The Keymaker, Opening the Door to Energy Data Possibilities

Kai Zhou, Opinion Dynamics

The ability to flexibly create, combine, or subset any dataset relies on a unique identifier; otherwise known as a unique key or primary key. In many Energy Efficiency Evaluations, evaluators work with utility customer and program tracking data. The utilities or program implementer typically set up these datasets such that they can be linked using account numbers. However, when evaluators report results or draw samples, they often do so in terms of sites, households or business facilities. This results in a disconnect in the unit of analysis because account level information does not perfectly capture households or facilities. This is especially true in commercial or multifamily facilities that often contain more than one utility account.

This poster presents a method for creating unique keys at the site level from utility customer data, leveraging tools such as regular expressions, address unit parsing, geocoding, word weighting, logical criteria grouping and sorting. This method is very fast because it minimizes comparisons between records to only those that are likely to match. The end product is a key mapping that shows how every original record in the starting data corresponds to the generated unique key.

In a step-by-step by workflow process, the posters outlines some general information on how to visualize and build a tool for creating the key, the steps include the following:

1. Name de-aggregation
2. Address de-aggregation
   - Notes on using geocodes
3. Leveraging additional data commonly found in utility or program tracking data
   - Account numbers
   - Phones
   - Emails
   - Mailing addresses
   - Contacts
   - Handing multiple versions of the same data field
   - Handing missing values within the data field
4. The looping algorithm of: sort, flag, apply rules and repeat

In addition, the poster provides information on creating and applying the final mapping key to the final or to multiple datasets. It also includes a section to how to check and fix different types of errors in the resulting mapping key.

Applications for this technique go beyond rolling up utility account numbers to site level, it can also be applied to a wide range of tasks that used to be difficult or time consuming. With this rigorous approach to creating unique identifiers, we open the door to new analyses and improved data connections while at the same time providing quality checks of the data.