UNIVERSITY OF LEEDS

MOSL-YW-LIDA Non-Household Benchmarking Project.

Water Use Benchmarking

October 2023

Tamara Garcia del Toro

Gordon Mitchell

Andy Newing

Water Use Benchmarking is a joint project between Yorkshire Water Services Ltd and the University of Leeds' Leeds Institute for Data Analytics (LIDA). It is sponsored by Market Operator Services Limited (MOSL) and funded by the Market Improvement Fund.

The project was conducted through the LIDA Data Science Development Programme at the University of Leeds. We gratefully acknowledge the support of Mitchell Yeoman-Boldry (YW), Liz D'Arcy, and Milo Halford (MOSL).

LIDA Data Scientist	Tamara Garcia del Toro bs20tgdt@leeds.ac.uk
Academic Supervision, School of Geography, University of Leeds	Gordon Mitchell g.mitchell@leeds.ac.uk Andy Newing a.newing@leeds.ac.uk

Abbreviations

CMOS	Central Market Operating System
IDBR	Inter Departmental Business Register
DfE	Department for Education
IQR	Interquartile range
LIDA	Leeds Institute for Data Analytics
MOSL	Market Operator Services Limited
MSOA	Middle Layer Super Output Area
NHH	Non-household
NHS	National Health Service
SIC Code	Standard Industrial Classification Code
SM	Smart Meter
SPID	(Water) Supply Pipe ID
SPIDCORE	Supply Pipe ID Core
SWW	Southwest Water
TW	Thames Water
UPRN	Unique Property Reference Number
YW	Yorkshire Water Services Ltd

Contents

List of Figures

List of Tables

Executive Summary

1. During 2017 the non-household (NHH) water market was reformed, with water utilities becoming wholesalers and the market opened to intermediary retail companies. As the NHH market operator, MOSL seek to ensure smooth market operation (supported by their CMOS database of transactions), and has interest in supporting water demand reduction, through targeted action and advice on water use and leakage.

2. The project was funded by the Market Improvement Fund. The Market Improvement Fund was set up to fund innovative projects that would benefit the non-household water market and its customers. The fund is overseen by the Strategic Panel and administered by MOSL. It was delivered via the LIDA Data Science Development Programme at the University of Leeds. This programme is focused on data analytics primarily using extant datasets, and significant collation or generation of new datasets is outside the normal scope of the programme. The project is a follow on from an earlier LIDA-MOSL data science project (Hulse et al., 2022) and ran for six months (April - Sept 2023).

3. Data anticipated for this project include large volume–low resolution (i.e., coarse) data on metered demand for all customers on the MOSL CMOS database (i.e., all NHH customers nationally, but monthly meter reading), plus a low volume of high resolution (15 minute meter reading) smart meter (SM) data courtesy of three major business water providers (YW, TW, SWW). The analysis draws on pre Covid CMOS data, but the smart meter data is more recent and could be pandemic affected.

4. The project aims and objectives were:

- Water use benchmarking. For the 6 SIC sectors that account for 50% of NHH demand, identify typical water demand (by customer type and activity rate).
- Enhance benchmarking via leakage analysis using smart meter data.
- Develop a standard definition of a leak for use in subsequent analysis.
- Benchmarking report and recommendations.

5. Our work against the above aims and objectives was significantly hampered by data availability. We had access to the CMOS data in month four (of six), allowing to do some data work on benchmarking using national external datasets of business activity. SM data was not accessible until month six when it was too late to proceed with further data processing, hence we were unable to address the leakage objectives.

6. We were able to investigate the potential for linkage with national, all market datasets potentially indicative of customer activity rates. These comprised the ONS Business Structure Database and Inter Departmental Business Register (IDBR)(ONS, 2023) from which we obtained aggregated business counts for employment and turnover size bands at MSOA level, the Department for Education Schools, pupils and their characteristics: January 2022 (DfE, 2023) where we extracted number of staff and pupils as well as school characteristics data, and NHS Digital Hospital Admitted Patient Care Activity, 2021-22 (NHS, 2023) which contained hospital outpatient and admissions for all hospitals in England. We were then able to match these data to their water usage by MSOA and postcode data.

