125 years of data

EEA Conference 2013
SKYCITY Convention Centre, Auckland, New Zealand, 19–21 June 2012

David Hume
Electricity Authority, Level 7, 2 Hunter Street, Wellington
Email: david.hume@ea.govt.nz

***Abstract*—**
The power industry today generates vast volumes of data. Much of this data is hidden in complex relational databases and data warehouses such as those used in the Electricity Authority's Centralised Data Set (CDS), and Data Warehouse. Other data is available online but more often than not can be difficult to interpret, prepare, and analyse.
This paper visually explores a small and interesting selection of this industry data; from the growth in total New Zealand generation nameplate capacity (with beginnings at Reefton in 1888), to the electricity consumed every six seconds at the author's home in Wellington. From real-time spot market price data, to past historical transmission outage data, to the movements in future planned generation outage data.
The paper also discusses modern techniques utilising open source software to help aid the preparation and presentation of data.

I. Introduction

This paper provides results and analysis from an erratic smorgasbord of both well known, and lesser known data-series within the NZ electrical power system/market. Most of these data series are either openly available on-line, or through the published Centralised Data Set (CDS). As part of the Market Performance team at the Electricity Authority, the author has been fortunate enough to be involved with some of the data wrangling required to get some of these data sets to "conform" before analysis and then visualisation. The process often leads to further questions like, why? and how?, leading to yet further analysis and visualisation!
In line with the theme of the conference (but out by two years...) the paper begins with a very small data set representing the growth of generation name-plate capacity in New Zealand. This series begins, not 125 years ago at Reefton, but 127 years ago at a private gold field near Queenstown.
Jumping ahead to the modern day, Section 2 looks at the electrical demand of the authors home in Wellington. This data is recorded every 6 seconds to a database running on a Raspberry Pi computer. This dataset represents, perhaps, a small part of what future "Big Data" may look like on the residential demand side of the industry.
Section 3, explores a small sample set of energy price data including real-time 5 minute prices and briefly illustrates the differences in price series used within the market.
Using data openly published on the Centralised Data Set, Section 4 explores, and attempts to identify any trends present in forced transmission outage data.
Section 5 briefly explores the Planned Outage Co-Ordination Process (or POCP). This database, run by the System Operator is used by Transpower (as asset owner), and most Generators to publicise planned outages. Finally, Section 6 provides a brief over-view of the open source software used by the author for all analysis and presentation used in this paper.

II. Growth in New Zealand generation capacity since 1886

. . . the powerful light of the arc-lamp burst forth, like a flash of a mighty meteor, a murmur of admiration rose from the spectators, and there was an immediate scampering of feet towards the screen of the display. As on the former occasion, the light throbbed a good deal, but at its maximum brilliancy illuminated the town over a very wide area, with its cold, cheerless, phosphorescent rays. . . But if the arc-light was an attraction outside, the interior of the Oddfello's Hall was infinitely more so. Rows of lamps were suspended fromn the building, encased in a variety of fantastically shaped shades of different colours, and the whole scene was one of stricken splendour. It was indeed a "Hall of dazzling light. . . "

Inangahua Herald, 6/9/1888

Over the last several years Nicky McClean from the Electricity Authority has put together a very small, but interesting hisotrial record of electricty generation in New Zealand. The series, though likely not fully complete (especially in the early period) still gives a good indication of the growth in generation capacity in New Zealand throughout the 20th Century.
The data includes the estimated nameplate capacity, approximate commissioning and decommissioning dates as well as useful comments. Sometimes a start date has a precision down to the hour, at other times it is down to just the year (in which case it is assumed to be commissioned/decommissioned at the start of the year). When generation is known to have an end date, or has been temporary decommissioned, a negative of the nameplate value is used, allowing a computer program to correctly sum up total nameplate capacity over time.
The raw data file used in this paper is publically available as a simple ASCII text file and provided as part of the Centralised Data Set (CDS). As some care has been taken over the format of the file, it is able to be read (parsed) with a simple computer program. In this case, a small python script has been used to parse the data into memory from where the data can then be examined for further analysis and visualisation (see Figure 1).
We see that perhaps the first recorded use of electricity in New Zealand was 127 years ago on the 3rd of February, 1886, at the private Bullendale gold mine (two years before the public power supply started at Reefton). Interestingly, the Bullendale scheme was DC and included New Zealand's first DC transmission link from the power house to where the power was required at a nearby quartz stamping battery. Two years later, on the 1st of August, 1888, Reefton demonstated the first public supply of electricity in New Zealand. And the rest, as they say, is history. . .

