125 years of data

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

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 name-plate 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

Data analysis often requires significant time and effort wrangling (or 'munging') the data into a required form suitable for analysis and then visualisation. The output, often a shared visualisation, can lead to further feedback and questions such as why? or, how? This can lead to yet further analysis and updated visualisation. Although this process is often rewarding, without the appropriate tools it can take significant time and effort[1].

This paper provides some examples of analysis and visualisation from a smorgasbord of data-series within the NZ electrical power system/market. Most of this data is either openly available on-line, or through the Centralised Data Set (CDS).

Section II presents a very small historical data set representing the growth of name-plate generation 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 today, section III investigates the electrical demand of the author's Wellington home while section IV, explores a small sample set of energy price data including real-time 5 minute prices.

Using data published in the CDS, section V explores, and attempts to identify any trends present in forced transmission outage data, while section VI 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 VII provides a brief overview of the open source software used by the author for all analysis and presentation used in this paper.

[1]The authors tools of choice, all open source and freely available, include: Python and the "iPython notebook", plus the Pandas,NumPy and Matplotlib Python modules, among many other Python modules.

II. Growth in New Zealand generation capacity

'Very little is yet really known about electricity'

Mines Inspector commenting on the new electric dynamos at the Bullendale, 1887.

Over the last several years Nicky McLean from the Electricity Authority has put together a very small, but interesting historical record of electricity generation in New Zealand. The series, though continually evolving, is enough to illustrate the growth in generation capacity in New Zealand over the past 125 years[2].

The data includes: 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 beginning of the year).

The raw data file is publicly available as a simple text file and provided as part of the CDS. Some care has been taken over the format of the file so 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).

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 prior to the public power supply 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 demonstrated the first public supply of electricity in New Zealand. The rest; as they say, is history…[3]

…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 from 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

[2]Or 127 years, depending on your preference!
[3]Additional historical events will be talked about in the presentation.

