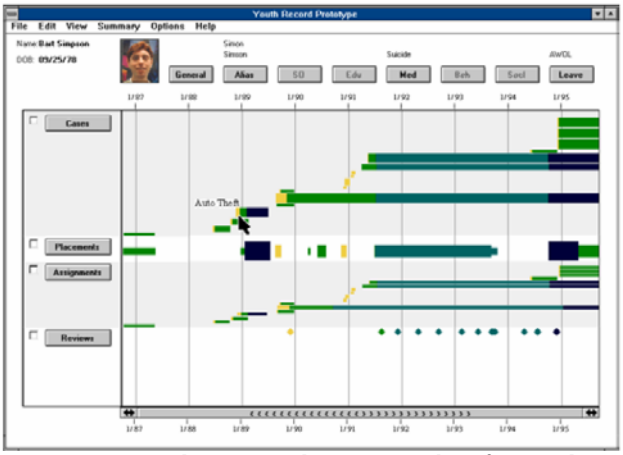


Figure 2.15: Visualization similar to Gantt chart for visualizing sequence of data, coexistence of data, and duration of data existence

3 DESCRIPTION OF SOLUTION: INFOVIS TECHNIQUES

3.1 The criteria of choosing information visualization tools

How we choose information visualization techniques can be approached form management aspects and technology aspects.

From the management point of view, since our ultimate objective is to convince the industry that there are already many valuable insights hidden in routinely collected data which can greatly assist in many construction management functions, the visualization tool needs to have comprehensive capability of interpreting data, be easy to use for general public, and have potential to integrate with physical data visualization (CAD drawings, digital pictures, electronic documents). So the solution to enhance construction industry's ability of visualizing data, from the management aspect, is to find mature visualization system as opposed to the novelty but infant solution which is only for specific tasks. The solution needs to be comprehensive enough but not necessarily to be the most effective. To this end, we aim for commercialized software that incorporates as many techniques mentioned in the section two as possible.

From the technology point of view, the first priority feature of the visualization solution is the ability to retrieve and present multi-dimensional data more intuitively and quickly. Although it is unlikely that the construction industry would build databases systems that deal with hundreds of thousands records like the bank industry does, the concept of retrieving and visualizing construction data is still similar to the database query and database visualization. So firstly we find the "visual query" feature is important and the "brush and query" technique is the most intuitive for general public to query data. Secondly, since there is at most 3D space in the computer screen while there are many dimensions of data needed to be presented, we think the tools that

can show different dimensions of data at users demand is also important. The feature of showing data of all dimensions in one single view like parallel coordinates or table lens is also intuitive for users to comprehend the dimensionality of data. Or users can create multiple views representing different dimensions, and then link them together. The coordinated multiple views techniques ca n support this functionality. Another feature is that users can instruct which dimensions of data to show, and then the system can automate the generation of the optimum images. By "optimum" we mean that the visual encodings of the image are expressive and effective. Therefore, the feature of Polaris project would suffice to this end. Lastly, all the visualization of abstract data will need to be associated with corresponding concrete data like 3D drawing of building or electronic documents ultimately; therefore the tools that support concept of "coordinated multiple views" again draw our attention.

As to visualizing time dependent data, we rule out techniques emphasize the cyclic time dependent data since any occurrence of things related to a construction project rarely reoccur dependent to the cyclic time. For example, we may enforce that each floor be finished weekly if the building is symmetrical, but too many uncertainties just make the industry difficult to maintain that. Or we may work faster so beat the cycle. It is not the case that bus driver will start earlier. Even being able to undergo the same activity weekly, unlike that bus routes and schedules remains unchanged every weekdays, doing work on second floor is never the same as doing work on 30th floor. Also we found out the frequency of construction data is low, but the data of a certain time point or time period is heterogeneous. Therefore we think vertically juxtaposing variety of data on the same time axis is good enough to visually answer the most of the questions posed by. For example, if we juxtapose Theme River with Gantt Chart of events, we may identify which events in which timing contribute to the change of trends. Although it is unlikely a comprehensive visualization system would implement the exact feature of Theme River or Gantt Chart, we can utilize the vertical bar chart and horizontal bar chart to simulate the similar idea.

3.2 The Choices of Visualization Tools

Based on the criteria analysis, we narrow our choices down to some major information visualization software vendors. The Tableau system is best in terms of automating the generation of optimum images. Users can simply select whatever dimensions of data and then the system will create the images for users. Rather than selecting range of data and assign chart types step by step in an ad hoc manner using Excel, the Tableau let users retrieve and present data relatively much more efficiently and effectively. Another system we favor is the Advisor system. The features of "linked data views" and "query by brush" could be very useful. Although users need to select chart types from Advizor's 15 standard chart, those charts has built-in restriction so that different chart types can only present certain combinations of data type. For example, since it is a bad practice to use length to convey categorical measurement scale of data, the Advizor restricts users from visualizing non-quantitative data by bar length of bar charts. Besides, it also only needs users to select data dimensions in order for generating charts of data. We think these two systems altogether do meet most of our required criteria of the visualization techniques and both of them are available, therefore we decide to test the desired visualization techniques on our construction data by using them.

4 SCENARIOS OF USE

We believe users' knowledge background decisively governs how users interact with and value a new system. A user who has some databases and information visualization knowledge would explore systems' functionalities faster and generate as many interesting images as possible. But he might not be able to interpret images' underlying insights if he is lack of basic domain knowledge. A seasoned construction manager may have better sense in interpreting images, but his lack of information technology knowledge and his stereotype domain knowledge would inhibit him from utilizing the tools in more depth. Therefore, one of the goals of this project is to establish systematic principles for domain experts to exploit the information visualization technology.

In terms of principles, we try to establish the formalism for scenario of use in addition to visualization techniques themselves. This formalism need to be systematic and simple so that data can be interactively retrieved and visualized in a mechanical way. Otherwise, the domain experts would just try data queries and images differently each time they encounter different scenarios even those scenarios might be the same in nature. For example, one scenario is: how many sub-trades are affected if the change orders occur between January 2004 and July 2004 and the reason of change is "design change"? Another scenario is: When did change orders involving main floor and interior construction occur? They look different at the first glance, but they are the same scenario if thinking about the nature of data. Those two scenarios both are to see how filtered subsets of data distributed along a certain dimension. The procedural steps of retrieving and visualizing data should be the same. Currently, we have observed 4 patterns of scenario uses, which are definitely not exhaustive. They are: find data distribution; compare data distribution; compare occurrence and trend of time dependent data; find associations between data. They are explained in detail in the following subsections

4.1 Find quantity distribution of data

"Data cubes" is a different concept of storing and organizing data that was first proposed by Microsoft a decade ago [28] Using the terminology of data cubes, data can be divided into two types of dimensions: Dimension and Measure. If values of raw data of a certain dimension are quantitative, that dimension is of Measure type. If values of raw data of a certain dimension are categorical or ordinal, that certain dimension is of Dimension type. **Here we define "quantitative dimension" and "non-quantitative dimension" as the Measure type dimension and Dimension type dimension referred in data cubes concept**. Now, take the simpler car sales data for example. The sales data has six dimensions: sales number, car models, year of manufacture, car colours, car sales, and profit. The first four dimensions are non-quantitative while the last two dimensions are quantitative. Each data value of a non-quantitative dimension has a corresponding data value of all other quantitative dimensions if they are one to one; each data value of a non-quantitative dimension has a statistical data value (total, average, mean, etc) of all other quantitative dimensions if they are one to many**. For the aforementioned values or statistical values, we define them as quantity measurement**. So in the car sales data example, the original relational raw data table (Figure 4.1.a) can be mapped to four separate tables (Figure 4.1.b). Or we can say the data are viewed from four different angles. Each table records the data values of different non-quantitative dimension (black texts) and their corresponding quantity measurements (red texts).

Number	Model	Year	Color	Sales	Profit	record count
N1	Chevy	Y1990	red	5	$10,000	1
N2	Chevy	Y1990	white	87	$15,000	1
N3	Chevy	Y1990	blue	62	$13,000	1
N4	Chevy	Y1991	red	54	$11,000	1
N5	Chevy	Y1991	white	95	$9,000	1
N6	Chevy	Y1991	blue	49	$8,700	1
N7	Chevy	Y1992	red	31	$7,600	1
N8	Chevy	Y1992	white	54	$9,450	1
N9	Chevy	Y1992	blue	71	$20,000	1
N10	Ford	Y1990	red	64	$12,000	1
N11	Ford	Y1990	white	62	$9,000	1
N12	Ford	Y1990	blue	63	$8,700	1
N13	Ford	Y1991	red	52	$7,600	1
N14	Ford	Y1991	white	9	$15,000	1
N15	Ford	Y1991	blue	55	$13,000	1
N16	Ford	Y1992	red	27	$11,000	1
N17	Ford	Y1992	white	62	$12,000	1
N18	Ford	Y1992	blue	39	$9,000	1

Figure 4.1.a: Data table of car sales data

Model	Total of Sales	Average of Sales	Total of Profits	Average of Profits	Total of Record Count
Chevy	508	56	$103,750	$11,528	9
Ford	433	48	$97,300	$10,811	9

Year	Total of Sales	Average of Sales	Total of Profits	Average of Profits	Total of Record Count
Y1990	343	57	$67,700	$11,283	6
Y1991	314	52	$64,300	$10,717	6
Y1992	284	47	$69,050	$11,508	6

Color	Total of Sales	Average of Sales	Total of Profits	Average of Profits	Total of Record Count
red	233	39	$59,200	$9,867	6
white	369	62	$69,450	$11,575	6
blue	339	57	$72,400	$12,067	6

Number	Sales	Profit	record count
N1	5	$10,000	1
N2	87	$15,000	1
N3	62	$13,000	1
N4	54	$11,000	1
N5	95	$9,000	1
N6	49	$8,700	1
N7	31	$7,600	1
N8	54	$9,450	1
N9	71	$20,000	1
N10	64	$12,000	1
N11	62	$9,000	1
N12	63	$8,700	1
N13	52	$7,600	1
N14	9	$15,000	1
N15	55	$13,000	1
N16	27	$11,000	1
N17	62	$12,000	1
N18	39	$9,000	1

Figure 4.1.b: Data tables grouped by “car model”, “year of manufacture”, “car colour”, and “sales number” respectively

Many questions we have about the data are how they quantitatively distribute along certain dimensions. By **quantity distribution of data,** we mean the distribution of quantity measurement along a non-quantitative dimension. So in the previous car sales example, for the “Model” dimension, the quantity distribution along it can be that “Total of Sales” of blue cars, red cars, and white cars are 233, 369, and 339 respectively; or it can be that “Average of Profits” of them are $9,867, $11,575, and $12,067 respectively. The visual distribution of the example can be seen in Figure 4.2. From the visual distribution, we see the original data table can be turned into systematic information nicely. Apply the same concept we also can try to find the quantity distribution of subset of data. The subset of data is obtained by filtering. Since the filtering is done by filtering out some unwanted values on some dimensions, we can only observe the quantity distribution of data along dimensions other than the filtered dimensions.

