Journal of Sustainable Development of Energy, Water and Environment Systems

http://www.sdewes.org/jsdewes

ISSN 1848-9257

Journal of Sustainable Development of Energy, Water and Environment Systems

http://www.sdewes.org/jsdewes

Year 2019, Volume 7, Issue 3, pp 466-481

Smarter Cities: Cleaning Electricity, Gas and Water Metered Consumption Data for Social and Urban Research

Mishka Talent
Fenner School of Environment and Society, Australian National University, 46 Sullivans Creek Road,
ANU, Canberra, ACT, 0200, Australia
e-mail: mishka.talent@anu.edu.au

Cite as: Talent, M., Smarter Cities: Cleaning Electricity, Gas and Water Metered Consumption Data for Social and Urban Research, J. sustain. dev. energy water environ. syst., 7(3), pp 466-481, 2019,
DOI: https://doi.org/10.13044/j.sdewes.d6.0242

ABSTRACT

It is claimed that 'Big Data' could help cities become 'smart', utilise resources more efficiently, and improve inhabitants' quality of life. Metered consumption data of electricity, gas and water are collected and stored for each household in Australia and provide a valuable source of information for researchers hoping to understand the patterns of consumption and improve resource utilisation. This work tests the assumption that these datasets are sufficiently 'clean' to allow interrogation and details the common errors encountered. An inspection of 1-7 years of meter readings of electricity, gas and water for ~60,000 households in Canberra, Australia as well as all datasets of the local government's energy and water efficiency programs, found this not to be the case. Common errors found include: missing data, biases in erroneous data, errors generated by the data custodians, duplicate entries, the effect of different custodian objectives and of poor data constraints (free flowing text).

KEYWORDS

Data cleaning, Data errors, Duplicate data, Unique dwellings, Electricity, Water.

INTRODUCTION

As the world population increases, with close to 4.2 billion now living in urban areas [1], the increasing demand for scarce resources, such as potable water and energy, have focused attention on improving efficiency of consumption and the 'sustainability' of the city. While total city-wide consumption of such resources are relatively straight forward to measure, the variation and determinants of this consumption at the household level are poorly understood despite more than 40 years of investigation [2].

'Big Data' can provide insights into the patterns of resource consumption [3] and be used to reduce wasteful consumption and improve the sustainability of cities [4]. In this work, 'Big Data' is defined as a dataset that has more than 1 million rows [5]. Resource consumption may be reduced through structural changes to city developments during the planning process or though real-time operational improvements. In real-time, for example, washing machines or electric vehicle chargers, throughout a city, can be controlled in such a way that minimises peaks in electricity consumption [6]. This 'smart grid' is an active area of research by computer scientists [7].

The recent increased volume of data and improvements in data mining techniques have the potential to progress the understanding of the determinants of resource consumption. Early work in this field by Newman and Kenworthy [8] explored patterns of consumption at the city level. Exploration of the patterns of water consumption in Australian households was advanced by Beal *et al.* [9] who analysed end-use consumption through household surveys of 252 homes in South East Queensland, as part of a AUD 50 million project focusing on water security and water recycling. In Australia, there are surprisingly few examples using large and detailed datasets derived from household-level billing data. Troy *et al.* [10] was the first to use utility-held metered water consumption data to explore patterns in residential consumption. Their work uncovered a variety of factors affecting consumption, such as leaking water meters and pipes, which were not apparent from surveys alone.

'Smart cities'

Large datasets of metered water and energy consumption are potentially of great value in understanding the factors of consumption but only where those datasets are sufficiently 'clean' to allow detailed examination. Many are unstructured, or poorly structured and without standard formatting require human management [6]. These issues were identified by Bellini *et al.* [11] in their work on cleaning datasets for 'smart city' applications in Italy: "Presently, a very large number of public and private data sets are available from local governments. In most cases, they are not semantically interoperable and a huge human effort would be needed to create integrated ontologies and knowledge base for smart city". The lack of public information on the cleanliness of such datasets is an impediment to those who wish either to analyse or use such datasets or provide tools and methods of cleaning such data. Bellini *et al.* [11] claimed that the main technical obstacle to realising 'smarter city' solutions is related to data aggregation.

Considerable work has been invested to reduce the errors in utility and government services data in the European Union (EU). The EU Joint Research Centre (JRC) through the Covenant of Mayors Initiative [12] provided statistical methods of cleaning abnormal data, such as energy consumption data. The Council of Europe (CoE) provided guidance on methodologies for data cleaning for developers of Statistical Data Warehouses [13]. They suggested, for large administrative datasets, that 'automatic cleaning' was required. Furthermore, Infrastructure for Spatial Information in Europe (INSPIRE) Directive provided technical guidelines and specifications on utility and government services data [14]. Data cleaning may not be required where these guidelines are followed. For example, in their analysis of hourly energy consumption for heating of 8,293 households in Denmark, Gianniou *et al.* [15] stated, that cleaning involved "removing the data with missing values, and adjusting the reading time stamp to align winter and summer time". McLoughlin *et al.* [16] did not report any cleaning of smart meter data in their analysis of domestic electricity consumption in Ireland. Where specifications for utility and administrative data have not been followed there is little practical literature on the nature of the errors in these datasets.

Real-world examples

There are few examples in the literature on the cleaning process that use real administrative or utility data. Missier *et al.* [17] showed examples from an Italian government tax submissions database and the types of errors they encountered. Lee *et al.* [18] were primarily concerned with removing duplicate records from matched name and address fields in Singapore. Bellini *et al.* [11] developed a system to ingest and reconcile 'smart city' data using datasets from Italy. Matching street addresses were noted as problematic in their datasets, 5.75% of their location data was wrong and not reconcilable. Typically, little information is provided on the cleaning process.