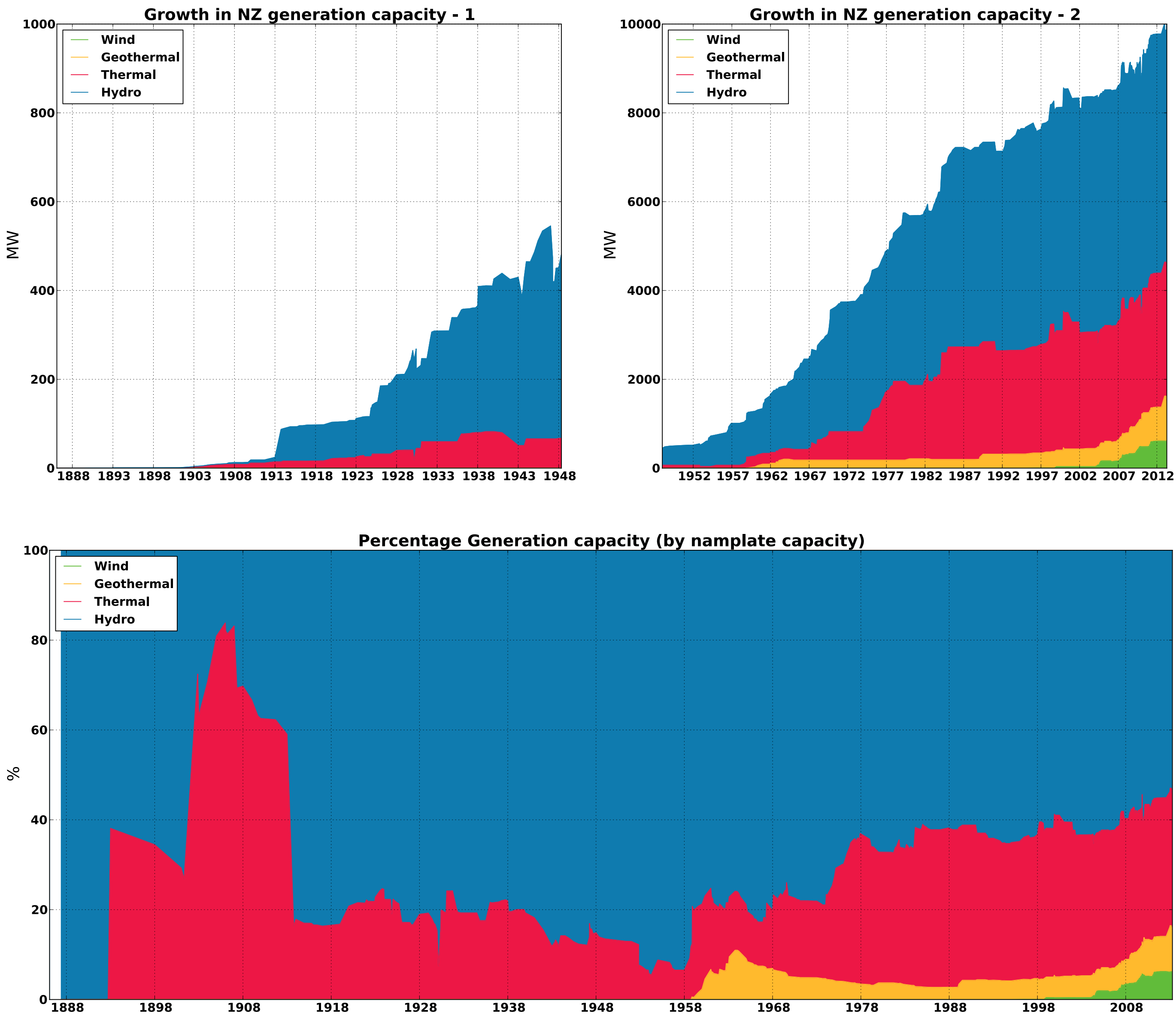

Fig. 1. From the *"Hall of dazzling light"* to present day. 125 years of data illustrating the approximate historical generation capacity growth in New Zealand.

III. Home energy monitoring

In complete contrast to the previous section we jump 125 years to the present day and the data, collected at the author's home, of total house-hold electrical consumption. At the time of writing, ≈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 similar datasets could exist[4].

Through analysis of the data, owners (of the data) can benefit from its use. For example, aside from an obvious awareness of consumption, and hence savings, one could imagine a future where individual house-hold data could help retailers target suitable customers for their portfolio[5].

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 home energy monitoring system.

[4]Perhaps not at this resolution?

[5]Perhaps this is already happening to some extent?

The system described here consists of:

1) a Raspberry Pi computer; and,
2) an off-the-shelf 'Current Cost' energy monitoring kit.

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

Figure 2 illustrates a typical high-usage day. The house-hold consists of two adults, and two young children and so should represent the energy usage of a typical young family living in New Zealand. Other characteristics 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) located in Wellington, ≈200m above sea-level.

Observing such time series data one can start to identify different appliance loads by eye![6] For example, the washing

[6]Note the occasional inductive in-rush current.

