Longitudinal data exploration of upper respiratory tract infection disease in Bandung city

B Tantular, I G N M Jaya, Y Andriyana
Department of Statistics, Faculty of Mathematics and Science, Universitas Padjadjaran, Sumedang, West Java, 45363, Indonesia

Corresponding author: bertho@unpad.ac.id

Abstract. Longitudinal data have a complex structure. Unlike cross-sectional data, there are more associated features with longitudinal data. Associated features to longitudinal data including the heterogeneity between subjects, the serial correlation and the characteristics of changes over time. Therefore exploring longitudinal data is more complicated than that of cross-sectional data. Various methods were developed by many researchers to explore longitudinal data. In general, there are at least three things that could be described from longitudinal data: mean structure, variance structure, and correlation structure. Visual methods and subject-specific profile were studied for such datasets to detect these structures. Methods were applied to data sets from upper respiratory tract infection data in Bandung City. From the result, we can describe the growth patterns at both group and individual level and then selecting suitable statistical models which contain within and between groups variations.

Keywords: longitudinal data; exploratory analysis; upper respiratory tract infection data; visual methods; subject-specific profile.

1. Introduction
Longitudinal data have a complex structure. Unlike cross-sectional data, there are more associated features with longitudinal data. Associated features to longitudinal data including the heterogeneity between subjects, the serial correlation and the characteristics of changes over time. Therefore to find any associated features in longitudinal data, exploratory analysis can be useful. In general, to describe these features we can visualize the patterns of that data [4]. There are much information can be described from this visualization like summarizing data with underlying data patterns simultaneously and selecting the statistical model. According to Verbeke & Molenberghs [6] exploratory analysis for modelling longitudinal data modelling not often be done, because the most effort is focused on model building and inferences.

The longitudinal data usually have non-linear because it has consisted of different growth rate at different periods. Hence, time points must be considered and lead to more complex models [3]. The longitudinal data exploration must pay attention to nonlinear patterns, heteroscedasticity, dependence, and between-subjects variability.
Therefore, there are at least three things that could be described from longitudinal data are mean structure, variance structure, and correlation structure. Visual methods and the subject-specific profile were studied for such datasets to detect these structures. Visual methods are useful to describe the longitudinal pattern, between subject variability and dependencies. Subject-specific profiles are useful to describe the mean structure and variance structure. Besides those suitable statistical models are suggested for this data. This paper shows that the exploration methods applied to data sets from Upper Respiratory Tract Infection (URTI) data in Bandung City were used by Jaya et al. [10]. Obtaining the growth patterns at both group and individual level and then selecting suitable statistical models which contain within and between groups variations.

2. Methods for Longitudinal data exploration

Data exploration is a technique to describe the pattern of data. The structured of the data determines the methods of the data exploration. Individual growth patterns, overall growth patterns, within-subject and between-subjects patterns under different subjects and variables should be examined. Besides that, separate plots for every individual level can be produced. According to these plots, the first notion of possible functional association, the variability of the data and the effects of every level. Therefore the growth pattern, interaction between variables and the variance of within and between subject must be considered [4].

According to Verbeke & Molenberghs [6] and Ker [4] visual method and individual specific-profile have been used to longitudinal data exploratory. These methods include: (i) investigate the distributions of data (i.e., the evolution of mean, the variance structures, and the correlation structures); (ii) exploring subject-specific profiles (i.e., measuring the overall goodness-of-fit, testing and plotting the individual, overall profiles and possible model forms (e.g., investigate interaction between variables under different subject levels)); (iii) exploring within-subject and between-subjects behaviors. The three steps mentioned above for single-level data, and multilevel or random coefficients analysis is needed when between-subjects variations present.

3. Data Analysis and Result

Upper respiratory tract infection (URTI) is a serious problem for public health caused by boca virus. Both of Children and adult can be infected by Bocavirus. URTI is not only health but also an economic problem. It has a cost to society for treatment and covers for absenteeism from school and works [11]. URTI is one of the infectious diseases in Bandung, West Java-Indonesia that has become a serious concern of the government due to the rapid transmission and high numbers of incidence of the disease [12]. In 2016 is found 12,579 much higher than the other city in West Java, Indonesia. The high incidence of URTIs in Bandung is caused by the high population density and high mobility of the population. Develop an early warning system is needed to reduce the negative impact of URTI and controlling the spread of this disease [10].

URTI data was measured repeatedly over time from 2012 until 2016 under different conditions for different 30 district in Bandung City. This data has longitudinal structures; it means that repeated measurements are nested within the districts. Structures like this also called hierarchical structures or multilevel structures [8], [9]. This structures cause dependence and contain heteroscedasticity among observation. This dependency may be caused by a spatial structure of the data, because of the complexity the spatial dependence not noticed in this paper.