Data cleaning

Data cleaning, also known as data ‘cleansing’ or ‘scrubbing’, deals with the detection and removal of errors and inconsistencies to improve data quality [19]. Cleaning may be required for a number of reasons including, data-quality procedures have not been followed [20], data was sourced externally, it was generated at different times by different people using different conventions [21], or was collected for a different purpose.

There are often persistent errors in any datasets. According to Hankins [22], “Most organizations are plagued by a state of denial as to the level of quality they have”. A founding father of the data warehouse, WH (Bill) Inmon, observed that, “It’s after the user starts to use what’s out there that they start to scream” [22]. The cleaning of data is one of the three largest problems in data warehousing, but attracts little discussion [20]. John Ladley, a research fellow at the Meta Group, suggests that 80% of data-warehousing effort goes into cleaning, extracting and loading data [22]. The cleaning process is time expensive and there is no single program to automate it [23].

Errors in data reduce the ability to discover genuine patterns of interest and so must be resolved. “Without clean data, everything that is based on or results from the information loses credibility” stated Leonard Dubois [22] – the proverbial ‘rubbish in is rubbish out’. Data that is used for decision making must be of a good quality to avoid wrong conclusions [19]. Only by cleaning can the data be rigorously analysed and justifiable error bounds on results be reported.

Usually, the number of errors and the ‘best’ cleaning method are unknown [24] and in only a few cases can cleaning methods be checked against a known clean dataset [17] or by hand [18]. Custom computerized cleaning algorithms are widely used because of the difficulty of generalising domain-specific problems [21].

While some commercial tools have been developed for specific data types, these are of limited use for other types of data. Names and addresses are the most common types of data that these tools handle [20]. The Extract, Transform and Load (ETL) software typically provide limited tools to transform data and only allows conversions between standard formats [23].

Cleaning methods and frameworks

A large body of literature deals with common types of unclean data. Such literature often proposes general approaches for resolving such errors. These errors include different field entry formats:

- Spelling mistakes;
- Missing values;
- Extra spacing (white space);
- Invalid entries;
- Unknown characters (e.g. hand writing);
- Different levels of aggregation;
- Duplicates;
- Embedded values (e.g. multiple values in free-form attributes);
- Varying representation (e.g. Y, y, yes);
- Departures from business rules;
- Different versioning (legacy data);
- Non-standard representation [25].

There are three main components to data cleaning:

- Identification of erroneous data;
- Analysis or auditing of the data to determine the correct value;
- Correcting the erroneous values.

In practice these are not always distinct steps, but part of an iterative data preparation process.

There are two common approaches to identifying erroneous values: inference-based, where patterns in the data are identified and are used as rules for cleaning and data-based, where pre-defined rules are used to find cases that match either exactly or with varying degrees of probability.

Inference-based pattern recognition. Inference-based error detection looks for similarities within a data field and uses the patterns to identify outliers. Words with unusually high frequency, for example the word “Unknown” found in data can be identified and removed easily using this method. However, not all data mining patterns are useful and may be an artefact of the type of data analysed [20]. Guyon *et al.* [26] suggested that a pattern is often only informative if it is difficult to predict by using a model based on previously seen data. Numeric outliers, for example, are easy to identify, but in some data, such as income, extreme values may not be erroneous.

The data often need to be aggregated in different ways to identify erroneous values. Having more data sources increases the possibility of identifying discordant data, either by cross-checking with complementary data or by looking for patterns in small sub-sections of the data. For example, grouping all “John Smith” entries together and cross checking street name and suburb may identify duplicate entries. Equally, grouping by street name and suburb may identify duplicate entries with misspelt names (“John Smith” and “Iohn Smith”) where “Smith” is a pattern in a subset of a data field.

Cleaning using pattern recognition is often required for non-standard types of data, but requires the involvement of a human being (usually an expert) because many errors and inconsistencies cannot be resolved automatically [21]. Manual editing of identified errors may also be required but is usually expensive.

Data-based probabilistic matching. Probabilistic matching is where pre-defined rules are used to estimate the probability that ‘objects’ from one datasets corresponds to the same ‘object’ in another. It is also known as ‘fuzzy matching’. Much research and many commercial tools have focused on matching text-based data, typically in the form of personal names and property addresses. Methods of matching text-based data include using ‘wild cards’, ‘character frequency’, ‘edit distance’, keyboard or typewriter ‘distance’ and ‘phonetic similarity’, see for example [27].

Probabilistic matching was used and evaluated by to match names and addresses. With a ‘clean’ training dataset they were able to ‘tune’ the matching algorithm to allow 0.4% false positives in this dataset and thereby estimate the number of false positives in the test data. The Australian Bureau of Statistics’ (ABS) ‘gold standard’ is defined as a match rate between datasets of 80-90% with an acceptable false positive match rate of at most 3% [28]. Dataset matching is checked manually on a subset of the data to ensure this standard is met.

Cleaning efficiency. The efficiency of different cleaning algorithms is important because large datasets and poor programming may mean that some cleaning methods take days to complete [20]. On the difficulty of efficiently identifying and repairing erroneous data, Missier *et al.* [17] suggest that ad-hoc solutions were often sought to balance the computational load and the predictive accuracy. No complete methodology to clean data is so far possible. This work does not consider computational efficiency although it is important for some large datasets.

Duplicated data. An important aspect of joining datasets is the identification of duplicate records that have occurred during merging [29]. A general approach to identifying duplicate records is to cluster similar items together and then compare items in each cluster [24].

Research question

This research aimed to investigate the state of cleanliness of utility-held metered electricity, gas and water consumption data as well as local government-held (administrative) energy and water efficiency programs data. Such data is vital in understanding patterns in energy and water consumption. This work also aimed to identify the typical errors in these datasets as well as tools and methods required to 'clean' them.