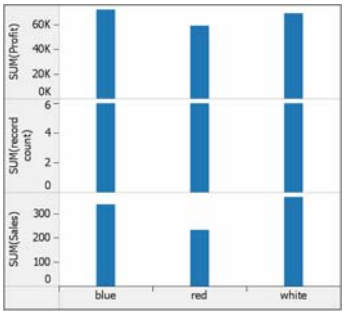


Figure 4.2: Quantity distribution along dimension “car colour” in terms of totals of “profit”, “record count”, and “sales”

One thing deserves notice is that not all non-quantitative dimensions are always in the same relational table. Although we can join them, non-quantitative dimensions that are crossing tables can not share quantitative dimensions. Anyway, in the case of dimensions crossing tables, the only quantity distribution considered is record counts.

Applying this formalism, now users can retrieve and visualize data to answer many questions with the same procedure using visualization tools. In Advizor, users can generate several bar charts in which data values of non-quantitative dimensions are positioned on the x axis while quantity measurements are represented by bar lengths. So now we have images of complete data’s quantity distribution on many non-quantitative dimensions. In Advizor, the bars of the bar charts can be coloured. Therefore if we filter out some data by subtract-brushing some bars, the “linked data views” feature will synchronize the change of coloured bars’ length because now we are visualizing a subset of data. Detailed screen shots and the accompanied scenario descriptions are illustrated in the section 5.1.1. In Tableau, we can drag a non-quantitative dimension along which we want to see the quantity distribution onto column shelf, and drag an interested quantitative dimension to row shelf. If we want to filter out some data, drag the other dimensions you desire to filter out to filter shelf, then filter out some values of those dimensions by unselecting values in the pop out dialogue box. Because of the “dragging dimensions to shelf” feature of the Tableau, users can be extremely fast in changing dimensions they desire to observe. Detailed screen shots and the accompanied scenario descriptions are illustrated in the section 5.1.2.

4.2 Compare quantity distributions of data

It is meaningless to compare quantity distributions of data on different non-quantitative dimensions. For example, it is no point to compare how car sales distributed along the “Car Color” dimension and the “Year of Manufacture” dimension because there is no common ground for comparisons. So we always compare quantity distributions of data on a same non-quantitative dimension. There could be two kinds of comparison. One is to compare how different quantity measurements distribute along a same non-quantitative dimension. For the car sales example, we can have quantity distribution along “Sales Number” dimension in terms of the “Profits” and “Sales”. The visual comparisons of these two quantity distribution is seen in the Figure 4.3. This kind of comparison helps us find correlations between quantitative dimensions. The other type is to compare quantity distribution of different subsets of data. These different subsets of data only differ in values of one non-quantitative dimension. For example,

we have a subset of previous car sales data whose car colour is red; we also have another subset of data whose car colour is blue. Now we put these two subsets of data along the same non-quantitative dimensions like “Year of Manufacture” or “Model” to compare quantity distributions of these two different subsets of data. The visual comparison is seen in Figure 4.4.

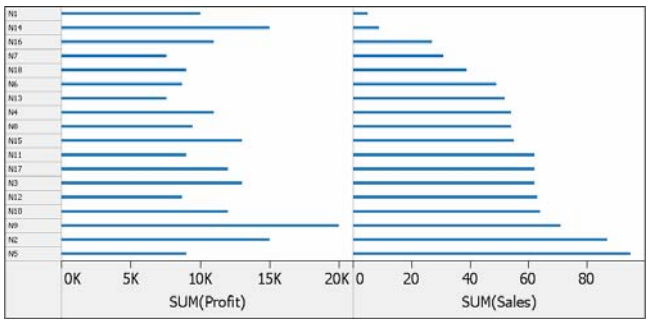
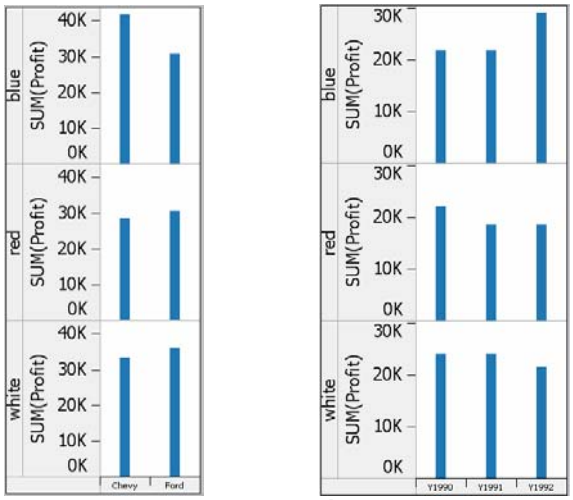


Figure 4.3: Quantity distributions comparison along the dimension “sales number” in terms of between totals of “profit” and “sales”



(a) (b)

Figure 4.4: (a) Quantity distributions comparison along the dimension “car model” in terms of totals of “profit” between different subsets of data whose car colours are blue, red, and white respectively (b) Similar to (a) except comparing along “year of manufacture”

In the Advizor, we can use the table sheet similar to Table Lens to do the first type of comparison extremely easily. On the other hand, it is not that effective to do the second type of comparisons by 3D charts (called Multiscape chart in the Advizor) though it can work to this end. Detailed screen shots and the accompanied scenario descriptions are illustrated in the section 5.2.1. In the Tableau, users can drag one non-quantitative dimension to row shelf and as many quantitative dimensions to column shelf for accomplishing the first type of comparison. For the second type of comparison, the procedure is similar to the task of “find quantity distribution of data”. We can simply add the non-quantitative dimension users desire to “compare” in front of the non-quantitative dimension which is already placed in the row shelf during the steps of finding quantity distributions. Detailed screen shots and the accompanied scenario descriptions are illustrated in the section 5.2.2.

4.3 Compare Trend and Occurrence of Time Dependent Data

In the construction management application, questions regarding time dependent data include: does data exist at a specific time (time window identification)? When does data exist? How long does data exist (activity duration)? How does data change along time? What is the sequence of data (activity sequence)? Which data coexist in a certain time point or period? (causal effect identification). For the question of how does data change along time, it is the same as finding quantity distribution of data along time dimension, which is non-quantitative. Therefore bar charts or line charts are both good enough to answer the questions. For the rest of the questions, we believe the traditional Gantt chart is still the best solution. If we juxtapose these two types of charts together, we can answer all kinds of questions just mentioned simultaneously. Even more, we even can observe how data coexist in a certain time affect the change of data along time.
Both of the Tableau and Advizor have dedicated chart types to simulate the Gantt chart. And both of them have bar charts. However, it is more difficult to juxtapose different types of charts in the Tableau, so we only use the Advizor to first create a Timetable which shows how different physical locations affected by change orders occur along time. Then we generate a bar chart to show how change order projected cost changes along time. By putting these two charts together, we see the sequence of physical locations affected by change order, the change of change order projected cost along time, the coexistence of physical locations in a certain time, and how the physical locations exist in certain time period affect the cost change. This kind of charts combinations are not innovative idea, it exists in the construction industry for a long time. What the information visualization technology enhance that is the powerful interaction functionality. In Advizor, we can quickly change dimensions we desire to see. For example, a few clicks, the previous physical location distribution on Timetable change into physical component distribution on Timetable. Not to mention the “filter by brush” functionality that let user focus on a certain period of time. Detailed screen shots and the accompanied scenario descriptions are illustrated in the section 5.2.2.

4.4 Associate Data of Different Dimensions

There is also another important task of examining construction data: find “what data has something to do with what data”. This is important because the indexes, drawing numbers of design drawing for example, in the log files can help locate the physical data like documents, drawings, video clips, etc. And after conducting exploratory data analysis, we might find some data interesting and want to know their related data, or even the physical data. For example, when we find out main floor has lots of change order during a certain time period, we will be eager to know what components of that floor cause the changes, the design drawing associated with those components, and related correspondences or meeting minutes. From the scenario formalism point of view, this kind of task is: given certain values of dimensions, find associated values of other dimensions. This scenario formalism is similar to the "Find quantity distribution of data along a non-quantitative dimension", but the difference is now we do not concern quantitative measurement of associated data. For example, we now only want to know the red cars sold were manufactured in which years rather than how many red cars manufactured in which years were sold. The reason we differentiate this scenario from previous scenario is that this task is relatively easy than other tasks. There is no need to do quantity comparison in order to understand the quantitative distribution. Therefore some dedicated type of images and scenario should be

identified so that users can focus their attention on finding "where the data is" if his real intention of examining data is to just to locate data. Otherwise, users might be distracted by quantitative property of data and get lost.