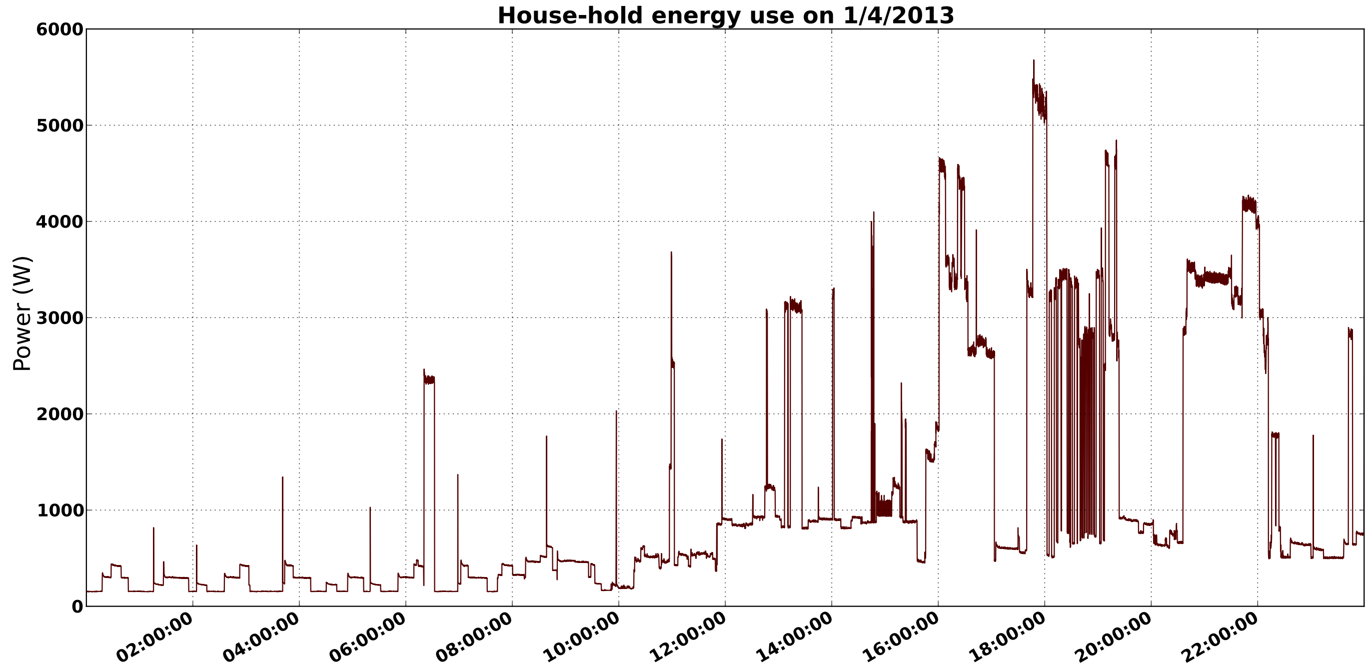

Fig. 2. Typical house-hold electrical energy use.

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 delicious quiche!

In terms of the demand profile, it is interesting to observe how our demand compares with the average demand in our region. This sort of analysis could help retailers in the future profile customers demand patterns. Figure 3 illustrates the mean daily profile of the author's home, compared with the mean daily profile of the closest GXP (Wilton 33kV) 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 $\approx$1 hour compared with the regional demand. 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.

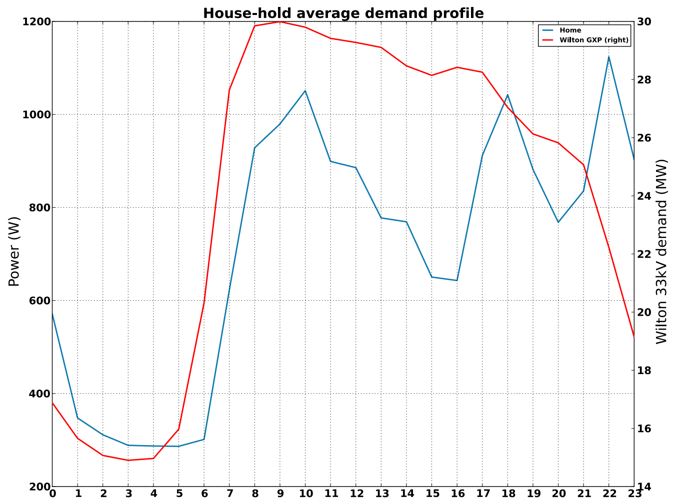

Fig. 3. Average daily demand profiles (between November 2012 to April, 2013) of the author's home vs Wilton GXP in Wellington.

IV. Real-time spot market price data

The market performance team at the Electricity Authority has a role to play in monitoring the market. As a tiny part of this, an automatic real-time 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 Javascript and viewed in a web-browser;
6) sends SMS alert text messages under certain conditions.

Figure 4 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[7]. This type of visualisation is ideal for monitoring large data sets. In the case of the NZ electricity market, all prices over the 250+ GXPs for the past week can be inspected within a matter of seconds[8]. The bands of decreasing contrast indicate increasing prices, until a threshold is reached where the colour switches from shades of green to red. Here, at 7:10pm on Monday the 5th of April, prices reached around \$400/MWh at Fernhill in Hawkes Bay (FHL0331). This was likely caused by a temporary constraint on the network in the region, perhaps in this case the interconnecting transformers at Redcliff. On

[7] This uses an open-source Javascript library called d3.js, along with a module called cubism.js developed by Mike Bostock and freely available on-line.