The Methods section of this paper describes the dataset employed and outlines the general approach taken to cleaning. In the Results section, general results are presented in tabular form. Detailed examination of the typical data quality issues encountered is also presented through four examples.

METHOD

The approach used in this paper was to clean all the data despite diminishing returns. Known errors were exhaustively rectified. The priority was to minimise discarded data to avoid a reduction of the sample size and to minimise introduced sample biases. Many iterations of cleaning and joining were required before the data could be used for analysis.

Datasets

The datasets used in this work relate to analysing the physical determinants of energy (electricity and gas) and water consumption in residential dwellings in Canberra, Australia [30]. The local power and water utility (ActewAGL) provided quarterly meter readings of electricity, gas and water consumption data for dwellings. Consumption data was requested for 59,781 dwellings out of a population of approximately 200,000 dwellings (see Example 3 below). Dwellings were selected through a stratified random sampling technique [30]. The utility could not match 5,993 requested addresses to their database. Values were based on utility meter consumption readings. Generally readings of all three meters (electricity, gas and water) at a dwelling were taken on the same day. Meter readings for different dwellings were spread over the year. Estimated readings (for example, due to lack of meter access) at some sites were verified at least every year by actual readings. All dwellings had between one and seven years of consumption data (2006-2012). The local Government provided administrative data for all energy and water efficiency programs over the same period (2006-2012) [30]. There were eight water and energy efficiency programs over that period.

Automated and manual cleaning algorithms were written in the 'R programming' language [31].

General approach to cleaning

The general approach to data cleaning and joining involved six steps:

- Data was cleaned using a specially written nearest neighbour spelling match algorithm (fuzzy matching) in combination with a list of known correct data values where possible, e.g. property addresses;
- Known errors, for example addresses that could not be matched to a list of all known addresses, were manually inspected and corrected where possible;
- Duplicate records were identified by generating unique 'keys' using combinations of variables within each dataset. Records that were identical on all criteria were automatically removed. Those that matched on some criteria but not all were manually inspected;

- In joining multiple datasets, only a one-to-one match was performed. Any unmatched values were manually inspected, compared to potential (probabilistically) matching data and corrected where possible, to enable one-to-one matching;
- Previously unidentifiable errors in each individual dataset were uncovered by cross-validating data in the joined datasets. These were either manually repaired or by use of an algorithm. In some cases, it was possible to determine a 'master' dataset, but in many cases it was not. Corrected values were derived from other data within the same data set. For example, the reported length between electricity bills could be verified from the difference between meter reading dates;
- Outlier values were identified. These were only removed after manual inspection, and in the case of energy or water consumption, only after satellite (Google Earth™ or Google Street View™) or physical inspection of the site. For example, commercial use of residential dwellings could be identified in this way.

It is worth noting that errors in measured values, such as energy consumption, were not considered, statistical outliers that were confirmed to be residential dwellings were not removed and missing values were not imputed in this work.

Address cleaning

Dwelling address was an important matching 'key' used for joining some datasets and required special attention during the cleaning process. All address values were first run through a basic address cleaning algorithm.

Extra spaces (white space) were removed and all text converted to upper-case values. Single text-string addresses were broken down into components of unit number, street number, street name and street type and suburb. Some datasets were supplied with the address broken into these components. Each was cleaned independently and in combination with the other components of the address string by comparing those with a list of known possible addresses from the cadastre. Text string 'distance' was used to find the closest match where an exact match could not be found (fuzzy matching). A stringent string distance algorithm was used for matching to minimize false positive matches.

Suburb name values were also cleaned using a distance string text function to identify spelling errors. These types of errors were infrequent. Often, however, the incorrect suburb had been entered. This was discovered by comparing the full address reported to a list of known possible addresses from the cadastre. Such errors could also have been identified by spatially comparing the block and section number for each block and the corresponding suburb from the cadastre with the entered data values.

Abbreviated street types were often misspelt, e.g. CR and CRES abbreviations of 'crescent'. All street types were compared with a list of known possible street types. Many streets were found to have the correct street name but wrong street type. These were repaired automatically where possible. On many occasions, two streets have the same name but different types and were spatially joined e.g. Barry Street and Barry Place. Errors identified in these addresses were repaired by combining the street number with street name to check whether a dwelling existed with that number on both streets. Dwellings with the same street number and name but missing street type could not be differentiated. These addresses were discarded.

Other errors

Errors that could not be cleaned by standard or well defined methods were recorded. Four examples were chosen that show typical errors in these datasets. These four examples contribute to the second aim of this work, to identify the typical errors in these datasets as well as tools and methods required to 'clean' them.

RESULTS

All datasets contained a proportion of errors and missing values. The combination of exact and fuzzy matching successfully cleaned almost all address values and allowed joining of all datasets. Less than 1% of all address values in any of the datasets required human intervention in cleaning.

A more prominent problem was missing data. Table 1 shows an overview of steps in cleaning the metered electricity consumption data. The number of rows in the dataset after each major cleaning step is also shown. The original raw baseline dataset provided by the electricity utility contained 1,006,155 rows. Duplicate rows accounted for 6.3% of the original dataset. Solar generation values were removed from 1,382 dwellings (consumption values were retained). Electricity consumption values that covered the same reporting period (same beginning or end dates) and were for the same dwelling accounted for 20.0% of the original dataset. In these cases, consumption values were added together so each row represented consumption over a single time period. After cleaning there were 604,981 rows, 60.1% of the original dataset.