Both the Tableau and the Advizor have features for users to find which data correspond to what data. However, the "linked data views" feature of the Advizor is more intuitive for users to associate data horizontally. The linked data views also promise the possibility to link abstract data values to their corresponding physical data like drawings archives. In addition, since the construction project is mostly modelled in a hierarchical way, which means the construction data has inherent hierarchical structure in addition to relational structure, the tree structure charts (Heatmap charts) that the Advizor provides are perfectly useful for users dig data association vertically. In the Advizor, first we have a bar chart to show which change orders that the client disagrees with the contractors in terms of the extra cost. Then we target change orders have relatively large differences between the projected cost and approved cost. Then we can quickly know the affected physical components (granular and individual), affected sub-trade, and indexes of sub trades' revision documents in the Heatmap charts.

5 SCREENSHOTS OF SOFTWARE IN ACTION

In this section, we use step by step screenshots of software in action to demonstrate how to follow scenario formalism for operating visualization tools. In subsection 5.1 the scenario of finding quantity distribution of data is demonstrated by using the Advizor and the Tableau. In subsection 5.2 the scenario of comparing quantity distribution of data is demonstrated also by using the Advizor and the Tableau. In subsection 5.3 the scenario of comparing trend and occurrence of time dependent is demonstrated only by using the Advizor

5.1 Quantity distribution of data

In subsection 5.1.1, we will use the Advizor to show how total of records count (e.g. number of change orders, a statistical value) distribute along major group element, location, reason of change, initiated document, and approved date, which are all non-quantitative dimensions. Also shown are how total of trade's change order amount and projected change order cost, which are statistical values of quantitative dimensions, distribute along sub-trade and issued date that are non-quantitative dimensions respectively. Then we filter out some data by limiting the major group element being "interior construction" and location being "main floor", and then examine the previously described data distributions again. In 5.1.2, we will use the Tableau to try the similar tasks.

5.1.1 The Advizor's screen shots of quantity distribution

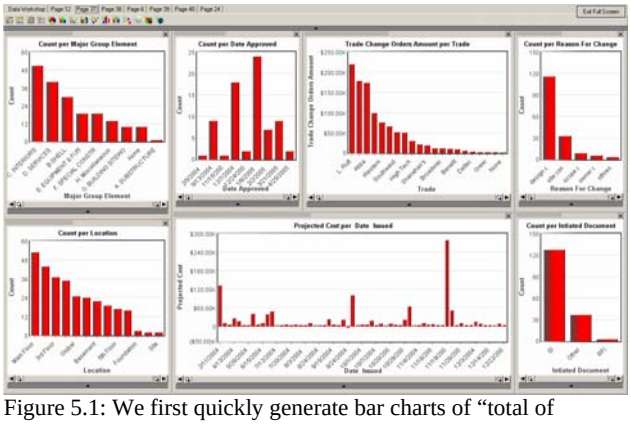


Figure 5.1: We first quickly generate bar charts of "total of records count versus major group element", "total of records count versus location", and so on.

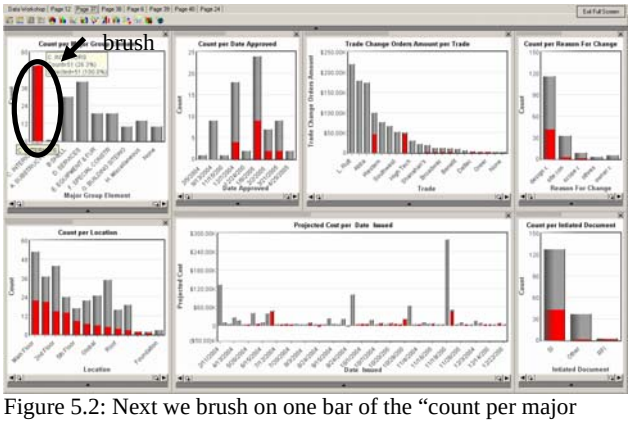


Figure 5.2: Next we brush on one bar of the "count per major group elements" bar chart to filter out data whose major group elements are not interior construction element.

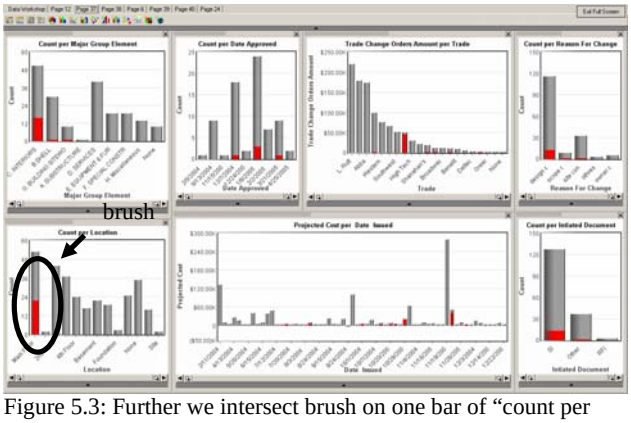


Figure 5.3: Further we intersect brush on one bar of "count per location" to filter out data whose locations are not main floor.

5.1.2 The Tableau's screen shots of quantity distribution

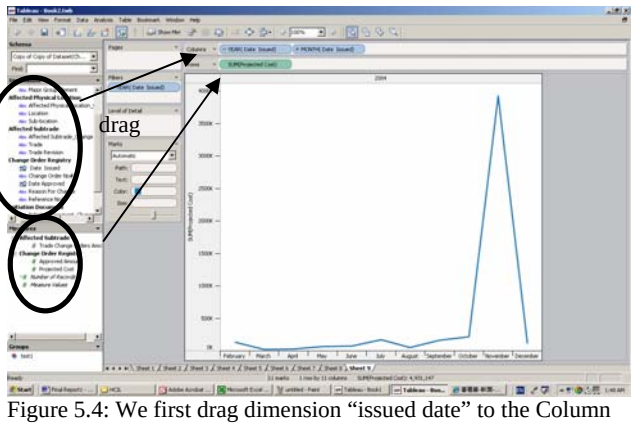


Figure 5.4: We first drag dimension "issued date" to the Column Shelf and dimension "projected change order cost" to Row Shelf.

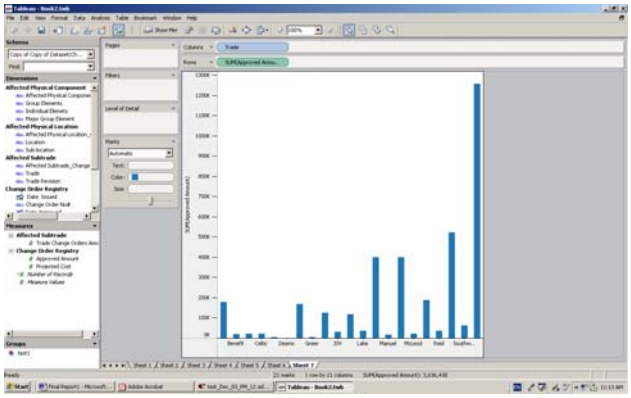
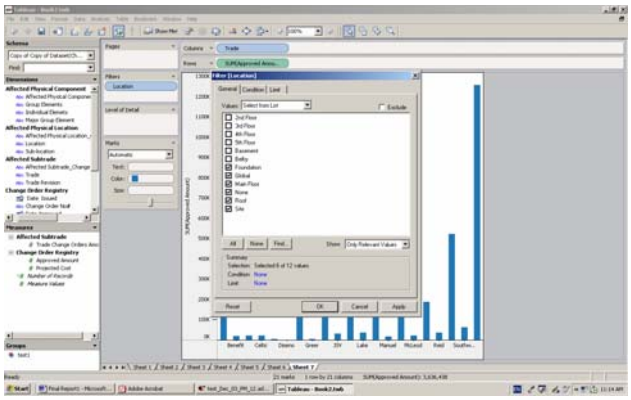
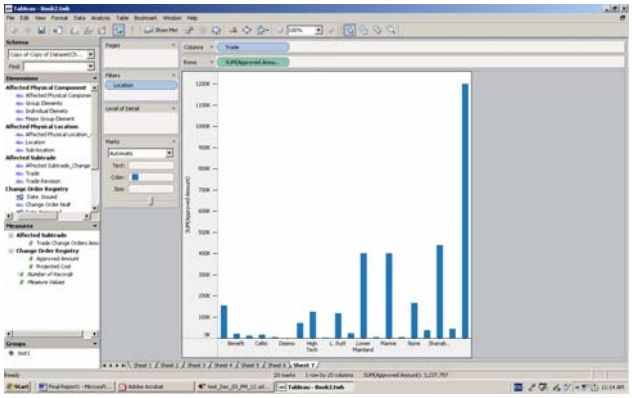


Figure 5.5: Now we quickly drag the dimension "issued date" and dimension "projected change order cost" off the shelves and replace them by dimension "sub-trade" and dimension "trades' change order amount"

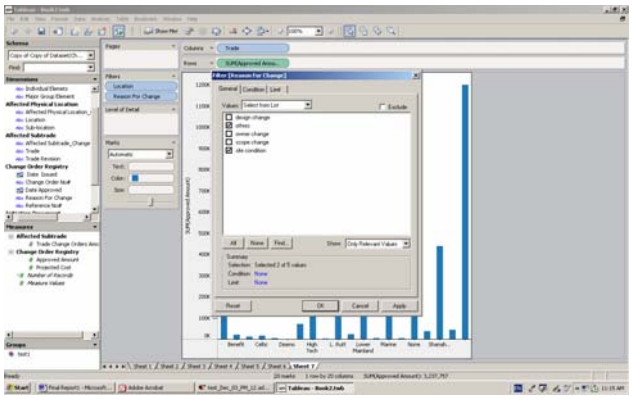


(a)

