Similarity-Based Forecasting with Simultaneous Previews:

A River Plot Interface for Time Series Forecasting[1]

Paolo Buono[1], Catherine Plaisant[2], Adalberto Simeone[1], Aleks Aris[2,3], Ben Shneiderman[2,3], Galit Shmueli[4], Wolfgang Jank[4]

[1]Dipartimento di Informatica, Università di Bari, Bari, Italy
[2] Human-Computer Interaction Laboratory, [3]Computer Science Dept. University of Maryland
[4]Robert H. Smith School of Business, University of Maryland
buono@di.uniba.it, adalberto.simeone@gmail.com; {plaisant, aris, ben}@cs.umd.edu;
{gshmueli, wjank}@rhsmith.umd.edu

[1] To appear in the 11th International Conference on Information Visualisation. Zurich, Switzerland; 2-6 July 2007.

Abstract

Time-series forecasting has a large number of applications. Users with a partial time series for auctions, new stock offerings, or industrial processes desire estimates of the future behavior. We present a data driven forecasting method and interface called Similarity-Based Forecasting (SBF). A pattern matching search in a dataset of historical time series produces a subset of curves similar to the partial time series. The forecast is displayed graphically as a river plot showing statistical information about the SBF subset. A forecasting preview interface allows users to interactively explore alternative pattern matching parameters and see multiple forecasts simultaneously. User testing with 8 users demonstrated advantages and led to improvements.

Keywords--- **Forecasting; time series; river plot; simultaneous previews; visualization; user interfaces.**

1 Introduction

Forecasting is important in our everyday life. We need to predict the likelihood of success when we sign contracts, make investments or buy products. Whenever the future is uncontrollable and uncertain, forecasting is needed. Planning and forecasting are related in that plans may need to be altered according to forecasts. Armstrong contrasts their difference as "planning is about how the future should look like, while forecasting is about how the future will look like" [1].

Among the different types of forecasting, time-series forecasting is the most common and has the largest number of applications [1]. Researchers have paid much attention to improve forecasts and their accuracy; however, little attention has been given to the visualization and user interaction. Interactive exploration of forecasts has the potential to enable understanding of interesting phenomena that are hard to attain with the existing practices and approaches.

The visualization and exploration of time series has been studied extensively but challenges remain [15]. A recent innovation uses the focus+context technique [12] particularly useful for spotting (ir)regularities.

Classic statistical approaches for time series forecasting are either model-based (e.g., ARIMA models) or data-driven (e.g. exponential smoothing). Each approach has its advantages, and they are both popular in practice. In both cases the focus is on a single time series, and forecasting is typically based on extrapolation. An alternative approach, called econometric models, assumes a causal relationship between the time series of interest and a set of other related time series (e.g., forecasts of daily traffic conditions may use information on the time-series of temperature). Previous work has not used historical time series that measure the same phenomenon (which can be considered replications).

We propose a similarity-based data-driven forecaster. Users conduct a pattern matching search in a dataset of historical time series, and generate a subset of curves similar to the partial time series to be forecasted. The forecast is displayed graphically as a *river plot* showing statistical information about the similarity-based subset. A preview interface allows users to interactively explore the effect of the pattern matching parameters and see multiple forecasts simultaneously. This new interactive forecasting interface was built on top of TimeSearcher [5], a time series visualization tool.

We begin with a review of the original TimeSearcher interface and an auction dataset example used throughout the paper. We then describe the river plot and the forecasting interface. We introduce an interactive simultaneous preview interface that assists users in understanding the impact of the parameters used in the forecasting. Finally, we describe the changes we made based on user testing.

2 Original TimeSearcher Interface

TimeSearcher is a time series visualization tool that allows interactive exploration of time series data [9], [5]. Examples of data used in TimeSearcher include weather or air quality measures, oil well production, online auctions, or stock prices. For each item (e.g. an auction) TimeSearcher displays multiple time series representing multiple variables (e.g. price, price velocity, and price acceleration). TimeSearcher can also associate each item with a set of attribute data (metadata) that remain constant over time (e.g. seller rating or auction start day).

