RLCA
CONSULTING

Data for Mining & Analysis

Panels are often the most useful data arrangement for data mining and other analytic tasks. Yet assembling EIA data into panels is time consuming. Because 70 percent or more of a mining/analysis project is taken up by data assembly and conditioning, a relatively small investment in a data assembly tool would pay dividends by reducing the cost of data and policy analysis projects, whether cost is denominated in staff time or dollars.

The API does not deliver panel data. For the most part, it delivers time series in the form of xml or JSON, either of which can be used to build part – but not all – of a panel. Some data must come from sources not covered by the API. This may represent a gap in EIA's in-house tool stack.

Data Panel Anatomy

A typical data panel might look like Figure 1, in which each data row, or "instance" contains property values recorded at the same time.

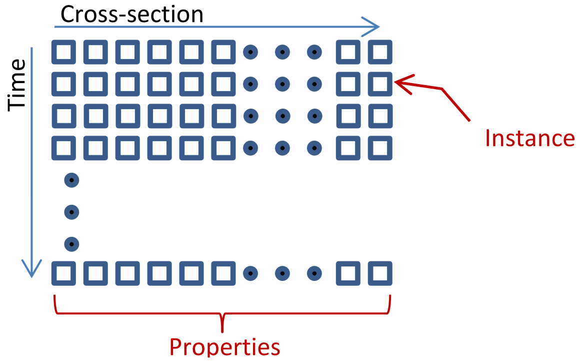

Figure 1

Time series delivered by the API comprise one column of the panel, which may contain just a few or perhaps many columns. Where several API calls/series are used to build the panel, each column must have the same temporal granularity (hour, day, month, quarter, *etc.*) and each row, or instance, must report data for the same time. Ensuring that this is so is yet another step in the panel assembly process.

Figure 2 is a more detailed look at a typical panel. Here, the panel contains both numeric and non-numeric data. Panels like this are both common and important in data mining.

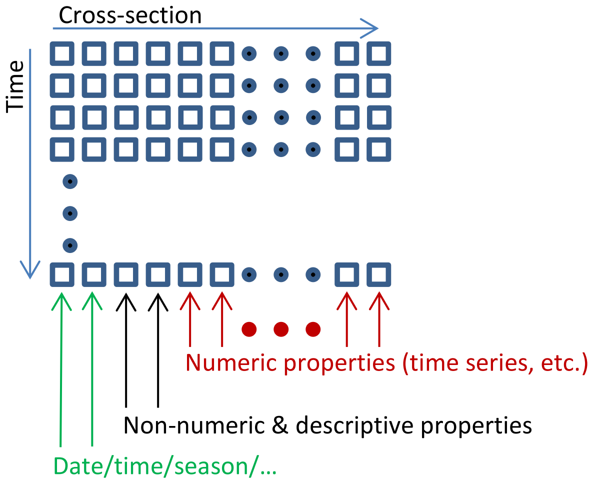
Cross-section
Time
Numeric properties (time series, etc.)
Non-numeric & descriptive properties
Date/time/season/…

Figure 2

Non-numeric and descriptive properties may include things like fuel codes and fuel properties, generator sizes, RTO and balancing authority acronyms, *etc*. that come from sources like EIA-860, EIA-923, *et al*. (Although electric power examples are used here, the same principles apply to other parts of the energy space.) Each instance (row) in such a panel is said to contain “mixed” data.

Once assembled, a panel can be manipulated to meet analytic needs with relative ease. It can be subdivided, sorted, filtered, normalized, transformed, *etc*. Most such operations are more difficult – and error prone – if data are not already in panel form.

Example: Panel Data Used for Visualization

EIA-930 data for Florida during January – May, 2016 provide a small, simple example of how panel data were used in a data visualization exercise using a data mining platform (Weka Version 3.8). Each of the 31,000+ instances (vectors) is comprised of the following columns (properties):

- Date/time (date/time)
- Balancing authority acronym (character)
- Demand in MWh (numeric)
- Generation in MWh (numeric)
- Net interchange in MWh (numeric)

Data were conditioned by splitting interchange into two variables, imports and exports, and expressing each as a fraction of demand. Zero values were replaced by extremely small numbers and base-10 logs were taken of demand, generation, imports, and exports to reduce skewness.

Figure 3 is a screen shot of a scatter plot showing demand on the x-axis and imports on the y-axis. Each balancing authority is shown in a different color.

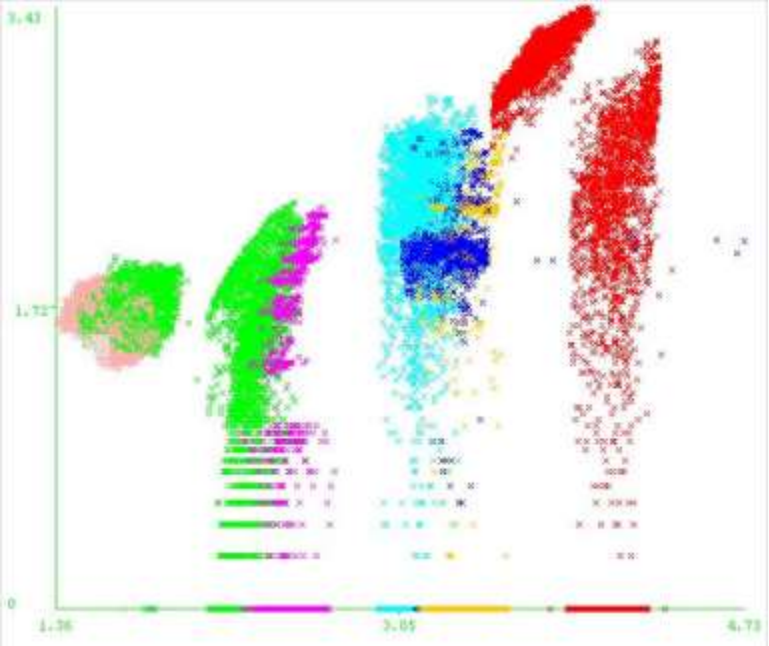

Figure 3

The figure illustrates several things: First, mixed data are important when mining for information. Second, mining mixed data makes patterns perspicuous and they may turn out to be important. Certainly, patterns within the data show up in this example. Third, data conditioning done here was very easy once a panel was built.