International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:2, 2010

Visual-Graphical Methods for Exploring Longitudinal Data

H. W. Ker

***Abstract*—**Longitudinal data typically have the characteristics of changes over time, nonlinear growth patterns, between-subjects variability, and the within errors exhibiting heteroscedasticity and dependence. The data exploration is more complicated than that of cross-sectional data. The purpose of this paper is to organize/integrate of various visual-graphical techniques to explore longitudinal data. From the application of the proposed methods, investigators can answer the research questions include characterizing or describing the growth patterns at both group and individual level, identifying the time points where important changes occur and unusual subjects, selecting suitable statistical models, and suggesting possible within-error variance.

***Keywords*—**Data exploration, exploratory analysis, HLMs/LMEs, longitudinal data, visual-graphical methods.

I. Introduction

EXPLORATORY analysis is a technique to visualize patterns of data. Visualization implies that the data information is processed into a visual display (Cleveland, [3]). Graphic representations of summarizing data have the advantages over numerical summary because graphics can describe the numerous values and underlying data patterns simultaneously (Behrens [2]). Furthermore, visualization of data is important and useful in model building because it can guide predictions (i.e., identify the potential significant parameters, especially when there is no theoretical background knowledge to guide model formulation), improve the understanding of the fact hidden behind the summary statistics and significant tests, stimulate new hypotheses to be tested, and detect the possible underlying patterns. Visualizing data should be the first step in data analysis. Behrens [2] suggested that all initial work should be published. Stoolmiller [10] also noted that visual-display is especially useful in analyzing growth curve data. However, Verbeke & Molenberghs [13] mentioned that most research in statistical modeling has been largely focused on model formulation, inferences, and software implementation. Other important aspects such as exploratory analysis have received less attention.

Longitudinal data typically represent the measurements of a response over time under different experimental conditions for different subjects. They possess hierarchical structures in the sense of repeated measurements are nested within individuals. Repeated measurements are the first level and individuals are the second levels (Hox [6]; Daper [5]; Wang [12]). The errors in longitudinal data often exhibit heteroscedasticity and dependence. Longitudinal patterns are often nonlinear because there are different growth rate at different periods (Cudek & Klebe [4]). In the developmental process studies, researchers are often interested in the time points where the important changes occur. The exploration of longitudinal data is thus more complicated than cross-sectional data (Bates & Pinheiro [1]). However, traditional graphical methods in most of statistical texts such as stem-and-leaf plots, histograms, or box plots are not enough for longitudinal data exploration because of the characteristics of longitudinal data (i.e., nonlinear patterns, heteroscedasticity, dependence and between-subjects variability). Graphical techniques for longitudinal data include:

1. characterizing the nature of data and identifying the time points where the important changes occur;
2. describing the form of growth at group and individual level (e.g., linear or nonlinear);
3. identifying a set of parameters that can help to interpret growth curves;
4. suggesting suitable statistical models (i.e., standard multiple regressions or hierarchical linear models/linear mixed-effects models, HLMs/LMEs); and
5. identifying longitudinal patterns and unusual subjects.

Because of these advantages, data exploration should be performed first before any model buildings or statistical tests. Although graphical techniques are helpful in statistics practices, there are no principles to guide the exploration of longitudinal data. Conventional visualization methods and new graphical techniques have not been integrated into a systematic procedure to suit longitudinal data exploration. This paper attempts to organize/integrate the visual-display techniques and then demonstrate a series of plots that illustrate how to be used for exploratory analysis of longitudinal data.

II. Graphical Methods for Longitudinal Data Exploration

The types of graphics for data exploration depend on the nature of the data. Most longitudinal data analyses address individual growth patterns over time. If different treatment conditions or different background variables (i.e., gender, school sectors, and social-economic-status) are involved, the overall growth patterns as well as the within-subject and between-subjects patterns under different treatments and

H. W. Ker is with the Chihlee Institute of Technology, Taipei, Taiwan (e-mail: hker@ mail.chihlee.edu.tw).

background variables should be examined, too. Separate panel plots can be produced for each of the treatment/background levels. These plots can provide initial impressions regarding possible functional forms, data variability as well as individual behaviors on treatment effects. In other words, longitudinal data explorations should include exploring growth patterns, patterns of experimental conditions, interactions of variables as well as the within-subject and between subjects variability.

