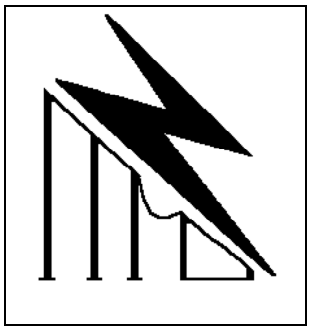
THE NRS LR PROJECT:
NOTE ON DATA HANDLING & PROCESSING

97/11/12
Ref.: 71112lrp



M Dekenah
Marcus Dekenah Consulting cc
(012) 6671347
25 Maluti Ave
Doringkloof
0157

CONTENTS

1. INTRODUCTION

The purpose of this document is to describe data errors encountered during the course of the NRS LR project, and present tactics developed to prevent the errors getting onto the database, and circumvent any systematic residual errors on the database.

Data-handling processes affect database-operator errors. Data-base handling processes will also be described towards this end.

All tactics were concieved by the author and Mr. M. Berchowitz of TLC Software CC.

This document is important for two reasons:

- Data processing methods are central to the aim of the NRS project, and must therefore be transparent.
- Follow-on work by other researchers must be enabled. This document is the first attempt to provide guidance for this.
- The author wishes to stimulate comment in this area.

2. OVERVIEW OF DATA PROCESSING

Data is collected in the field from loggers. The data is normally loaded onto a lap-top, transferred to the project manager and later loaded onto the database.

After this, the data is analysed. The process is shown in the figure below. The analysis phase of the figure has been expanded somewhat to show derivation of important parameters.

There are some pitfalls to analysis of this load data. It was recently learnt that Eskom TRI has recorded about 120 “problems” associated with data from loggers.