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

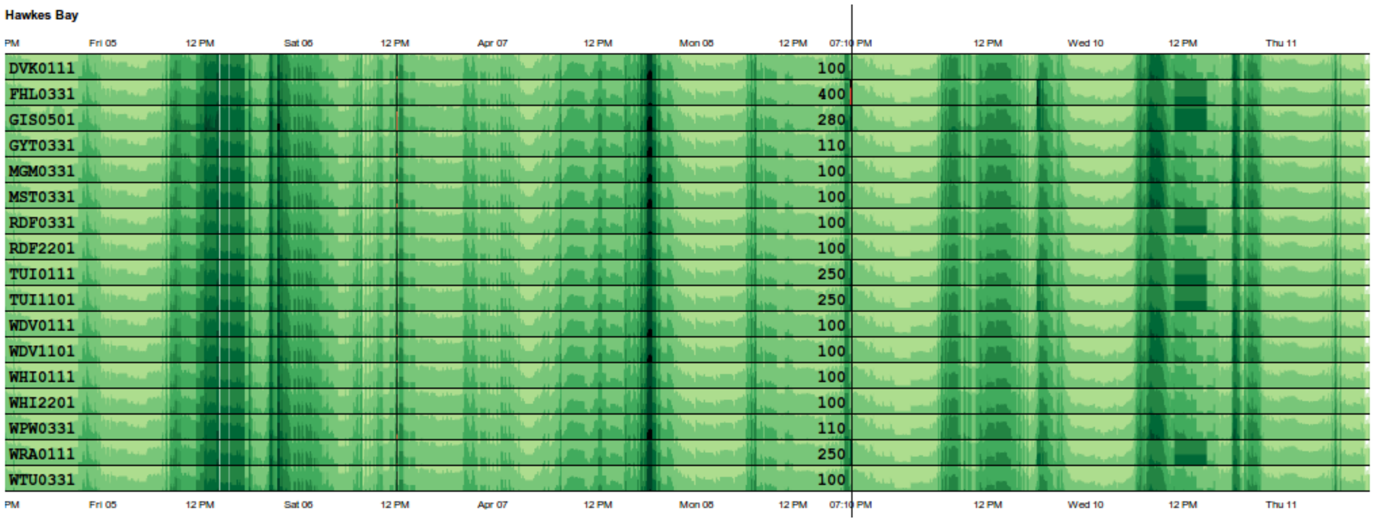

Fig. 4. Example of visualization of spot prices using horizon charts. Prices observed in Hawkes Bay over a week in early April, 2013.