Graphical approaches for longitudinal data exploratory can include:

1. exploring the distributions of raw data: the first step in data exploration is to get an initial impression regarding the central tendency, the spread and the skewness of variables. Plots can be used are histograms, dot plots, bar charts, stem-and-leaf displays and box plots.
2. plotting numerical descriptive measures to illustrate individual profiles, overall profiles and possible model forms: Graphics for these purposes are mean plots, standard deviation plots and median plots.
3. exposing the interactions between/among variables under different treatment/background levels: the usual graphics for exploring interactions are interaction plots and Trellis approaches (Pinherio and Bates [8]; Venables & Ripley [11])
4. exploring within-subject and between-subjects behaviors: the three guidelines mentioned above are for single-level data exploration. Longitudinal data have hierarchical structures that repeated measurements are within individuals. It is crucial to examine within-subject and between-subjects variations. Graphics used are empirical growth plots (Singer & Willett [9]) and Trellis approaches.
5. determining the need of multilevel analysis: multilevel analysis is needed when between-subjects variations present. When there is no individual difference, ordinary regressions are enough for statistical modeling.

III. Data

The visual-search data set (Peterson & Kramer [7]) is used to demonstrate longitudinal exploratory analysis. The visual-search experiment was aimed at examining the interaction of top-down and bottom-up influences on visual search. Bottom-up influences, such as sudden appearance of a new objector with its own motion or form, are to capture attention without intention. Top-down influences such as the knowledge that a dropped item is more likely found on the floor than on the ceiling, may or may not be under intentional control (Peterson & Kramer [7]). There were eighteen subjects (4 males and 14 females) in this study. The experiment was divided into 45 blocks with 24 repeated and 24 new configuration trials in each block. The distractors are 0° and 90° rotated and mirror-imaged L's. Each stimulus presentation has 11 distractors and one target. There are three sessions for this experiment and each session contained 15 blocks of 48 trials within a total of 2,160 trials. Each session has 20 practice trials. Response times (RTs) were accurate to 2 milliseconds. In this analysis, variable block is redefined as variable time (equals to block -1).

IV. Data Analysis

A. Exploring "Growth" Patterns

The first step and perhaps the best way to get a sense of a new data set are to visualize or plot the data. Most longitudinal data analyses address individual growth patterns over time. Thus, the first useful exploratory analysis is to plot the response variable against time including individual and overall mean profiles. Individual mean profiles, which summarize the aspects of response variable for each individual over time, can be used to examine the possibility of variations among individuals and to identify potential outliers. The overall mean profile summarizes some aspects of the response variable over time for all subjects and is helpful in identifying unusual time when significant differences arise.

Fig. 1 shows the lines connecting the dependent variable (mean response time (RT)) over time for each subject. A number of patterns could be observed from Fig. 1. First, most subjects have higher mean RTs, which mean slower responses occur at the beginning of the observation period, and they tend to decrease over time. Secondly, the spread among the eighteen subjects is substantially larger at the beginning than that at the end. Subject 153 is an exception because it has very high mean RT at the end of the observation period. Third, there exist noticeable variations among subjects.

The overall mean growth curve over time is presented in Fig. 2. The overall mean RTs are larger at the beginning and decrease over time; and there is some variation roughly after time 16. Fig. 3 shows the mean RTs for each subject on each block for three days. The lower panel is day 1 (Block 1~15); the middle panel is day 2 (Block 16~30); and the upper panel is day 3 (Block 31~45). Day-to-day variation in mean RTs is also noticeable. For example, the overall trend in day 1 has higher mean RTs compared to those in day 2 and day 3. In day 1, higher mean RTs or the slower responses tend to occur in the first block as shown with open circles. Within each day, subject-to-subject variability is greater than block-to-block variability within a subject.