3.1. Exploring the marginal distribution

3.1.1 Individual profile plot. The individual profiles are displayed in Figure 1. This plot shows the pattern of the data over time for every subject. These plot also indicate the variability of the data at given times. In general trend of this data over times are displayed in Figure 2. This plot shows smoothed line of trend. These plot indicate the pattern of the data that URTI decreasing by time.
3.1.2 The evolution of mean. The mean evolution is displayed in Figure 3. These plots describe the profile of a number of relevant population evolves over time. The results shows that mean of URTI decreasing by time. This exploration is needed for choosing a fixed-effects structure for the linear mixed model.

Figure 4 display the individual plot of means for every district over time. These plots shows different pattern of mean evolution among district. This plot indicate a random-effects structure for the linear mixed model. There are many districts have similar pattern but there is some district have different pattern with others. Beside that in general there are districts have low data fluctuation but there is any district have high fluctuation than others.
3.1.3 Variance Structures. Mean evolution is not enough for building longitudinal model. Therefore, the variance evolution is another important component needed in order to choose the best longitudinal model. The means profile are showed in Figure 1, and the corresponding variance structured is plotted in Figure 5.

![Figure 4. Mean profiles plot by District](image)

![Figure 5. Variance Structured plot](image)

The variance structured seems to be unstable. It means that the different variance model as a starting point for building longitudinal model.

3.1.4 Correlation Structures. The correlation structure is used to describe the correlation within subjects. The correlation depends on the pair of times and simplify to the time lag under the stationarity assumption.
Correlation structured plot including the scatter plot, histogram, and correlation coefficients were displayed in Figure 6. It is also called as scatter plot matrix [6]. The figure under diagonal elements presents the correlation structured between residuals which is obtained from pairs of measurement occasions. The main diagonal elements show the histograms of the residuals. The coefficients correlations shows on upper-diagonal elements for every pairs of time points.

The result shows that the correlation is decreasing by increasing distance. Similar scatter plots under main diagonal implies stationarity. On the main diagonal, histogram shows that the variance structures are different and have skew to the left curve.

3.2. Exploring Subject-specific Profiles

3.2.1 Individual coefficient of the model plot. Response variable of URTI data in Bandung city is a count so the model is used on this paper is generalized linear model with log-link [1].

\[ \log(\mu) = X\beta \]  

where \( \mu \) is the expectation of response and \( \beta \) is the regression coefficients. For time as explanatory variables on model (1) and we model for every district. The results of intercept and slope with interval confident are displayed in Figure 7.
These figure shows that both intercept and slope are different among district. This means that the model contains random effect.

3.2.2 Measuring the overall goodness-of-fit
The determination coefficient $R^2$ is used to measure the overall goodness-of-fit of a linear regression model. On Generalized linear model coefficient determination is calculate through the deviance of the model. Based on likelihood function deviance is defined as

$$D = -2 \log \left( \frac{\lambda_0}{\lambda_1} \right)$$

where $\lambda_0$ is likelihood function under null condition and $\lambda_1$ is likelihood function under alternative when convergence. The fit of the model is shows by small deviance. According to Faraway [13] that psuedo R-square for Generalized Linear Models is

$$psuedo-R^2 = 1 - \frac{Residual\ Deviance}{Null\ Deviance}$$

(3)

To evaluate the performance of a candidate model describes observed longitudinal profiles, the pseudo determination coefficient $psuedo-R^2$ can be calculated for each subject. Based on [6] the overall measure for goodness of fit is meta psuedo-R$^2$

$$psuedo-R^2_{META} = 1 - \frac{\sum_{i=1}^{n} Residual\ Deviance_i}{\sum_{i=1}^{n} Null\ Deviance_i}$$

(4)

where $i = 1, 2, \ldots n$ is a subject which is district. This $psuedo-R^2_{META}$ is used to explain the proportion of the total within-subject variability. The result of URTI data in Bandung City is presented in Table 1.
Table 1. Pseudo R-square and meta psuedo R-square of the model

| Pseudo R-square | Value |
|-----------------|-------|
| psuedo-R²       | 0.2526|
| psuedo-R²META   | 0.3261|

Table 1 shows that psuedo-R²META is larger than psuedo-R². It means that overall coefficients determination better than without longitudinal structure. Therefore the model must contain the longitudinal structure.

