21, rue d'Artois, F-75008 PARIS
http : //www.cigre.org

CIGRE-148

CIGRÉ Canada
Conference on Power Systems
Vancouver, October 17- 19, 2010

Load Analyzer – a Software Tool for Load Data Analysis

J. CHEN[1], A. LAU[1], W. LI[1], K. WANG[2]
BC Hydro[1] (CAN), Simon Fraser University[2] (CAN)

SUMMARY

This paper presents an overview of the motivation, design, and features of the Load Analyzer, an application software tool for load data analysis at BC Hydro.

The Reliability Decision Management System (RDMS) at BC Hydro consists of data models such as a load curve model, which comprises minute-based metering data of individual loads and hourly load curve data of each delivery point. The data quality in the RDMS is extremely important to BC Hydro. To enhance the data quality and to facilitate the data analysis, a customized software tool is proposed and developed. The main functionalities of this application include *data exploration* and *data manipulation*.

The data exploration focuses on retrieving and examining the observed data. The Load Analyzer provides the following features to users. First, the complicated data structure in the RDMS is well organized so that the data can be selected and retrieved through an intuitive human-machine dialogue. Second, a flexible data visualization tool is integrated, with which the user is able to look into the data in a graphical manner. Third, algorithms are developed for identifying the problematic or suspectable data.

The data manipulation focuses on modification of data according to user's experience or predefined rules. It provides users the flexibility of changing the data directly on the graph or through a program with predefined algorithms. Besides, the Load Analyzer is designed to be highly extensible. When a new algorithm is developed later, it can be easily plugged into the Load Analyzer.

The rest of the paper is organized as follows. In Section 1, the background of this application development is introduced. The underlying motivation and the significance of this Load Analyzer are explained. In Section 2, the functionalities and design of this application as well as some intrinsic algorithms for data analysis are discussed. A simple study case is introduced in Section 3 to explain how to use the Load Analyzer.

KEYWORDS

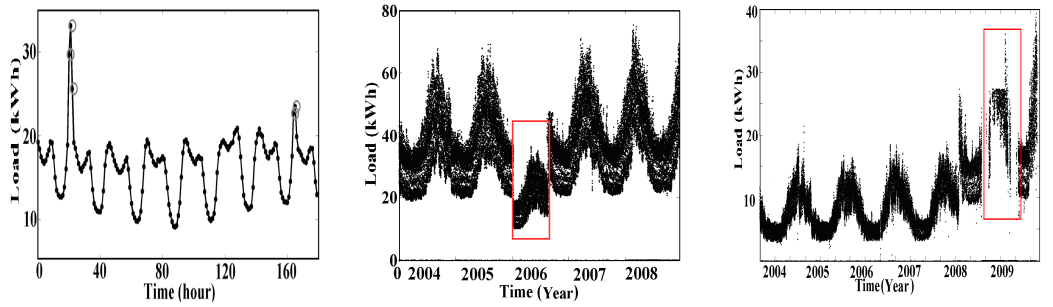
Load data, Data analysis, Time series, Smoothing methods

jiyi.chen@bchydro.com, adriel.lau@bchydro.com, wen.yuan.li@bchydro.com, wangk@cs.sfu.ca

1 Background

The Reliability Decision Management System (RDMS) [1] at BC Hydro consists of data models such as a load curve model, which comprises minute-based metering data of individual loads and hourly load curve data of each delivery point. The quality of the data is crucial in calculating accurate energy-related reliability indices such as the Delivery Point Unreliability Index (DPUI), the Expected Energy Not Supplied (EENS), and the Load Coincident Factors (LCF), which are important component in building power system base cases and in supporting the system state estimation. The load data quality must be assured.

There always exist *corrupted data* and *missing data* in any load data collection system. Corrupted data refer to the ones that significantly deviate from their regular patterns. Missing data are the ones that are not recorded in meters or lost due to various reasons including meter problems, equipment outages, etc. Figure 1 shows some examples of corrupted data and missing data. It is necessary to conduct load data analysis to identify corrupted and missing data and correct them.