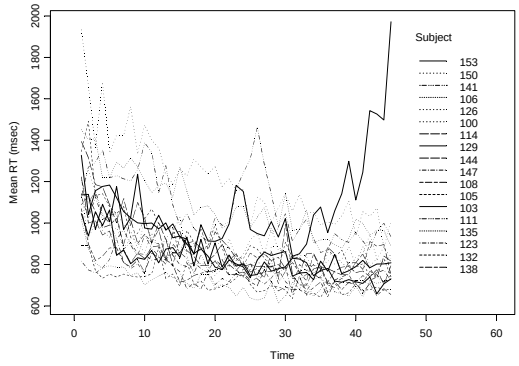

Fig. 1 Mean RT (over 48 trials) for each subject plotted versus time

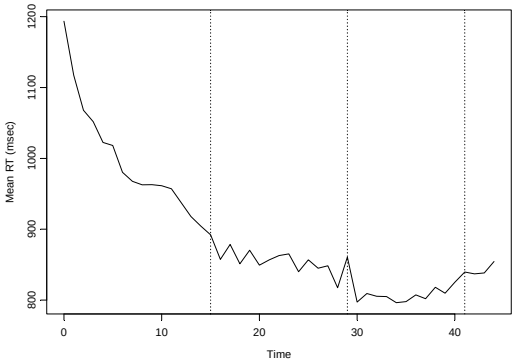

Fig. 2 Overall mean RT (over subjects and trials) plotted versus time

Fig. 4 shows the overall mean RTs ordered from the smallest to the largest against time. Note that overall mean RTs decrease from time 0 to time 15; some main shifts of overall mean RTs occur at time 15, 29, and 41. Fig. 5 shows the patterns of mean RTs for each subject over time. All of the patterns are similar for all subjects except subjects 153,150 and 141. Subject 153 has higher mean RTs at the end, subject 150 has higher mean RTs at the beginning, and subject 141 has higher mean RTs at the middle of the study.

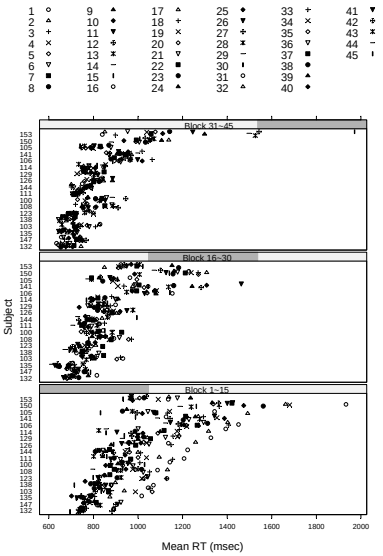

Fig. 3 Mean response time for each subject for each day

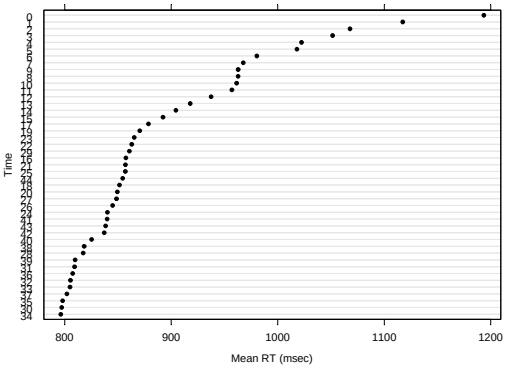

Fig. 4 Ordered overall mean RT against time

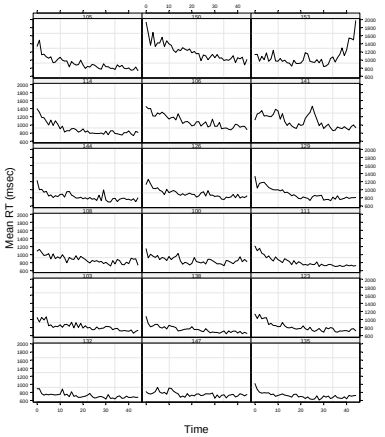

Fig. 5 Mean RT for each subject over time

B. Exploring the Patterns of Experimental Conditions

In addition to time (which is equal to block-1), two major experimental conditions including treatment configuration (new or old) and distractor presence (present or absent) are

considered. This exploratory analysis is intended to discover the overall and individual patterns of each experimental condition and their interactions on mean RTs. Subsequently, the patterns of mean RTs for each subject and the patterns of overall mean RTs on each experimental condition and their interactions over time are investigated.

Fig. 6 is a plot of overall mean RTs for two configuration levels (new and old) over time. The mean RTs for these two configuration levels are decreasing over time. However, the mean RTs for new configuration are consistently higher than for old configuration. The differences between both configurations become larger roughly after time 8.