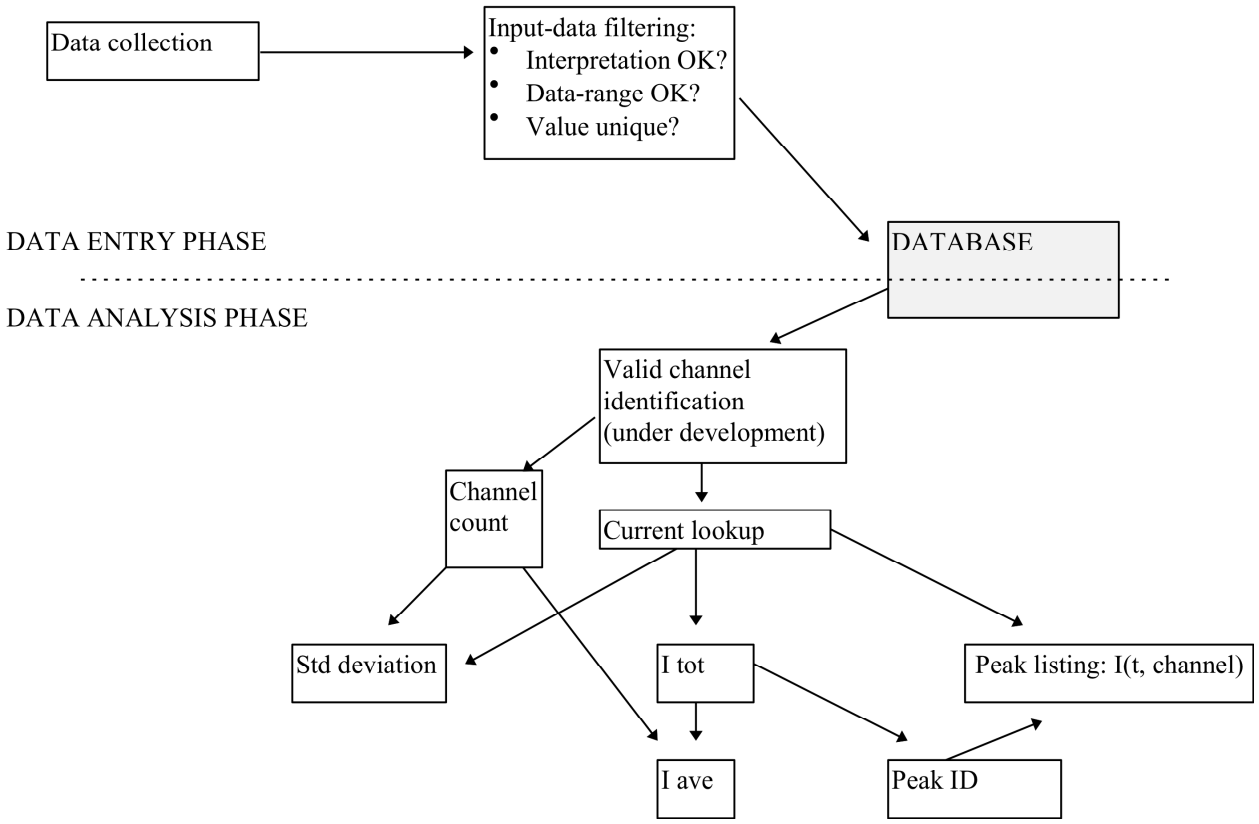


Figure 2: Data processing applied in NRS LR Project

3. CLASSIFICATION OF ERRORS BY SOURCE

Data errors arise from several generic sources:

- Data handling & data transfer
- Data logger handling (software & hardware)
- Data logger operating software (design ?)

3.1 Data handling & transfer

Corruption of source-data normally occurs for several reasons.

3.1.1 Data handling:- Corruption of file-format

The file-format has been changed. Since file-filters need to recognise some key characters (for synchronisation), they can be "tripped up" by this.

This may happen by someone viewing the file and saving in another format with the same file-name & extension.
It is for this reason that the "DATAVIEW" package was developed.

3.1.2 Data transfer:- Electrical noise

Electrical noise introduces spurious characters into the data-stream during downloading of data to the lap-top. This can change both formatting and data characters, and make it very difficult to tell if two adjacent fields are separate.

Normally “FF’s” are inserted into the data fields, and this is a sign of a communication-cable problem. The normal cause of comms-cable problems are:

- Comms cable-length too long
- Connection/continuity not satisfactory

The following areas may be corrupted:

1. Logger ID code
2. File header
3. Date/time stamps
4. Data values
5. All separator characters

Corrupted characters may be graphic, invisible, or numeric or alphabetical. Therefore a decent system is needed to isolate errors.

3.2 Logger handling:- Software-initialisation

Loggers may be set-up incorrectly, particularly in the following areas:

- Logger ID incorrect or modified to incorrect value
- Date & time for real-time clock
- Logger averaging interval
- Number of active channels

It is important to systematically reduce this risk.

3.3 Logger handling:- Hardware-initialisation

Hardware set-up problems are physical.

- A measurement channel may fail in some way. Normally channels “fail” to either a high value, a low value, or “float”.
- A measurement channel may be disconnected, either accidentally (i.e. by disconnection of plug) or intentionally (systematically). Normally accidental disconnection of a channel results from closing the door on a logger-cubicle, straining the connections. Systematic disconnection results when the logger channels are not connected sequentially, from 1-7, but are connected with “holes” in-between.

3.4 Logger-intrinsic problem

Data collected from some loggers displays a data-repetition problem. In this case the data is not strictly corrupted, but the same data is downloaded several times

in a single download. Days of data may be missing at the same time, whilst the expected volume of data collected during the session may be correct.

This problem has been noted at one site where the logger memory-size is 64 KB. The problem is random in nature, and is still being investigated.

4. ERROR-CHECKING: DATA-ENTRY PHASE

Several techniques are applied in order to minimise the effect of all of the above errors.
The following concepts are applied:

1. *Organise database application to reduce operator-errors.*
2. *Maintain traceable data-input process .*
3. *Reject most bad-input data by design.*
4. *Analyse data to minimise effect of residual errors (covered in section 5).*

4.1 Organisation of database

The data base is organised with the load-readings of a unique data-logger all inserted into its own database-table.

A load-monitoring project is thus represented by a directory containing the same number of database tables as there are loggers in a township.

All logger tables are keyed on date/time, and channel number.
Thus it is ensured that each reading is unique, and may be inserted into the logger table once only[1].

4.2 Routing of incoming data

Data-storage is organised so that location of a project (i.e. what directory is associated with a particular township-name and year)[2].

This is used to direct new data loaded into the database to the correct directory and database table, minimising systematic errors associated with routing of data.