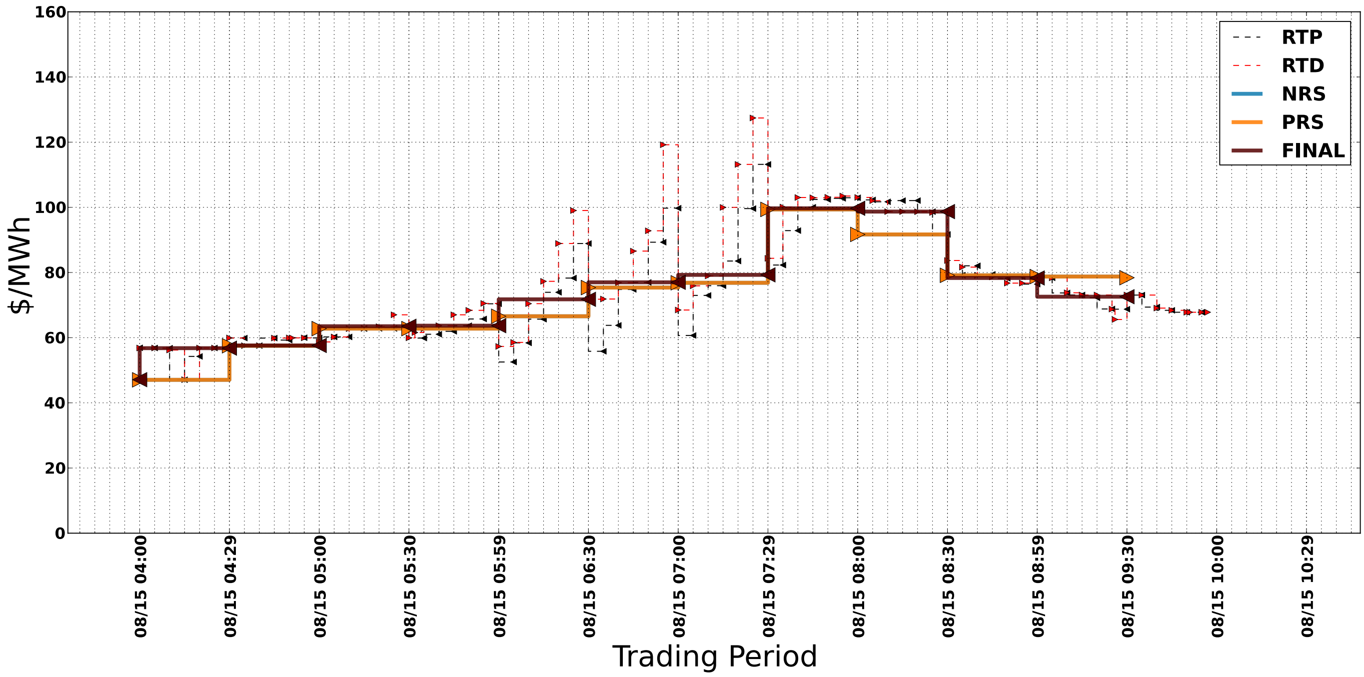

Fig. 5. Example of the difference that can be observed it different forward looking and backward looking price series in the NZ electricity market.

Wednesday afternoon the same transmission constraint bound resulting in high prices in the same region.

One issue with real-time prices is the differences observed between real-time and final prices. These differences may be due to a number of factors including: inaccurate demand forecasts or bids, generator ramp rates, intermittent generation, among others. Figure 5 illustrates a fairly typical example of the differences that can occur during a morning peak period, in this case on the 8th of August, 2012. The Non-Responsive and Price-Responsive Schedules (NRS/PRS) indicate the final scheduled prices 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 service. 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 that 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 by the primary outage.

Start and End times
: Indicating the start and duration of all outages.

For example; at 11:20am on 2nd of July, 1999, the Livingston–Naseby–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[9].

[9]Note: 1999 and 2012, at the time of writing, had only a half-year of data. Counts for this data are multiplied by 2 in the OLS regression analysis.

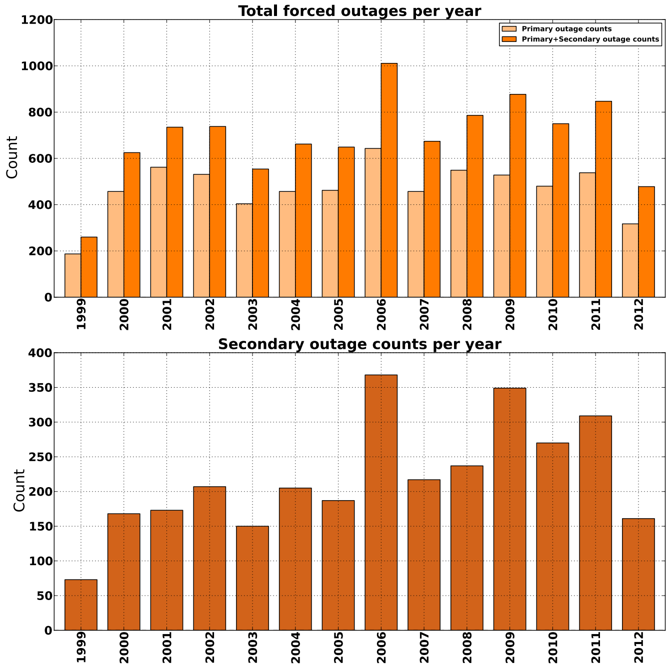

Fig. 6. Annual counts of total forced HVAC outages.