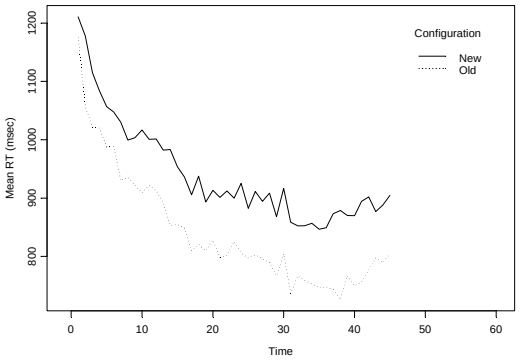

Fig. 6 Overall mean RT on two levels of configuration over time

Fig. 7 is used to examine the patterns of mean RTs on the two levels of configuration for each subject. The patterns of subjects 150, 153 and 141 are much different from other subjects. These three subjects have more peaks either at the beginning (subject 150), at the middle (subject 141) or at the end (subject 153) of the observation periods. The mean RTs for new configuration are higher than for old configuration for most subjects over time.

Fig. 8 depicts the patterns of mean RTs for two levels of treatment configuration plotted separately for each subject. Fig. 7 is helpful in examining the individual differences of mean RTs on configuration levels while Fig. 8 is useful in visualizing the patterns of both configuration levels separately for each subject.

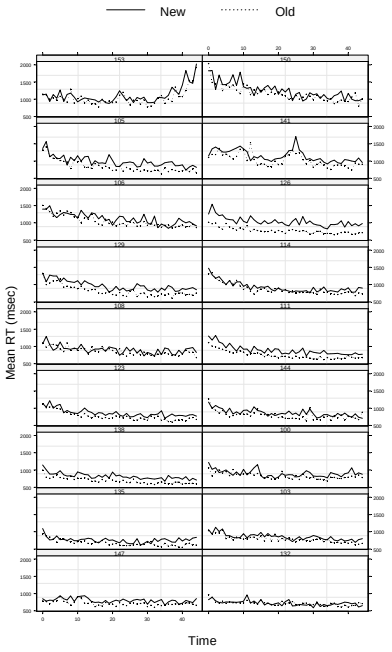

Fig. 7 Mean RT on two levels of configuration over time by subject

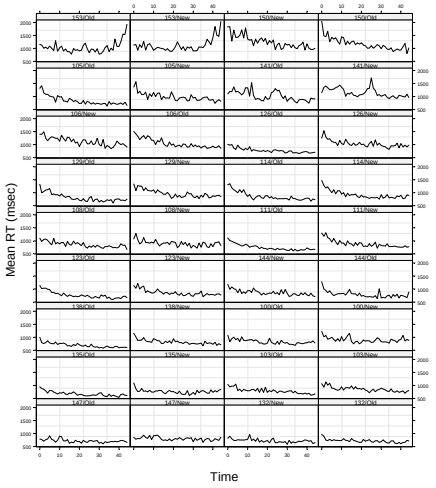

Fig. 8 Mean RT for new configuration and old configuration for each subject

Likewise, a similar approach used for the exploratory analysis of treatment configuration is applied to that of onset distractor presence. Fig. 9 shows that onset presence has slower response (higher mean RTs) at earlier time (e.g., before roughly time 7), and has some variations afterwards. The patterns of the two levels of onset presence are relatively similar and their differences are small. Fig. 10 and Fig. 11 show the patterns of mean RTs on two onset presence levels for each subject. The differences between two levels of onset presence are small for each subject. Subjects 153, 150 and 141 have different patterns from others.