(a) Corrupted data 1 (b) Corrupted data 2 (c) Corrupted and missing data

Figure 1 – Examples of corrupted and missing data

1.1 Complexity of Data

In the RDMS database, there are multiple levels of load data, starting from metering, to delivery points, to buses. The data are also categorized into sub-levels such as unfiltered, filtered and adjusted data, each of which indicates the status of data after certain processes. The basic hierarchy of the load data is shown in Figure 2.

In the data model of the database, there are many rules for efficiency or other purposes. Unfortunately, these rules may confuse data users. The metering data include the PI Tag data and EMeter data. The mappings between meters and delivery points and between delivery points and buses are not one-to-one mapping relations. Also, there are multiple terms for delivery point identification, such as *delivery point name*, *delivery point version number* and *delivery point id*. However, data users only know a delivery point name. Besides, there are many meta-data for attributes of meters, delivery points and buses. These rules and meta-data complicate the data model considerably. It is desirable to develop a tool which can automatically retrieve any data from the database.

1.2 Data Visualization and Manipulation Tool

One of the most important steps in load data analysis is to look into load data in a graphical manner. The graph provides much information that cannot be observed by just looking at data values. Although some commercial software tools such as SPSS, Minitab and MS EXCEL can be used, these tools have limitations. First of all, they cannot be used to automatically retrieve data from the database based on users' requirements. Secondly, the plotting and data analysis functionalities in the tools are also limited. For example, MS EXCEL cannot easily perform a natural zoom-in and zoom-out in a plotting, which is often required by users in conducting the data analysis. Moreover, users often want to manipulate the data directly in the graphs, which the commercial software cannot provide.

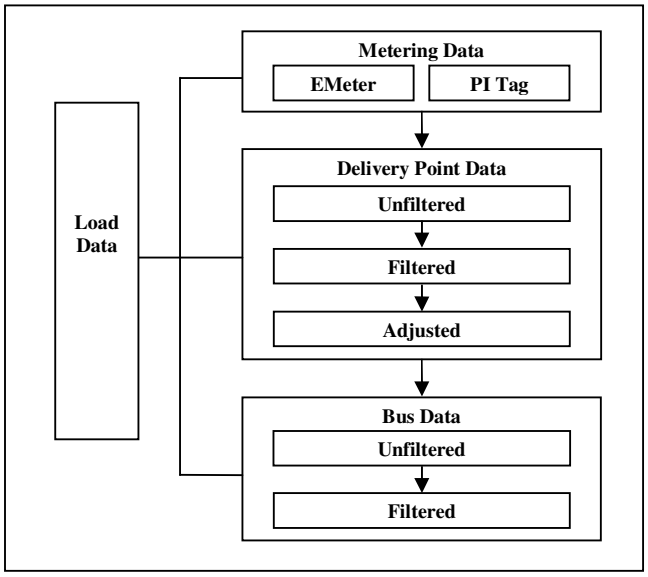


Figure 2 – Data structure in the RDMS

2 The Load Analyzer

In this section, the major features and functionalities of the Load Analyzer are described. The essential goal of the Load Analyzer is to offer the flexible data analysis functionality.

2.1 Data Exploration

To provide the convenience of load data observation, the Load Analyzer organizes the load data, incorporates an interactive visualization tool and some data analysis algorithms.

2.1.1 Data Organization

As mentioned in Subsection 1.1, the conceptual hierarchy of the load data model is simple but its physical organization is relatively complicated. A *business object* [2] is designed to bridge the gap between them. A concept of "Databook" is introduced into the business object. In the "Databook", the three original levels of load data, namely, meters (EMeter and PI Tag), delivery points and buses, are kept. Another hierarchy is the bus data. A group of buses forms a station; a group of stations forms an area; a group of areas forms a region; and a group of regions forms a division. All the information, together with the meta-data, is encapsulated into the object "Databook". A simplified "Databook" object is shown in Figure 3. With this "Databook" object, the Load Analyzer provides the user with an interface to access the database with the minimum effort. The user only needs to specify the names of delivery points (or any other data which they are interested in) to retrieve the load data and other related information.