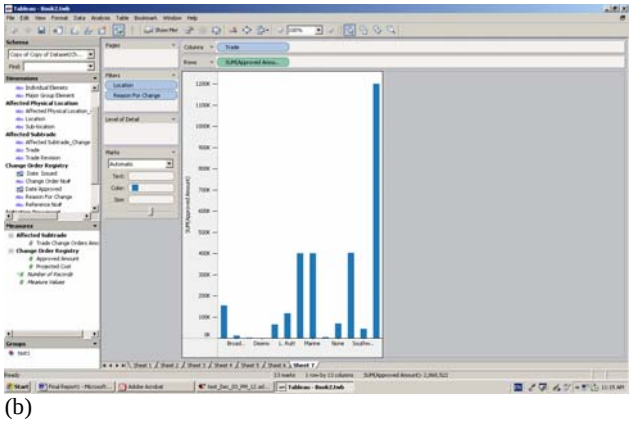


(b)

Figure 5.6: Next we filter out some data values of dimension "location" by dragging dimension "location" to Filter Shelf and ticking in the pop-up dialogue box. (a) Filter in action (b) Filter result



(a)



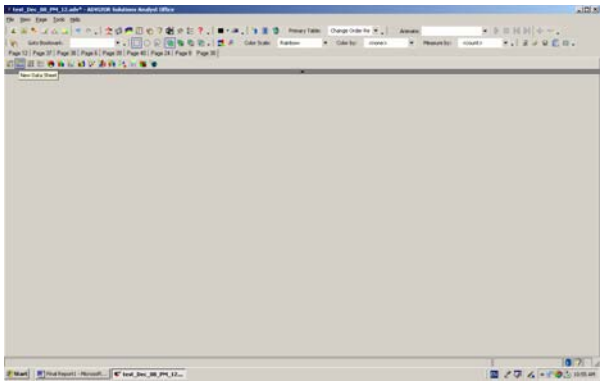
(b)

Figure 5.7: We further filter out some data values of dimension "reason of change". (a) Filter in action (b) Filter result

5.2 Compare quantity distributions of data

In subsection 5.2.1, we will use the Advizor to show the visual effect of correlation between two quantitative dimensions that are "projected change order cost" and "approved change order cost" by using data table where values of data are represented by length of bars instead of the actual numbers. This is the first type of quantity distribution comparison mentioned in section 4.2. Also we will use interaction-enabled 3D plot to examine the records count distribution along dimension "issue dates" for each data value of dimension "location", and vice versa. This is the second type of quantity distribution comparison mentioned in section 4.2. In subsection 5.2.2, we will use the Tableau to first compare different quantity distribution along dimension "sub-trade" in terms of total of "projected change order cost", total of "approved change order cost", and total of "trade change order amount". We also quickly do another comparison along dimension "issued date". This is the first type of quantity distribution comparison mentioned in section 4.2. Then we will use it to demonstrate how to visualize the second type of quantity comparison of data mentioned in the section 4.2 by using data of dimension "sub-trade", "trade change order amount", and "reason of change". This is the second type of quantity distribution comparison mentioned in section 4.2.

5.2.1 The Advizor's screen shots of quantity distribution comparison



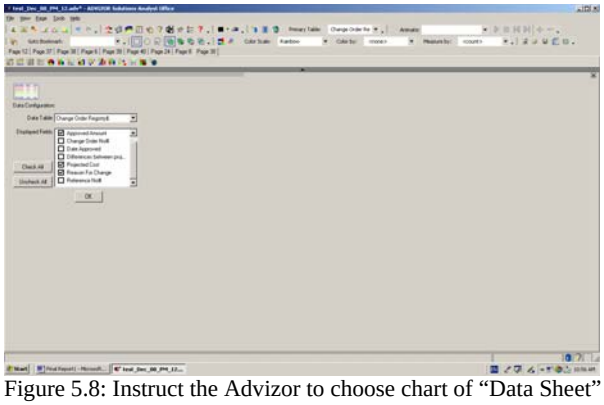


Figure 5.8: Instruct the Advizor to choose chart of "Data Sheet" type, and then check we want to show dimensions of "projected change order cost", "approved change order cost", and "reason of change"

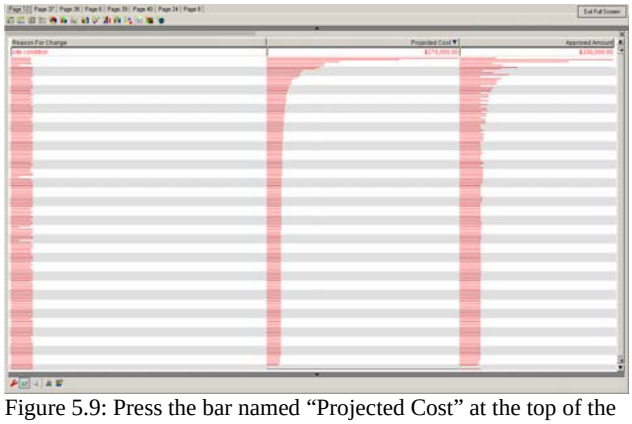


Figure 5.9: Press the bar named "Projected Cost" at the top of the sheet to order these three dimensions by projected change order cost

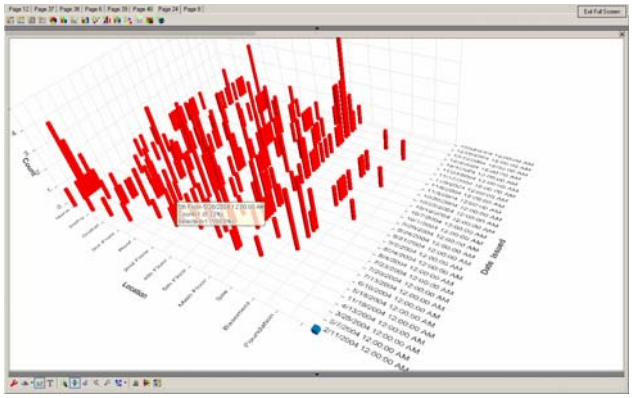


Figure 5.10: Now we change to do the second type of comparison by generating a Multiscape chart in which dimension "location" falls on x-axis, dimension "issued date" falls on y-axis, and record counts falls on z-axis.

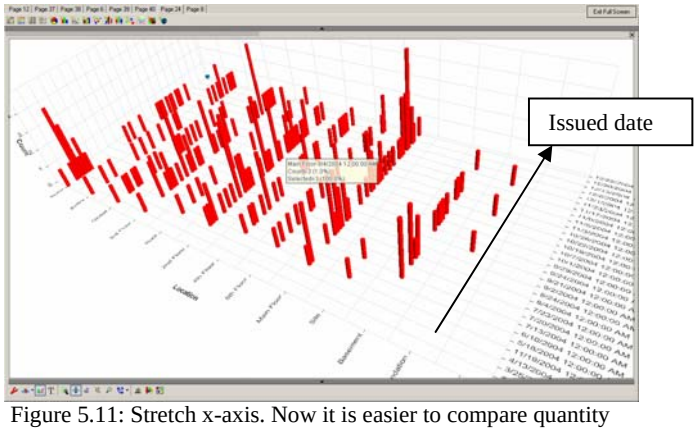


Figure 5.11: Stretch x-axis. Now it is easier to compare quantity distribution along dimension "issued date"

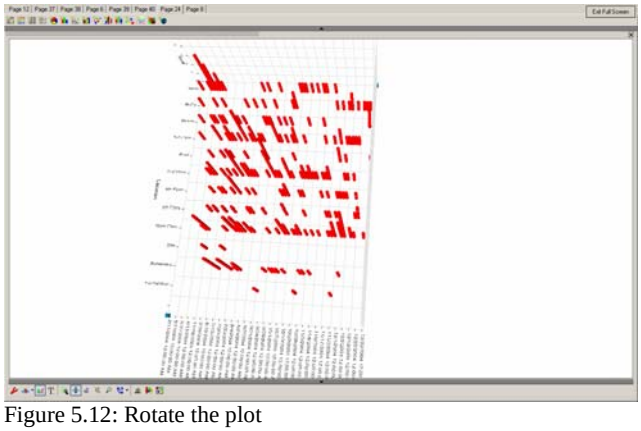


Figure 5.12: Rotate the plot

5.2.2 The Tableau's screen shots of quantity distribution comparison

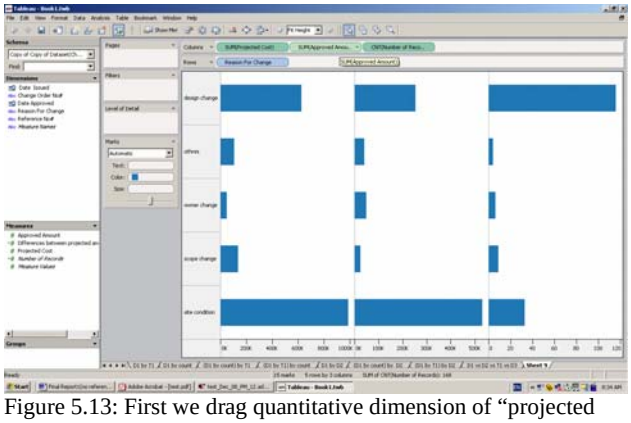


Figure 5.13: First we drag quantitative dimension of "projected change order cost", "approved change order cost", and "record count" to Column Shelf, instruct the system to total these three dimensions, and drag the non-quantitative dimension "reason of change" to Row Shelf

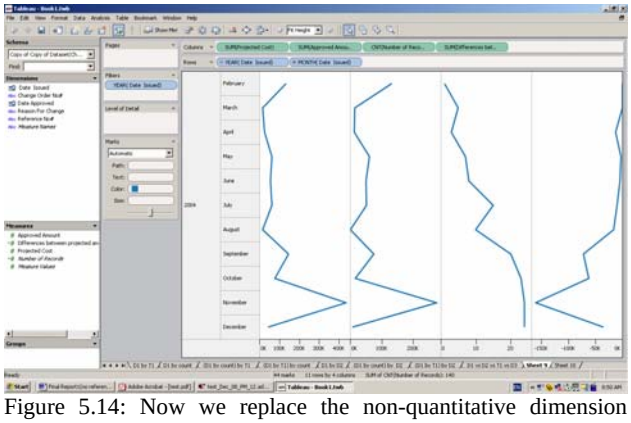


Figure 5.14: Now we replace the non-quantitative dimension "reason of change" on the Row Shelf by "issued date". Also add another quantitative dimension "difference between projected and approved cost"