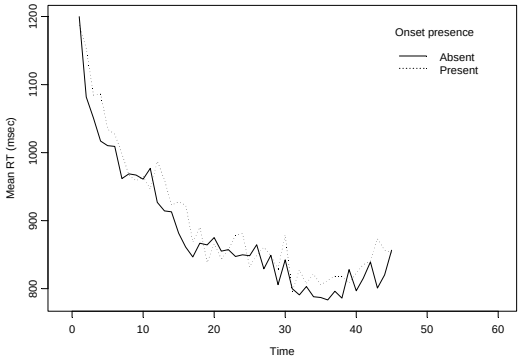

Fig. 9 Overall mean RT on two levels of onset presence over time

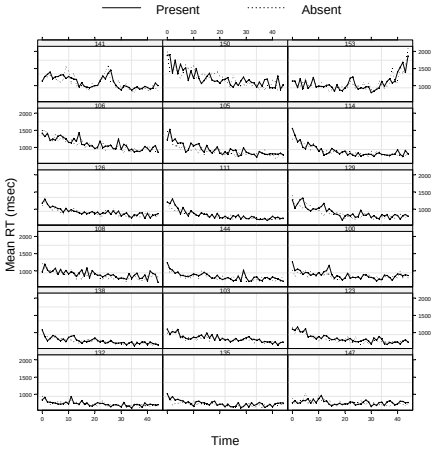

Fig. 10 Mean RT on two levels of onset presence over time by subject

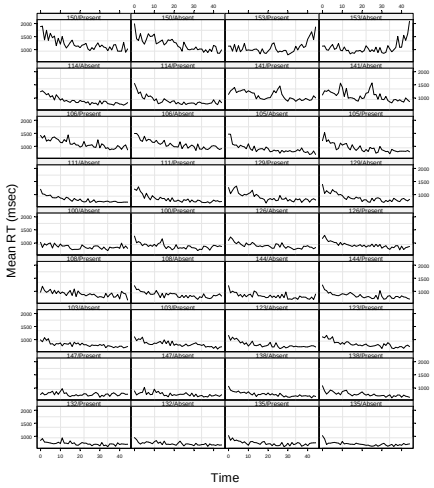

Fig. 11 Mean RT for onset presence and onset absence separately for each subject

Similarly, the exploratory analyses for the interactions of onset presence and configuration are displayed in Fig. 12 to Fig. 14. The conclusions are almost the same as those given in the exploratory analysis of onset presence. Fig. 12 depicts that the patterns of four experimental conditions are similar. New configuration (distractor present or absent) has higher mean RTs than old configuration (distractor present or absent) has most of the time. The difference in new configuration with distractor present and with onset absent is small, so is the difference in old configuration with distractor present and with distractor absent. The patterns of these four interaction conditions for subjects 153, 150 and 141 are much different from others.

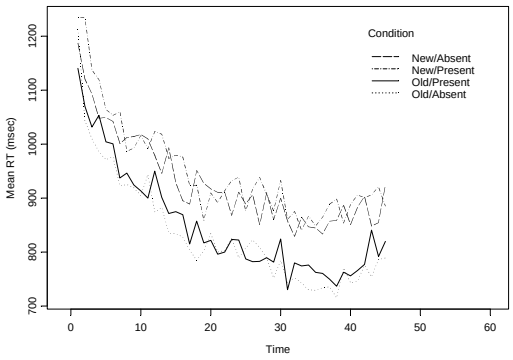

Fig. 12 Overall mean RT on the interaction of onset presence and configuration over time

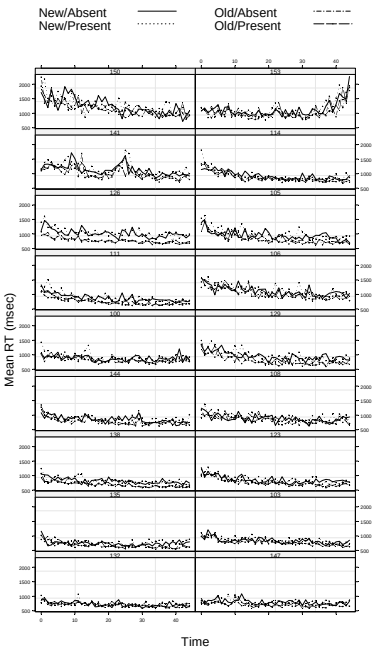

Fig. 13 Mean RT on the interaction of onset presence and configuration over time by subject

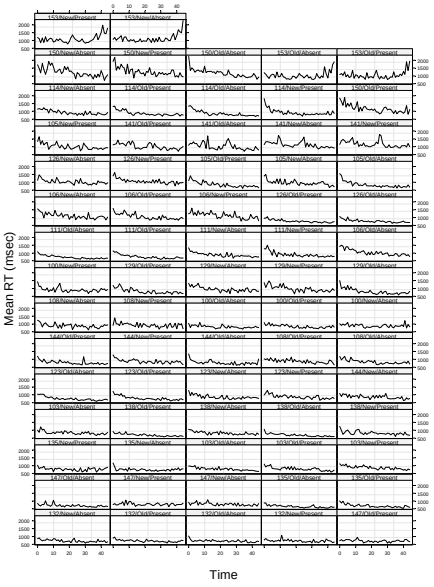

Fig. 14 Mean RT on the interaction of onset presence and configuration over time by subject

C. Selecting a Preliminary Mean Structure

The curve in Fig. 2 indicates that this visual-search data set may have quadratic functional form. The visual-search data has three explanatory variables: configuration, distractor presence, and time. From Fig. 9, the curves for two levels of onset presence have no apparent difference and are approximately parallel with each other. Hence, there may be no interaction between distractor present and time.