Table 1. Overview of steps taken to clean electricity data and resulting number of rows after each cleaning stage

Description	Number of rows in the dataset
Electricity only dataset	1,006,155
Negative and corresponding positive values removed	975,149
Zero consumption values removed	911,263
Semi-duplicates removed (duplicated on some but not all rows)	876,951
Houses with unidentifiable solar transactions removed	861,530
Solar generation values removed	806,624
Combine consumption for bills with same billing start dates	606,628
Combine consumption for bills with same billing end dates	604,981

The remainder of this section of the paper outlines the more unusual cleaning problems encountered. The following four examples show typical errors in the data but those that could not be cleaned by standard or well defined methods. The examples reveal missing data, biases in erroneous data, errors generated by the data custodians, cross-validation, classification type errors, duplicate data entries, the effect of different objectives of different data custodians and the effect of poor data constraints (free flowing text) during data collection.

Example 1: Free text string cleaning

Free-flowing text often occurs in datasets because it allows for easy data collection without the need to define ‘valid’ data prior to collection. Free-flowing text is prone to a large number of errors and can be very difficult to use in subsequent analysis because it is irregular. Sophisticated data cleaning methods are required. The following example was derived from the analysis of a government energy efficiency rebate program that ran between July 2005 and June 2013.

The government energy efficiency program comprised 22 eligible residential upgrades for which the government would provide a financial rebate. The aim of analysis was to link the dataset to energy consumption to identify efficiency measures that lead to the greatest energy savings. One of the rebateable items was “Cavity Wall Insulation”.

A variety of probabilistic (fuzzy) joining methods from the OpenRefine™ tool set were used to group similar free text strings. All grouped values had to be manually verified because of the large variety of possible values. Expert experience of the energy efficiency program helped in classifying the data. It would not have been possible to use a pre-defined lookup table.

There were found to be 221 descriptions that appeared to relate to "Cavity Wall Insulation". More unusual examples include: "wal insulation", "wall cavity ibnsulation", "Wall Batts", "Unsulation to external walls", "Rockwall Insulation", "Power circuits for cavity rockwool no: 6416253", "Inwall insulation – upstairs", "Install inwall Cavity insulation", "Electrical modifications for inwall insulation", and "Cavity wall". Figure 1 shows the number of different spellings as a proportion of the database. The most common 28 descriptions made up 80% of the entries in the dataset. Table 2 shows the descriptions of "Cavity Wall Insulation" that occurred multiple times in the dataset. The frequency is shown in brackets.

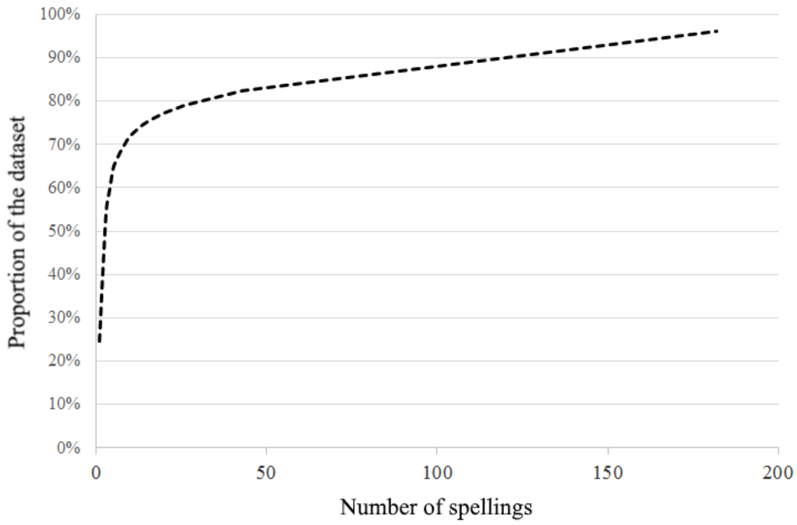

Figure 1. Number of spellings of "Cavity Wall Insulation" as a proportion of the dataset

Table 2. Common descriptions of "Wall Insulation"

Cavity Wall Insulation (248)	Installed Cavity Wall Insulation (3)
Cavity Wall Insulation (166)	Rockwool Cavity Wall Insulation (3)
Cavity Wall Insulation (140)	Rockwool to External Walls (3)
Install Cavity Wall Insulation (49)	Wall & Ceiling Insulation* (3)
Wall Insulation (49)	Cavity Wall Insulation (2)
Wall Insulation (20)	Cavity Wall Installation (2)
Wall Insulation (17)	Ceiling and Wall Insulation* (2)
Cavity Wall Insualtion (14)	Electrical Works for Wall Insulation* (2)
Install Wall Insulation (14)	Install Ceiling and Wall Insulation* (2)
Wall Cavity Insulation (11)	Insulate External Walls (2)
Supply and Install Cavity Wall Insulation (8)	Insulation – Cavity Wall (2)
Wall Cavity Insulation (7)	Rock Wool Wall Insulation (2)
Wall Insualtion (7)	Rockwool Cavity Wall (2)
Install Wall Cavity Insulation (6)	Rockwool Cavity Wall Insulation (2)
Cavity Wall Insulation (5)	Supply and Install Cavity Wall Installation (2)
Rockwool Insulation* (5)	Supply and Install Cavity Wall Insulation (2)
Cavity Wall Insulation to External Walls (4)	Supply and Install Rockwool Cavity Wall Insulation (2)
Ceiling & Wall Insulation* (4)	Supply and Install Wall Insulation (2)
Installation of Cavity Wall Insulation (4)	Supply and Installation of Cavity Wall Insulation (2)
Supply and Install Wall Insulation (4)	Wall Batts (2)
External Cavity Wall Insulation (3)	Wall Insulation to External Walls (2)
Install Cavity Wall Insulation (3)	

Perfect cleaning was not possible. Probabilistic matching failed for description that did not include the word "wall" or "insulation". These included preparation for wall insulation, such as upgrading electrical cabling. Probabilistic matching also failed in cases where descriptions among categories were similar. For example, "insulation", could not be attributed to one of the 27 rebateable categories because it was not possible to assign it to either 'ceiling insulation', 'wall insulation' or 'floor insulation' categories. Approximately 4% of the data could not be attributed to one of the categories (indicated by '*' in Table 2). Consequently, it was not possible from the information provided to determine the exact number of recipients of wall insulation. It was estimated that 1,011 households received a rebate for "Cavity Wall Insulation".