2.1 Auction Dataset

The example dataset used to illustrate our work in this paper contains 158 eBay auctions for a new Palm m515. The raw auction data consist of the bid history, which describes the temporal sequence of bids placed over time, and auction metadata (seller ID, buyer ID, seller's rating, item sold and its characteristics, etc.). Forecasting the closing price of an auction based on the time series of bids placed at the beginning of the auction is of interest. This time series consists of individual bid values placed at irregularly spaced time intervals. Managing irregularly spaced times series is challenging, and many options exist to visualize them [3]. For these data we chose to move from unequally spaced points into a continuous price curve by using penalized smoothing splines, which can be sampled at evenly spaced intervals [14]. This functional approach readily yields other important measures of the price curve dynamics, like its velocity or acceleration, giving us three variables to display in TimeSearcher 3. Price dynamics have been shown to be important predictors when forecasting the auction price [11], [17]. All auctions in this dataset are 7-day auctions so no rescaling is needed. After smoothing the raw data, the curves are sampled at 100 time points.

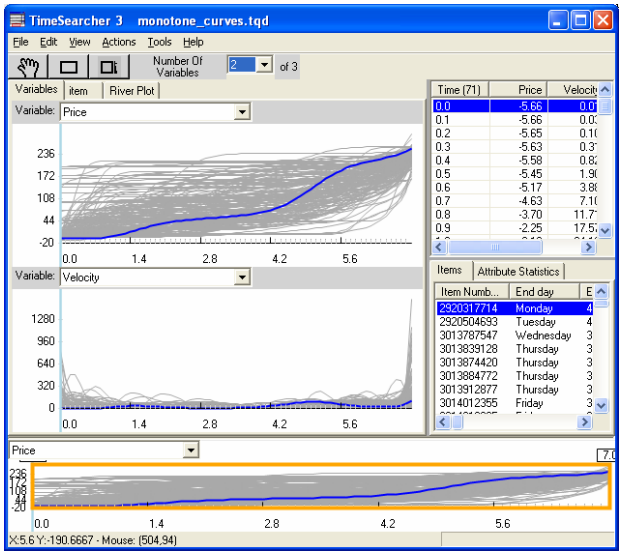

Figure 1 TimeSearcher 3 displaying 158 Palm eBay auctions with price and velocity curves, values table and auction attributes

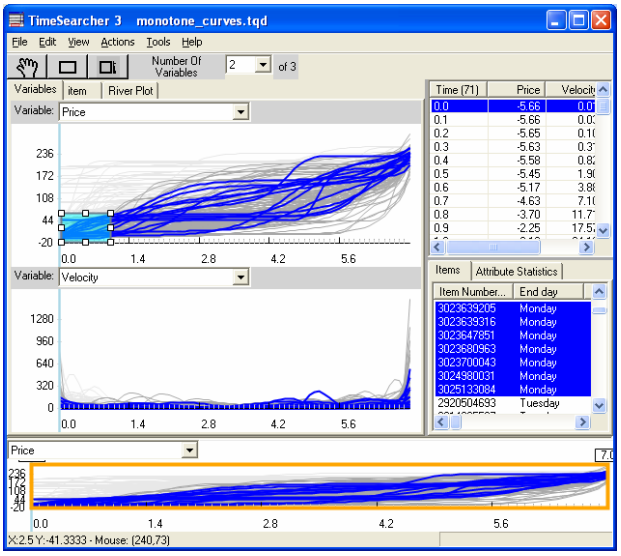

Figure 2 Filtering to see only auctions that started at a low price and ended on a Monday

Figure 1 shows the main view of TimeSearcher 3, i.e. the "Variables" view. The top panel shows the evolution of price for each auction over time. The second panel shows the velocity, that is, the first derivative of price curves. Users can add other panels for other variables. The selected auction appears blue in the timelines and in the table of attributes (bottom right). The top right table shows the current price of the selected item (auction) at each time point.

TimeSearcher 3 allows users to zoom in the areas of interest and review attribute values and time series values. The exploration is facilitated by highlighting as well as interactively filtering the time series. Users can filter items by drawing one or more TimeBox widgets directly on the lines (Figure 2). Only items whose time series passes through the TimeBoxes are kept and the others are filtered out [9]. The items table is updated accordingly. Another widget, called SearchBox, allows users to mark a pattern in one line and find items with similar patterns at any time in the series [5], [14].