The model could have the form:

$$\overline{RT_{ij}} = \beta_{0j} + \beta_{1j}(Onset)_{ij} + \beta_{2j}(Configuration)_{ij} + \beta_{3j}(Time)_{ij} + \beta_{4j}(Time)^2_{ij} + \beta_{5j}(Onset)_{ij}(Configuration)_{ij} + \beta_{6j}(Configuration)_{ij}(Time)_{ij} + \text{Quadratic Function of Time} + R_{ij} \quad (1)$$

where the quadratic function of time is the time points where may have important changes

D. Selecting a Statistical Model

The second step for longitudinal modeling is to select suitable model to interpret data. An appropriately specified statistical model is helpful in interpreting the random variation in the data, achieving the efficiency of estimation, as well as obtaining valid inferences of the parameters in the mean structure of the model. The possible choices are standard multiple regression and hierarchical liner models/linear mixed-effects model (HLMs/LMEs). Standard multiple regression technique is used for single-level data and the within-subject errors are homoscedastic and independent. HLMs/LMEs are suitable for modeling hierarchical data. However, it should be noted that if a hierarchical data set does not show between-subject variability, and the assumptions of homogeneity and independence for residual errors are tenable, standard multiple regression is a better choice.

A useful tool to explore the random-effects structure is to remove the mean structure from the data and use ordinary least square (OLS) residuals to check the need for a linear mixed-effects model and decide which time-varying covariate should be included in the random structure.

Fig. 15 shows the boxplot of residuals by subject corresponding to the fit of a single linear regression by using the same form of the preliminary level-1 model. This is the case when grouping structure is ignored from the hierarchy of data. The residuals are not centered around zero. There are considerable differences in the magnitudes of residuals among subjects. For example, subject 153 and 150 have wider magnitudes while subject 123 and 138 have narrower magnitudes. There also exists high variability among subjects. Fig. 15 indicates the need for subject effects, which is precisely the motivation for using linear mixed-effects model. Separate linear regression models were employed to fit each subject to explore the potential linear relationship. Fig. 16 depicts the standardized residuals versus the fitted values for the preliminary level-1 model fitted to each subject. The points are not scattered evenly around the mean of the standardized residuals but exhibit a funnel shape. The band of the residuals widening to the right shows the characteristics of nonconstant variance, i.e., the assumption of homoscedasticity is violated.

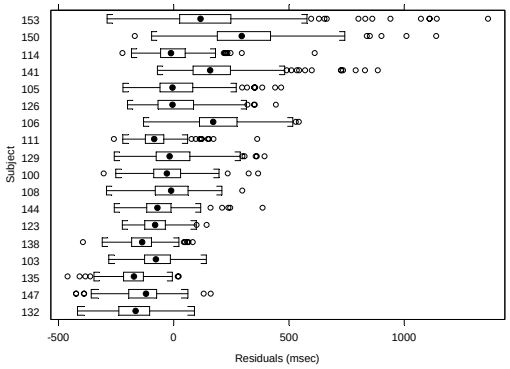

Fig. 15 Boxplot of residuals by subject to a single preliminary level-1 model

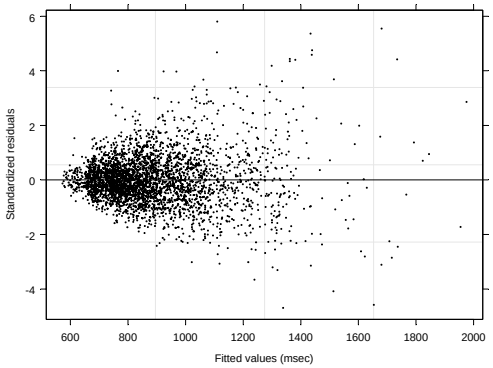

Fig. 16 Standardized residuals versus fitted values for the preliminary level-1 model fit to each subject

Fig. 17 is the mean of raw residuals by subjects from the preliminary level-1 model fit to each subject. The means are

computed for each subject on each time point. It is hard to see any systematic patterns from this figure. However, there does appear variability among subjects. Fig. 18 is the variance of raw residuals by subjects from the preliminary level-1 model fit to each subject. The variances are computed for each subject on each time point. Subjects 141, 150, and 153 have unusual pattern because there are some very high peaks at some time points for these three subjects. Fig. 19 is the variance of raw residuals from preliminary level-1 model on each time point. This plot shows that the variance of raw residuals varies as a function of time. It is also possible that if subject 153 is removed, the curve will become smoother.