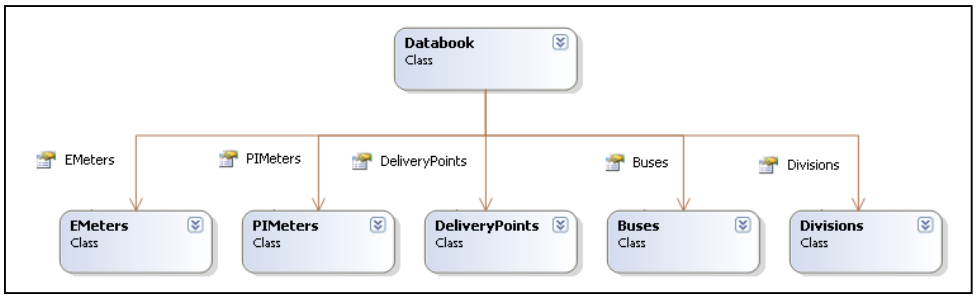


Figure 3 – Databook Object

2.1.2 Visual Inspection

A powerful and flexible visualization tool is very helpful in the process of data analysis. Many problematic data could be easily found by a user when he/she sees the graphs of data. For example, the corrupted data in Figure 1(b) are so obvious that a not-experienced user would be able to identify the problem just by eye observation. This example indicates the possibility of finding corrupted data using regular visualization. However, if the corrupted data are much less obvious, like the "locally corrupted data" described in [3], then "zoom-in and zoom-out" and "panning" functionalities are needed. The Load Analyzer provides powerful visualization functionalities including "zoom-in and zoom-out", "panning", "showing data values", "synchronizing with the data source", "printing", etc. The Load Analyzer also provides direct links among multiple levels and sub-levels of data in a visual manner. A study case in Section 3 demonstrates the visualization in the Load Analyzer.

2.1.3 Statistics

Statistics are often representative indices for the status of the load data. When a user selects a certain time interval, the Load Analyzer shows basic statistics, such as sample mean, sample variance, max and min values, etc. The Load Analyzer is also designed to show some other self-developed indices, such as the *anomaly index*.

Anomaly index is a set of indices under development to represent the healthiness of load data in a selected period. For example, a simple index named *Yearly Average Deviation Index* (YADI) could be used to measure the deviation of load data in a selected period from the yearly average value in the same period of other years. It is assumed that 5 years' data are under consideration, and that the data between *t_start* and *t_end* (*within one year*) is the selected period. The YADI of the time interval (*t_start*, *t_end*) is defined by

$$YADI = \frac{AVG(time\_period)}{AVG(reference)},$$

where the *AVG* is the average value of load data in the selected time period, *time_period* is the time period between *t_start* and *t_end*, and *reference* is the time interval used for reference. In this case, it is the same time period between *t_start* and *t_end* in other years of the 5 years under consideration.

2.1.4 Identification of Corrupted Data

The goal of exploring the data is to check if the data are in good quality or not. Although visualization provides valuable intuitive information to users, it cannot systematically identify and report corrupted data areas. To assist users in making better judgment about the healthiness of data, some algorithms have been developed while others are still under development.

Currently, the nonparametric smoothing methods with confidence intervals have been developed to identify two types of corrupted data [3], namely *locally corrupted data* and *globally corrupted data*. The locally corrupted data is defined as the data deviating significantly from the local trends and the globally corrupted data is defined as the data deviating significantly from the global trends. The B-Spline smoothing [4] and Kernel smoothing [5] are used to model the trends of the load curve. A smoothing parameter is utilized to control the smoothness of the modeled curve and modeled trends (local or global). The confidence intervals are created for user's reference. The data outside the confidence intervals are identified as corrupted data as shown in Figure 4.