7. We were unable to associate CMOS customers to external activity at an individual ('atomic') level, due to the lack of any common field/variable. The UPRN introduced by ONS should in principle allow such data linkage (UPRN is a field in the CMOS database) but UPRN is not used in any external data sets we explored. A process of geographical analysis using post code matching allowed us to

constrain the data linkage and match small area water use and business activity. At this scale we found no correlation between customer size and CMOS water usage data for any of the business activities explored, for any of the different size metrics used. The reasons for this are uncertain and could relate to poor accuracy in the CMOS data water use data, or possibly that potential correlation is swamped by 'noise' from leakage. These results are clearly not accurate enough, and benchmarking mark should in future be carried out using the SM data.

7. Recommendations for further work are:

- focus benchmarking activity on activities most likely to benefit (high demand/potential for effective benchmarking), which we suggest are (2007 SIC) section G (retail and wholesale), section I (accommodation and food service activities), section E (education), section R (arts, entertainment and recreation);

- carry out customer size and water usage correlation tests using smart meter data. Preferably this should include a focus on c. 30 smart metered customers per SIC sector, where detailed primary data on customers (i.e., not held in national databases) can be obtained directly from them;

- implement further CMOS data cleaning to ensure the accuracy of SIC matching; and

- identify additional sector specific size metrics (e.g., as appear in the Thames Water dashboard) and availability at national level.

1. Introduction

MOSL are seeking to identify how to reduce water demand in the NHH sector, to meet a water efficiency goal set by government. This is addressed principally via: (1) a 'benchmarking' exercise, where they hope to identify customers that have atypically high-water use, and so who are prospects for water conservation interventions / guidance; and (2) identification of leakage, and guidance on leakage reduction strategy. MOSL ultimately want to operate a data dashboard so customers can compare their demand to that of similar customers. This raises problems as to date, there is a lack of individual level customer data that can be linked to their water meter records.

Previous work to realise these aims was carried out in a preceding MOSL-YW-LIDA project (Hulse et al., 2022) who also directed efforts into a leak identification algorithm. This project reported that 6 Standard Industrial Classification (SIC) divisions account for half of non-household water demand:

- Division 1: Crop and animal production, hunting and related service activities.
- Divisions 10 & 11: Manufacture of food products & Manufacture of beverages
- Division 20: Manufacture of chemicals and chemical products
- Divisions 55 & 56: Accommodation & Food and beverage service activities
- Division 85: Education
- Divisions 86, 87 & 88: Human health activities, Residential care activities & Social work activities without accommodation.

The SIC is a code used by the business regulating authorities to describe a type of economic activity. The first classification was released in 1980 and it has since been updated three times, with the most recent classification released in 2007. The hierarchy of the SIC classification varies from one year to another. The 2007 SIC highest category is an alphabetical section, which is then broken down into two-digit divisions, three-digit groups and four-digit classes. The further down the hierarchy, the more specialised the business activity becomes within each section. For this project we initially aimed to use division and group hierarchies, in analysis of both the entire NHH market (CMOS database) and a sample of customers monitored using water smart metering.

2. Project aims and objectives

The project aims and objectives were:

- **Water use benchmarking**. For the 6 SIC sectors that account for 50% of NHH demand, identify typical water demand (by customer type and activity rate). To do this, the tasks were to:
  - collate the SM data and produce statistical distributions of demand. SIC division is a high-level class, and there are lower tiers where this can be repeated to identify demand distribution in the more specific activities;
  - repeat using the CMOS data, to provide a check on the representativeness of the sample of SM records relative to the whole NHH market;

- seek data at the national level that can describe the size / activity rate in these SIC groups (e.g., size distribution of secondary schools based on number of pupils; or of an activity based on number of employees). In assuming water use and activity size/rate are correlated, we identify customers who are above the expected consumption for their size/activity, and so could identify priority candidates for a water efficiency investigation and potential intervention.