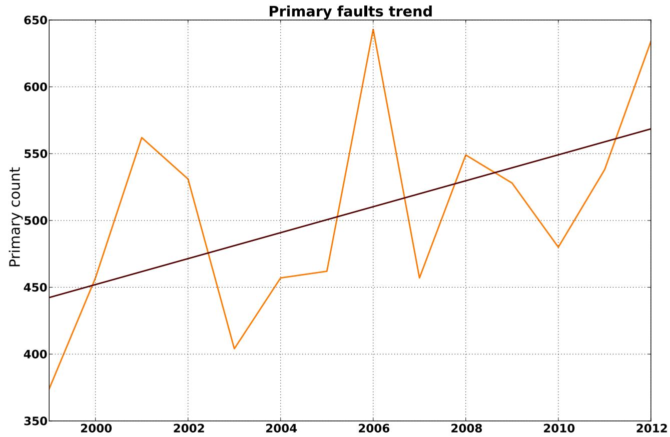

Fig. 7. A straight line fit of primary outage counts

As illustrated in Figure 6, 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 can also output various coefficients and statistical indicators, including measures such as confidence intervals.

Testing this result statistically indicates the results are borderline conclusive and so we can't reject the null hypothesis or conclusively state that the primary forced outage rate is growing[10].

[10]As the p-values indicate. A p-value is always between 0 and 1. It indicates the probability of the difference in the 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.

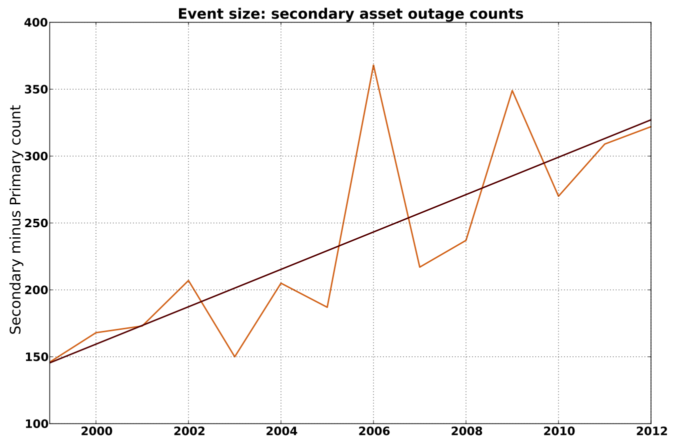

Fig. 8. A straight line fit of secondary outage counts.

```
<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
==============================================================================
"""
```

Although the total count of primary outages can not be conclusively proved to be growing, the same can not be said about the increase in secondary outages associated with each primary outage. As illustrated in Figure 6, it appears that the size of forced outage events, in terms of how many secondary items are tripped, is increasing.