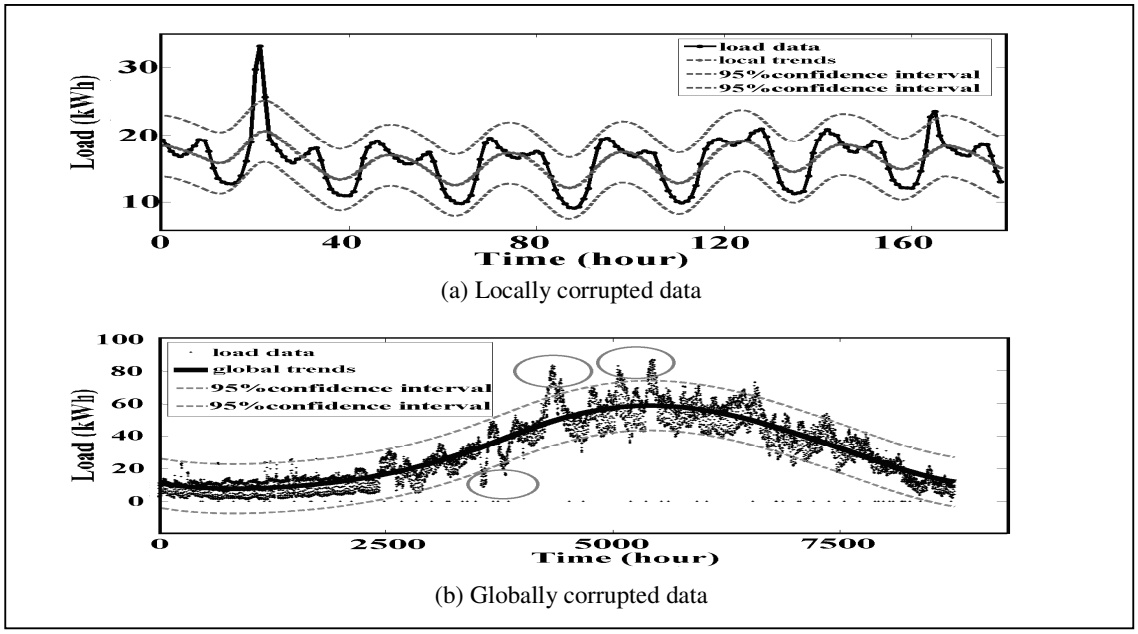
2.2 Data Manipulation

After data exploration, users most likely want to modify corrupted data or any data in question. The data manipulation functionality is important to users. The idea of manipulating data includes two aspects: either modifying some data manually or using a program.

2.2.1 Manual Manipulation – Modification on the Graph

Traditionally, when a user finds any problematic data in the graph, he/she has to trace back to the database from the graph, and then modify the data manually. This method can work if only a few data need to be modified. The Load Analyzer also allows users to modify any data directly on the graph.

For example, as shown in Figure 5, if all the data in the rectangle is considered problematic, a user can use the mouse to select that part, and drag it vertically to an appropriate place. The changed data in the graph are automatically saved in the database.