Users may also filter using attribute values. They can sort the items list and select the items they want to keep and filter out the unselected items, or they can use the filter attribute window, accessible from the menu, to specify the desired range of attribute values to be kept, and filter out the rest of items. In Figure 2, auctions are sorted according to the "End day" attribute. Then, auctions that end on a Monday are selected and all the remaining auctions are filtered out using "Filter Unselected" in the menu (their color is light grey).

3 The River Plot View

As the number of items increases, overlapping becomes a problem so we included a summary visualization called river plot, accessible in a separate tab below the toolbar. It shows summary statistics of the time series values (Figure 3). The four regions visible are bounded from bottom to top by the minimum, 25th percentile, median, 75th percentile and the maximum value curves. A separate river plot is shown for each variable. The river plot can be seen as a continuous box plot representation (like those used to show medical data [4]). It is also related to other trend summarizations, such as the topic river [8], but we use it for a different purpose and enable interactive user control.

The river plot view of the filtered view in Figure 2 is displayed in Figure 3. The statistics for attributes of the items, such as minimum, standard deviation, etc. are visible by switching to the attribute statistics tab at the bottom right corner. The river plot provides a rapid understanding of the distribution of time series values of the whole or part of the dataset. For instance, Figure 3 shows that variability is low at the beginning and at the end, while high in the middle. This suggests that Palm m515's all pretty much sell at the same price although the price in the middle varies greatly.

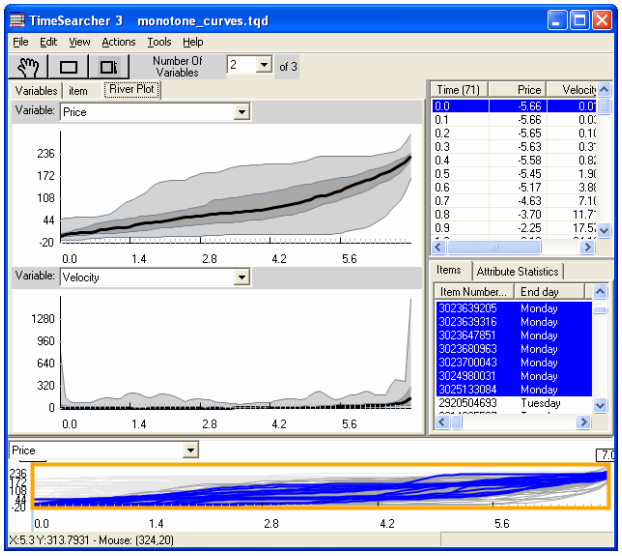

Figure 3 River plot of filtered items of Figure 2

4 Forecasting

Having a database of similar historical time-series (auction price curves, medical records, meteorological data, etc.) and a partial time series, it is possible to base a forecast on the historical data at hand. As stated in the introduction, there are basically two approaches for forecasting: model driven and data driven. The model driven approach is based on fitting a model to the time series, based on the specific domain. The data driven approach identifies patterns in the current time series and uses those to extrapolate into the future. We take a data driven approach, but rather than extrapolating the partial time series using its past, we compare it with historical time series in order to find similar behavior in the database. This assumes that the partial time series behaves similarly to the historic time series. This approach requires the adoption of similarity algorithms and it assumes that exceptional events are excluded and that all possible events are represented in the historic database. Since a data-driven approach requires larger datasets than model driven methods, the historic dataset must include a sufficient number of records. The data driven approach is domain independent, and enables automated forecasting.

To obtain a forecast, users first select a partial time series to be forecasted (the source). Users then decide if the entire database should be used for forecasting, or a subset of it. In our auction example, a user trying to forecast a 7-day auction that ends on a Monday might choose only 7-day auctions that end on a Monday. Figure 2 shows the filter made by the user.

