Citation: Toussaint, Wiebke (2020). Domestic Electrical Load (DEL) Study datasets for South Africa [Data Descriptor]. University of Cape Town, Energy Research Centre. Preprint. https://doi.org/10.25375/uct.11774691.v1

The Domestic Electrical Load (DEL) Study Datasets for South Africa

Wiebke Toussaint

January 30, 2020

Energy Research Centre, University of Cape Town

Abstract

The South African Domestic Electrical Load (DEL) study datasets were collected during the National Rationalised Specification Load Research Programme from 1994 to 2014. The datasets contain metered household electricity consumption data, and socio-demographic survey data for a diverse sample population spanning urban, informal and rural environments, five climatic zones, a large spectrum of income groups, newly to long-term electrified households, and different dwelling structures in South Africa and Namibia. The DEL datasets comprise the original SQL database, and five extracted datasets that facilitate data access. These are DEL Metering (5 minute electrical current, Voltage, frequency, real and reactive power), DEL Metering Hourly Data (DEL Metering current data aggregated to one hour), DEL Survey (de-identified socio-demographic surveys), DEL Survey Secure Data (includes personal identifiers and GPS coordinates) and DEL Survey - Key Variables (a harmonisation of DEL Survey data for frequently used variables). The DEL datasets are the largest, longest and richest collection of residential electricity consumption data in Africa. They will be of value to electricity planners, energy researchers and development agencies.

Background & Summary

The National Rationalised Specification (NRS) Load Research Programme is a multi-party, joint academic-public-private research effort that was launched to inform South Africa's electrification strategy. Initiated in 1994, the NRS Load Research Programme aimed to provide inputs towards policy development and technical design guidelines for the domestic electricity distribution business in South Africa. The programme was overseen by the NRS 034 Working Group at Eskom, South Africa's power utility. Under this programme (frequently referred to as the Domestic Load Research Project) a comprehensive data collection effort was designed and managed to collect electricity meter readings and conduct

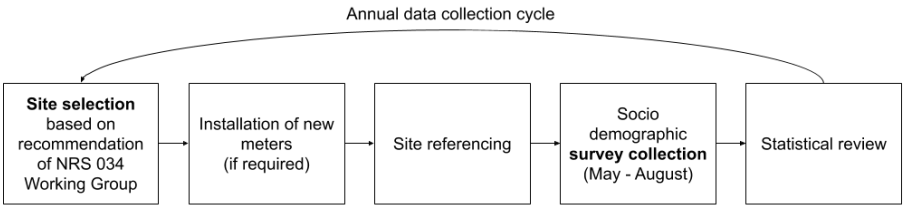

Figure 1: Annual data collection cycle

annual socio-demographic surveys of metered households. Figure 1 shows the annual data collection cycle performed for the duration of the programme.

The data outputs that have resulted from the research programme have been archived and published as the Domestic Electrical Load (DEL) study datasets.The DEL data includes granular electricity meter readings taken at 5-minute intervals and household surveys that capture socio-economic characteristics of metered households and some non-domestic entities in South Africa and Namibia for the period from 1994 to 2014. As the largest and longest study of residential energy consumers in Africa, the DEL datasets provide a unique insight into energy usage behaviour across a diverse demographic of households, spanning multiple climatic zones. The map in Figure 2 visualises the geographic and temporal extent of the DEL collection in South Africa.

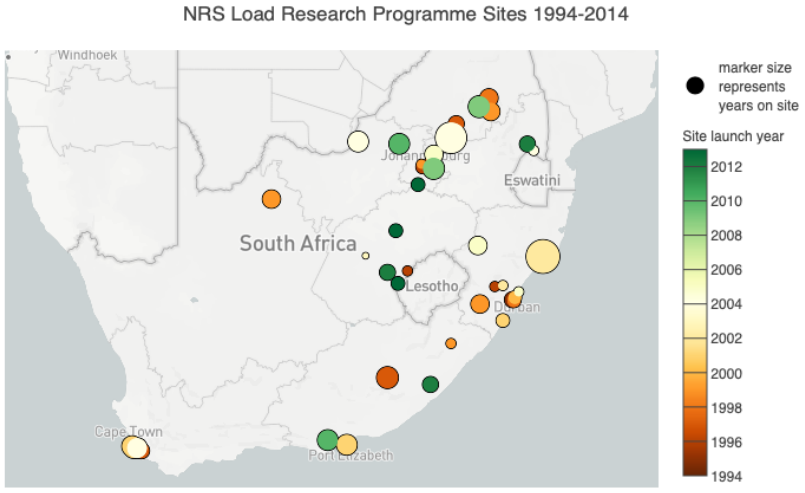

Figure 2: Map view of DEL study data collection sites

Initiated by Dr Ron Herman (Stellenbosch University) and Prof. Trevor Gaunt (University of Cape Town), the study was promoted by the NRS 034 Working Group established within Eskom for this purpose. Early funders and

collaborators included the Department of Minerals and Energy Affairs (now Department of Energy), the Council for Scientific and Industrial Research, as well as Stellenbosch, eThekwini and Nelson Mandela Bay Municipalities. From 1994 to 2009 eight municipalities contributed to data collection. Eskom Research, Testing and Development became actively involved in the study in 1997. From 2001 onwards Eskom was the major data contributor and funder of the study. Prior to 1994, the National Energy Council and Development Bank of Southern Africa funded the development of the data loggers used in the study, as well as early research efforts by Dr Ron Herman and J.J. Kritzinger that influenced the study.

Figures 3 and 4 show the number of households metered and surveyed per year by municipalities and Eskom. On average about 750 households were metered per year, amounting to a total of 14945 household-years metered over the 20 year period. In total 10001 household surveys were collected.