- **Enhance benchmarking via leakage analysis**. Using the smart meter data, we:
  - Apply the leak detection algorithm (from Hulse et al., 2022) to estimate leakage rates for different activities by SIC). Determine if some SIC activities are more prone to leakage than others. If an SIC appears to have an above average rate of leakage, this adds a further measure to the analysis above, when considering audit/intervention (i.e., if a customer is smaller than average but an above average consumer *and* is in a SIC group that overall appears to have above average leakage, there is a further rationale to investigate that customer.
- **Benchmarking report and recommendations**. Report on the work above and reflect on requirements to develop a benchmarking capability. Identify barriers and opportunities to developing a national NHH water use dashboard.

- **Leak definition** (secondary objective). The water industry has no standard definition of a leak, so when analysing SM data, the analyst must define the two key parameters (peak flow above night-time minimum, and duration of that flow). Our leak detection algorithm (Hulse et al., 2022) allows SM data to be analysed under different leak definition parameters. There is then the possibility of developing an industry standard definition of a leak, as follows:
  - YW have a zone (DMA) in Hadfield, Sheffield where every property is on a SM (1000 NHH and 1000 households). Aggregate input to the zone is monitored so it is possible to identify whole DMA leakage using the DMA input record.
  - Select a leak period identified in the DMA record and assess for leaks in individual properties in the DMA, with different leak algorithm settings (peak flow/duration parameters). Identify the combination of these settings that returns a leak volume closest to the DMA based estimate of leakage.

3. Benchmarking

3.1. Data

MOSL wish to identify customers that have an abnormal water usage via a benchmarking exercise to understand normal water usage for each business activity according to their size/activity rate and then develop a dashboard that will allow customers to understand their water usage relative to those benchmarks. This process requires (1) identification of national level business size databases with SIC and identifiers; (2) benchmarking using SM data; (3) check the representativeness of the SM data using CMOS data.

3.1.1. National level Datasets

Several datasets were explored to complete this objective. We found that there is still poor use of the UPRN in national public data sets, as well as a lack of individual business data availability due to privacy laws. Therefore, for non-state funded business activities we found aggregate counts at

different geographic levels from the IBDR of turnover and employment band size data for all business activities.

The smallest geographical area available was MSOA which will have a minimum population of 5000 and a mean population of 7200. Aggregates are also rounded to the nearest five in order to avoid reidentification. MSOA level data can be linked to each business activity water usage data using postcode data and calculating an average business size and average water usage per MSOA. For state funded business activities such as Education and Health we found premise specific size data, some of which could be matched to specific premises using postcode data.

Education specific size datasets were obtained from the DfE school census. This data was composed of number of pupils per school, school name and address, number of girls registered, number of boys registered, number of stuff and number of full-time equivalent staff.

3.1.2. Smart meter data

The initial aim for this objective was to create statistical distribution of the smart meter water data to be compared to the statistical distributions of size metrics data. Due to time delays in the arrival of the smart meter data, we were not able to use it to create the distributions, and instead we relied entirely on the CMOS data to obtain daily water usage distributions.

3.1.3. MOSL CMOS data

This dataset consists of premises average daily water consumption (yearly total / 365) as six separate data extracts by region (North, East, Southeast, West, West Country, and Wales) comprised of 1,244,609 unique premises supplemented with ancillary identifiers such as SPID and SPIDCORE, SIC codes, Valuation Office Agency (VOA, the government agency that values properties for the purpose of council tax and business rates in England and Wales) reference numbers, customer/business names, postcodes, and UPRN fields. Ancillary data is only provided for premises with annual consumptions more than 500m^3; premises with annual consumption below this threshold are potentially sole traders and as such only SPID, SIC and average daily consumptions are provided for these premises for confidentiality reasons.

We have used CMOS 2019 average monthly water usage data. Older data is more reliable due to the billing and reporting process (e.g., more recent data may be based on estimated use). Data pre-processing was carried out in order to ensure data was appropriate for use and as error free as possible. Data analysis was conducted using the programming language Python4, and the dataset manipulated and analysed making use of Pandas5, an open-source data analysis and manipulation tool built on top of the Python language, and analogous to Microsoft Excel.