Example 2: Errors common to manually collected water efficiency program data

The Toilet Replacement Program Part 1, which ran between 2008 and 2010, provided a rebate to dwelling owners who replaced older toilets with the low-flow variety (less than 6 litres per flush). This example shows the biasing of the sample where rigorous data cleaning was not performed. Missing values were common in many variables of the dataset. This reduced the possibility of using many of those variables to understand the change in water consumption that resulted from replacing toilets. Validation of data was possible in cases where multiple instances of the same variable, but from different sources, were provided. This allowed for the identification of previously unknown errors and the repair of known errors in the data.

Missing values. The complete dataset contained 5,393 rows of data. Most variables had some missing (null) values. Missing data was particularly common for some variables, which limited later analysis. Fortunately the 'Intervention Date' had no missing values, but 'Application Date', 'Date Completed' and 'Issue Date' were missing in 91%, 92% and 9% of the data respectively. 'Full Flush' and 'Rebate Amount' were both missing in 91% of the values. A unique meter number was provided for 91% of the data. Only 'Number of Toilets' had sufficient data (> 10%) to be used in water consumption analysis. Missing data was not imputed.

Cross-validation. By cross-validating Address and Block, Section, Division (BSD) dwelling identifiers with a list of known possible addresses and BSD values, 99% of the missing and identifiably incorrect values were repaired. There were 441 rows of data with no street name and 141 with no block number. Missing BSD values could be generated with reference to the 'Address' field. Only four entries had no street number or BSD and could not be identified. These were removed from the dataset.

In some cases, identified erroneous values could only be estimated from other variables. For example, a value of '39090' for the 'Application Date', where most other values were in a day/month/year format, indicated an erroneous value. It was not possible to cross-validate such values, but other variables, such as, 'Intervention Date or 'Completed Date' were used to estimate a correct value. This method provided an estimated month but not a precise day, this was sufficient for analysis.

Custodian-generated errors. In preparing and storing the data, the data custodian had attempted to clean the original dataset. However, this cleaning had introduced new errors into the dataset, some of which were detectible. Two types of custodian-generated errors were evident in this dataset.

The address of each dwelling, in the form of a single string of text, was potentially useful as a 'key' to match to other datasets. However, the address string provided had been generated by the data custodian from the individually collected data records (street number, name and type). These individual data records had not been sufficiently cleaned. The full address therefore contained the combined errors from all those individual components and so it was harder to clean than the individual address components. Instead of attempting to clean the full address string, the individual components of each address were cleaned and joined to provide a new cleaner address text string. Had the individual components of the address string not been provided, the address-cleaning process would have been much more difficult.

The dataset contained a second type of custodian-generated error. It appeared that the BSD values had been generated by the data custodian from the address field. But an error in their algorithm was identified. Street number had been used to identify BSD values without considering the unit number. In large apartment complexes, which spanned several blocks, the BSD values were therefore incorrectly allocated. A new unique identifier was generated by combining the original BSD value with the unit number (UBSD) and comparing it to a master list of ‘valid’ UBSD values from the cadastre. Before cleaning, 67 dwellings could not be matched to the cadastre using UBSD as a unique ‘key’. Of these, 40 were repaired using an algorithm to repair the block number based on the unit number and an additional 14 were repaired manually. Only 13 dwellings were found to be un-matchable to the cadastre.

Bias in erroneous values. It should not be assumed that ‘cleaned’ data is error-free, and where possible it was valuable to quantify the bias of any erroneous values, especially where those values could not be repaired and were removed. A sample can be biased during the cleaning process where errors do not occur randomly.

Errors in the BSD described above were found to occur almost exclusively in high-density dwellings. All but 1 of the 67 dwellings that could not be matched to the cadastre were apartments. Higher-density dwellings (more than 1 dwelling per block) represented only 398 of the 5,393 records in the total dataset. The unmatchable data (67 records) represented 16.8% of high-density dwellings but less than 1% of stand-alone dwellings. Without rigorous cleaning, analysis of water consumption (or savings) in different dwellings would have been highly biased because a greater proportion of higher density dwellings could not be matched to consumption data and would have been removed.

Example 3: Duplicate removal – identifying unique residential dwellings

For the analysis of the determinants of residential energy and water consumption it was important to be able to uniquely identify and categorize each dwelling, such as a free standing house or apartment. Duplicate records had to be removed because duplication of dwellings in the datasets would result in an under-estimation of consumption, or conversely missing dwellings would result in an over-estimation of consumption in the sample and the general population. The requirement to ensure each dwelling was included once and only once in the dataset was more difficult to achieve than anticipated.

In an attempt to create a master list of unique dwellings in Canberra from data easily accessible, the Geocoded National Address File (G-NAF) [32] was combined with the Australian Capital Territory (ACT) Government Cadastral dataset (which identified only parcels of land, not dwellings), the number of electricity records at that address, the counts from the National Census Meshblocks data [33], and site inspection. G-NAF had to be combined with other data sources because it did not differentiate between commercial and residential dwellings.

The G-NAF is the authoritative list of addresses in Australia. It combines cadastral data from territory governments of Australia with registered addresses from the Australian Electoral Commission and Australia Post. G-NAF provided a confidence rating for each address, coded as 0, 1 or 2 according to the number of sources that contain that address, either 1, 2, or 3, respectively. A score of −1 is given to any address that no longer exists in any dataset, but did previously. In February 2012, G-NAF contained 202,511 unique addresses for the ACT (Table 3).