A list of logger-ID against table-name is maintained for each township. Each time data from a logger is loaded, the logger-ID is checked against the list. If the logger-ID is new, a new table is created, and the data is directed into it. The new Logger-ID and table-name is then added to the list.

4.3 Database: input-operations logging

All “add” type operations on a particular logger-table are noted in a unique log-file (1 per logger data-table), which records the following:

[1] A second attempt causes a “key violation”, and the operation fails.
[2] This association is set up only when a new township is created.

1. Name of data source-file
2. Progress information
3. Report on nature & type of errors encountered (if any)

This allows the make-up of any logger-table to be traced back to a particular source-file (which mostly have unique names).

Errors which are logged fall into the following categories:
- Interpretation (of source file)
- Range
- Insertion (into database table)

The input filtering process happens in 3 stages as shown in figure 1.

4.3.1 Interpretation-errors

Interpretation errors are logged when the input-filtering software cannot decipher the source-file data-line which has been presented to it (i.e. the result of corruption).

Normally this is a result of insertion of non-numeric characters into the data on one source-file line.

In all such cases the particular data-line is discarded (i.e. all data values at that date-time ignored), and the line number and error type is written to the log-file.

4.3.2 Range-errors

This error is encountered if any values interpreted are outside a predetermined range. This normally represents the case where data has been corrupted, but only with numerical characters.

The following ranges are normally used for range checking:
Date-time: Year in range 1991-2000
Data-values: Current in range 0-100 Amp

All lines of data containing range-errors are discarded. The occurrence of any error is logged with its source-file line-number.

In the case where a date range-error is encountered, the entire data for the day is discarded (for that logger).

4.3.3 Insertion-errors: key violations

Key violations occur when attempting to write a particular date/time stamped data-point to a data-table more than once.

The data point is discarded and the error is logged with its source-file line-number.

5. CATEGORISATION OF RESIDUAL DATA ERRORS

The process described above filters a good proportion of "bad" data out of the system, controls database operator error and allows error-tracking.

It is recognised that the following errors *do* get onto the database at present[3].

Table 1: Data-table errors

Error type #	PROBLEM	DESCRIPTION
1	a) Channel-failure [high] b) Channel-failure [low]	Channel fails to a high or low value which is relatively constant. A hardware failure which progresses to a final value over 1-2 months. Only the failure-to-high case has been identified in NRS LR data. Failure-to-low is hypothesised.
2	Empty channel	Logger is picking up data from a channel which is not connected. The logger has been incorrectly initialised.
3	Floating channel	Channel which fails to a random value which is random over time. Consumer load is not correlated to group-load. This error has not yet been identified in NRS LR data, but has been identified by ESKOM TRI.

Incidence of errors

The incidence of type 1a errors is not very high[4].
Type 1b errors have not been noted yet. However, analysis methods presently applied handle type 1b and type 2 errors equally well.
Type 2 errors are noted in most projects at some time[5].

It is not possible to distinguish type 1b and type 2 errors from the situation where a consumer is connected but never takes any load at all[6].

[3] Software algorithms are refined each year based upon the most recent period of learning. The structure of load-data storage will however not be modified.
[4] Four type 1a logger-errors have been noted over the entire period of data-collection. All of these occurred at Durban projects in 1997. This is roughly 0.3% of the sample.
[5] This is a problem we are trying to overcome with training. Normally 8% of sample or less.
[6] The electrification "nightmare", which may be more common in poor areas and is very dependant upon reigning connection policy.

The following sections describe a method developed to deal with type 1 and type 2 errors, presently applied in the NRS LR project analysis.

Alternatives are also suggested.

No method has yet been developed to deal with type 3 errors.

6. ERROR CHECKING: DATA ANALYSIS PHASE

The following sections deal with the process, its outputs, control & correction.

6.1 Overview of analysis process

Figure 1 shows how LR data is analysed to extract peak demand information once data is in the database.

A “township” load-time profile is constructed, which represents the group-load of the community from the very beginning to the very end of the analysis period.

For each time-mark in the profile, the scalar-sum (I tot) of all the loads and sums-of-squares of all loads (and number of unique readings collected at that instant (*i.e. the channel-count*)) is extracted into a list.

The list is sorted and the five highest instants (Peak ID) of group-load are noted.

For each of these instants, the Admd (I ave) & standard deviation of the individual load is calculated and presented.