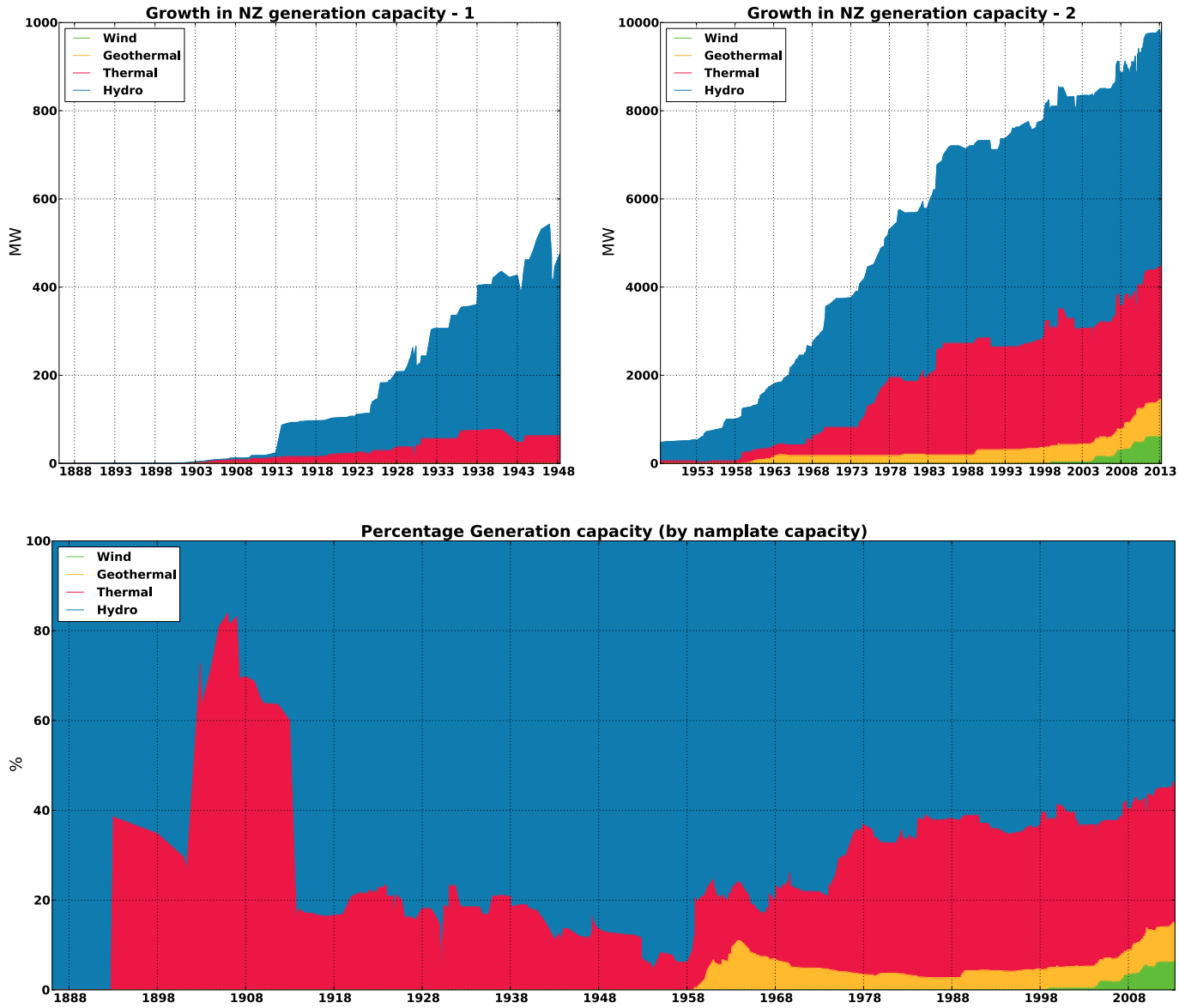


Fig. 1. 125 years of data illustrating the approximate historical generation capacity growth in New Zealand.

III. Home energy monitoring

In complete contrast to the previous section we jump 127 years to the present day and the data, collected at the authors home, of total house-hold electrical consumption. In this case, around 5 months of data has been recorded at a 6 second resolution into a database containing over 1.7 million rows, and growing. This is a single reading for a single household. The increase of smart meters in New Zealand means many such datasets must exist. Through analysis of the data, holders of such data should be able to benefit from its use. For example, one could imagine that in the future, individual house-hold data such as this could be used to help tailor individual customer tariff plans? Perhaps this could already be happening to some extent?

Even without a smart meter, it is relatively straight-forward (with the aid of various internet blogs) for an electrical engineer to set up a low cost house-hold energy monitoring system.

The system described here consists of:

1) a low cost computer (a small Linux based computer called a Raspberry Pi); and,
2) an off-the-shelf energy monitoring kit (in this case, a Current Cost energy monitor (+ USB data cable) purchased on TradeMe for $90.

Every 6 seconds the Current Cost logs the energy use at the mains switch-board of the authors home. This is recorded to a MySQL database running on the Raspberry Pi, from which the data can be downloaded and then analysed.

The figure below illustrates a typical day of electricity usage of the authors house-hold. Our house-hold consists of 2 adults, and two young children and so should represent the energy usage of a typical young family living in New Zealand. Other characteristics of the house-hold include:

1) electric hot water;
2) solid fuel (wood) burner;
3) electric oven;
4) gas hob;
5) one fridge-freezer and one small chest freezer;
6) house is located in the wind in Wellington, 230m above sea-level.