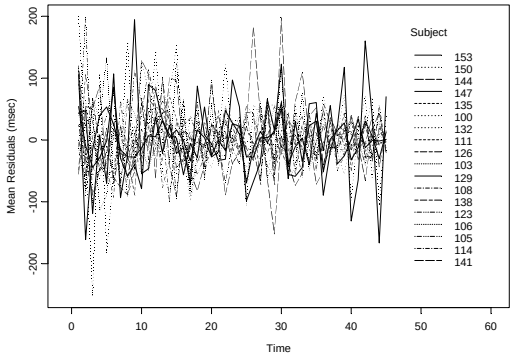

Fig. 17 Mean raw residuals from the preliminary level-1 model fit to each subject

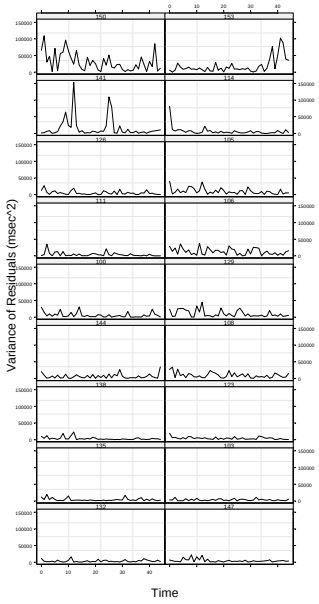

Fig. 18 Variance of raw residuals from the preliminary level-1 model fit to each subject

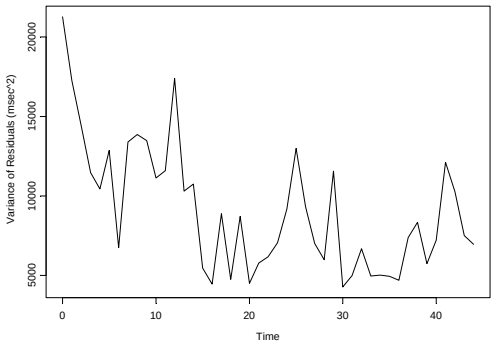

Fig. 19 Variance of raw residuals from preliminary level-1 model on each time point

Fig. 20 is the variance of raw residuals from level-1 preliminary model for the two levels of configuration. The variances are computed on the two levels of configuration on each time point. This figure shows that the variances of new configuration are higher than old configuration most of the time. It is possible that there is more variability in the new configuration condition than in old configuration condition. Different level-1 error variances for the two levels of configuration may be needed. Fig. 21 is the variance of raw residuals from level-1 preliminary model for the two levels of onset presence. The variances are computed on the two levels of onset presence on each time point. The differences between the two levels of variances of raw residuals of onset presence are small.

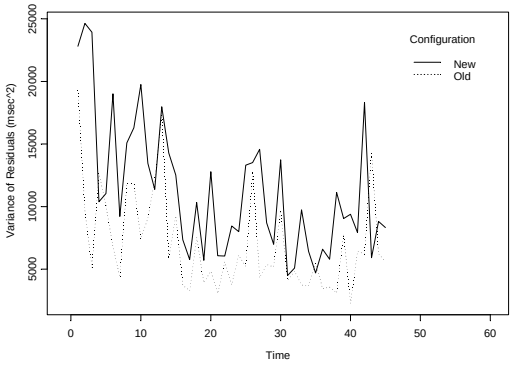

Fig. 20 Variance of raw residuals from preliminary level-1 model for configuration levels