Many errors still exist in G-NAF. This was highlighted during initial construction of the National Broadband Network (NBN), which was required to provide an internet connection to every dwelling (in certain areas). The accuracy of data for detached dwellings was found to be 95%, but only 30% for apartment blocks [34]. G-NAF over-estimates the number of units where the Body Corporate has a mailing address (confidence 0), and under-estimates the number of units where some unit owners do not receive mail or are not enrolled to vote at that address (confidence 0). The number of addresses with a confidence rating of zero was high because all corner blocks were designated an address for both the adjacent roads. The difficulty in identifying high-density apartments resulted in most having a confidence rating of 1.

Table 3. G-NAF Estimated Unique Dwellings in the ACT, Australia

G-NAF score	Occurrences
−1	4,207
0	41,298
1	20,698
2	136,308
Total	202,511

In some cases, G-NAF facilitated the identification of dwellings not reported in the cadastre. Unofficial dwellings could be identified where residents installed a mail box and received mail to that address or enrolled in the national election register even if that address did not have a separate land title.

Cadastral data for ACT provided by the Territory Government in February 2012 contained 142,334 blocks, of which 19,199, were duplicates according to their ‘unique’ identifiers. These apparent duplicate records were mostly apartment blocks or corner blocks, which had a designated address for each of the adjacent roads on which they were located. The number of blocks in the cadastre was lower than the number of addresses in G-NAF because each block may have multiple dwellings (e.g. apartments).

The cadastre also under-estimated the number of dwellings because of unofficial dwellings, such as second residences (“granny flats”). Such an error is biased towards older parts of the city where larger block sizes allow for second dwellings to be constructed, or where separated garages had been converted into dwellings.

It was difficult to identify all commercial buildings from the cadastre, especially in high density multi-use buildings, which were typically commercial on the first 1-2 floors and residential above. In these cases, a site visit was often necessary to differentiate the two.

By combing G-NAF with the cadastre, a more accurate identification of unique residences could be achieved. The number of electricity account holders for each block and a site inspection were combined to improve accuracy. Census Meshblock counts were also used because in high density apartments a single apartment block corresponds to a single Census Meshblock. The Census provided an estimate of dwellings in that apartment block.

G-NAF, cadastre and Census Meshblocks were spatially joined to match the unique BSD identifiers of the cadastre with the unique address identification value of G-NAF and the unique Meshblock ID’s. A new unique ‘key’ was developed to help identify unique residential dwellings. A unit number (U) was prefixed to each BSD.

The UBSD was particularly useful for identifying duplicated corner blocks within the cadastre. Corner blocks were duplicated on the UBSD ‘key’ but only 1 of those addresses had a confidence rating of 1 or 2 in G-NAF. The entry with the lowest confidence rating was removed.

It was not possible to rely on a single data source in determining the number of dwellings in medium and high density developments. Table 4 **Error! Reference source**

not found.shows the differences in dwelling counts from the 5 sources of data for two examples of medium density developments. Block P1 was found to have 801 addresses in G-NAF, 242 addresses in the cadastre, 241 according to the 2011 census Meshblock and 81 identified by the electricity utility based on the number of account holders at that address. The exact number of unique dwellings was confirmed as 242 only by a site visit. Block B1 Table 4 in was found to have 118 addresses in G-NAF, 1 address in the cadastre, 38 estimated from the 2011 census Meshblock data and 0 identified by the electricity utility. A site visit revealed 58 dwellings.

Where a list of all possible addresses is required, for example for mailed surveys, G-NAF may be well suited, as incorrectly labelled correspondence will be returned to the sender. Where an accurate count of total residential dwellings is required, then no single data source could be relied on. Without a site visit, the correct number of dwellings in these two examples would have been impossible to estimate, and not available to an automated computer algorithm.

Table 4. Example showing dwelling counts using four methods to identify the total number of dwellings at each development

Block Identifier	B1	P1
Description	4-5 storey apartment complex (built in 2004)	Dispersed large 1-2 storey development (built in 1970)
G-NAF dwelling count	118[b]	801[c]
Cadastre dwelling count	1	242
ABS census dwelling count (Meshblock)	38[a]	241
Number of electricity accounts	0[d]	81
Site visit count	58	242

[a] ABS Meshblock dwelling count also included an adjacent apartment complex. Number of dwellings was estimated as a proportion of the total Meshblock count: total site visit count was 143 for both apartment blocks and 94 in the ABS Meshblock count.

[b] Confidence of 1 = 58, Confidence of 0 = 60.

[c] Confidence of 1 = 1, Confidence of 0 = 560, Confidence of 0 = 2, Confidence of −1 = 240, 560 had no unit number.

[d] Electricity provider could not match any addresses to their billing system. Dwellings had electricity connected on inspection.

Example 4: Differing custodian objectives

Database custodians often have different objectives to the researcher analysing their data, and these differences may result in different definitions of ‘clean’ data. An energy or water utility, for example, may require very little information to bill their customers. This may include a postal address to receive the bill and a functioning electricity, gas or water meter to measure their consumption. For the utility, even the address of the premises may not have to be perfectly accurate. As long as the bill arrived to a paying customer, the address file is considered sufficiently ‘clean’ for their purpose. This may occur, for example, with unapproved ‘granny flat’ dwellings where there may be a separate letter box to receive bills, even though the dwelling is not recognised by the government-held cadastre. This makes linking to other datasets often impossible. Similarly, a separate letter box may not exist even though the ‘granny flat’ is separately metered. To join the bill data to other data relating to the dwelling, such addressing issues must be resolved.