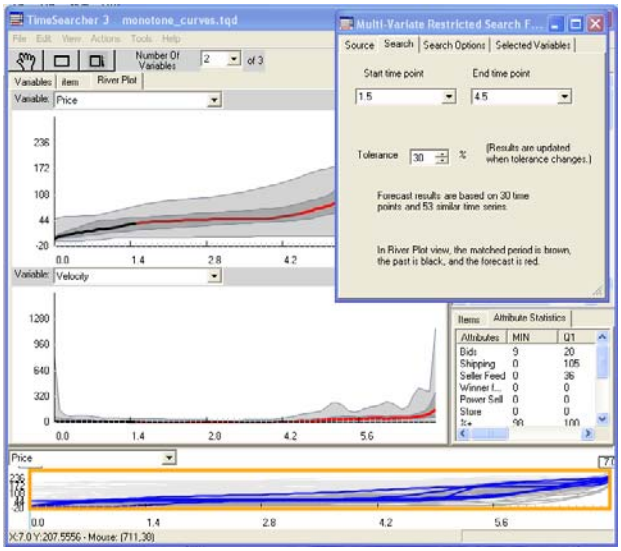

Figure 4 Forecasting interface: red represents the forecast. The median of the subset of matched items during the period used for matching is brown, while the median before this period is black

Running the similarity search produces a subset of similar items, the SBF subset, which is displayed to the user. It can be reviewed either in the variable view or the river plot view. The median of the river plot of this SBF subset is the forecast, while the minimum, maximum, the 25th and 75th percentile indicate the statistical distribution of the SBF subset, which visually indicates the forecast uncertainty.

Figure 4 shows TimeSearcher 3 with the forecasting window overlapping on the top right. Users can select a file as the source and set time bounds for the forecast by specifying the time interval within the time series that should be used. The river plot in Figure 4 is based only on the SBF subset. The median of the SBF subset is colored in black, brown and red. The red part is the forecast. The brown part is the median of the SBF subset during the period used for matching, while the black part is the median before this period. Similarity algorithms have many parameters that affect which items will be found similar and remain in the SBF subset and those parameters need to be selected.

Users can select one or more variables of the dataset for the search algorithm to consider. When more than one variables are selected, they are used conjunctively. In other words, for a time series to be in the SBF subset, it must match the pattern in all the variables selected. In addition, users can dynamically change the setting of the other pattern matching parameters, which are the similarity algorithm, the time interval to be considered for matching, the tolerance, and the data transformations applied. (See [5] for the details of the transformations and similarity algorithms.). The effect of those changes is reflected immediately on the display, which facilitates experimenting with choices of parameters. Users can see how the forecast changes as they increase the tolerance

(i.e. acceptable distance from the source); or change from one search algorithm to another; or change transformations and normalizations applied.