3.2.3 Testing longitudinal structures. Model is used for longitudinal data with count response is generalized linier mixed model (GLMM) [6]. Mathematical model has a form

\[ \log(\mu) = X\beta + \delta \]  

(5)

where \( \delta \) is random component which is longitudinal structure of the data. In general \( \delta \) was assummed has mean zero and different variances. Let we have two models that is Model (1) dan Model (5). There is two possible relationships betwee both model: Model (1) is updated from Model (5) or in ither hand Model (1) is nested in Model (5) dan Model (1) is a different model with Model (5). Comparing these both model using different of deviance

\[ \text{diff} = D_1 - D_2 \]  

(6)

was Chi-square distributed with degree of freedom \( k = p_1 - p_2 \). \( p_i \) is a number of parameter of \( i \)-th model. Signification of the test means Model (5) mode fit then Model (1) [1]. The Result of URTI data in Bandung City is

| Model | Deviance | df | Diff  | df | p-value |
|-------|----------|----|-------|----|---------|
| D1    | 41604    | 149|       |    |         |
| D2    | 40830    | 148| 733.59| 1  | < 2.2 e-16|

According to Table 2 has a very small p-value, so the test is significant. Therefore model (5) is more fit than the model (1), so the model with longitudinal structure is suitable for URTI data.

3.3. Selecting a Statistical Model

Base on the result showed before, there are three possible statistical models can be used to analyze this URTI data in Bandung City. First is the fixed-effect model, is based on exploring the marginal distribution through mean evolution pattern. The second is the random-effect model, which is based on variance structure pattern and correlation structure plot. The third is a multilevel model, is based on exploring subject-specific profiles through measuring the goodness of fit and testing longitudinal structures.

4. Conclusion

Based on the result was describe on the analysis before that URTI data of Bandung city have mean structures pattern decreasing by time, the variance function seems to be relatively unstable, and the correlation is decreasing by increasing distance. Similar scatter plots under main diagonal implies stationarity. Beside that URTI data of Bandung city have intercept and slope are different among the district, the model contains the longitudinal structure and model with longitudinal structures more suitable.
Statistical model will be used for URTI data in Bandung City is model that contain longitudinal structures such as fixed effect model, random effect model or multilevel model.

5. Acknowledgement
This paper is supported by the Basic Research Grant of the Directorate General of the Higher Education, Indonesia (PDUPT, SP DIPA-042.06.1.401516/2018).

References
[1] Agresti A 2002 Categorical Data Analysis 2nd edition (New York: John Wiley & Sons)
[2] Ilk O 2004 Exploring Mean Structure In Longitudinal Data With Graphics: Linked Brushing Approach
[3] Ilk O and Cook D 2011 Graphical Methods for Exploratory Multivariate Longitudinal Data Analysis
[4] Ker H W 2010 Visual-Graphical Methods for Exploring Longitudinal Data International Journal of Computer, Electrical, Automation, Control and Information Engineering 4 2.
[5] Faraway J J 1999 A Graphical Method of Exploring the Mean Structure in Longitudinal Data Analysis Journal of Computational and Graphical Statistics 8 1 pp 60–68.
[6] Verbeke G and Molenberghs G 2000 Linear Mixed Models for Longitudinal Data, Springer-Verlag (New York) chapter 4 pp 31-40
[7] Diggle P J, Heagerty P, Liang K-Y, Zeger S L 2002 Analysis of Longitudinal Data (Oxford: University Press) chapter 3 pp 33–53
[8] Hesketh S R et. al. 2004 Generalized Latent Variables Modelling: Longitudinal and Structural Equation Models (Oxford: University Press) chapter 3 pp 33–53
[9] Snijder T A B, Bosker R J 2012 Multilevel Analysis: An introduction to basic and advance multilevel modelling 2nd edition. (London: SAGE Publications)
[10] Jaya I G N, Tantular B, Zulhanif 2018 Spatial Clustering of Upper Respiratory Tract Infections in Bandung City by Means Local Moran’s I and Spatial Scan Statistics International Journal Of Engineering Sciences & Management 8 4 pp 42-49
[11] Cotton I, Jaspan M, & Rabie. 2004 Management of upper respiratory tract infections in children S Afr Fam Pract 50 2 pp 6-12.
[12] Ratnadewi Y 2015 Waspadai ISPA, Meningkat Kasusnya di Kota Bandung. Article retrieved from Pikiran Rakyat: http://www.pikiran-rakyat.com/bandung-rayar/2015/08/04/337026/waspadai-ispa-meningkat-kasusnya-di-kota-bandung
[13] Faraway J J 1999 Extending the Linear Model with R (Boca Raton: CRC Press) chapter 1 pp 1-10