During the evening period the blips have the characteristic of an induction motor switching on. In several instances the inductive inrush current can be observed as a spike. These

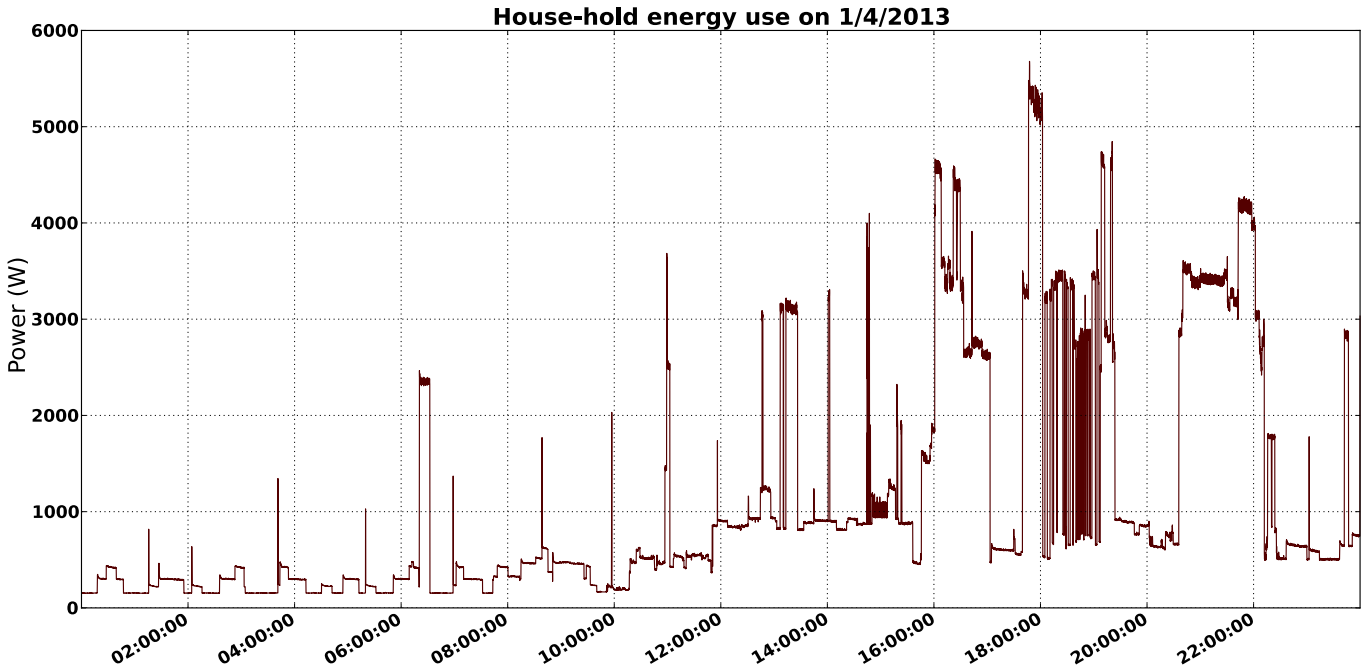


Fig. 2. Typical house-hold electrical energy use.

blips are caused by the fridge and freezer loads. On this day it appears the washing machine was switched on just prior to the 3pm School pickup, the kids had a bath before dinner and the oven was turned on just prior to 6pm to cook a lovely quiche!

In terms of the demand profile, it is interesting to observe how our demand compares with the average demand for our region. This sort of analysis could help retailers in the future profile customers demand patterns and use this information to inform appropriate tariffs or for other competitive advantages perhaps. Figure X illustrates the mean daily profile of the authors home, compared with the mead daily profile of the closest GXP (Wilton) which supplies the region.

One could deduce from this that the occupants of this home tend to sleep in! The morning peak electricity usage has a clear lag of 1 hour compared with the regional demand (the Wilton 33kV GXP). Additionally, there is a clear after school/dinner peak. The author is a little unclear of the cause of the bump at 10pm. This may be caused by hot-water and running the dishwasher.

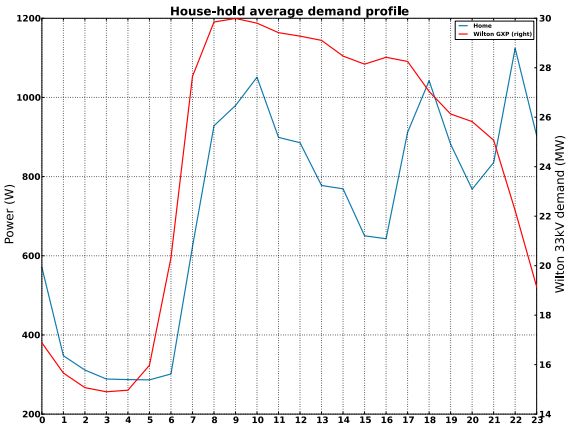


Fig. 3. Average daily demand profiles (between November 2012 to April, 2013) of the author's home vs Wilton GXP in Wellington.