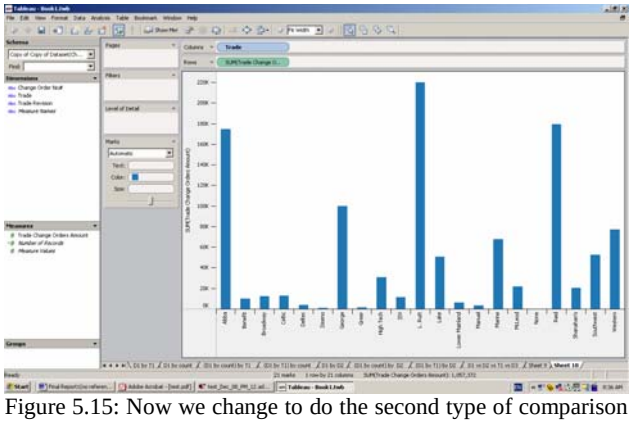


Figure 5.15: Now we change to do the second type of comparison by dragging non-quantitative dimension "sub-trade" and quantitative dimension "trade change order amount" to Column Shelf and Row Shelf respectively. Also we instruct the system to do total on the quantitative dimension, which is "trade change order amount" here

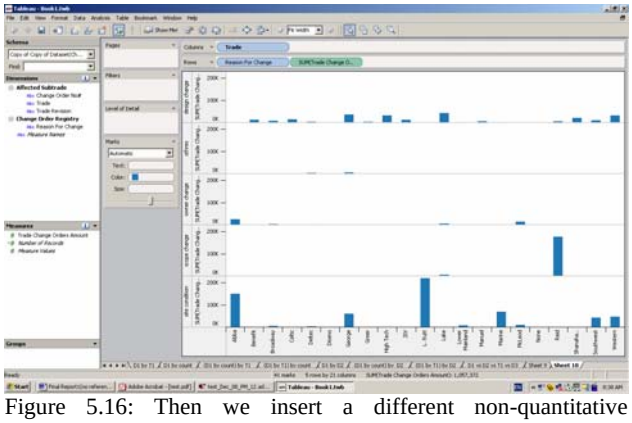


Figure 5.16: Then we insert a different non-quantitative dimension "reason of change" in front of the quantitative dimension "trade change order amount" on the Row Shelf.

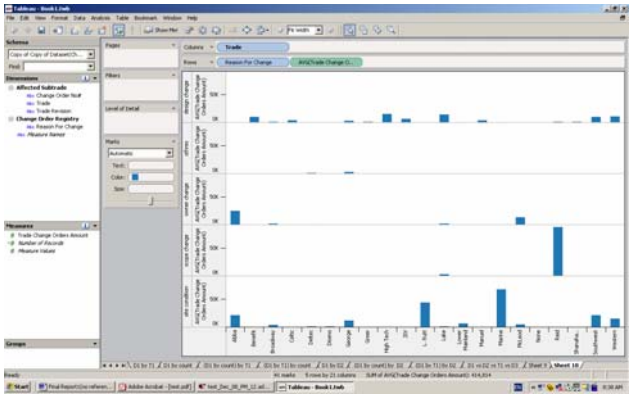
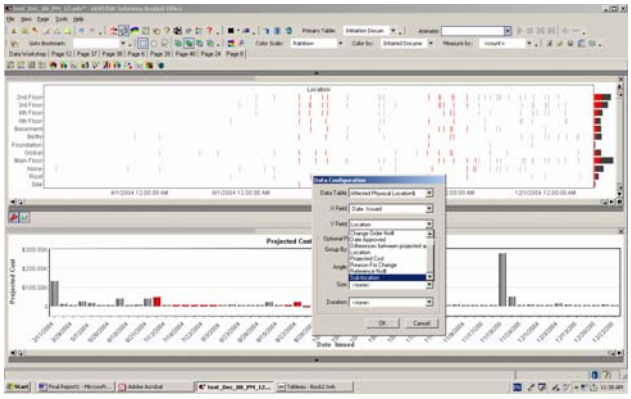


Figure 5.17: Simply instruct the system to average the data values of the quantitative dimension "trade change order amount" to change to compare quantity distribution in terms of the average of trade change order amount.

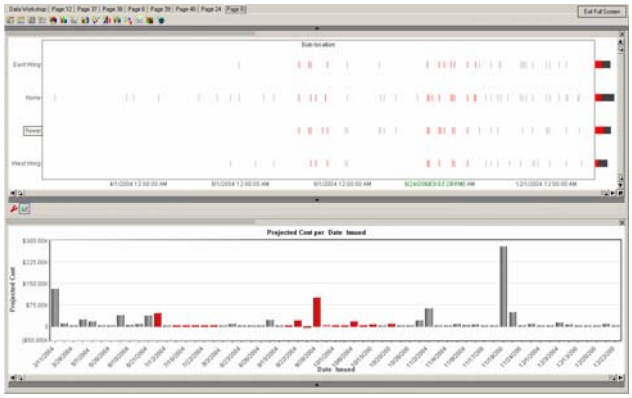
5.3 Trend and Occurrence Comparison of Time Dependent Data

In this scenario formalism, we will only use the Advizor to visualize time property of data. We will use the Advizor's Timetable chart (similar to Gantt chart) to visualize the location where change orders are needed and when they are encountered. Juxtaposed with the Timetable is the bar chart showing the trend of values of dimension "projected change order cost" along dimension "issued date". We also show how the interaction ability let us quickly change dimensions for observation.

Figure 5.18: Generate the aforementioned charts using the Advizor's GUI. Brush the clustered visual encodings in the TimeTable, corresponding data in the bar chart are highlighted too.

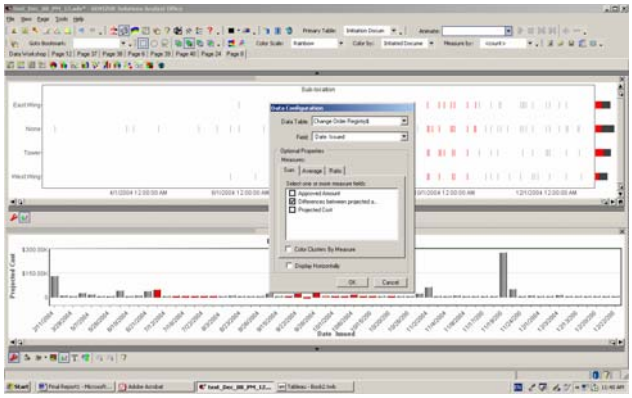


(1)

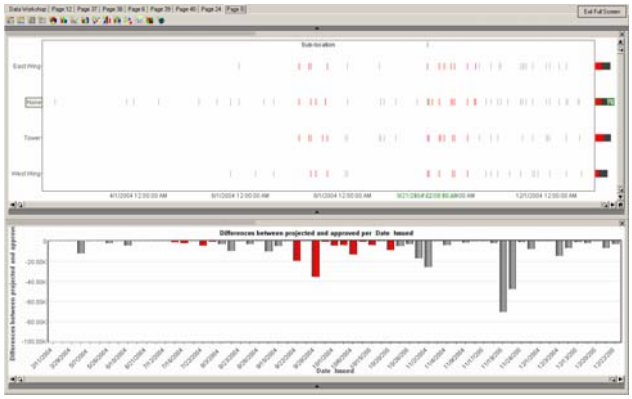


(2)

Figure 5.19(a): Change the dimension "location" in the TimeTable to dimension "sub-location". (1) Change in action (2) Change result



(1)



(2)

Figure 5.19b: Change the dimension "projected cost" in the bar chart to dimension "differences between projected and approved cost". (1) Change in action (2) Change result

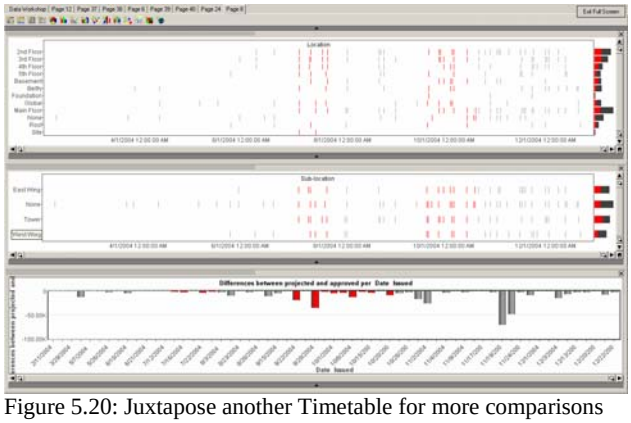


Figure 5.20: Juxtapose another Timetable for more comparisons

Figure 5.21: Replace the bar chart by line chart for visualizing trend of time dependent data

5.4 Association between Data of Different Dimensions

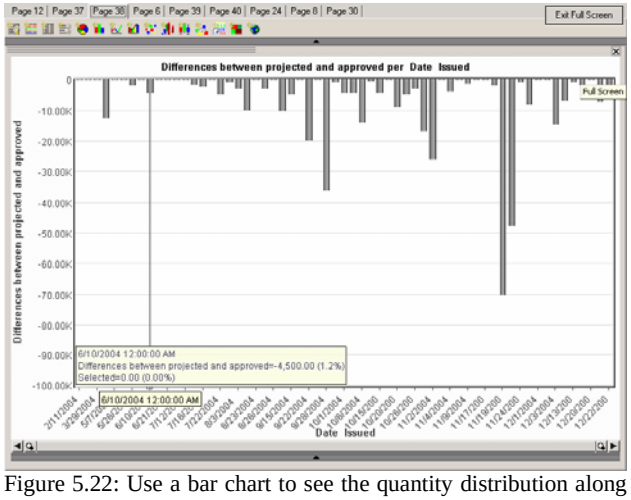


Figure 5.22: Use a bar chart to see the quantity distribution along "issued date" in terms of "differences between projected and approved cost".