(a) Locally corrupted data

(b) Globally corrupted data

Figure 4 – Example of locally and globally corrupted data

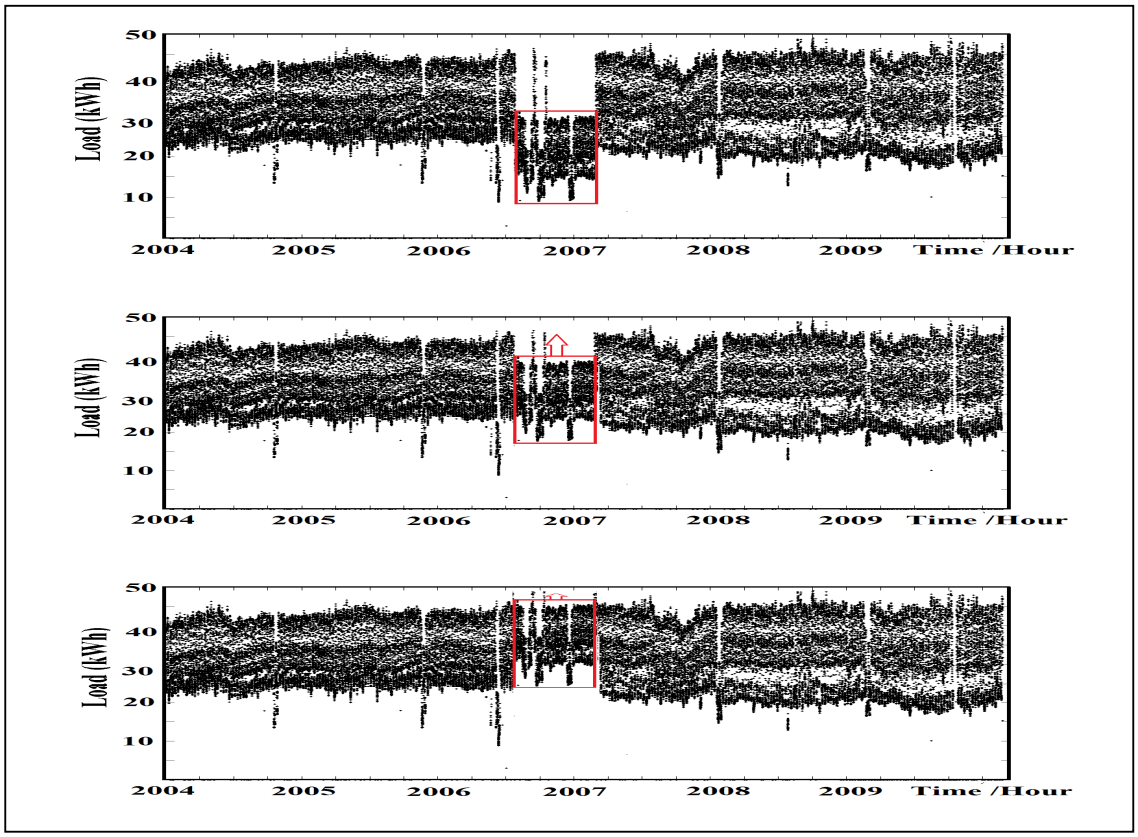


Figure 5 – Modification on the graph

2.2.2 Programmatic Manipulation – Corrupted Data Correction

Manual manipulation on the graph is handy and intuitive. However, it is labor intensive and lack of rigorous measures. Moreover, when there are missing data, some effective algorithms are needed to fill in the holes (missing data).

Currently, a multiplicative model based hole-filling algorithm has been developed. In this algorithm, the load data can be decomposed into three components: long-term trend factor, annual seasonality (periodicity) factor and irregularity factor. Only the long-term trend factor and annual seasonality are considered currently. The long-term trend are modeled as global trends as shown in Figure 4(b). The B-Spline smoothing and Kernel smoothing are typically robust even when there are relatively huge holes in the load data. The seasonality factor is the deviation of the load data from the long term trends. The seasonality factor of the data in the holes can be estimated using the average of the seasonality factors during the same time period in the previous year and the next year. Thus, with the long-term trend and seasonality factors, the load data in the holes can be estimated. Experiments have been done on most of the delivery point data and the results are very good. Figure 6 shows an example for the effectiveness of this method.