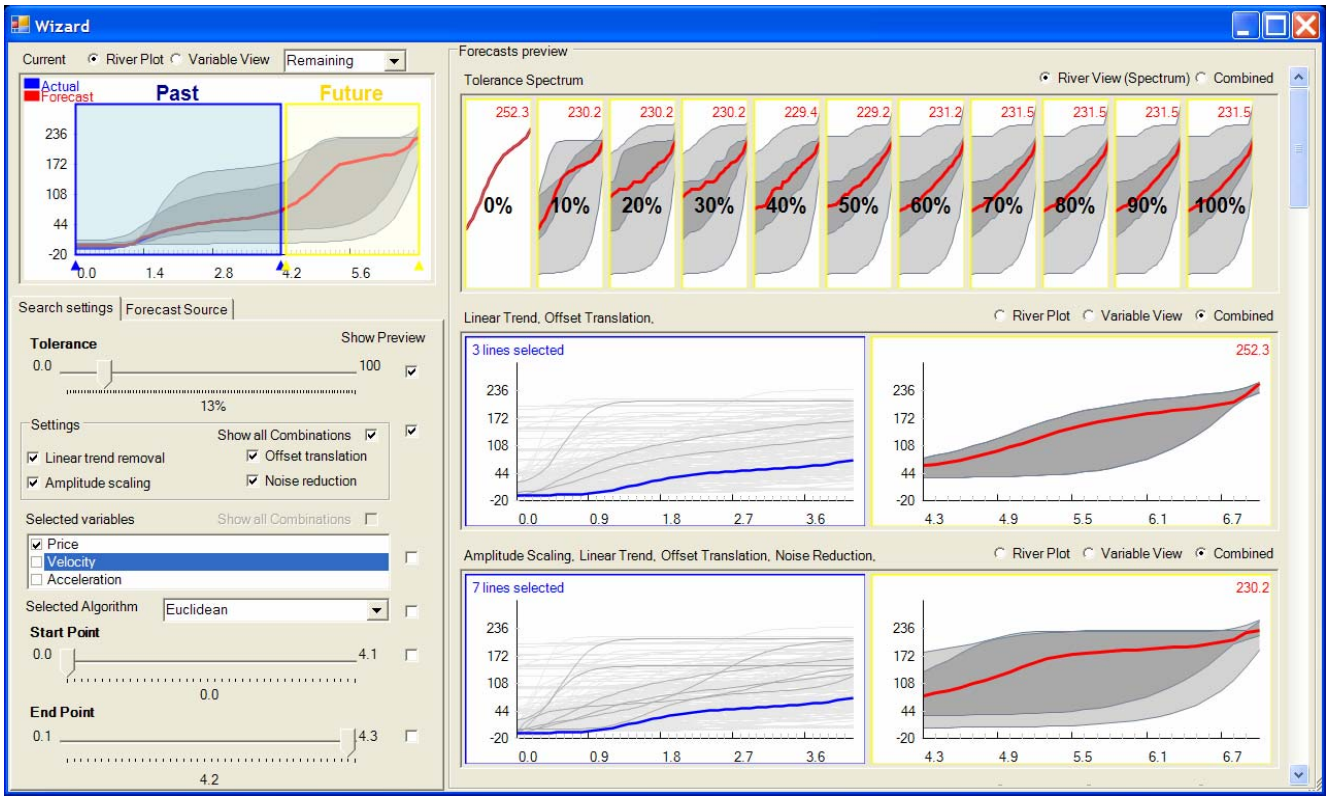

Figure 5 – The simultaneous preview interface. The top left area shows the current selection The bottom left area contains the options the user can select. The right area is the preview area, that contains the Tolerance Spectrum Preview (on the top) and the preview of binary parameters (in this figure is shown the combined view, composed by both the variable view and the river plot)

This feature was always greatly appreciated by users who experimented with the interface. We believe that users are more likely to get a sense of the forecast variability and robustness by comparing forecasts produced with different choices of parameters. Some users commented that they believed the results of the initial forecast (usually run with the default parameter values) but later realized the richer possibilities when seeing how sensitive the results were. This process gives users a more realistic understanding of the uncertainty involved in forecasting.

Observing users interacting with TimeSearcher 3 revealed that exploring all possible choices of parameters is a challenge. Tests are done one at a time, and there is no easy way to compare different forecasts to estimate differences. Unless users are experts in the similarity algorithm used, they opportunistically try variations and are rarely able to predict how different parameters will influence the forecast. Those observations led to the design of a forecasting preview interface, which allows users to systematically see the results of multiple parameter variations at once.

5 Simultaneous Preview Interface

Our initial inspiration for the simultaneous preview interface was a photo retouching tool. Adobe Photoshop has a “Variations” interface [2], which allows users to see multiple previews in order to choose parameters. In addition, the ability to view multiple previews of different forecast settings should allow users to select better parameters. Early research on Side Views [16] introduced the idea of a context sensitive set of previews, also in the context of graphics editing. When users browse the menu entries, a “side view” appears in a tool-tip like fashion, showing what will happen if the highlighted menu item is selected. In [10] and [13] other novel ways of presenting dynamic previews are shown. We implemented a similar approach for forecasting.

The interface shown in Figure 5 is divided into three main areas: the option panel (bottom left), the current selection (upper left) and the preview area (right). In the option panel, users can select values and choose which parameter to vary in the previews. For each parameter there is a “Show Preview” checkbox. When clicked, the corresponding preview panel is added in the preview area. The preview can handle continuous parameters

(such as the tolerance - that is the similarity factor, or the start and end point of the pattern to be matched) and sets of binary parameters (such as all possible transformations). In the case of continuous parameters the panel shows several previews taken at various steps in the full range that the parameter can assume. In the case of binary parameters, the preview panel shows two series of panels for all possible combinations of the binary parameters. For each combination users can see either the river plot or the variable view or both at the same time. In Figure 5, the Tolerance Spectrum preview panel and the Search Settings preview panels are activated. The interface allows users to quickly glance over different forecasts with different settings. In the Tolerance Spectrum panel, users can see the forecast at different tolerance values. The simultaneous preview panel of Figure 5 shows that the biggest difference is between 0% and 20% tolerance (so users might decide to zoom on 10% to see what exactly is changing in further detail). In the two panels below, users preview different combinations of search settings (using the "Show all Combinations" checkbox – the other parameters used for those forecasts are those currently selected in the option panel). With all transformation applied, the forecasted final price is $230.2, whereas with only two of those settings (in the second panel), the forecasted final price is $252.3, a substantial difference.