3.2. Data Pre-processing

3.2.1. Filtering the desired business activities

Previous work (Hulse et al., 2022) identified six business activities responsible for 50% of all NHH water demand. For this project we focussed activity on those six types. Hulse used the 1980 SIC classification as this is the classification used for most customers in the CMOS data base (Figure 1).

This introduces inconsistency in the database but creates a more substantive problem in that most external business datasets use the newer SIC 2007 classification. To address these issues, we made some assumptions about the reconciliation of the 1980 and 2007 SICs for a small number of affected customers.

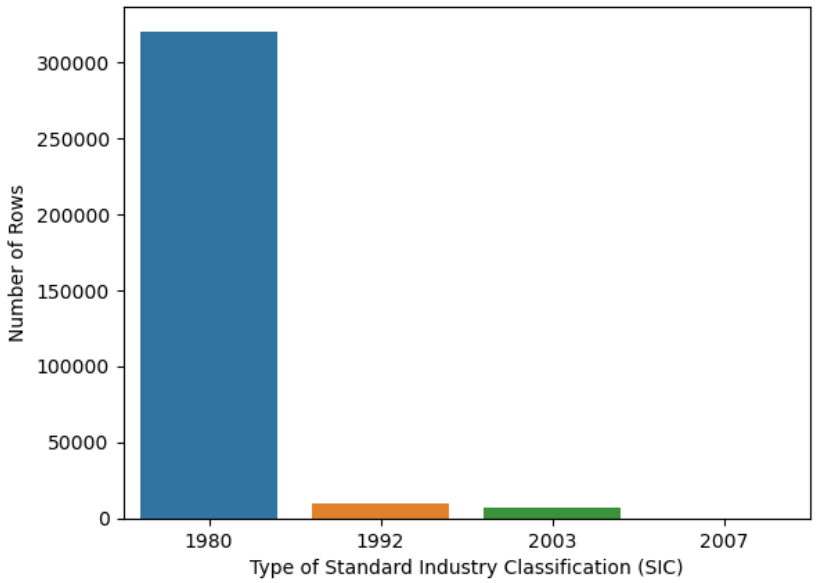

Figure 1**.** *SIC Classification in the CMOS database.*

The CMOS data contains water usage data for all business activities. There are three business activity classifying variables in the data. *SIC type* defines the type of SIC classification used for that row (customer), which can be one of four (1980, 1992, 2003, or 2007). The 2007 SIC Classification is the latest available classification and has significant changes relating to the grouping of different groups compared to the 1980s classification. Next, *SIC Class* identifies a customer's specific business activity and its general type (SIC division). Finally, the *Master Division* variable was created when MOSL attempted to harmonise the division, and this division is the prevailing division name for the row.

Data exploration showed that neither SIC code not Master Division accurately describe the rows and instead there is a large volume of data that is misclassified to the wrong SIC. By way of example, Figure 2 shows that different two-digit divisions are present in the Health Master Division rows (top) and that there is a large presence of non-education SIC and master division labels within the education based on the SIC and master division variables, highlighting the data misclassification issue.