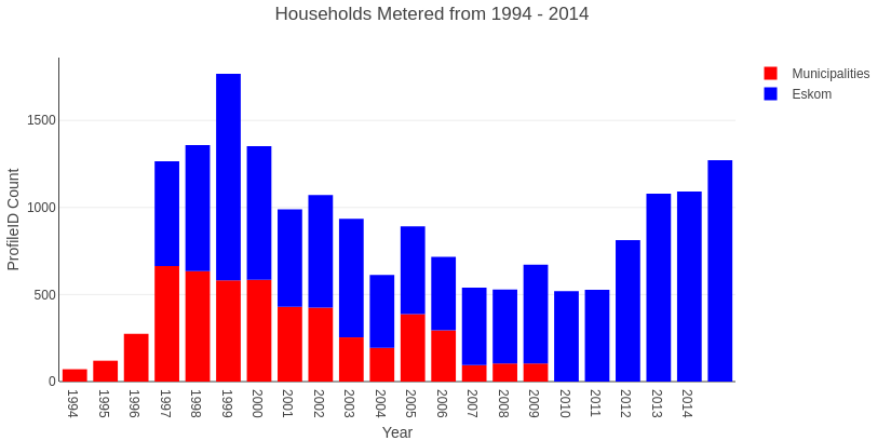

Figure 3: Number of households **metered** per year from 1994 - 2014

The NRS Load Research programme and the DEL study datasets made a major contribution to the electrification of South African households [1]. The research outputs and standards that were developed based on analysis of the data, such as the Hermann-Beta algorithm [2][3][4][5], the Electricity Distribution Guidelines for the Provision of Electricity Distribution in Residential Areas [6] and the South African Geo-based Load Forecasting Standard [7] influenced the design of South Africa's power system. They enabled research to improve design load specifications, to develop new technologies and decision making tools that supported Eskom and municipalities to accurately forecast and right-size new power transmission and distribution infrastructure [8].

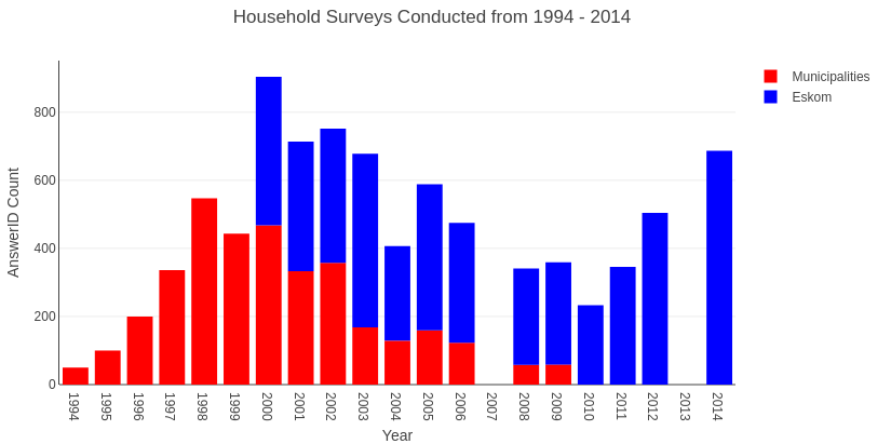

Figure 4: Number of households **surveyed** per year from 1994 - 2014

Methods

The NRS Load Research programme produced a complex and comprehensive collection of datasets. A complete overview of the available data that has been archived in DataFirst's data repository is shown in Figure 5. The annual DEL data collection process consisted of two primary activities. Firstly, electricity meters captured electrical current, power, voltage and frequency readings for a sample of selected households. Secondly, an annual door-to-door survey was conducted to capture detailed household attributes. After rigorous data validation, the original DEL data was captured in a Microsoft SQL database that connects household metering *ProfileID*s to survey *AnswerID*s in a *Link table*. The surveys were updated in 2000 and the data loggers for electricity metering were replaced in 2009. The data in the SQL database is thus not continuous over these years, as encodings changed.

While the intention of the programme was to capture households, early stage data validation indicated that some of the assumed households were actually small businesses, typically shops in informal settlements or townships. Understanding the electricity consumption of clinics and schools was considered important, and some facilities of this nature were also monitored. A small sample of shops, clinics and schools have thus also been captured as non-domestic entities. In addition to the South African households that were monitored, the study captured Namibian households for three years from 2000 to 2002.

From 2000 onwards, a Microsoft Access database is available with GPS coordinates referencing the location of individual households. A geo-spatial csv file contains the GPS coordinates of each measurement site, which includes 60 or more households. A GLR enumeration database is available as a separate Microsoft SQL database to provide the data encodings for survey responses. Each

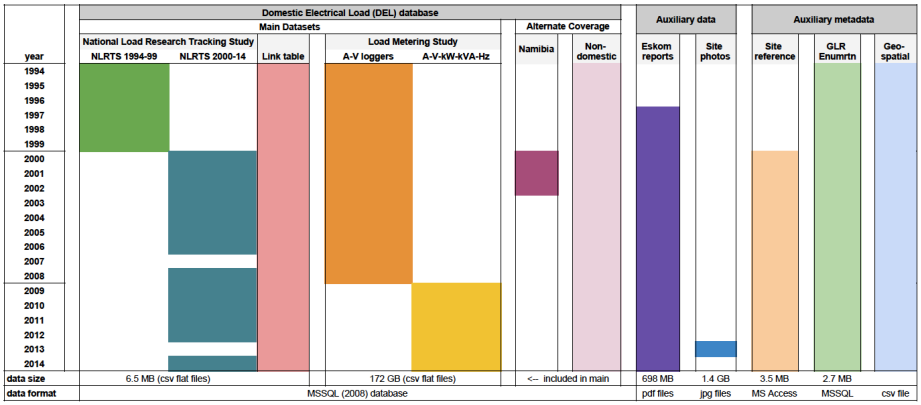

Figure 5: The NRS Load Research programme data

annual survey cycle closed with a report that provides a detailed description of the data collection process for that year. These reports are the property of Eskom and have been released alongside the datasets for the period from 1997 to 2014.

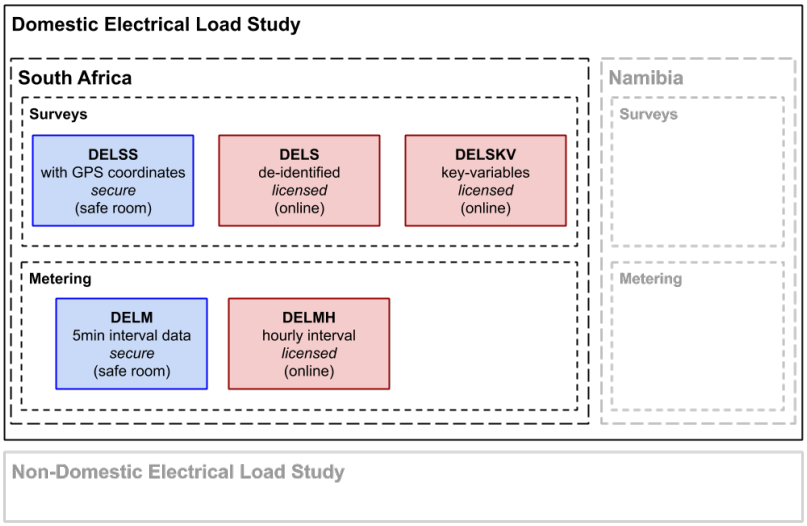

Figure 6: Overview of archived DEL datasets

The original SQL database, including the non-domestic entities and Namibian households, is accessible in DataFirst's safe room. To increase access to the data, 5 datasets have been extracted from the database and released for academic use. Figure 6 shows which data has been included in these 5 datasets. All 5 datasets include only South African, domestic data. Two datasets, DELM

[9] and DELMH [10] datasets are variations of the metered data. The DELSS [11], DELS [12] and DELSKV [13] datasets are variations of the survey data. Blue datasets are accessible in DataFirst's safe room. Red datasets are available online. Grey areas contain data that has not been released in one of the five datasets.