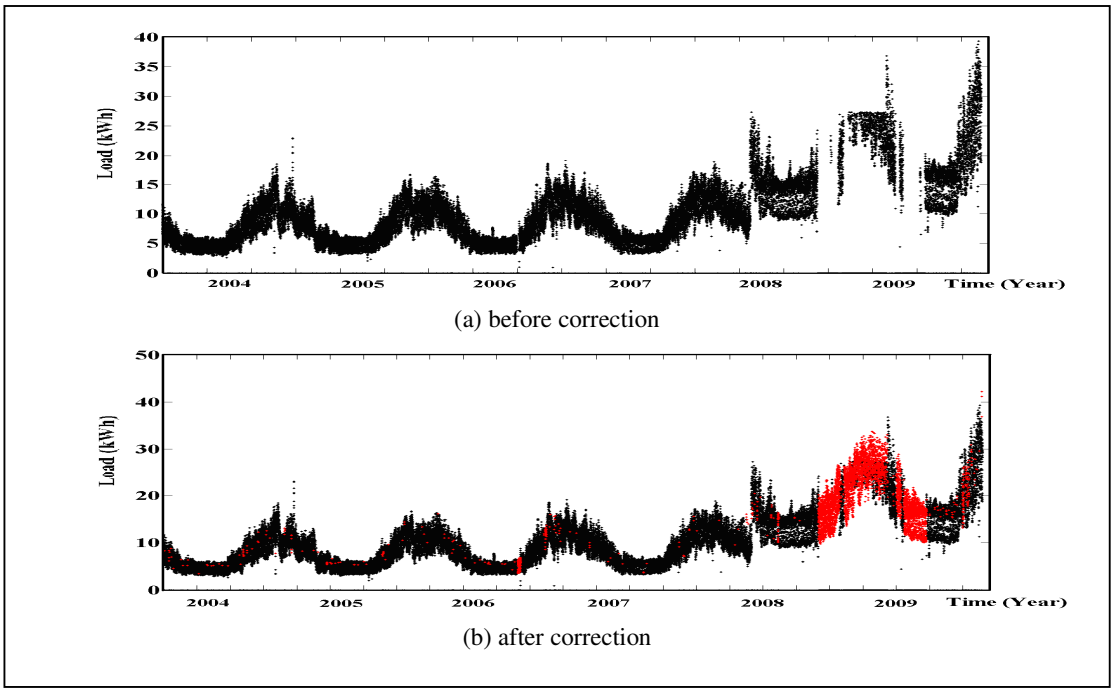


Figure 6 – Hole filling result

3 A Study Case

In this section, a simple example of conducting the data exploration using the Load Analyzer is illustrated. Assume that the delivery point named "DELIVERY POINT1" in the BC Hydro RDMS database between fiscal year 2005 and fiscal year 2009 is interested.

The whole data hierarchy is constructed as a tree structure and listed on the left-hand side of the Load Analyzer interface. To retrieve the data, the user only needs to

(1) Select the delivery point (or other data) name in the tree-view on the left-hand side;

(2) Choose the starting date and ending date on the main screen;

(3) Choose the "sub-level (unfiltered, filtered or adjusted)" the user is interested in;

(4) Click "Add" button on the main screen.

Then, the data of "DELIVERY POINT1" will be loaded into the data grid on the right-hand side and plotted on the main screen. The "unfiltered" data and the graph are shown in Figure 7. The related information such as the delivery point class, delivery point type, delivery point forecasts, and delivery point mapped meters and buses can be seen when the related tabs on the right bar are clicked, as shown in Figure 7.