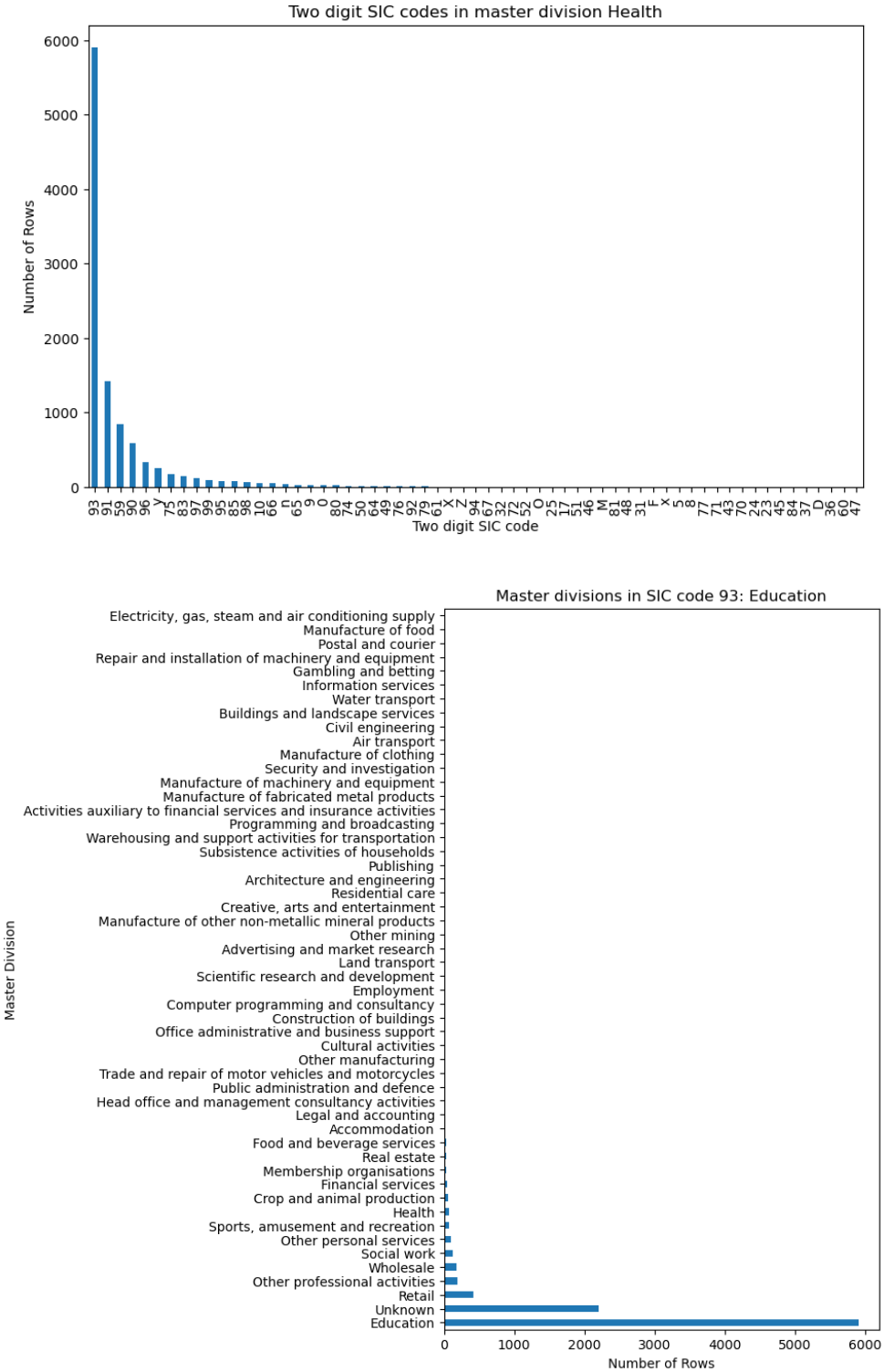

Figure 2. *Data misclassification in Health and Education sectors*

Following initial data exploration, we built a protocol for data cleaning in order to return the cleanest dataset for each SIC. This protocol differed for each SIC and can be seen in Table 1. We retained every row that had the expected Master Division and SIC code for each SIC group and built a separate dataset for each of the six SICs of interest to this project. The reason for this approach was that it returned the cleanest data set per SIC with the lowest time investment. Manually

inspecting the data would result in a much cleaner database for analysis but might take 6 person months or more. Clearly, more accurate customer classification / SIC reconciliation would be useful and could lead to improved quality of analysis as it would allow for more detail analysis at SIC group and class level. We suggest working towards creating additional SIC group and SIC class variables and validating the accuracy of the SIC code in the data.