Because the instants of highest group load have been noted, it is a relatively trivial process to look up those instants and obtain a breakdown of individual loads at any of the peak-periods.

6.2 Outputs of analysis

The analysis described makes the following information available over the period of consideration:

- Group-load (I tot) vs. Time profile
- Average load (I ave) vs. Time profile
- Standard deviation vs. Time profile
- No. channels (N) vs. Time profile
- Distribution of individual consumer loads (any time instant)
- % Average load vs. % time breakdown

All of this information is available for export to other researchers, but at present these outputs include type 1b & type 2 errors.

6.3 Identification of valid channels

Channels are at present assessed by channel-counting which takes place at three levels:

Documented sample size: This is the documented size of a project in terms of data returned from site documentation. It returns $\mathbf{N_{max}}$, the upper bound of number of channels monitored.

Channel count: This is the number of channels for which readings are found on this database at a particular time instant ($\mathbf{N_t}$). No lower-threshold is applied in the extraction of this data.

Active channel-count: The assessed number of channels which are actually active during the period of inspection (**N**).

A channel is defined as "active" if the current over the period of assessment ever exceeds 0.2 A (about 46 Watt).
The "period of assessment" is the period included between two successive downloads which captures the *instant* of interest[7].

6.4 Identification of inadequate project control

In a properly controlled project, $\mathbf{N_t <= N_{max}}$ and $\mathbf{N < N_{max}}$

$\mathbf{N_t}$ should very nearly equal $\mathbf{N_{max}}$ in "rich" areas.

All other situations represent either poor control, or a general power failure[8].

6.4.1 Identification of poor logger initialisation

If $\mathbf{N_t > N_{max}}$ then a project is probably not properly initialised.
Extra channels are initialised but logging no data.

6.4.2 Identification of poor site documentation

In a project with poor documentation, more active channels are detected than are documented. Thus $\mathbf{N > N_{max}}$.

6.5 Peak-event data-processing applied in NRS LR project reports

The following procedure is applied to extract peak-event data for NRS LR reports:

1. The number of loggers in a township is checked against the site documentation to ensure that each database-table is valid[9].

[7] This period is chosen because loggers normally malfunction due to human intervention, typically at a download.
[8] A general power-failure is represented by the situation where $N_t << N$.
[9] This problem can arise when a logger ID is tampered or corrputed. Facilities have been developed to list each logger,its associated ID and its volume of data, for a township.

2. A “township” load profile is created for the project in question.
3. The Std deviation vs. Time trace is inspected for type 1a errors. If found, the offending channel-data is removed from the relevant logger table over the period of the fault. A type 1a error is identified by a Std deviation which is large and fairly constant over time.
4. The date & time of the five highest peaks are noted.
5. The $\mathbf{N_t}$ vs. Time graph is inspected to detect whether any peak was the result of a reinstated power failure. If so the value is discarded.
6. $\mathbf{N_t}$ is compared to $\mathbf{N_{max}}$ to determined if the project displayed any control problem. If so, the mean & standard deviation at the peak are re-assessed based upon the lowest of **N** or $\mathbf{N_t}$.

Movements noted from this adjustment are normally less than 8% of the final Admd.

This procedure is time-consuming and is being refined as described in the following section.

6.6 Development in progress

A “valid-channel identifier” is being developed to establish **N**, on a per-channel basis[10]. The validator then presents the findings and allows the selection of channels which must be *excluded* from steps 2-6 above.

The result of this two-stage process will force $\mathbf{N_t}$ to be a subset **N**, resulting in exclusion of most *systematic* type 1b and type 2 errors[11]. All analysis outputs described in section 6.2 will be essentially correct.

The new approach has not yet been applied.

[10] The process is described in section 6.3, last paragraph.
[11] The situation where a logger does collect “empty” data for a short while because the instramentation plug is disconnected is *still* a problem.

7. CONCLUSIONS

GIGO (Garbage-in, garbage-out) is a recognised problem. Sustained development effort, based upon latest feedback is required to ensure data quality.

A primary process has been implemented to reduce most random errors in the source data, before insertion onto the database. A secondary process has been implemented to reduce effects of recognised errors in data on the data-base.

A refinement is in development which will enhance the secondary process further.

Processed data utilising any refinements will always be available to other researchers.