Small errors in the dates when water was consumed are largely inconsequential (as long as the customer does not complain). The only effect will be to allocated consumption to the wrong bill period, but over several bills, the total correct payment should occur. However, such misallocation of the time of the consumption will make analysis that includes a time-series component more dubious.

Not only may the billing dates be incorrect, but zero or negative consumption values exist in the billing ‘current file’. These exist for a variety of administrative reasons, but commonly they correspond to bill reversals. In extreme cases, up to 100 consumption values, many negative, positive or zero may exist even though a single consumption value would be sufficient to record water consumption for a single 3-month billing period. For the researcher interested in using such bill data, cleaning of the data is required. Negative consumption cannot simply be removed as there is often a corresponding positive value. Thus, a cleaning process must attempt to identify the positive value associated with the negative value and remove this pair of data points. Duplicated negative and positive values were also found and removed. Finally, any remaining negative values were subtracted from the positive values during that billing period, and all final tallies were positive amounts.

Customers with multiple values for each billing period consistently had such errors, while other customers were largely error-free. Ignoring the errors would introduce a new sample bias and therefore they must be cleaned by the astute researcher. The bias was not further quantified.

A number of papers state that corrections should be sent back to source [17, 19, 20]. But this may not be wise where the data is only considered ‘dirty’ by the researcher and not by the data custodian. The cleanliness of the data depends on its application.

DISCUSSION

The high matching rate achieved with address data is likely due to the relatively young age of the city of Canberra, the city was created in the year 1911 and most construction occurred after 1950 resulting in a well-defined address format. Furthermore, the addresses in the datasets were all believed to have electricity and water supply connected, reducing the number of unusual addresses, such as caravans.

The current literature provided valuable tools and methods for cleaning certain types of erroneous data, such as addresses. However, most erroneous values could not be cleaned with automated tools. The four examples in this work show the typical errors encountered in these datasets. They revealed examples of missing data, biases in erroneous data, errors generated by the data custodians, cross-validation, classification type errors, duplicate data entries, the effect of different objectives of different data custodians and the effect of poor data constraints (free flowing text) during data collection. A large human effort was required to clean all the datasets. The lack of public information on the cleanliness of such datasets is an impediment to those who wish to use them. This finding matches the findings of Bellini *et al.* [11] in their study of administrative data in Italy.

This work will be useful to practitioners who wish to use privately-held electricity, gas or water consumption data or government-held data derived from their energy and water efficiency programs.

CONCLUSIONS

The use of large datasets (‘Big Data’) may facilitate an understanding and reduction of urban resource consumption and improve the sustainability of cities but only where such datasets have been sufficiently ‘cleaned’ and are well structured to allow interrogation.

This work found that utility-held metered electricity, gas and water consumption data as well as local government-held (administrative) energy and water efficiency programs data required significant cleaning.

Some of the typical errors in the data could be cleaned using well defined tools, such as fuzzy matching of text strings. Many, however, were unique to the datasets involved.

The four examples in this paper demonstrate these common errors. Common errors included missing data, biases in erroneous data, errors generated by the data custodians, duplicate entries, the effect of different custodian objectives and of poor data constraints (free flowing text).

It was found that data considered by the custodian to be 'clean' may not be sufficiently 'clean' to allow interrogation by researchers. Complete cleaning of some datasets was not be possible, the identification of all dwellings retrofitted with ceiling insulation, for example, was not possible because of the rich but ambiguous free-flowing text data type employed in record keeping. Similarly, even with four of the most detailed datasets in Australia, the identification of 'every' dwelling within the city of Canberra remains an intractable problem, currently still solved only with a site visit.

ACKNOWLEDGMENT

This work was supported by the ACT Government Environment and Planning Directorate. The author is grateful to Patrick Troy for his invaluable comments and critique, Nadia Talent for her editing and the anonymous reviews for their constructive feedback.

REFERENCES

1. United Nations, World Urbanization Prospects: The 2018 Revision (P.D. Department of Economic and Social Affairs, ed.), New York, USA, 2018.
2. Tanverakul, S. A. and Lee, J., Decadal Review of Residential Water Demand Analysis from a Practical Perspective, *Water Practice and Technology*, Vol. 11, No. 2, pp 433-447, 2016, https://doi.org/10.2166/wpt.2016.050
3. Podgornik, A., Sucic, B. and Urosevic, L., The Concept of an Interactive Platform for Real Time Energy Consumption Analysis in a Complex Urban Environment, *Journal of Sustainable Development of Energy, Water and Environment Systems*, Vol. 3, No. 1, pp 79-94, 2015, https://doi.org/10.13044/j.sdewes.2015.03.0006
4. Morik, K., Bhaduri, K. and Kargupta, H., Introduction to Data Mining for Sustainability, *Data Mining and Knowledge Discovery*, Vol. 24, No. 2, pp 311-324, 2012, https://doi.org/10.1007/s10618-011-0239-5
5. Batty, M., Big Data, Smart Cities and City Planning, *Dialogues in Human Geography*, Vol. 3, No. 3, pp 274-279, 2013, https://doi.org/10.1177/2043820613513390
6. Al Nuaimi, E., Al Neyadi, H., Mohamed, N. and Al-Jaroodi, J., Applications of Big Data to Smart Cities, *Journal of Internet Services and Applications*, Vol. 6, No. 1, p 25, 2015, https://doi.org/10.1186/s13174-015-0041-5
7. Lässig, J., *Sustainable Development and Computing—An Introduction (Computational Sustainability)*, pp 1-12, Springer, Berlin, Germany, 2016, https://doi.org/10.1007/978-3-319-31858-5
8. Newman, P. G. and Kenworthy, J. R., *Cities and Automobile Dependence: An International Sourcebook*, Gower Technical, Aldershot, Hants, UK, 1989.
9. Beal, C. and Stewart, R. A., South East Queensland Residential End Use Study, Urban Water Security Research Alliance, Technical Report, Brisbane, Australia, 2011.
10. Troy, P. N., Holloway, D. and Randolph, W., Water Use and the Built Environment: Patterns of Water Consumption in Sydney, City Futures Research Centre, Sydney, Australia, 2005.
11. Bellini, P., Benigni, M., Billero, R., Nesi, P. and Rauch, N., Km4City Ontology Building vs Data Harvesting and Cleaning for Smart-city Services, *Journal of Visual Languages & Computing*, Vol. 25, No. 6, pp 827-839, 2014, https://doi.org/10.1016/j.jvlc.2014.10.023