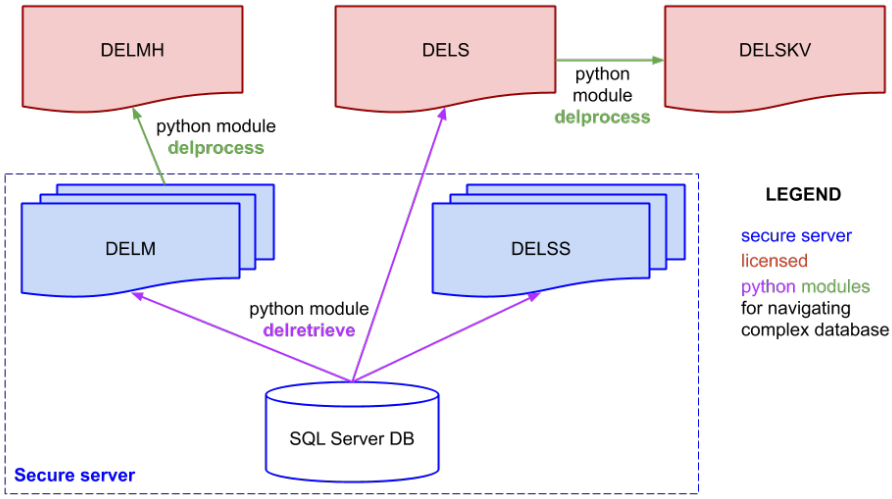

Figure 7: Data model for DEL datasets

The relation between the datasets and the original database is shown in the data model in Figure 7. DELM and DELSS are only accessible in DataFirst's secure server room. Making DELM available online was not possible due to the size of the dataset. DELSS contains sensitive, personal information and is thus only accessible in a controlled environment. DELS is the de-identified version of DELSS. DELMH and DELSKV have been derived from DELM and DELS respectively. The remainder of this section contains details pertaining to data collection and data processing for each of the DEL datasets.

Domestic Electrical Load Metering (DELM)

DELM [9] contains the electricity metering data from the NRS Load Research Programme collected at 5 minute intervals. The selection of a 5 minute metering cadence is important so that the collected data can be used for quality of supply analysis. Evidence for this was presented in early investigations predating the NRS Load Research Programme [2]. Meters installed at households from 1994 to 2008 electricity measured the voltage and current only. From 2009 to 2014 loggers were upgraded and the current, voltage, real and reactive power and power frequency of households were metered.

Start	End	Cycle
1994-06-17	1995-06-26	G1994
1995-06-04	1996-09-10	G1995
1996-02-09	1996-09-19	G1996
1996-08-04	1998-01-24	G1997
1997-12-14	1999-05-17	G1998
1998-10-28	2000-02-12	G1999
1999-09-09	2001-03-14	G2000
2000-10-20	2002-12-11	G2001
2001-11-28	2002-12-23	G2002
2002-10-02	2004-05-17	G2003
2003-12-01	2005-02-08	G2004
2004-08-17	2006-02-23	G2005
2004-07-16	2007-01-05	G2006
2006-10-24	2008-01-31	G2007
2007-04-23	2009-01-29	G2008
2008-11-12	2010-01-19	G2009
2009-11-12	2010-12-15	G2010
2010-12-01	2011-11-29	G2011
2012-01-01	2013-01-09	G2012
2012-11-01	2013-12-31	G2013
2014-01-01	2014-08-31	G2014

Table 1: DELM Data Collection Cycles

Sampling Procedure

The sampling procedure and sample design are described in detail in the NRS National Load Research Project Reports for 1994-1996 and 1997 in the Appendices titled *Load Data Collection Guides*. The reports are included as additional documents. The sample design was reviewed annually and updated from time to time as the need arose. The general characteristics of the sample population and the collection process are described below.

Sample population characteristics. Sampling communities were selected based on the following requirements outlined in programme reports: The target community should have a high degree of electrification, should be stable and willing to co-operate with the project. There should not be many gaps in connectivity. As first-time consumers require a period of adjustment to the use of electrical power, it was assumed that individual load patterns would be erratic for the first two years. Thus "newly electrified" communities should have had access to electricity for at least 24 months before being selected to participate in the study.

Sample size. 70 - 100 consumers (households) were deemed a sufficient sample population for statistically significant load metering.

Sample collection. A random systematic method was suggested and where possible used to select households to be monitored. In general sample selection was optimised to fully utilise data loggers, meaning that loggers were installed on electrical poles that had the most connections so that all logger channels could be utilised. The approach taken at the beginning of the study was as follows:

1. List all the dwelling stand numbers from the township plans.
2. Divide the number of stands by the number of available loggers (call the resulting number sl)
3. Select a random starting point, say at stand sp.
4. Add multiples of sl to sp to give the stand numbers at which to site the loggers.
5. Check (4) to ensure that all or most of the data channels can be used at the point under consideration. If necessary move one pole forward or backward to optimise logger utilisation.
6. Repeat the process until all the loggers have been sited. Meticulous attention must now be given to identifying each monitored dwelling with its logger and channel.

Data Collection and On-site Validation

Data collections from every month were gathered together into a feedback report where any problems with data collection at a site were communicated to the site manager and resolved. Site referencing was done on an annual basis just prior to the winter survey collection process to capture the site 'as is' with minimal likelihood of alteration due to maintenance interventions. During site referencing process the galvanic connectivity between a household or energy customer and the corresponding data logger channels was documented and updated in the database to associate a customer load with the correct questionnaire.

Data Processing

This dataset has been produced by extracting all electricity metering data from the original NRS Load Research SQL database using the $saveRawProfiles$ function from the *delretrieve* python package [14]. Details on using *delretrieve* to extract metering data are provided in the Usage Notes section.