IV. Real-time spot market price monitoring

The market performace team at the Electricity Authority has a role to play in monitoring the market. As a tiny part of this, an automatic realtime price monitor system has been set up using the python programming language. Currently, every 5 minutes the system does the following:

1) FTPs the WITS FTP server and downloads the most recent 5 minute prices;
2) parses the data into memory;
3) adds the data to an ad-hoc database;
4) generates various statistics, including max/min gxps and average regional prices;
5) interactively plots price trends at all GXPs for the last week using a javascript enabled web-browser;
6) sends SMS alert text messages under certain conditions.

The figure below illustrates the price monitor in action, in this case prices in the Hawkes Bay region for the Week ending 11am Thursday the 11th of April, 2013 are shown. The picture uses a concept called horizon charts which are able, with the help of contrast, to plot time-series data in a confined space (footnote: this uses a Javascript library call d3.js along with a module called cubism.js developed by Mike Bostock and freely avialble on-line.). This tpye of visualisation is ideal for monitoring large data sets. In the case of the NZ electricty market, all prices over the 250+ GXP's for the past week can be inspected within a matter of seconds. The bands of decreasing constrast indicated increasing prices, until a threshold is reached where the colour switches from shades of green to red. Here, at 7:10pm on the Monday the 5th of April, prices reached around $400/MWh at Fernhill in Hawkes Bay (FHL0331). This was likely caused by a constraint on the network in the region, perhaps in this case the interconnecting transformers at Redcliff. Clearly, a similar transmission constraint bound, causing high prices on the Wednesday afternoon in the same region.

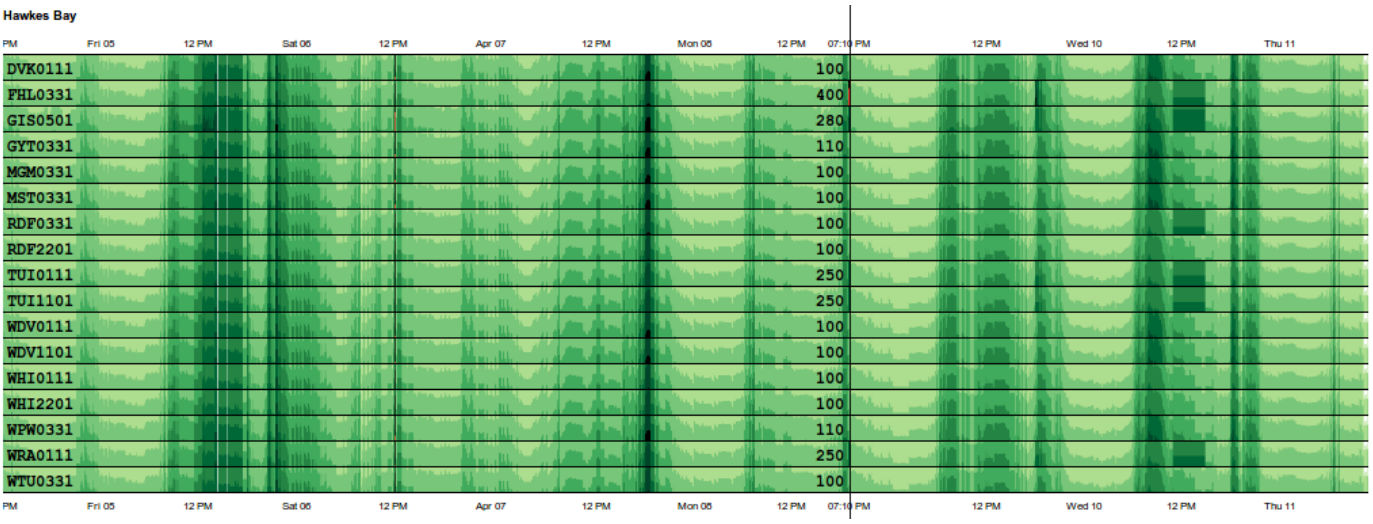


Fig. 4. Example of the difference that can be observed it different foward looking and backward looking price series in the NZ electricity market.