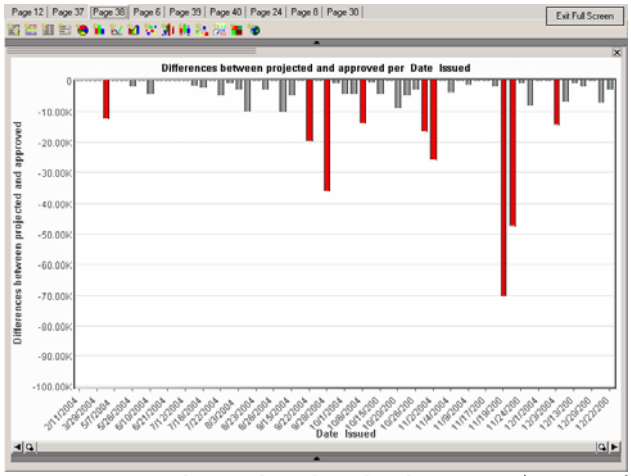


Figure 5.23: Brush to select data that have over $10,000 of differences of projected and approved cost

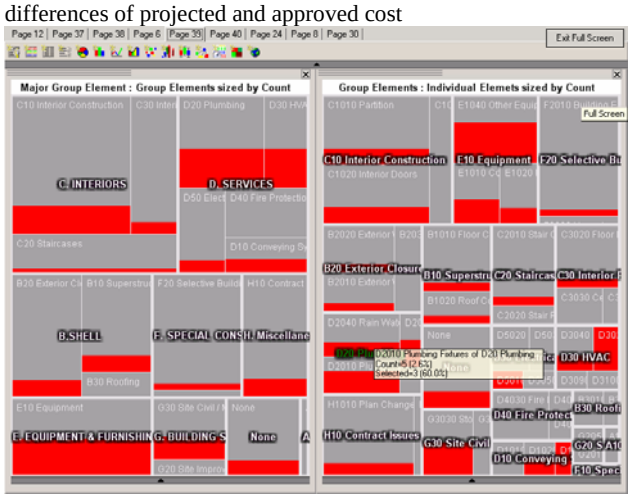


Figure 5.24: Switch to already generated Heatmap chart that visualizes data of dimensions of "major group element", "group element", and "individual element" (hierarchically structured).

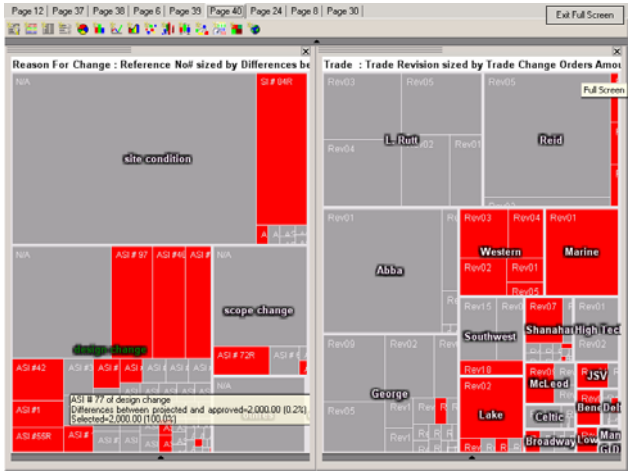


Figure 5.25: Switch to already generated Heatmap chart that visualizes data of dimensions of "reason of change" and "initiated documents"; data of dimensions of "affected sub-trade" and "sub-trade revision number".

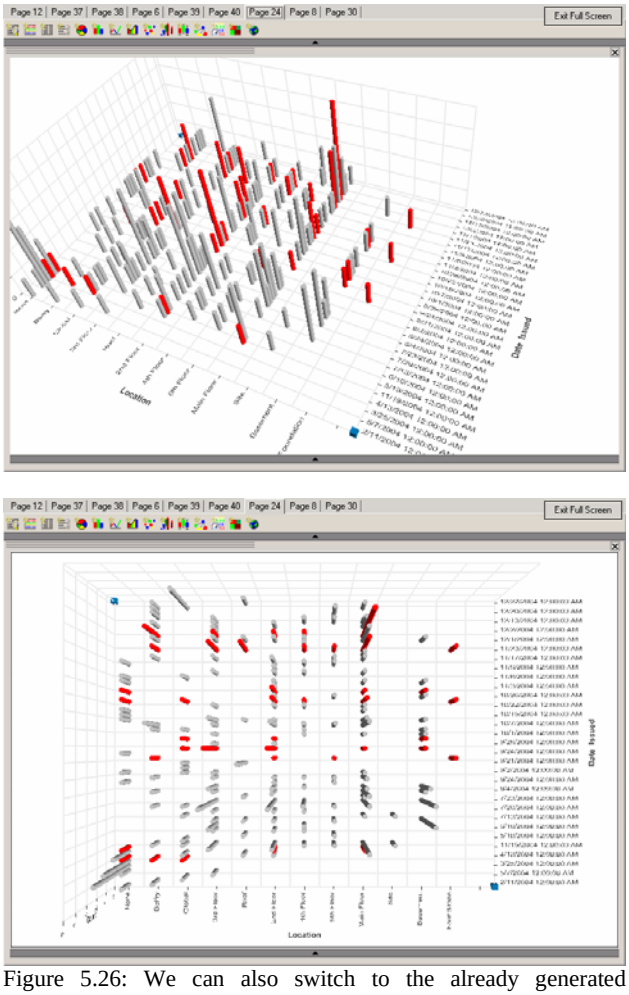


Figure 5.26: We can also switch to the already generated Multiscape chart that visualizes related data of combinations of two dimensions.

6 RESULTS

The outcomes of this project can be analyzed from two aspects. The first aspect is the validation of improvement in interpreting construction project data if state-of-the-art information visualization techniques are used. The second one is the findings of principles for efficiently and effectively utilizing information visualization techniques in retrieving and presenting multidimensional data. We will analyze them in the subsection 6.1 and 6.2.

6.1 Validation of Improvement in Interpreting Data Due To the Use of Information Visualization Techniques

In this section, we will first summarize some selected information extracted from the screenshots in the section five. And then we will compare these images with Excel images generated by other researchers. Lastly, we will conclude the validation.

6.1.1 Summary of Information Extracted

The screenshots of section 5.1 are visualization of quantity distributions of data, which provide us the following information:

- From Figure 5.1:
  - Interior construction and service construction encounter more change orders than other components of the building.
  - Main floor and second floor encounter more change orders than other locations of the building.
  - The client approved most of the requests of change order in the middle of the construction phase.
  - There are three to six requests of change order involving relatively high extra cost. They occur at the beginning, middle, and near end of the record time period.
  - Three sub-trades have relatively high extra cost due to change orders.
  - Change orders largely are caused by design changes and initiated by site instructions.
- From Figure 5.2: For the change orders that are related to interior construction:
  - Two affected sub-trades have relatively very high extra cost.
  - Each location of the building more or less is affected.
  - They are mostly caused by design change and initiated by site instruction.
- From Figure 5.3: For the change orders that are related to interior construction and main floor:
  - One sub-trade affected has relatively very high extra cost.

The screenshots of section 5.2 are visualization of comparing quantity distributions of data, which provide us the following information:

- From Figure 5.13: For different reasons of change order, the approved cost is somewhat proportioned to projected cost. However, those two have no correlation to number of change orders. That is, the reason of change that has more number of change orders does not necessarily have more extra cost.
- From Figure 5.14: With time elapses, the number of change order increases. However, the extra costs needed do not increase with time except an extraordinary increase in November. Also noted is that the client has no problem of approving requests of extra cost at the beginning. However, along the increases of change orders, the amount of disagreement starts to rise.
- From Figure 5.17: From sub-trades' perspectives, site condition caused change orders cost relatively much higher than rest of the other reasons of change. And two sub-trades are impacted most.

The screenshots of section 5.3 are visualization of trend and occurrence comparison of time dependent data, which provide us the following information:

- From Figure 5.18: We notice that there are two periods of time when the change orders involve almost all locations of the building. During those two time period, one sharply increase of projected cost is observed.

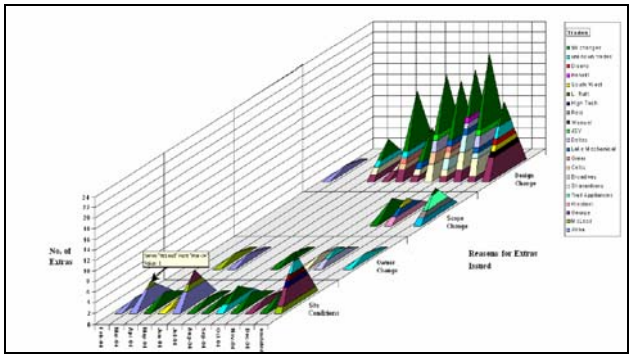
The screenshots of section 5.4 are visualization for locating associated data, which provide us the following information:

Of those change orders that the client strongly disagree with:

- From Figure 5.24: We identify the physical components associated with those change orders.
- From Figure 5.25: We identify the reasons of change, indexes to reference documents, affected sub trades, and indexes to sub-trades' revision documents associated with those change orders.
- From Figure 5.26: We find the coincidental time and spaces in which those change orders occur.

6.1.2 Comparing with Excel Generated Images

The research of visualizing construction data has been conducted by the Construction Engineering and Project Management Group of civil engineering department at University of British Columbia for some time. For the same dataset we use in this project, there are also several Excel generated images of those data and insight derived thereof in [29] (Figure 6.1). Only by comparing still images, it is difficult to conclude which image interprets data better than the other. Also visualizing middle sized dataset by charts is not beyond Excel's capability. However, the time the author of [29] spent on generating and interpreting images is much longer than the time we took when state-of–the-art information visualization techniques are used. In Excel, functionalities of data storage and data graphics are separated. Users need to find out what data needed and where they are located in the data tables. Then the users follow the processes of editing charts, no better than charting by hand if the hassle of editing legend is counted. After examining the chart generated, assuming the chart is meaningful; the users are inspired by the information and want to visually examine other aspects of the data. Now users need to repeat the process of searching in data sheet and graphing. Another issue of Excel is that users need imaginations to cram several dimensions into a single chart, which takes users lots of effort. And sometimes the images created are not intuitive. The notion here is that the key to effectively interpret data visually is not about still images themselves. The key is how users can minimize time spent on efficiently and effectively iterating the process of retrieving and visualizing data.