12. Kona, A., Melica, G., Koffi Lefeivre, B., Iancu, A., Zancanella, P., Rivas Calvete, S., Bertoldi, P., Janssens-Maenhout, G. and Monforti-Ferrario, F., Covenant of Mayors: Greenhouse Gas Emissions Achievements and Projections, EUR – Scientific and Technical Research Reports, 2016.
13. CoE on Data Warehousing, The S-DWH Design Manual (Methodology: Data Cleaning), EC, Centre of Excellence on Data Warehousing, 2017.
14. INSPIRE – Infrastructure for Spatial Information in Europe, D2.8.III.6 Data Specification on Utility and Government Services – Technical Guidelines, European Commission Joint Research Centre, 2013.
15. Gianniou, P., Liu, X., Heller, A., Nielsen, P. S. and Rode, C., Clustering-based Analysis for Residential District Heating Data, *Energy Conversion and Management*, Vol. 165, pp 840-850, 2018, https://doi.org/10.1016/j.enconman.2018.03.015
16. McLoughlin, F., Duffy, A. and Conlon, M., A Clustering Approach to Domestic Electricity Load Profile Characterisation using Smart Metering Data, *Applied Energy*, Vol. 141, pp 190-199, 2015, https://doi.org/10.1016/j.apenergy.2014.12.039
17. Missier, P., Lalk, G., Verykios, V., Grillo, F., Lorusso, T. and Angeletti, P., Improving Data Quality in Practice: A Case Study in the Italian Public Administration, *Distributed and Parallel Databases*, Vol. 13, No. 2, pp 135-160, 2003, https://doi.org/10.1023/A:1021548024224
18. Lee, M. L., Lu, H., Ling, T. W. and Ko, Y. T., *Cleansing Data for Mining and Warehousing* (*Database and Expert Systems Applications*) (Bench-Capon, T. J. M., Soda, G. and Tjoa, A. M., eds.), Springer, Berlin, Heidelberg, Germany, pp 751-760, 1999, https://doi.org/10.1007/3-540-48309-8_70
19. Rahm, E. and Do, H. H., Data Cleaning: Problems and Current Approaches, *IEEE Data Eng. Bull.*, Vol. 23, No. 4, pp 3-13, 2000.
20. Quass, D., A Framework for Research in Data Cleaning, Unpublished Manuscript, Brigham Young University, Provo, Utah, USA, 1999.
21. Galhardas, H., Florescu, D., Shasha, D. and Simon, E., Declaratively Cleaning your Data using AJAX (Journees Bases de Donnees), Citeseer, France, 2000.
22. Hankins, M. L., Cleansing Emerges as Trend in Data Warehouse Efforts, 1999, http://www.afcea.org/content/?q=node/932, [Accessed: 17-January-2018]
23. Raman, V. and Hellerstein, J. M., An Interactive Framework for Data Cleaning, Computer Science Division, University of California, Berkeley, California, USA, 2000.
24. Hernández, M. A. and Stolfo, S. J., Real-world Data is Dirty: Data Cleansing and the Merge/purge Problem, *Data Mining and Knowledge Discovery*, Vol. 2, No. 1, pp 9-37, 1998, https://doi.org/10.1023/A:1009761603038
25. Chaudhuri, S., Ganjam, K., Ganti, V. and Motwani, R., Robust and Efficient Fuzzy Match for Online Data Cleaning, *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pp 313-324, San Diego, California, USA, 2003, https://doi.org/10.1145/872757.872796
26. Guyon, I., Matic, N. and Vapnik, V., Discovering Informative Patterns and Data Cleaning, AAAI Technical Report WS-94-03, 1996.
27. Monge, A. E. and Elkan, C., The Field Matching Problem: Algorithms and Applications, *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp 267-270, Portland, Oregon, USA, 1996.
28. Australian Bureau of Statistics, Assessing the Quality of Linking School Enrolment Records to 2011 Census Data: Deterministic Linkage Methods, Research Paper 1351.0.55.045, Canberra, Australia, 2013.
29. Evermann, J., An Exploratory Study of Database Integration Processes, Knowledge and Data Engineering, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 1, pp 99-115, 2008.

30. Talent, M., Troy, P. and Dovers, S., Canberra Residential Energy and Water Consumption Baseline Report, Australian National University, Canberra, Australia, 2013.
31. R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015.
32. Paull, D., A Geocoded National Address File for Australia: The G-NAF What, Why, Who and When, PSMA Australia Limited, Griffith, ACT, Australia, 2003.
33. Harper, P., Information Paper, Draft Mesh Blocks Australia, Australian Bureau of Statistics, Canberra, Australia, 2005.
34. Hutchinson, J., NBN Co Corrects National Address Database, 2012, http://www.itnews.com.au/News/304064,nbn-co-corrects-national-address-database.aspx, [Accessed: 19-August-2018]

Paper submitted: 13.03.2018
Paper revised: 21.08.2018
Paper accepted: 01.09.2018