Data extraction and file structure. The $Profiletable$ in the database contains all the measurement data. The data is stored as a profile identifier, $ProfileID$, an associated $Datefield$, a record of the units read and a validity descriptor. The metadata for the observations must be retrieved from the

Profiles table. To manage data volumes, *saveRawProfiles* extracts meter readings in batches and by unit of measure (A, Hz, kVA, kW, V). Data extracts are stored in a hierarchy of csv files arranged by unit of measure and collection year (1994 - 2014).

File naming convention. Extracted csv files were named according to the following convention: collectionYear-collectionMonth_collectionCycle_unit.csv eg. 1996-03_G1993_A.csv
This file contains 5min current (A) readings collected in the year 1996 in March for the collection cycle G1993 (Group 1993).

Missing values. No post-processing was done after data extraction and all database records, including missing values, are stored exactly as retrieved.

Known Issues

For the DELM dataset, the following data collection issues are known:

- **Calibration of voltages and instruments.** Prior to 2009 data loggers were built in-house and only elementary calibration was done (insufficient for commercial standards). After 2009 all loggers were changed to commercial loggers with standard industry calibration of electricity meters.
- **Time Synchronisation.** Meter readings have date and time stamps. Every time data was downloaded from the logger, the meter clock was adjusted to the laptop clock, which was set before going into the field.
- **Logging Errors.** Early logging devices had a 6 week storage capacity. When this capacity was exceeded a "data buffer full" error would occur. Other common modes of technical failure included 'floating' data channels, readings failing to '0' load and readings failing to full scale Amps. A data marking table was generated to validate profile IDs on each day against a set of data quality rules (included as external resource). Based on these rules readings were marked as 'Y' (valid) or 'N' (invalid).
- **Sampling Sufficiency.** Sampling sufficiency was determined by calculating the standard deviation on customer behaviour at the time of annual peak demand (ie 60 or more customers were require to contribute to the annual peak demand, within an acceptable standard deviation).

DELM, Hourly Data (DELMH)

DELMH [10] is an aggregated subset of the 5-minute interval electricity metering data in DELM. The large volume and high metering cadence of the DELM data is unwieldy to access and process. Many applications that do not require highly granular data will be able to extract value more effectively and conveniently

from aggregate values. DELMH contains all current (Amps) observations aggregated to hourly values. It can be merged easily with DELSKV to link socio-demographic variables with household electricity consumption. DELMH and similar custom datasets can be produced from DELM with the python package *delprocess* [15].

Data Processing

This data has been produced by aggregating all current (Amps) metering data from DELM using the $reduceRawProfiles()$ function in *delprocess*. Details on using *delprocess* to aggregate metering data are provided in the Usage Notes section.

The following data processing steps were performed:

- **Invalid readings.** The $Valid$ data value was converted from 'Y' to 1 (valid) and 'N' to 0 (invalid). Missing $Valid$ values were filled with 0 values. See the section on Technical Validation for further details on how validity of readings was determined.
- **Missing values.** Missing readings were treated as per $pandas.dataframe.mean$ default: $skipna = True$; i.e. missing values are excluded when computing aggregates.
- **Data aggregation (study cycles).** Data was aggregated per year, across temporally overlapping study cycles.
- **Data aggregation (observations).** The following processing steps were performed to produce the aggregate dataset:
    1. $Datefield$ values were converted to integer values, rounded to 9 positions left of the decimal, and converted to a numpy datetime64 object with nano-second units. This was done to coerce the data to consistent time intervals.
    2. Readings were grouped by $RecorderID$ and $ProfileID$.
    3. The grouped data was resampled to hourly values (i.e. $Datefield$ column converted to H offset)
    4. The mean meter reading and $Valid$ values were calculated over the resampled intervals.
    5. Rows with all missing values were removed.
    6. The mean $Valid$ value was set to 0 unless it was 1 (i.e. if at least one reading in an hour was marked as invalid, the mean $Valid$ value would be less than 1 and the validity for that hour thus marked as invalid).

Power conversion. To convert the current readings (A) in DELMH to power values use the formula $A * 230/1000 = kWh$. This calculation is an approximation of power consumption, not the actual measured value. Power quality varied across households and the measured voltage was not always stable. For an accurate power calculation the voltage readings from the DELM table should be used. Note that the calculated power using measured voltage and current frequently corresponds with neither the measured real nor reactive power.

Domestic Electrical Load Survey Secure Data (DELSS)

DELSS [11] contains the survey data collected as a component of the NRS Load Research Programme. The survey, initially known as the National Load Research Tracking Study, was undertaken from 1994 to 2014 and covered households in South Africa and Namibia. The questionnaires was updated once in 1999 to incorporate lessons learned from the early study years. DELSS thus contains two subsets of survey responses. Surveys from 1994 to 1999 capture responses to the first questionnaire. Some of the questions were rephrased in the questionnaire used from 2000 onward. Additional questions were also added to the original questionnaire. Survey responses are consistent from 1994 to 1999, and again from 2000 to 2014. A separate survey was created for non-domestic entities (shops, schools and clinics). These surveys are not included here, as DELSS contains exclusively household data.

Sampling

Sampling procedure and response rate. The household surveys were collected to provide socio-demographic, dwelling, and economic information on metered households. The household survey was thus guided by the physical requirements and constraints of domestic electrical load metering and the sampling process described for DELM. Survey enumerators were instructed to obtain at least an 80% response rate within a particular location (suburb or settlement). Revisits would be done until this target was reached and individual homes were revisited up to 3 times.

Deviations from sampling design. Every year before the winter survey collection period, site referencing was done to ensure that data loggers were metering the correct households. In the case where errors were encountered (for example if a logger or wire had been moved and reconnected incorrectly after a routine municipal maintenance activity), the household to be surveyed was updated to correspond with the household that had been metered. Thus the target list for the household surveys was established and validated every year.

Data Collection

Surveys were done every winter between May and August (inclusive). When access was difficult, briefings with a body corporate, estate managers or the

traditional leaders in the area were facilitated to smoothen access to households. The survey was done with a representative household member who provided information about the household. Where possible, this was the head of the family living in the house. Where the family head was not available, this was a resident or visiting family member.

Training of enumerators. A briefing session was conducted prior to field surveys every year. Municipal enumerators did not receive special training , as surveys were usually done on the job.

Interview language and length. Interviews were conducted in the language of the household and translated to English by the enumerator during the survey interaction. Interview length depended on the type of customers and typically took 15 minutes.