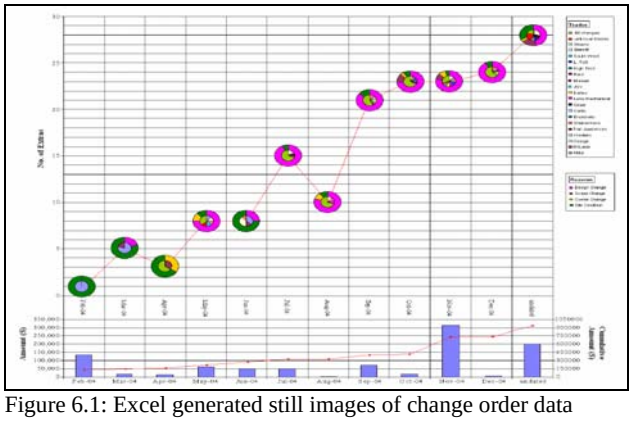


Figure 6.1: Excel generated still images of change order data

6.1.3 Concluding the Validation

The interpretation of data is the process of exploratory data analysis. And data visualization is one of the effective and intuitive approaches of conducting the data analysis. In construction management application in which the data is multidimensional, the number and variety of images needed for exploratory data analysis is much more than the dataset that has large amount of records but low dimension. For example, the data points for a CAD drawing of 3D building are very many, but only one 3D image can deliver the picture of the data. Therefore, the easiness of iterating the cycle of retrieving data, generating visualization of data, observing visualization of data, and then retrieving data again is crucial. The information visualization tools we choose and the scenario formalism of using tools we develop can let users retrieve data and create images of data on the fly so that users can focus their attention on observing images in hope of extracting valuable information from data. Also the "linked data views" or so called "coordinated multiple views" let users see all the dimensions of data in linked different views, which we think is the most effective way of browsing multidimensional data. Otherwise, we will need to put lots of thinking on how to consolidate nD data into one 2D or 3D space. Therefore, we conclude that the information visualization techniques we chose can improve construction industry's ability of interpreting heterogeneous construction data.

6.2 Finding of principles for retrieving and presenting data efficiently and effectively

By "efficiently", we mean users can quickly retrieve and present different dataset using minimum effort. By "effectively", we mean the iterative process of retrieving data, visualizing data, extracting information from visualization, and then retrieving data is optimized. Principles of these two purposes are explained in the subsection 6.2.1 and 6.2.2.

6.2.1 Principles to Efficiency

If using the Excel as a comparator, many of the principles for retrieving and presenting data efficiently have been incorporated in the tools we used. The principles and corresponding information visualization technique (actually most of them are interaction techniques) are:

- Alleviate users' burden of transforming data into images step by step required by the Excel. Time needed to produce a bar chart in Excel is several times what needed by the two tools we used. The key

technology to improve that is the "simply choose dimensions and then images are generated automatically". This means a lot to multidimensional data because the combinations of images increase exponentially when dimensions scale up. Users should not spend their time on generating images especially when lots of images are essential to conducting exploratory data analysis. For example, there are seven bar charts created by the Advizor in the Figure 5.1. It only took the principle investigator less than five minutes to generate and align them nicely.

- Alleviate users' burden of querying data, especially when complex queries are needed. Both of the tools can let users change data dimensions desired to see and filter data ranges on the fly. The key technology includes "query by brush" in the Advizor, "drag and drop dimensions onto shelves" in the Tableau, and traditional GUIs by which users can select or change data dimensions or data ranges. For example, the only effort to transit from Figure 5.13 to Figure 5.14 is 3 steps of dragging dimensions onto and off the Row Shelf in the Tableau.
- Alleviate users' burden of configuring and editing images. When conducting exploratory data analysis, users often need to compare multiple images. Therefore, users have the need to directly manipulate on visual encodings like zooming, panning, rotating, resizing views, and changing visual encodings. So the information visualization technology to this end is the computer graphic capability. The dedicated information visualization systems have much better graphics capability than the Excel. Take Figure 5.16 for example, the built-in visual encoding techniques automatically align different quantity distribution on a same dimension. Otherwise, users of Excel need to manually juxtapose them. Take Figure 5.11 and Figure 5.12 for another example, the computer graphic capability let users stretch scale of x-axis and rotate the orientation of the 3D chart freely.

6.2.2 Principles to Effectiveness

From the "effectiveness" point of view, the principles are partly embedded in the techniques themselves and partly come from the scenario formalism we develop in this project. The principles relating to techniques themselves are:

- Use optimum visual encodings to exploit human beings' visual perception ability. Although we can not say an image is wrong, some encodings are relatively less effective in utilizing human being's visual cognition ability. For example, users may find it difficult to understand the order of ordinal data if they are encoded by color hue since there is no inherent order in it. However, the color saturation inherently ranging from light to dark conveys the ordinal measurement scale effectively. Both of the tools we used build the knowledge of effective visual encoding into tool's specification. Take Figure 5.13 for an example; because the combination of dimensions is non-quantitative and quantitative, the Tableau system automatically generates a bar chart to present them. Also in the Figure 5.14, since the non-quantitative dimension is time in nature, the line chart is recommended not only because the position is the most effective encoding in representing quantitative data, also human beings' accustoming to see lines as trend in time also accounts for the system's choice.
- Use of linked data views. For different combinations of data dimensions or analytical tasks, there exist visualizations thereof that are more effective. For example, Gantt chart is best for visualizing the occurrence of time dependent data; line chart is best for visualizing trend of time dependent data. When trying to blend these two different analytical tasks into one image, one of the tasks will be compensated. Although we want to maintain best images for different data or visualization tasks, they could be useless if they stand alone. For example, if we select to see a subset of data, but only one image update itself to reflect the selection, users can not conduct exploratory data analysis because not all the views represent the same data. With the techniques of linked data views, now views being more effective can represent the same dataset by which users have common grounds to conduct visual data analysis. Therefore, now we have the flexibility to use different and better images for specific data types or analytical tasks.

The scenario formalism gives users a higher level way of thinking when "questioning data". Rather than posing questions in ad-hoc manner, many kinds of questions can be generalized into single scenario. And for each single scenario, a better methodology of retrieving and visualizing data can be identified. Take the finding quantity distribution along non-quantitative dimension for example; users can focus their attention on pairing a quantitative dimension and non-quantitative dimension. If they want to know more dimensions, choose other dual dimensions. In addition, since the "position" is good for representing categorical data; "position" and "length" are good for representing quantitative data, the "scatter plot" or "bar chart" are all effective for visualizing the dual dimensions. As to whether the information extracted useful and comprehensive, the results summarized in section 6.1 speak themselves. Therefore, the effectiveness comes from the procedure of:

- Users analyze which scenario formalism their questions fall into.
- Use that scenario's suggested way of querying and retrieving data.
- Expressive and effective images are generated.

The suggested procedures work only when we have all the aforementioned information visualization techniques. The formalisms and suggested procedures thereof were explained in the section four and section five.

7 LESSONS LEARNED

During the course of conducting this project, we found that data collection and organization is as important as visualization, and scenario formalisms are difficult to be complete and need further research. Although we successfully identify key information visualization technology that improve ability of analyzing construction data that is heterogeneous in nature, there remains room of improvement in terms of linking abstract data to physical data. On the contrary to providing the richness of color graphing capability, we minimize the use of color for our reasons. Those lessons learned are elaborated in the following subsections.

7.1 Importance of Data Model and Data Organization

Visualization can deceive people if data is mishandled. Some of the dimensions of the dataset used in this project are created from text descriptions. The dimension "reason of change", "affected physical components", and "affected location" are interpreted and categorized manually from the text description of the actual change order log. Therefore, we think the "honesty" and "granularity" of the data are partly responsible for the effectiveness of interpreting data. For example, if we divide the reason of change into more categories, more (or less) salient information could come out.

Another issue is that during the process we try tools on data, we let the tools join data table by their built-in joining table functionality. We first were excited about being able to associate affected physical components and affected physical location with extra cost of change order. That means now we can see which components and locations contribute how much of the change order extra cost. Unfortunately we immediately realize the images generated were wrong because of our sloppy way of designing data tables. For example, each change order may associate with many different affected physical components, but we do not have data as to the change order cost of each component. Therefore, there is no way we can know the portion of the change order cost each component contributes. However, the tools simply join tables by associating the same key in each table (change order number in our dataset), and duplicate data records. Form the Figure 7.1, we know the change order number 127 has a projected cost of $8,828, and it affected three physical individual elements. But after carelessly joining tables, now the joined data give an illusion that those three physical elements account for projected cost of $8,828 respectively, which can be seen in Figure 7.2. And then users will be tempted to generate unusual (seemingly insightful) but wrong (actually is outliers) data interpretation.

The lesson learned here is that how data modeled and organized is important. And visualization can be used as a tool of giving feedback about the data so that we can redesign our data model and data organization system.