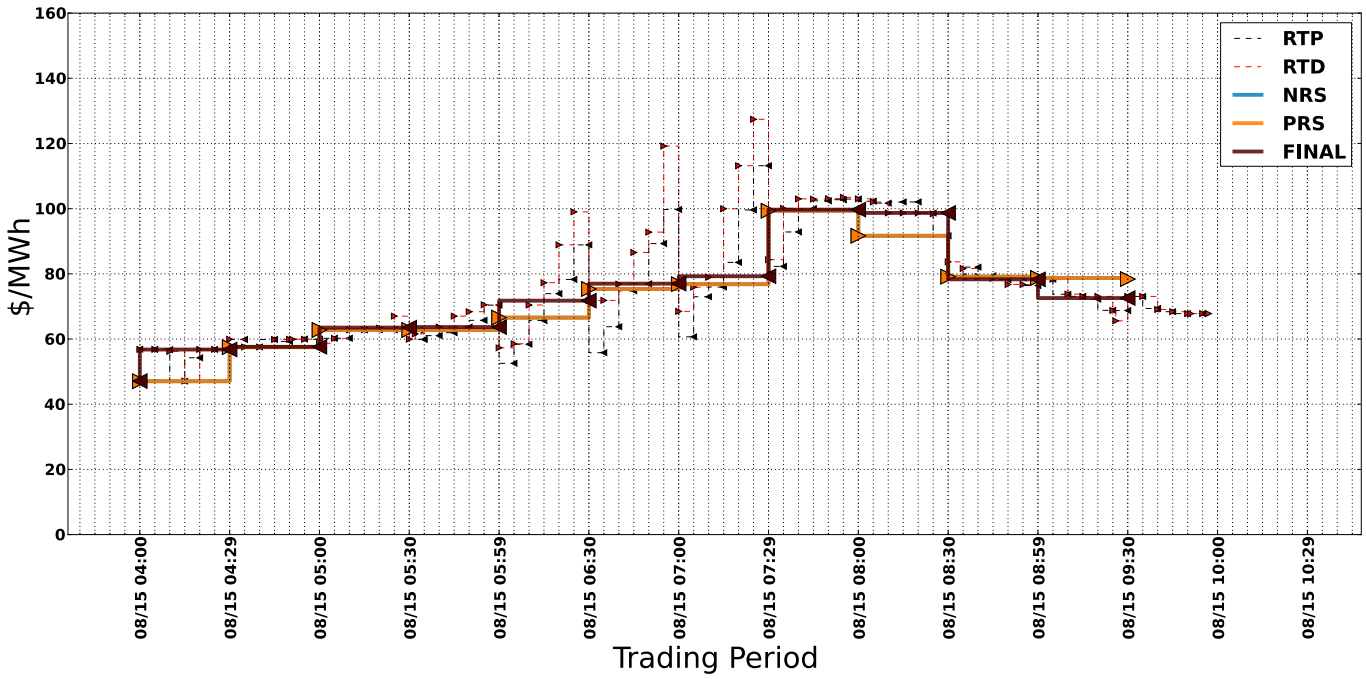


Fig. 5. Example of the difference that can be observed it different foward looking and backward looking price series in the NZ electricity market.

Although only one region is illustrated, all regions are available for visualisation.

There are currently issues with real-time prices, not-the-least of which is due to differences that can occur in realtime prices verses final prices. These differences may be due to a number of factors including inaccurate demand forecasts, generator ramp rates, intermittant generation, among others. Figure X below illustrates an exmaple of the difference in prices seen in the market over different time frames. The Non-Responsive and Price-Responsive Schedules (NRS/PRS) indicate the final scheduled prices that were indicated before the trading period started, i.e, they are forward looking or ex-ante prices. These schedules, along with final prices (which are backward looking, ex-post) are calculated every half-hour, or trading period. Additionally, there are two 5 minute price series. The 5 minute ex-post "real-time" price series which is published every 5 minutes, and a forward looking (ex-ante) "distpatch" 5 minute price series that is used by the System Operator and is not published. The dispatch is solved automatically every 5 minutes, or can be invoked by an operator at any time between.

V. Transmission unplanned outage data

As part of the CDS, Transpower provides two excel files that list, in cronological order, forced outages of Transpower's assets. One file contains AC transmission outages, while the other contains unplanned HVDC system outages. Each transmission forced asset outage has a unique identifier along with a code that describes the actual equipment that was removed from serive. This is called the primary outage. Associated with each primary outage is a list of additional secondary outages that were removed as a result of the primary outage. This section takes a brief look at the HVAC outages on Transpowers network.

The table below illustrates a slightly altered form of the data.

```
Ident   Flt_Item Rem_Item Start            End  \
SS99101 LIV-NSY1 LIV-NSY1 1999-07-02 11:20 1999-07-02 11:30
                 NSY-TF-1 1999-07-02 11:20 1999-07-02 11:40
                 NSY-TF-2 1999-07-02 11:20 1999-07-02 11:41
                 NSY-ROX1 1999-07-02 11:20 1999-07-02 11:30
SS99102 CML-FKN2 CML-FKN2 1999-07-02 22:13 1999-07-02 22:36
                 FKN-TF-4 1999-07-02 22:13 1999-07-02 22:38
SS99103 CML-FKN2 CML-FKN2 1999-07-02 22:50 1999-07-02 23:00
                 FKN-TF-4 1999-07-02 22:50 1999-07-02 23:16
SS99104 CML-FKN1 CML-FKN1 1999-07-02 23:04 1999-07-02 23:11
                 FKN-TF-2 1999-07-02 23:04 1999-07-02 23:12
SS99105 CML-FKN2 CML-FKN2 1999-07-02 23:09 1999-07-03 16:27
                 FKN-TF-4 1999-07-02 23:09 1999-07-03 16:27
```

where the fields are:

Ident
: Incident Identifier that links all the records associated with one fault. Note: the second letter indicates an 'S' for South or 'N' for North Island;

Flt_Item
: The primary forced outage identifier indicating the item of equipment which caused the forced outage. There is usually one of these per incident. The exception is for a double circuit fault;

Rem_Item
: The secondary outages forced out of service caused by the primary outage;