The combined view, displaying both the river plot and the variable view, shows users exactly on which items the forecast is based (those items that were deemed similar by the search algorithm), thereby allowing them to base decisions on the visible results. Users can zoom in on sections and change the beginning and end of the pattern to be matched (top left area of Figure 5).

5.1 User Testing

We collected feedback from a total of eight users in two rounds of testing. Users interacted with the preview interface without training for about 15-20 minutes using a think aloud protocol and then summarized the problems they encountered. All users were university students but only two were computer science students. Only one had experience with TimeSearcher. Several usability issues arose in the first round of testing. These were addressed and led to the version described in section 5. For example, in the initial version, instead of having boxes labeled "Past" and "Future" in the top left current preview panel, we had "Pattern" (enclosing only the brown part, i.e. the match period) and "Result" (enclosing the red part, i.e. the forecast) boxes, but this was not well understood. Some users also had problems recognizing the meaning of the two different colored lines in the current preview panel, so color was used more consistently in all panels and used for labels as well. Due to the high number of previews displayed simultaneously, some users asked for a better way to know which parameters led to a certain forecast, so we added a tooltip with this information.

Most of the users agreed that seeing many previews at once helped them to forecast time series more easily and accurately. Some users commented that they wanted to see how the parameters affected the results instead of having to experiment with many settings. Still, novice users had difficulties in noticing the often subtle differences between the various panels, so future work could focus on adding tools and features that assist users in judging the outcomes of the forecasts. More experienced users wanted future versions to display even more previews simultaneously and to filter the results so that the more relevant ones are more easily presented. Further work might support faster pattern search algorithms.

5.2 Discussion

When we presented the river plot interface with previews to users, we noticed that they were always surprised by the variations when parameters were adjusted. In fact, several users commented that seeing only one result from the first trial (using an initial set of parameters) could easily mislead them to believe that they saw the definitive forecast. However, being able to vary the parameters and rapidly see the changes allowed them to gauge the variability of the results and select better parameters. Nevertheless, it is clear that the simultaneous preview interface is complex and can overwhelm some users. Obtaining a forecast as a quick guess is fairly easy, but fully understanding the subtleties of parameter selection remains a challenge.

Currently, we are considering only the forecast for one data source. Future work could consider multiple (non-overlapping) sources. A further addition to this could be to assign weights to each time interval, which will affect the similarity measure.

The time series of items used here are smoothed curves, which introduces some uncertainty. In general, uncertainty may be introduced during the acquisition, transformation, and visualization [7], [18]. A further direction could be reducing uncertainty in visualization.

Finally, another future direction could be to combine forecasts [1], [6]. Using different search algorithms produces different forecasts. Combining the forecasts resulting from each algorithm may increase the accuracy.

Conclusions

Similarity-Based Forecasting (SBF) is a data driven forecasting method that uses the similarity of the series to be forecasted with a set of similar historical time series. The enhancement of TimeSearcher 3 with the river plot view and the forecasting tool shows the applicability and feasibility of this approach. The forecasting preview allows users to look at many combinations simultaneously. Users conduct a similarity search to select a subset of similar items from a time series dataset. The assumption is that the dataset represents the same phenomenon as the forecasted item. Alternatively, users may filter the dataset to arrive at a

relevant subset of similar time series and then continue with the forecasting process. The median of this subset becomes the forecast, while the minimum, maximum, 25th and 75th percentile indicate the statistical distribution of the dataset, which visually indicates the uncertainty associated with the forecast. The similarity search has several options and parameters that affect which items the subset will contain. Those options and parameters are the search algorithm, the subsequence of the forecast source to be used for similarity, transformations to be applied, the tolerance for matching, and the variables to be considered for the pattern search.

We believe that this work provides a basis for further research on building effective user interfaces for exploratory time series analysis and forecasting. We reinforce previous work that shows benefits if users can view multiple forecasts simultaneously.

Acknowledgements: We thank the Center for Electronic Markets and Enterprises and the R.H. Smith School of Business at the University of Maryland for providing support for this project.