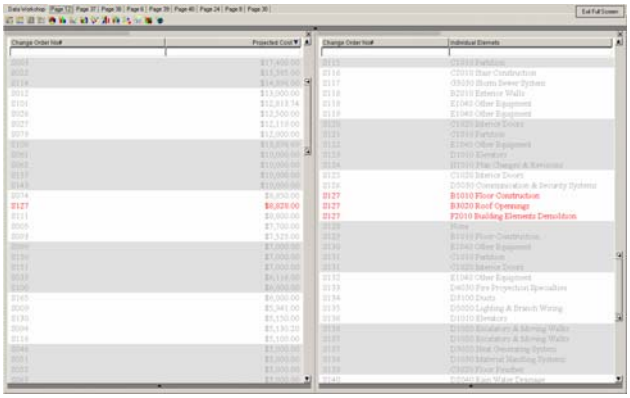


Figure 7.1: A record of "change order registry" links to three records of "individual elements"

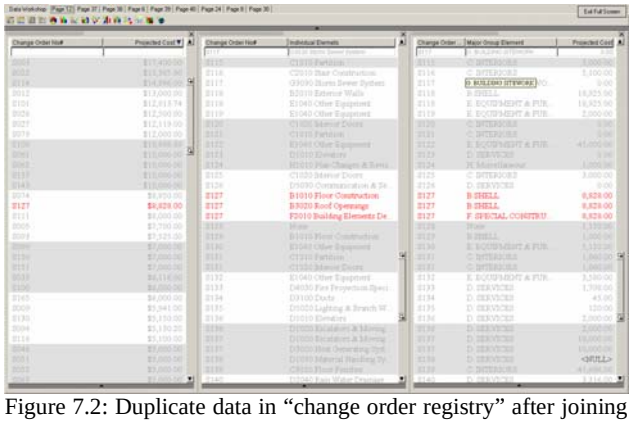


Figure 7.2: Duplicate data in "change order registry" after joining two data tables that have "1 to many" relationship

7.2 Research Area of Scenario Formalism

Scenario formalism is helpful but difficult to be complete. For example, although we thought the formalism of quantity distribution along one dimension seems to give us a systematic way to explore data, immediately we found in some case it is more meaningful to see distribution on combination of dimensions. Map configured by the combination of two spatial dimensions is an example. However, we still think the scenario formalisms help users effectively complete the iterative cycle of retrieving data, visualizing data, extracting information from visualization, and again retrieving data, and the search for scenario formalisms for visualizing multidimensional data deserves further investigations.

7.3 Current and Future Information Visualization Technology in Construction Industry Application

The most important technology in the visual analysis of multidimensional data is interaction which supports fast querying data, generating images, and manipulating images. The most powerful interaction techniques are "dragging data to shelves" features in the Tableau; "linked views and query by brush" in the Advizor; "visual zooming using scroll bars" in the Advizor. However, as to the still images, it is difficult to tell state-of-the-art information visualization technology generated one is more effective than one that are generated by Excel or even by hand. Both tools we used in this project do not support linking abstract data with scientific and document data. In this project, we categorize the physical components and locations or reasons of change from the text description by using our own naming language. Construction industry has the serious problems of naming things, even in the same language. In order to overcome this barriers, users should be able to link the purposely made abstract data (by "purposely", we mean we purposely translating scientific data into abstract data; collection of spatial points data versus "main floor" for example) back to their genuine data views in order to validate semantic perceptions of the data. In short, we want to link charts to 3D CAD drawing or electronic documents. Both of the tools we use in this project do not have this functionality that is specifically crucial to the construction industry.

7.4 Reduced Use of Color

Color is not necessarily needed. Coloring is very useful in the application of cartography or scientific visualization. In the application of visualizing multidimensional abstract data, color does not really help in differentiating categorical data especially

when number of values of categorical data exceed single digit. From Figure 7.3 and Figure 7.4, we observe that the difference of information effectiveness between them, images of the same data, is subtle. However, gray scale is still good for quantitative data. This could be good news for people who are color blind.

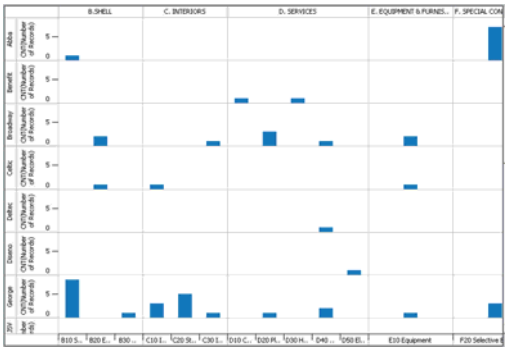


Figure 7.3: Visualization that uses positions to differentiate trades

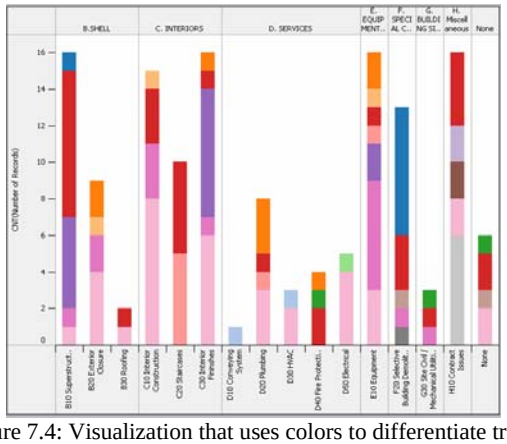


Figure 7.4: Visualization that uses colors to differentiate trades

Bibliography:

[1] A. D. Russell and A. Udaipurwala, "Using multiple views to model construction," in *CIB World Building Congress*, 2004,

[2] A. D. Russell and A. Udaipurwala, "Advantages of a multi-view representaion of project execution," in *Proceedings of Specialty Conference on Fully Integrated and Automated Project Processes*, 2001,2002,

[3] A. D. Songer, B. Hays and C. North, "Multidimensional visualization of project control data," in *Construction Innovation*, 2004, pp. 173-190.

[4] B. Shneiderman, "Tree Visualization with Tree-maps: 2D Space-Filling Approach," *ACM Transactions on Graphics*, vol. 11, pp. 92-99, 1992.

[5] K. W. Chau, M. Anson and J. P. Zhang, "Four-Dimensional Visualization of Construction Scheduling and Site Utilization," *J. Constr. Engrg. and Mgmt.*, vol. 130, pp. 598-606, August 1, 2004.

[6] D. P. Tegarden, "Business Information Visualization," *Communications of AIS*, vol. 1, 1999.

[7] D. A. James, "Interactive Data Analysis in a Manufacturing Setting-A Case Study," *Computational Statistics*, vol. 14, pp. 147-159, 1999.

[8] J. Bertin, *Semiology of Graphics*. Milwaukee: University of Wisconsin Press, 1983,

[9] W. S. Cleveland and M. Robert, "Graphical Perception: Theory, Experimentation and Application to the Development of Graphical Methods." *Journal of the American Statistical Association*, vol. 79, pp. 531-554, 1984.

[10] J. Mackinlay, "Automating the Design of Graphical Presentations of Relational Information," *Transactions on Graphics*, vol. 5, pp. 110-141, 1986.

[11] C. Stolte, D. Tang and P. Hanrahan, "Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases," *Transactions on Visualization and Computer Graphics*, vol. 8, 2002.

[12] S. G. Eick, "Engineering perceptually effective visualizations for abstract data," in *Scientific Visualization: Overviews, Methodologies, and Techniques* Anonymous Institute of Electrical & Electronics Engineer, 1997, pp. 191-210.

[13] C. Ahlberg and B. Shneiderman, "Visual information seeking:Tight coupling of dynamic query filters with starfield displays," in *Human Factors in Computing Systems*, 1994,

[14] A. Aiken, C. Jolly, M. Stonebraker and A. Woodruff, "Tioga-2: A direct manipulation database visualization environment," in *Proceedings of the Twelfth International Conference on Data Engineering*, 1996,

[15] H. Chen, "Compound brushing," in *IEEE Symposium on Information Visualization*, 2003, pp. 181-188.

[16] M. C. Chuah, S. F. Roth, J. Mattis and J. Kolojejchick, "SDM: Selective dynamic manipulation of visualizations," in *Proceedings UIST' 95 Symposium on User Interface Software and Technology*, 1995,

[17] C. L. North, "A User Interface for Coordinating Visualization based on Relational Schemata: Snap-Together Visualization," *Ph.D. Thesis*, 2000.

[18] G. Wills, "Linked Data Views," *Computing and Graphics Newsletter*, vol. 10, pp. 24, 1999.

[19] W. S. Cleveland and R. A. Becker, "Brushing Scatterplots," *Technometrics*, vol. 29, pp. 127-142, 1987.

[20] G. Wills, "Selection: 524,288 ways to say "this is interesting"," in *Proceeding of IEEE Symposium on Information Visualization*, 1996,

[21] A. U. Frank, "Different types of "times" in GIS," in *Spatial and Temporal Reasoning in Geographic Information Systems* Anonymous Oxford University Press, 1998,

[22] A. M. MacEachren, ***how Maps Work: Representation, Visualization, and Design*** . The Guilford Press, 1995,

[23] S. Havre, E. Hetzler, P. Whitney and L. Nowell, "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, 2002.

[24] W. C. Carlis and J. A. Konstan, "Interactive visualization of serial periodic data," in *11th Annual ACM Symposium on User Interface Software and Technology*, 1998,

[25] M. Ankerst, D. H. Jones, A. Kao and C. Wang, "DataJewel: Tightly integrating visualization with temporal data mining," in *International Conference on Data Mining (the 3rd International Workshop on Visual Data Mining)*, 2003,

[26] J. J. v. Wijk and E. R. v. Selow, "Cluster and calendar based visualization of time series data," in *IEEE Symposium on Information Visualization (INFOVIS'99)*, 1999,

[27] C. Plaisant, B. Milash, A. Rose, S. Widoff and B. Shneiderman, "LifeLines: Visualizing personal histories," in *ACM CHI '96 Conference Proceeding*, 1996, pp. 221-227.

[28] A. Bosworth, S. Chaidhuri, J. Gray, A. Layman, F. Pellow, H. Pirahesh, D. Reichart and M. Venkatrao, "Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals," *Data Mining and Knowledge Discover*, vol. 1, pp. 29-53, 1997.

[29] T. Korde, "Visualization of Construction Data," *M.Asc. Thesis*, 2005.