To conclude this, we assume the annual count of secondary outages as a proxy for event size. This seems reasonable; instead of being measured by lost electrical demand[11], we measure how many additional elements were 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
==============================================================================
"""
```

Figure 8 and the table above illustrate the results of the OLS regression analysis. This suggests, with reasonable certainty, that the size of event outages, in terms of the amount of secondary items tripped, has been growing.

The 95% confidence interval is between 7 and 21 additional outages per year with an average rate of growth of around 14 additional secondary outages per year, for each year, over the past 13 years.

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 (protection systems and settings?)
- increased human error;
- data/communication issues?
- actual reported dataset issues?
- lack of maintenance;
- lack of co-ordination 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 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 tripped 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/[12]

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

[11] Note: many of the forced outages will not necessary result is lost demand.

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

planning, and outage assessment, as required by the System Operator, but as it turns out, not so great for inspecting historic outages. The Market Performance team at the Authority are interested in both forward looking and historic outages. For historic outages there are a few 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 error prone making analysis of the data very difficult. For example;

- One generators generation outages are never confirmed. They are all marked as Tentative!;
- Outages can be entered multiple times with different identifiers and start/end dates;
- Generator outages are sometimes entered with the "MW loss" field blank;
- As generators can choose an outage ID, over the history of the dataset, there are outages whose IDs are not unique.

Despite these nuisances, the data can be made semi-useful by investigating the outages as they occur in time. This is achieved by time-serialising 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.

An output of this time-serialisation process is illustrated in Figure 9. This is an interactive data visualisation that is able to list the outages that have occurred, or are likely to occur. Outages are listed by Start data within a date range that can be selected with a mouse-click-drag. To achieve this, a Python script is used to download the database and manipulate the data into the required form. 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". Many improvements are planned for this system in the future.

VII. Open source software

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

This paper is written entirely with open-source, freely available software. All data analysis, visualisations, 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 development environment from within a web-browser. The notebook, and many Python packages such as Numpy and Pandas have been the authors daily tools now for two years, replacing both Matlab, Excel and R.

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

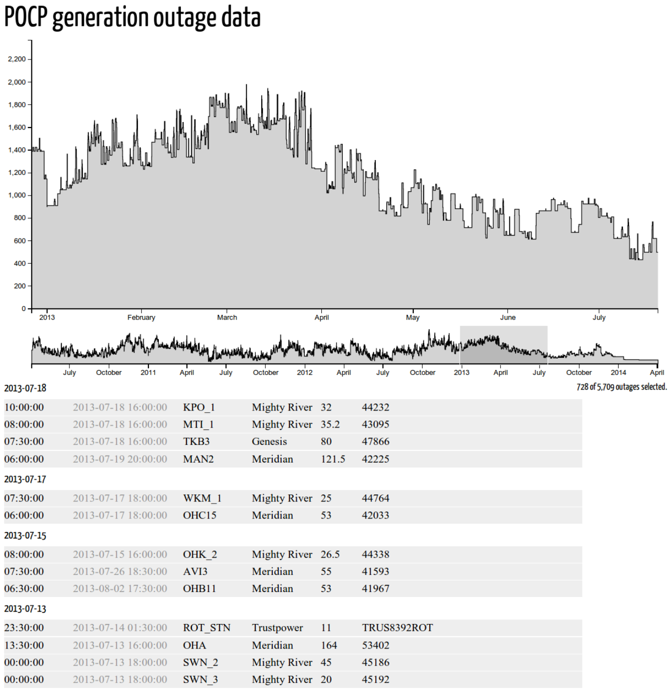

Fig. 9. Example of POCP monitoring

VIII. Summary

In an industry flush with data, data analysis and visualisation are becoming increasingly important. The exponential growth in electrical generation in the first half of the 20^{th} century 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 part, overtaken many commercial data analysis products. In particular, the Python programming language, combined with the iPython notebook, make for a formidable interactive workspace, ideal for both data analysis, visualisation and general Python code development.

Acknowledgments

The author would like to thank Nicky McLean, Ramu Naidoo, Greg Williams and Doug Watt for their various contributions.

References

[1] *Python for Data Analysis*, OReilly Media, Inc. , Wes McKinney, 2013.
[2] *Interactive Data Visualization for the web*, OReilly Media, Inc. , Scott Murry, 2013.
[3] *Version control with GIT*, OReilly Media, Inc. , 2nd. Ed, Jon Loeliger and Mattew McCullough, 2012.
[4] *JavaScript Enlightenment*, OReilly Media, Inc. , Cody Lindley, 2013.
[5] http://www.doc.govt.nz/Documents/science-and-technical/sap237.pdf
[6] http://www.reefcottage.co.nz/41_history.htm
[7] http://www.reeftongold.co.nz/morning-heritage-tours/electric-light
[8] http://www.TeAra.govt.nz/en/gold-and-gold-mining/12/2
[9] https://github.com
[10] http://ipython.org/

. . . On Wednesday the 24'th of November 1886, bright light had been brought to the bars of Dawson's, Kater's, Stevenson's and William's hotels by showman Walter Prince via underground cable through attaching a one kilowatt generator to the Oxley's brewery's steam engine. The test required regular visits of the spectators between the hotel and the brewery, and there was high demand at each point of supply. As a result, many were carrying an overload and it was not only the hotel that was lit up. . . . "

from 'Electrical development in New Zealand'
H.J. Beech