References

[1] Armstrong, J.S., Combining Forecasts, Principles of Forecasting: A Handbook for Researchers and Practitioners, J. Scott Armstrong (ed.): Norwell, MA: Kluwer Academic Publishers (2001),.417-439.

[2] Adobe Systems Inc.: http://www.adobe.com, last accessed (January 2007).

[3] Aris, A., Shneiderman, B., Plaisant, C., Shmueli, G., and. Jank, W, Representing Unevenly-Spaced Time Series Data for Visualization and Interactive Exploration, Proc. of INTERACT 2005, Rome (Sept. 2005), 835-846.

[4] Bade, R., Schlectweg, S., and Miksch, S.,Connecting time-oriented data and information to a coherent interactive visualization, Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems (2004), 105-112.

[5] Buono, P., Aris, A., Plaisant, C., Khella, A., and Shneiderman, B., Interactive Pattern Search in Time Series, Proc. of VDA'05, SPIE, Washington, DC (2005), 175-186.

[6] Clemen, R.T., Combining Forecasts: A review and annotated bibliography, International Journal of Forecasting, 5 (1989), 559-583.

[7] Djurcilov, S., Kim, K., Lermusiaux, F.J., and Pang, A., Volume Rendering Data with Uncertainty Information (2001), In David S. Ebert, Jean M. Favre, and Ronald Peikert, editors, Proc. of the Joint Eurographics - IEEE TCVG - VisSym-01, Wien, Austria, Springer-Verlag. (2001), 243–252,

[8] Havre, S., Hetzler, E., Whitney, P., Nowell, L. (2002), ThemeRiver: Visualizing Thematic Changes in Large Document Collections, IEEE Transactions on Visualization and Computer Graphics, Vol.8, No. 1 (January-March 2002).

[9] Hochheiser H., and Shneiderman, B., Dynamic Query Tools for Time Series Data Sets, Timebox Widgets for Interactive Exploration, Information Visualization 3(1) (Mar. 2004), 1-18.

[10] Igarashi, T., Hughes, J. F., A Suggestive Interface for 3D Drawing, 14th Annual ACM Symposium on User Interface Software and Technology, Orlando, FL (2001).

[11] Jank, W., Shmueli, G., and Wang, S., (2006). Dynamic, real-time forecasting of online auctions via functional models. In Proc. 12th ACM SIGKDD (Philadelphia, PA). ACM Press, New York, NY (2006), 580-585.

[12] Kincaid, R. and Lam, H. (2006), Line graph explorer: scalable display of line graphs using Focus+Context. In Proc. of the Working Conference on Advanced Visual interfaces (Venezia, Italy) AVI '06. ACM Press, New York, NY (2006), 404-411.

[13] Marks, J., Andalman, B., Beardsley, P., Freeman, W., Gibson, S., Hodgins, J., Kang, T., Mirtich, B., Pfister, H., Ruml, W., Ryall, K., Seims, J., and Shieber, S., Design Galleries: A General Approach to Setting Parameters for Computer Graphics and Animation, Proc. ACM SIGGRAPH 97, Los Angeles, CA (1997), 389-400.

[14] Shmueli, G., Jank, W., Aris, A., Plaisant, C., and Shneiderman, B., Exploring auction databases through interactive visualization, Decision Support Systems 42, 3 (2006), 1521-1538

[15] Silva, S. F. and Catarci, T. (2000), Visualization of Linear Time-Oriented Data: A Survey. Proc. of the 1st international Conference on Web information Systems Engineering (Wise'00) IEEE Computer Society, Washington, DC (2000), 310.

[16] Terry, M., Mynatt E. D., Side Views: Persistent, On-Demand Previews for Open-Ended Tasks, Proc. 15th Annual ACM Symposium on User Interface Software and Technology, ACM Press (2002), 71-80.

[17] Wang, S., Jank W., and Shmueli, G., Explaining and Forecasting Online Auction Prices and their Dynamics using Functional Data Analysis, Journal of Business and Economic Statistics (in press, 2007).

[18] Wittenbrink, C.M., Pang, A.T., and Lodha, S.K., Glyphs for Visualizing Uncertainty in Vector Fields, IEEE Transactions on Visualization and Computer Graphics 2,3 (September 1996), 266-279