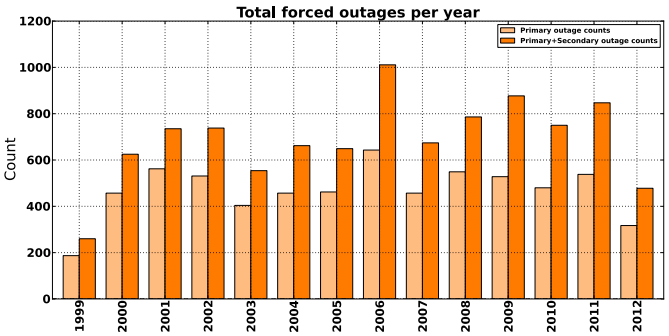
Island
: Indicator of where the outage occurred;

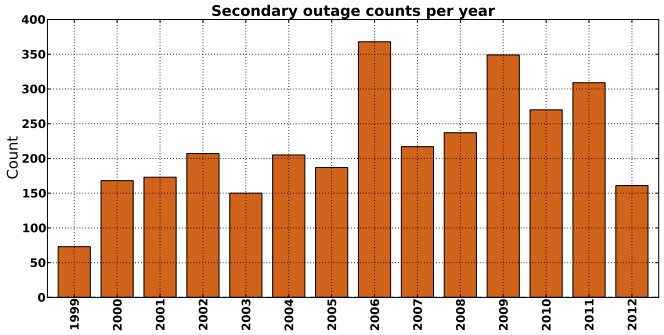
For example, at 11:20am on 2nd of July, 1999, the Livingston-Nasby-Roxburgh circuit tripped also taking out the supply transformers at Naseby. Although the LIV-NSY-ROX circuit was back in service 10 minutes later, it took 20/21 minutes before the Naseby supply transformers were back in service.

The following notes are included in the CDS.

1) Forced Outages are those for which the equipment was tripped or manually taken out of service within 24 hours of the fault occurring or being discovered.
2) A circuit is deemed out of service if any circuit breaker is open. In some cases, a forced outage may be recorded against a circuit section rather than the whole circuit, in particular for three terminal circuits. A transformer is deemed out of service if either the HV or LV circuit breaker is open (an open CB on the tertiary winding is generally not considered to constitute a transformer outage).
3) A trip – autoreclose-trip sequence is shown as one incident (with one identifier) if the autoreclose was unsuccessful because of a persisting fault.
4) If a second fault occurs or is discovered when attempting to return equipment to service, a second incident is recorded with a second identifier, e.g. failed autoreclose because of a protection fault; or a transformer cannot be returned to service because of a CB problem.
5) A number of outages within a relatively short space of time will generally be recorded under one identifier if they all had the same cause, e.g. a tree causes three trippings within 10 minutes. If the outages are caused by separate faults they will have separate identifiers, e.g. if a circuit trips for lightning 3 times in 10 minutes.
6) Unplanned outages caused by a delayed return-to-service after planned maintenance works are not included.
7) Whilst Transpower has used reasonable endeavours to ensure that this data is accurate, there is no guarantee that it is error free.

Lets take a look at the total forced outages per year. Note: 1999 and 2012 currently have incomplete data.





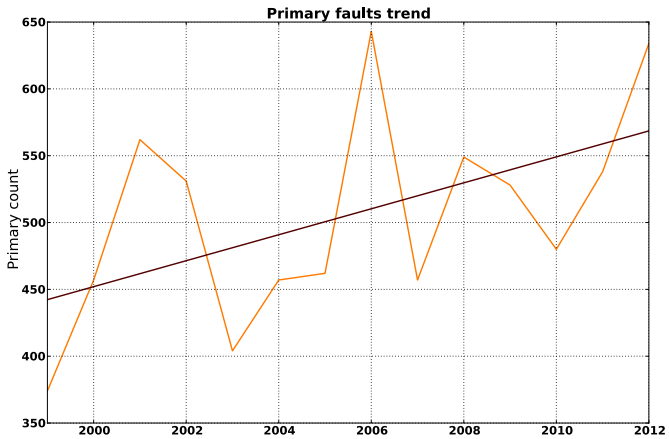
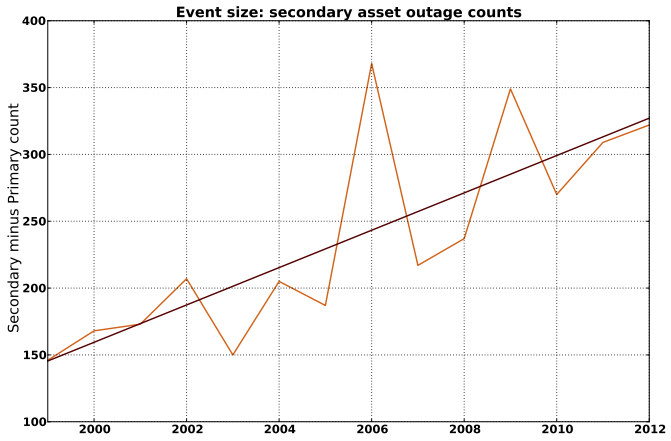


Figure x indicates that although the total count of primary "causer" faulted items is reasonably static, (with a mean of around 505 events/year), there appears to be an increase in the secondary outages associated with each primary outage. I.e., it appears that the size of forced outage event is increasing. To test this hypothesis, additional statistical analysis on the data is required.