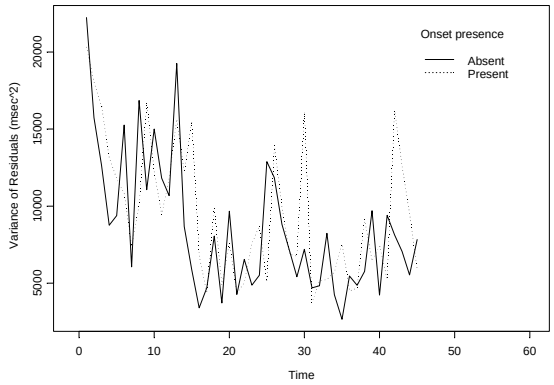

Fig. 21 Variance of raw residuals from preliminary level-1 model for onset presence

V. Results, Conclusion and Discussions

Longitudinal data are used in the search on growth, development and changes. To describe the patterns of individual growth, make predictions, and gain more insight into the underlying casual relationships related to developmental patterns requires studying the structure of measurements taken on different occasions. The advantages of utilizing visual-graphical techniques in longitudinal data exploration are demonstrated via visual-search data set. The expected outcomes would be to answer those research questions:

1. the time points where the important changes occur;
2. the patterns of developmental patterns over time (both at individual and group levels);
3. the forms of growth at group and individual level (e.g., linear or nonlinear);
4. suggestions for suitable statistical models (i.e., standard multiple regressions or hierarchical linear models/linear mixed-effects models); and
5. the potential significant predictors, and unusual subjects.

In this visual-search data set, the time points where the important changes occur are are at time 15, time 29 and time 41. The forms of growth are quadratic. The suitable statistical model is HLMs/LMEs. Several figures (e.g., Fig. 5, Fig. 7, Fig. 8, and Fig. 10) also present developmental patters for each individuals. The visual-search data has three explanatory variables: configuration, distractor presence, and time. The possible interaction terms are configuration and time, and configuration and onset presence. There may be no interaction between distractor present and time. The possible unusual subject is subject 153.

It should be noted that exploratory analysis can not replace the formal statistical tests. The model fit, significant predictors, and the functional form have to be recognized through formal statistical tests. The Through the information gathered from the visual-graphical methods at the preliminary stage, researchers then can do important checks on the adequacy of the summarization, bring researchers closer to the raw data, and provide the forms of possible initial models. Researchers can have a new way of "looking" at data via the visual-graphical methods.

Acknowledgement

The author is very grateful for Dr. Arthur Kramer and Dr. Matthew Peterson from University of Illinois at Urbana-Champaign generously provided the visual-search data set for this analysis.

References

[1] Bates, D. M., & Pinheiro, J. C. (1997). Software design for longitudinal data analysis. In T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren, & R. D. Wolfinger (Ed.), Modeling longitudinal and spatially correlated data: methods, application and further direction (pp. 37-48). New York: Springer-Verlag.
[2] Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. Psychological Methods, 2, 131-160.
[3] Cleveland, W. S. (1993). Visualizing data. Summit, NJ: Hobart Press.
[4] Cudek, R. & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. Psychological Methods, 7, 41-63.
[5] Draper, D. (1995). Inference and hierarchical modeling in the social science. Journal of Educaational and Behavioral Statistics, 20, 115-147.
[6] Hox, J. J. (2000). Multilevel analysis of grouped and longitudinal data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), Modeling longitudinal and multilevel data: Practical issues, applied approaches and specific examples (pp. 15-32). NJ: Lawrence Erlbaum Associates.
[7] Peterson, M. S., & Kramer, A. F. (2001). Contextual cueing reduces interferencee from task-irrelevant onset distractor. Visual Cognition, 8, 843-859.
[8] Pinherio, J. C., & Bates, D. M. (2000). Mixed-effects models in S and S-Plus. New York: Springer-Verlag.
[9] Singer, J. D., & Willett, J. B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. NEW YORK: Oxford University Press.
[10] Stoolmiller, M. (2002). Visual-graphical techniques for the analysis of growth curves: the shapes and predictors of growth in substance use for the male adolescents. Retrieved October 12, 2002, from http://www.oslc.org/users/mikes/sra2.html.
[11] Venables, W. N., & Ripley, B. D. (1999). Modern applied statistics with S-Plus (3rd ed.). New York: Springer-Verlag.
[12] Wang, J. (1999). Reasons for hierarchical linear modeling: A reminder. The Journal of Experimental Education, 68, 89-93.
[13] Verbeke, G. & Molenberghs, G. (2000). Linear mixed models for longitudinal data. New York: Spring-Verlag.