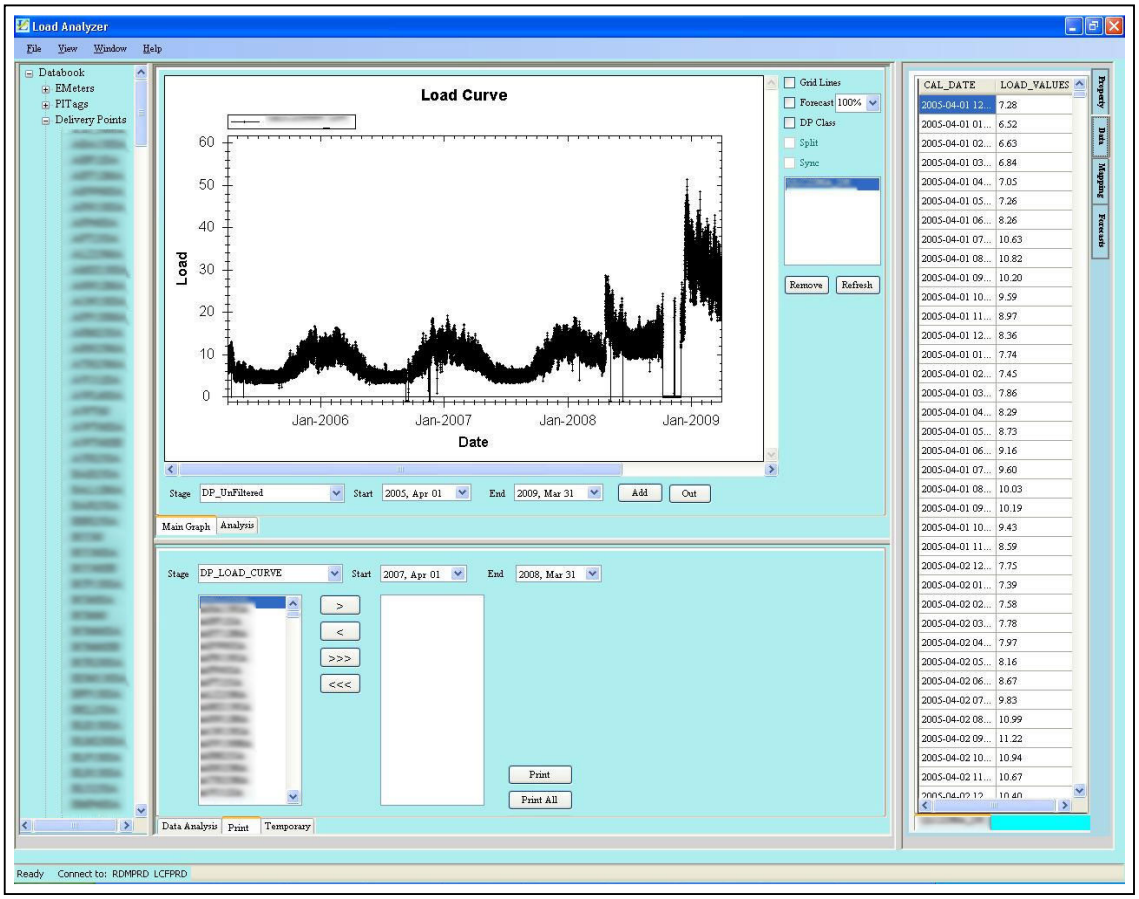


Figure 7 – Basic data retrieval and plotting of "DELIVERY POINT1"

Figure 7 shows the data before filtering and adjusting. It is obvious that there is a relatively long interval of missing data around Jan 2009. What do the data after filtering and after adjusting look like? To see the comparison, all a user needs to do is to follow the steps given above except selecting the sublevel "DP Filtered" or "DP Adjusted" instead of "DP Unfiltered". The results are shown in Figure 8.

In Figure 8, the top load curve shows the "unfiltered" data, the middle one shows the "filtered" data and bottom one shows the "adjusted" data. The bold horizontal lines in all the graphs represent the load peak forecasts.

After "filtering" and "adjusting", the holes have been filled. However, from the graph comparison, it can be seen that the result data are still not that satisfactory. The data since Jan 2009 changed a lot comparing to the data before "filtering" and "adjusting". Moreover, the trends of these data do not follow the previous years' trends. Thus, the judgement can be made that the data of this delivery point needs further analysis and adjustments.

This simple data exploration study shows how easily the problematic data can be found and how the load analyzer can assist the user to conduct data analysis.