Feedback of field teams. Feedback was received after data was captured at every site. The field data collection team was responsible for data capturing and bug fixing based on the output of an automated quality checking process.

Corrective actions. At the end of the winter collection period problems and improvements for the following year were discussed and where possible incorporated. Resampling was not done, but households were visited up to 3 times to service households where no member was at home. No official pilot survey was done, but as the study underwent continuous improvement over its entire duration, early surveys can be viewed as pilots.

Site referencing. Every year the site references were gathered at each site, captured and verified (February - May). This was used to produce the target list for the socio-demographic surveys. Once socio-demographics from surveys were returned, they were used together with the target list to link them to instruments. The link between survey responses and observational profiles in the links table was validated annually.

Quality control during survey collection. Throughout the study during the course of the data collection process data editing of survey responses was done in a number of stages before data entry into the database. Surveys were collected one site at a time, typically with 4 enumerators and one controller. The supervisor ensured standardisation of data collection, manage logistics and liaison and to visually verify all returned surveys on the day of data collection, thus identifying any 'illegal' data entries. Typically the work at each site would take 2 - 3 days to collect up to 60 surveys. Upper management visited every site over a number of years. Thus a portion of sites was visited annually. These visits were planned to coincide with survey collection.

Data Processing

This dataset has been produced by extracting the survey responses from the SQL database using the *saveAnswers* function from the *delretrieve* python package. Details on using *delretrieve* to extract survey data are provided in the Usage Notes section.

Partial de-identification Partial de-identification was done in the process of extracting the data from the SQL database with the delretrieve package. Only the names of respondents and home owners have been removed from the survey responses by replacing responses with an 'a' in the dataset. Documents with full details of the variables that have been anonymised are included as external resources. Other than partial de-identification no post-processing was done and all database records, including missing values, are stored exactly as retrieved.

Known Issues

For the DELSS dataset, the following applies:

- Survey questions and encodings changed in 2000.
- No survey data was collected in 2007 and 2011.
- **Missing surveys.** Eskom got involved in the study in 1997, but Eskom survey data is only available from 2000 onwards.
- Some anecdotal events that could be of interest have been recorded in Eskom Annual Reports.

Domestic Electrical Load Survey (DELS)

DELS [12] is the online version of DELSS, where sensitive data has been removed. In contrast to DELSS, this dataset contains fully anonymised survey responses as well as additional tables that are necessary to interpret the data. Like its secure version, DELS has been retrieved and anonymised from the original SQL database with the python package *delretrieve*.

Data Processing

DELS has been produced by extracting all tables other than the $Profiletable$ from the original NRS Load Research SQL database using the $saveTables$ and *saveAnswers* functions from the *delretrieve* python package. Further details on this are discussed in the Usage Notes section.

De-identification. De-identification was done in the process of extracting the data from the SQL database with the *delretrieve* package. Personal information has been removed from the survey responses by replacing responses with the value 'a' in the dataset. De-identified variables are names of survey respondents and household owners, all street and postal addresses, erf numbers and all telephone numbers. The following variables have been anonymised:

Groups table. The parent-child hierarchy of the groups table was deconstructed into levels and reconstructed as a 4-level multi-index table with the following levels:

1. Level - domestic / non-domestic
2. Level - Survey type : Eskom LR, NRS LR, Namibia [domestic] / Clinics, Shops, Schools [non-domestic]
3. Level - Years
4. Level - Locations

Other than de-identification and reshaping the groups table, no post-processing was done and all database records, including missing values, are stored exactly as retrieved.

DELS - Key Variables (DELSKV)

DELSKV [13] is a harmonisation of DELS. The DELS 1994-2014 questionnaires were changed in 2000. Subsequently survey questions and enumeration vary between the year ranges from 1994-1999 and 2000-2014. This makes data processing complex, as survey responses first need to be associated with their year of collection and corresponding questionnaire before they can be correctly interpreted. DELSKV is a user-friendly version of the original DELS dataset. It contains responses to the most important survey questions, as well as geographic and linking information that allows for the households to be matched to their respective electricity metering data. DELSKV and similar custom datasets can be produced from DELS with the python package *delprocess*.

Data Processing

The *delprocess* python package takes the complexities of DELS into account and makes use of spec files to specify the processing steps that must be performed. To retrieve data for all survey years, two separate spec files are required to process survey response from 1994-1999 and 2000-2014. The spec files used to produce DELSKV are explained in further detail in the Usage Notes section. They can be used as templates for new custom datasets.

Spec files specify the following processing steps:

1. List of search terms for which survey questions will be searched, and variables returned
2. Transformations (addition, subtraction, multiplication) of variables retrieved from search output
3. Bin intervals for variables (requires numeric data)
4. Lables for bins (requires binned data)
5. Details of bin segments

6. Replacement (encoding) of coded variable values
7. Higher level geography detail

In particular, the DELSKV dataset has been produced by specifying the transformations in Table 2 and the replacements in Table 3.

Variable	**Year Range**	**Transformation**
monthly_income	1994 - 1999	variable returned by the **income** search term
monthly_income	2000 - 2014	calculated as the sum of the variables returned by the **earn per month**, **money from small business** and **external** search terms
Appliance numbers	1994 - 1999	count of appliances (no data was collected on broken appliances)
Appliance numbers	2000 - 2014	count of appliances [minus] the count of broken appliances (except for TV which included no information on broken appliances)
total_adults (new)	1994 - 2014	sum of the number of all occupants (male and female) **over 16** years old
total_children (new)	1994 - 2014	sum of the number of all occupants (male and female) **under 16** years old
total_pensioners (new)	1994 - 2014	sum of the number of pensioners (male and female) over 16 years old
total_unemployed (new)	1994 - 2014	sum of the number of unemployed occupants (male and female) over 16 years old
total_part_time (new)	1994 - 2014	sum of the number of part time employed occupants (male and female) over 16 years old
roof_material, *wall_material*	1994 - 1999	value + 1
water_access	1994 - 1999	4 [minus] the **watersource** value

Table 2: DELSKV data value transformations

Appliance usage	**water_access**	**roof_material & wall_material**
0 = never	1 = nearby river/dam/borehole	1 = IBR/Corr.Iron/Zinc
1 = monthly	2 = block/street taps	2 = Thatch/Grass
2 = weekly	3 = tap in yard	3 = Wood/Masonite board
3 = daily	4 = tap inside house	4 = Brick
		5 = Block
		6 = Plaster
		7 = Concrete
		8 = Tiles
		9 = Plastic
		10 = Asbestos
		11 = Daub/Mud/Clay

Table 3: DELSKV data value replacements

Monthly income was adjusted for inflation by baselining it against values from Statistics South Africa for December 2016. The code for this can be found in *delprocess.surveys.py* lines 346 - 351.