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 counts   R-squared:                       0.268
Model:                            OLS   Adj. R-squared:                  0.207
Method:                 Least Squares   F-statistic:                     4.403
Date:                Thu, 11 Apr 2013   Prob (F-statistic):             0.0577
Time:                        12:45:40   Log-Likelihood:                -78.224
No. Observations:                  14   AIC:                             160.4
Df Residuals:                      12   BIC:                             161.7
Df Model:                           1
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        442.3143     35.394     12.497      0.000       365.198   519.431
index          9.7099      4.628      2.098      0.058        -0.373    19.792
==============================================================================
Omnibus:                        1.625   Durbin-Watson:                   2.141
Prob(Omnibus):                  0.444   Jarque-Bera (JB):                1.281
Skew:                           0.630   Prob(JB):                        0.527
Kurtosis:                       2.220   Cond. No.                         14.7
==============================================================================
"""
```

As illustrated, 2006 was a particularly bad year with a maximum of almost 650 forced outages, or, on average, one every 14 hours. In contrast, 2003 had only 404 primary outages or, on average, one every 22 hours. The average



over the entire data series is 505 outages per year, or one every 17.3 hours. On visual inspection there appears a concerning upward trend. However, with such varying data it is often difficult to test this statistically. One method is to attempt to fit a straight line. This is achieved using an ordinary least squares, or OLS, linear regression model. OLS models also output various coefficients and statistical indicators, including measures such as confidence intervals. In this instance it appears that it is statistically likely (to around a 95% confidence interval) that the number of primary forced outages over the last 13 years may have, at the very least, been increasing. On average, the Optimal Least Squares (OLS) regression results indicate a growing number of around 10 forced outages per year.

These results are borderline conclusive as the P-values indicate (footnote: A p-value is always between 0 and 1 and it tells you the probability of the difference in your data being due to sampling error, the p-value should be lower than a chosen significance level (0.05 for example) before you can reject your null hypothesis.) and so we can't conclusively state that the primary forced outage rate is growing, although it is likely that it is.

As illustrated in figure x, in addition to total primary forced outages it is interesting to analyse the secondary outage counts per year. Annual counts of secondary outages provides a very good proxy for the size of the primary forced outage events. In this instance, the size of any forced outage event is not measured by lost electrical demand, but instead measured by how many additional elements are lost due to the primary causing outage, i.e., the number of additional transformers, circuits or line sections lost when a primary forced outage occurs.

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                 counts   R-squared:                       0.611
Model:                            OLS   Adj. R-squared:                  0.578
Method:                 Least Squares   F-statistic:                     18.83
Date:                Tue, 09 Apr 2013   Prob (F-statistic):           0.000962
Time:                        12:01:37   Log-Likelihood:                -73.147
No. Observations:                  14   AIC:                             150.3
Df Residuals:                      12   BIC:                             151.6
Df Model:                           1
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        145.4571     24.628      5.906      0.000        91.798   199.117
index         13.9736      3.220      4.340      0.001         6.958    20.989
==============================================================================
Omnibus:                       11.111   Durbin-Watson:                   2.908
Prob(Omnibus):                  0.004   Jarque-Bera (JB):                6.867
Skew:                           1.466   Prob(JB):                       0.0323
Kurtosis:                       4.782   Cond. No.                         14.7
==============================================================================
"""
```

This suggests, with reasonable certainty, that the size of event outages, in terms of the amount of secondary items effected, has been growing. The 95

Assuming the data is to be believed, there could be many reasons for this increasing trend, including:

- increased system size;
- increased peak demand;
- increased complexity, for example, complex protection systems and settings?
- increased human error;
- data/communication issues?
- actual reported dataset issues?
- lack of investment;
- lack of maintenance;
- lack of co-rodination in planned outages?

This section has illustrated the results of an analysis on the forced outage data that is openly published as part of the Centralised Data Set (CDS). It suggests that primary forced outages may be on the increase, although this is borderline statistically. It does however appear that the size of forced outages, in terms of the number of secondary outages resulting from an initial primary outage, is on the increase, at least over the last 13 years.

Many more analyses have and can be made on this dataset. For example, grouping the data at the island level can give island wide information and we can also identify those assets who trip most often, etc. If interested, contact the author for more information or for a copy of the IPython Notebook, and data files used in this analysis.

VI. POCP DATA

The Planned Outage Co-ordination Process (POCP) and the POCP database are run by the System Operator and provide a voluntary platform through which industry participants can publish, or broadcast, their intended planned generation and transmission outages.

The current system was jointly developed by participants, including Transpower, through a series of joint industry workshops during 2002/2003. The WAG/System Operator are currently reviewing the POCP process.