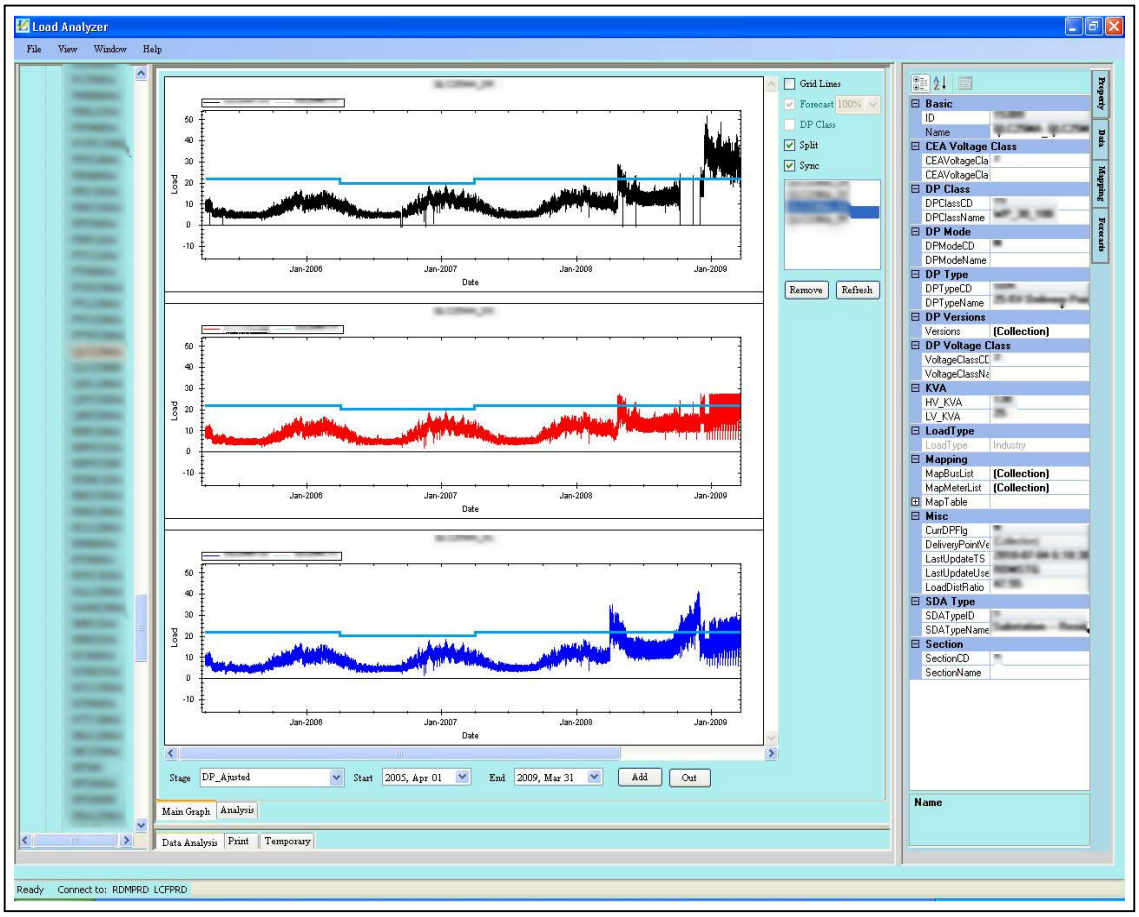


Figure 8 – Comparison among different sub-levels of "DELIVERY POINT1". The three sub-levels are "unfiltered", "filtered" and "adjusted"

4 Conclusion

This paper presents an overview of the motivation, design, and features of the Load Analyzer, an application software tool for load data analysis at BC Hydro. The Load Analyzer performs re-organization of load data in the RDMS, and provides flexible exploration and manipulation functionalities, which significantly enhances the load data analysis to ensure the load data quality at BC Hydro.

BIBLIOGRAPHY

[1] Li, W., Jonas, H. C., Yan, S., Corns, B., Choudhury, P., and Vaahedi, E. Reliability Decision Management System: Experiences at BCTC. Electrical and Computer Engineering, 2007. CCECE 2007. pp.409-412, 22-26 April 2007.

[2] Lhotka, R. Expert C# 2005 Business Objects, Second Ed. Apress, 2006.

[3] Chen, J., Li, W., Lau, A., Cao, J., and Wang, K. Automated Load Curve Data Cleansing in Power Systems. To appear in IEEE Transactions on Smart Grids, 2010.

[4] Ramsay, James O. and Silverman, Bernard W.. Functional Data Analysis, Second Edition. Springer, 2005.

[5] Wand, M. P. and Jones, M. C.. Kernel Smoothing. Chapman & Hall, 1995.