Known Issues

For the DELSKV dataset, the following applies:

- The 2000 - 2014 survey questions contain no variable for 'number of females: 50+', which goes against the pattern of other occupant age categories.
- Spacing in the original questions is irregular and can cause challenges when specifying transformations (eg. 'number of males: 16-24' and 'number of males: 25 - 34', 'part time' and 'parttime').
- Spelling mistakes in the original questions can cause challenges when specifying transformations (eg. 'head emploed part time').
- Appliance usage information was only collected after 2000.
- No binning was done to segment survey responses for this dataset.
- Missing values have not been replaced and are represented as blanks except for imputed columns (*total_adults*, *total_children*, ...) and appliances after 2000, where missing values have been replaced with a 0.

Data Records

The original NRS Load Research SQL database, the DEL datasets and all associated documentation and reports have been archived at DataFirst at the University of Cape Town. Table 4 summarises the data files and other sources used as data input and the processing performed to create the output data files for each dataset. The remainder of this section describes the data records associated with each dataset, and provides further details on the scope of each dataset.

Data input	**Other input**	**Processing**	**Output data files**
MSSQL database		`delretrieve.saveRawProfiles (1994, 2014,"csv")`	DELM
DELM data file hierarchy		`delprocess.loadprofiles. saveReducedProfiles (years, "H", "csv")`	DELMH
MSSQL database	delss_part_anon_var_type_blob delss_part_anon_var_type_char	`delretrieve.saveAnswers (anon=False)`	DELSS
MSSQL database	dels_anon_var_type_blob dels_anon_var_type_char	`delretrieve.saveAnswers()`, `delretrieve.saveTables()`	DELS
DELS data file hierarchy	appliance_00, appliance_94, behaviour_00, dist_base_00, dist_base_94	`delprocess.genS ([specfiles], 1994,2014)`	DELSKV

Table 4: Summary of DEL dataset inputs, processing and outputs

DELM

From 1994 to 2008 only current and voltage were recorded. Up to 16 households could be connected to one data logger. Each household had its own channel

for current readings, and all hoseholds connected to the same logger shared the voltage channel. Data loggers were changed in 2009 to facilitate wireless logging. This brought about a change in the data logging format. From 2009 onwards 13-channel loggers were used to connect up to 3 households. Each household had its own channel for current, voltage, real and reactive power readings, and all households connected to the same logger shared the frequency reading. The units of measure and collection intervals are listed below.

- **Current:** Amperes (A), 5 minute cadence.
- **Voltage:** Volt (V), 5 minute cadence.
- **Real power:** Kilo-Watt-hour (kWh), 5 minute cadence (2009 onwards).
- **Reactive power:** Kilo-Volt-Ampere (kVA), 5 minute cadence (2009 onwards).
- **Frequency:** Hertz (Hz), 5 minute cadence (2009 onwards).

All data records for DELM are listed in Table 5. The file hierarchy and naming of DELM data files follows a strict convention, as described in the methods section. The data files are saved and named by unit of measure and by collection year.

Category	File name	File type	Description
external resources	The NRS Load Research Project 1994-1996	pdf	Consultant's report on the first 2 collection cycles.
	The NRS Load Research Project 1997	pdf	Consultant's report on the third collection cycle.
	The NRSLR Project: Note on data handling and processing	pdf	Documentation describing the setup of data processing and error handling during early collection cycles.
	Analysis of National LR Project: Load and sociodemographic data: 1998	pdf	Analysis of data collected until 1999.
	Power quality rules	xlsx	Data marking rules to identify invalid observations.
data files	naming convention: Year-Month_Cycle_unit	csv	Meter readings in structured file hierarchy.

Table 5: DELM data records

DELMH

DELMH contains only current readings in Amperes (A) from DELM, aggregated over a 60 minute interval. The first daily interval is from 00:00:00 - 00:59:59. The data records for DELMH are listed in Table 6.

Category	File name	File type	Description
data files		csv	Hourly current readings for.
		csv	Hourly current readings for.
		csv	Hourly current readings for.
		csv	Hourly current readings for.

Table 6: DELMH data records

DELSS

The lowest unit of geographic aggregation in DELSS is street address, and household GPS coordinates from 2000 onwards. The GPS coordinates are taken at front gate or door of the dwelling. The GPS data can be obtained through an auxiliary data file, the Site Reference database. Table 7 lists the data records included in the dataset. The scope of DELSS includes the following variables:

Household variables. Household characteristics, occupant demographics of gender and age, education, employment, income, income sources, number of household occupants with income, dwelling ownership, gender of head of household

Dwelling variables. Dwelling characteristics, wall material, roof material, floor area, ceiling, insulation, rooms, water source, out-buildings, circuit breaker, business use

Appliance variables. Appliance characteristics, stove, hotplate, kettle, heater, iron, geyser, washing machine, TV, HiFi radio, lights, fridge, freezer, microwave, tumble dryer, broken appliances, appliance usage frequency

Energy behaviour variables. Cooking, heating, alternative fuels, paraffin, gas, wood, charcoal, coal

Electricity variables. Time electrified, power quality, lights dimming, power outages, power trips

DELS

The scope of DELS includes exactly the same variables as DELSS. However, all personally identifying information has been anonymised. Details on the anonymised variables are in the program files *dels_anon_var_type_...csv* The lowest unit of geographic aggregation is on the suburb or settlement level. Table 8 summarises the data records contained in this dataset.