The data, although available to participants, is not openly available, requiring registration and a password. The POCP database is available at: http://www.pocp.redspider.co.nz/ with a dashboard available at: http://nzeb.redspider.co.nz/[1]

Under the POCP:

- Asset owners provide information about their planned outages;
- The System Operator assesses any potential conflict with its Principle Performance Objectives, based on the information provided, and publishes the results of the assessment;
- Any interested party may view both the planned outage data, and the results of the System Operator assessments, so that they are aware of planned outages that may require their consideration or action;
- The POCP database supports the POCP by receiving, collating and publishing the planned outage data from asset owners, and is also used to publish the results of the System Operator assessments.

The System Operator provides a very nice dashboard through http://nzeb.redspider.co.nz/ which provides a more graphical interface with custom alerts etc. This is great for future planning, and outage assessment, as required by the System Operator, but as it turns out, not so great for

[1]A guide is also avialable at: http://pocp.redspider.co.nz/doc/guide.php

inspecting historic outages. The Market Performance team at the Authority are interested in both forward looking and historic outages. There are several issues with the way the current POCP database is setup:

Generators often cancel an outage after the outage window has finished

This makes if difficult to tell, retrospectively, weather an outage actually happened at all. A way around this is to inspect the Last Modified field and make the assumption that as the outage was modified after the "End time" field, the outage did in fact occur. This is not ideal. Ideally, outages should never be cancelled if they occurred. Outage's should only be cancelled if they do not occur. If an outage ends early, then the End Time can be modified to reflect this.

Inconsistent data

The current database allows generators quite a lot of freedom when entering data. This is problematic and makes analysis of the data very difficult. For example;

- None of Trustpowers generation outages are ever confirmed. They are all marked as Tentative;
- Outages can be entered multiple times with different identifiers;
- Outages can be indicated with no MW loss of generation;
- Over the history of the data, there are Outages whose IDs are not unique.

Despite these problems, it is useful to investigate outages as they occur in time by time-serializing the data and constructing a time series of outage MW magnitude. To get around some of the issues discussed above some ad-hoc time dependent logic is applied, in particular to determine if a cancelled outage did in fact occur. Due to the inconsistencies in the data, there are probably issues with some of the logic used. However, an output of this time-serialisation process is illustrated below. This is an interactive data visualisation that lists the outages started within the selected window. To achieve this a Python script is used to download the database and manipulate the data into a required form, including the construction of the time series data. A local webserver is then set-up to plot the data using, in this case, the excellent Javascript d3 library combined with a d3 module called "cross-filter".

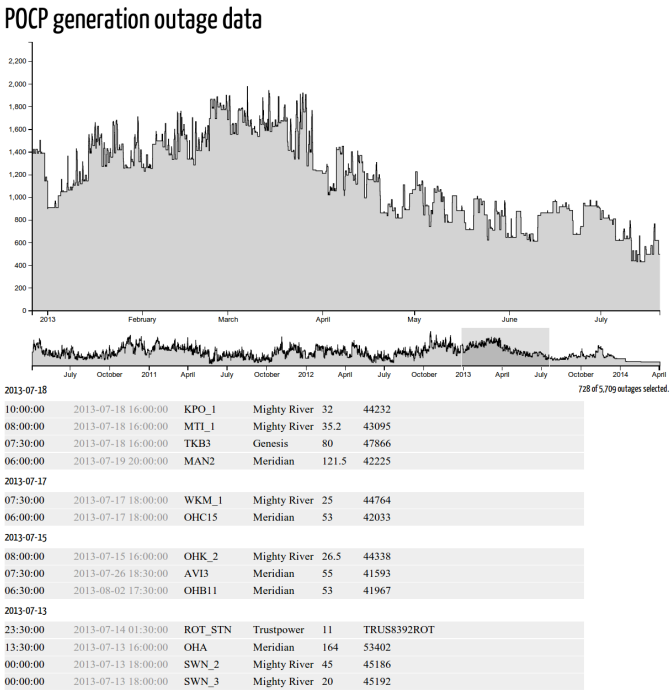


Fig. 6. Example of POCP monitoring at the EA.

VII. Open source software

The development of quality open source software has taken off in recent years due mainly to the ease of software colabaration through the use of the Git version control software and Github.

This paper is writen entirely with open-source, freely availble software. All data analysis, outputs and in fact the entire text of this paper have been written within a single iPython notebook. The iPython notebook allows an interactive python code devleopment environment from within the web-browser. The Notebook and packages such as Numpy and Pandas have been the authors daily tools now for over 18 months, replacing both Matlab, Excel and R.

The use of client-side (browser based) visualisation for interactivity is also growing extremely quickly. Javascript/d3 and it's various modules appear to be leading the pack.

VIII. Conclusions

In an industry flush with data, data analysis and visualisation are becoming increasingly important. The expoential growth in electrical generation in the first half of the 20th centry has in certain ways been replaced by an exponential growth in data. A single hour of data logging from a single home energy monitoring system will contain more data than the total historical generator name-plate data series used in this paper. For the engineering/data analyst, it is easy to become engulfed with too much data. Tools (and knowing how to use them efficiently) are important. In recent years the rise of powerful open-source software has in ways overtaken many commercial products. In particular, the Python programming language combined with the iPython notebook make for a formidable interactive workspace, ideal for both data analysis and visualisation.