Category	File name	File type	Description
external resources	Description SiteRef database	pdf	Documentation for MS SQL database
program files	delss_part_anon_var_type_blob	csv	long text variables to anonymise
	delss_part_anon_var_type_char	csv	short text variables to anonymise
data files	answers_blob_part_anonymised	csv	long text survey responses
	answers_char_part_anonymised	csv	short text survey responses
	answers_number_part_anonymised	csv	numeric survey responses
auxiliary data files	SiteRef DB	mdb	GPS site location data
filled questionnaires	NRS Tracking Study Survey Scans	pdf	completed surveys

Table 7: DELSS data records

Category	File name	File type	Description
external resources	National Load Research Tracking Study 1995	pdf	Questionnaire (1994 - 1999)
	National Load Research Tracking Study 2005	pdf	Questionnaire (2000 - 2014)
	Datamap for variables 1994-1999	xlsx	maps variables to data type, question and column number
	Datamap for variables 2000-2014	xlsx	maps variables for 2000 - 2014
program files	dels_anon_var_type_blob	csv	long text variables to anonymise
	dels_anon_var_type_char	csv	short text variables to anonymise
data files	answers_blob_anonymised, answers_char_anonymised, answers_number_anonymised, answers, groups, links, profiles, profilesummary, qconstraints, qdtype, qredundancy, questionaires, questions, recorderinstall	csv	extracted database tables of anonymised survey responses and measurement metadata
auxiliary data files	dels-1994-2014-codebook	bak	MSSQL database backup file with enumeration data

Table 8: DELS data records

DELSKV

DELSKV is derived from DELS and includes a subset of its variables. The lowest unit of geographic aggregation is on the suburb or settlement level. The variables in DELSKV are listed below and a summary table of data records is available in Table 9.

Household variables. Household characteristics, occupant demographics of gender and age, employment, income, number of household occupants with income

Dwelling variables. Dwelling characteristics, wall material, roof material, floor area, water source

Appliance variables. Stove, hotplate, kettle, heater, iron, geyser, washing machine, TV, fridge, freezer, microwave

Energy behaviour variables. Frequency of appliance usage

Electricity variables. Time since electrification, circuit breaker

Category	File name	File type	Description
program files	appliance_94, appliance_00	txt	json formatted spec files for appliance variables
	dist_base_94, dist_base_00	txt	json formatted spec files for household demographic variables
	behaviour_00	txt	json formatted spec file for energy behaviour related variables
data files	DELSKV	csv	harmonised dataset with values for all households

Table 9: DELSKV data records

Technical Validation

The mechanisms for checking data records prior to and after addition to the database during early study cycles are described in detail in the technical document titled *The NRSLR Project: Note on data handling and processing*, which is archived as an additional resource of DELM. The data validation process after database entry was refined and automated. All meter readings were checked against power quality rules, executed as a set of SQL queries. Based on the outcome of this validation process, a 'valid' column was created in the database and populated with 'Y'/'N' values. The power quality rules, descriptions and queries are included as an additional document with the DELM dataset and summarised in the list below.

- Voltage should always be between 47V and 276V
- Maximum value for the mean current should be less than 40A
- Check consistency between kVA, V and I readings
- Check consistency between kVA and kW readings
- Identify and test for missing data

- Identify spikes and floaters
- Identifies time periods in the current channel where the monthly mean is less than 0.2A or the monthly max is less than 0.15A

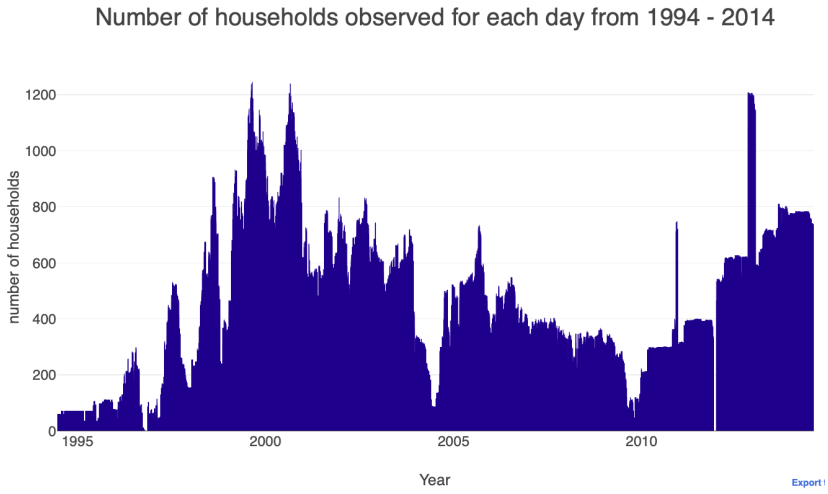

Figure 8: Daily count of valid metered households

Figure 8 visualises the number of households with valid readings captured per day. In 2004 a project was undertaken to match transactions of pre-paid electricity purchases with meter numbers to further validate household energy consumption on a high level. Survey response checking was done manually before and after data was captured in the database. The Data Collection section about the DELSS dataset describes the quality control measures taken during survey collection in detail.

Usage Notes

The DEL datasets are available from DataFirst at the University of Cape Town in Rondebosch, Cape Town, South Africa. Data access can be requested online on DataFirst's data portal. DELMH, DELS and DELSKV can be downloaded once this access has been granted. They have been released under an attribution non-commercial (CC BY-NC) license. Due to the sensitive, personally identifying information contained in DELSS, it can only be accessed on site in DataFirst's safe room. Due to the large size of DELM and associated system constraints, it also requires on site access in DataFirst's safe room.

Code availability

Two python packages have been released publicly on github to facilitate access to the DEL datasets. Both delretrieve and delprocess require python 3 and some associated packages. delretrieve also requires access to an MSSQL Server Database Engine with the original database. Further details are provided below.

delretrieve

This package is available online at `https://github.com/wiebket/delretrieve` [14]. The following python packages are required: *pandas, numpy, pyodbc, feather-format, plotly, pathlib. deltretrieve* was tested with the database file exported in SQL Server Version 8 and restored to a MSSQL2008 Database Engine.

delprocess

This package is available online at `https://github.com/wiebket/delprocess` [15]. The following python packages are required: *pandas, numpy, pyodbc, feather-format, plotly, pathlib, pyshp, shapely. delprocess* furthermore requires access to the hierarchy of csv or feather files extracted from the database with *delretrieve*, or the same files downloaded from the DataFirst, where they are archived. It is important that the naming convention and file hierarchy described in the Readme file of the package are adhered to.

Data processing

The Readme files of the *delretrieve* and *delprocess* packages explain in detail how to extract data from the SQL database and how to aggregate metering data respectively. Below we briefly show how to use *delprocess* to recreate DELMH and DELSKV.

DELMH

command line

```
delprocess_profiles -i H -s 1994 -e 2014 -c
```

python

```
from delprocess.loadprofiles import saveReducedProfiles
for year in range (1994, 2014 + 1):
    saveReducedProfiles(year, interval="H", filetype="csv")
```

DELSKV

command line

```
delprocess_surveys -f "base,appliance,behaviour" -s 1994 -e 2014
```

python

```
import delprocess
genS(spec_files="base,appliance,behaviour",
        year_start=1994, year_end=2014)
```

Spec files

Spec files are the easiest way to create custom datasets, like DELSKV, from DELS. The spec file is a json file that contains a dictionary of lists and dictionaries. All inputs must be strings, with key:value pairs separated by commas. A spec file must contain the following keys:

Key	Value
year_range	list year range for which specs are valid; must be `["1994", "1999"]` or `["2000","2014"]`
features	list of user-defined variable names, eg. `["fridge_freezer","geyser"]`
searchlist	list of database question search terms, eg. `["fridgefreezerNumber" ,"geyserNumber"]`
transform	dict of simple data transformations such as addition. Keys must be one of the variables in the features list, while the transformation variables must come from searchlist, eg. `{"fridge_freezer" : "x['fridgefreezerNumber'] - x['fridgefreezerBroken']"}`
bins	dict of lists specifying bin intervals for numerical data. Keys must be one of the variables in the features list, eg. `{"floor_area" : ["0", "50", "80"]}`
labels	dict of lists specifying bin labels for numerical data. Keys must be one of the variables in the features list, eg. `{"floor_area" : ["0-50", "50-80"]}}`
cut	dict of dicts specifying details of bin segments for numerical data. Keys must be one of the variables in the features list. right indicates whether bins includes the rightmost edge or not. include_lowest indicates whether the first interval should be left-inclusive or not, eg. `{"monthly_income":{"right":"False", "include_lowest":"True"}}`
replace	dict of dicts specifying the coding for replacing feature values. Keys must be one of the variables in the features list, eg. `{"water_access": {"1":"nearby river/dam/borehole"}}`
geo	string specifying geographic location detail (can be `"Municipality","District"` or `"Province"`)

Table 10: Required keys and permitted values for spec files

Full instructions on how to use the spec files to process the data are in the Readme file contained in the *delprocess* package.

Acknowledgements

Started by municipalities with funding from the National Energy Council and the Development Bank of Southern Africa, the two-decade NRS Load Research study was subsequently largely funded by South Africa's power utility, Eskom, with some initial funding input from the Department of Minerals and Energy Affairs. The publishing of the DEL datasets was funded by the South African

National Energy Development Initiative (SANEDI). Academic contributions to the data collection have been led by academics at Stellenbosch University and the University of Cape Town. Other collaborators and implementers include the Centre for Scientific and Industrial Research (CSIR), eight South African municipalities, and consultants.

Special acknowledgement goes to Marcus Dekenah, who managed data collection and sampling, wrote annual reports and reviews that have been archived as external resources, and provided verbal input and explanations on data collection and sampling processes that were invaluable to the writing of this data descriptor. Trevor Gaunt is one of the study initiators and lead many research efforts that emerged from the data collection. His historic accounts of the project provided deep insights into the process of reconstructing the dataset and the research efforts associated with it.

Author contributions

Wiebke Toussaint curated data, wrote software for data access and wrote this manuscript.

Competing interests

The authors declare no conflict of interest in their contributions to this paper.

References

[1] Gaunt, C.T., Herman, R., M. Dekenah, R.L. Sellick, Heunis, S.W. Data collection, load modelling and application to probabilistic analysis for LV domestic electrification. In: Proceedings of CIRED 15th International Conference. Nice: France (1999)

[2] Herman, R., Kritzinger, J.J. The statistical description of grouped domestic electrical load currents. Electric Power Systems Research. 27 (1993) p43-48

[3] Herman, R., Maritz, J.S., Enslin, J.H.R. The analysis of voltage regulation in residential distribution networks using the beta distribution model. Electric Power Systems Research, 29(3), 213-216 (1994).

[4] Herman, R., and J. S. Maritz. "Voltage regulation algorithm for a bi-phase distribution system feeding residential customers using a beta pdf load model." Electric Power Systems Research, 43(2), 77-80 (1997).

[5] Herman, R., Heunis, S.W. General probabilistic voltage drop calculation method for LV distribution networks based on a beta pdf load model. Electric Power Systems Research, 46(1), 45-49 (1998).

[6] South African Bureau of Standards. Electricity Distribution - Guidelines for the provision of Electrical Distribution Networks in Residential Areas Part 1: Planning and design of distribution networks (SANS 507-1:2014). Pretoria: South Africa (2014)

[7] Eskom. Geo-based Load Forecasting Standard (Ref. 34-1284: 2012). Johannesburg: South Africa (2012)

[8] Bekker, B., Eberhard, A., Gaunt, T., Marquard, A. South Africa's rapid electrification programme: Policy, institutional, planning, financing and technical innovations. Energy Policy, 36(8), 3125-3137 (2008).

[9] Eskom, Stellenbosch University, University of Cape Town. Domestic Electrical Load Metering Data 1994-2014 [dataset]. version 1. Johannesburg: Eskom, Cape Town: UCT, Stellenbosch: US [producers], 2014. Cape Town: DataFirst [distributor], 2019. https://doi.org/10.25828/p3k7-r965

[10] Toussaint, W. Domestic Electrical Load Metering, Hourly Data 1994-2014 [dataset]. version 1. Johannesburg: SANEDI [funders]. Cape Town: UCT [producers], 2014. Cape Town: DataFirst [distributor], 2019. https://doi.org/10.25828/56nh-fw77

[11] Eskom, Stellenbosch University, University of Cape Town. Domestic Electrical Load Survey-Secure Data 1994-2014 [dataset]. version 1. Johannesburg: Eskom, Cape Town: UCT, Stellenbosch: US [producers], 2014. Cape Town: DataFirst [distributor], 2019. doi:10.25828/7jtv-mc31

[12] Eskom, Stellenbosch University, University of Cape Town. Domestic Electrical Load Survey 1994-2014 [dataset]. version 1. Johannesburg: Eskom, Cape Town: UCT, Stellenbosch: US [producers], 2014. Cape Town: DataFirst [distributor], 2019. doi:10.25828/kzer-gd88

[13] Toussaint, W. Domestic Electrical Load Survey - Key Variables 1994-2014 [dataset]. version 1. Johannesburg: SANEDI [funders]. Cape Town: UCT Energy Research Centre [producers], 2014. Cape Town: DataFirst [distributor], 2019. https://doi.org/10.25828/mf8s-hh79

[14] Toussaint, W. delretrieve: Data Retrieval of the South African Domestic Electrical Load Study, version 1.01. Zenodo. https://doi.org/10.5281/zenodo.3605425 (2019).

[15] Toussaint, W. delprocess: Data Processing of the South African Domestic Electrical Load Study, version 1.01. Zenodo. https://doi.org/10.5281/zenodo.3605422 (2019).