Table 1. Criteria used during SIC cleaning

Division	**SIC Code**	**CMOS Master Division**
Division 1: Crop and animal production, hunting and related service activities	SIC 10	Crop and Animal Production
Division 10 & 11: Manufacture of food products & Manufacture of beverages	SIC 41	Manufacture of food
	SIC 41	Wholesale
	SIC 42	Manufacture of food
	SIC 42	Manufacture of beverages
	SIC 42	Wholesale
	SIC 42	Food and beverage services
	SIC 41	Manufacture of food
	SIC 41	Wholesale
Division 20: Manufacture of chemicals and chemical products	SIC 25	Manufacture of Chemicals
	SIC25	Wholesale
	SIC 66	Food and beverages services
Divisions 55 & 56: Accommodation & Food and beverage service activities	SIC 66	Retail
Division 85: Education	SIC 93	Education
Divisions 86, 87 & 88: Human health activities, Residential care activities & Social work activities without accommodation.	SIC 95	Health

Note. Table shows inconsistency in CMOS economic activity classification and assumptions used in the study to add CMOS master division data to a coherent activity division.

3.2.2. Duplicate SPIDCORES

We first ensured there were no duplicate SPIDCORES. SPIDCORES are water meter identifiers and should be unique per premise. We found 6746 duplicate SPIDCOREs, with duplicate rows having missing data. We deleted rows with the least amount of information. Figure 3 shows how water volume and number of rows (customers) varies between the raw and cleaned data (duplicates removed). After removing the duplicate data, water volume across the 6 SICs is reduced by 29% whereas only 8% of rows are lost.

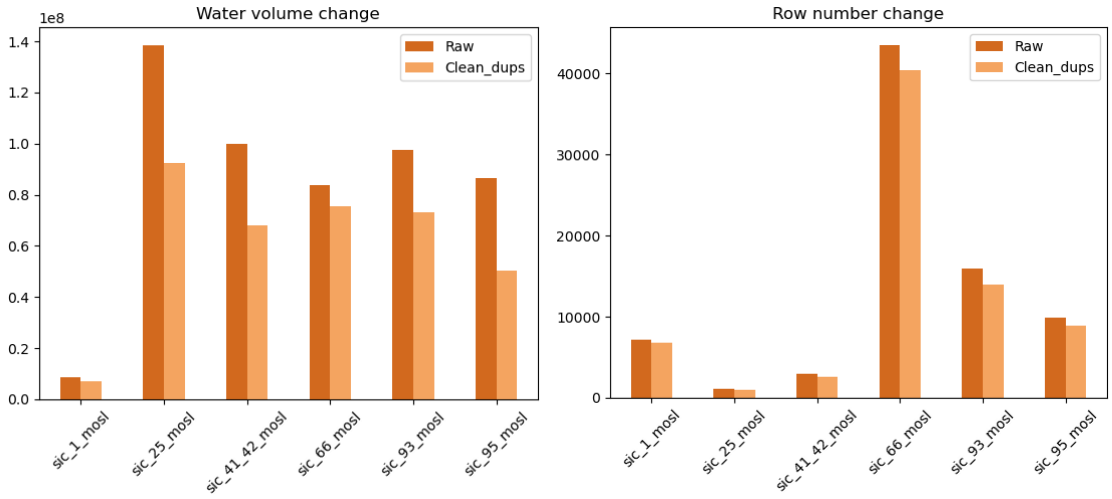

Figure 3. SPIDCORE duplicate processing

3.2.3. Negative and nil water usage

A total of 0.4% of rows are recorded as negative or having nil water usage. These are either due to: an incorrect meter read which is being lower than the previous months read, supplied by retailers to CMOS; or when a meter read is eventually submitted by retailers after a prolonged period of estimations, and a premises has actually consumed less water than estimated, the latter being the most common occurrence. As reads of this type are not reflective of a premises actual consumption behaviour, and may lead to misleading conclusions, any monthly rows of data containing negative volumetric consumptions are excluded from analysis. Within the aforementioned, we remove any rows (monthly consumptions) from the dataset that contain zero (is an estimate) or negative (correction in CMOS) values for actual consumption. Total water volume after negative and 0 water usage removal increases by 0.4%.

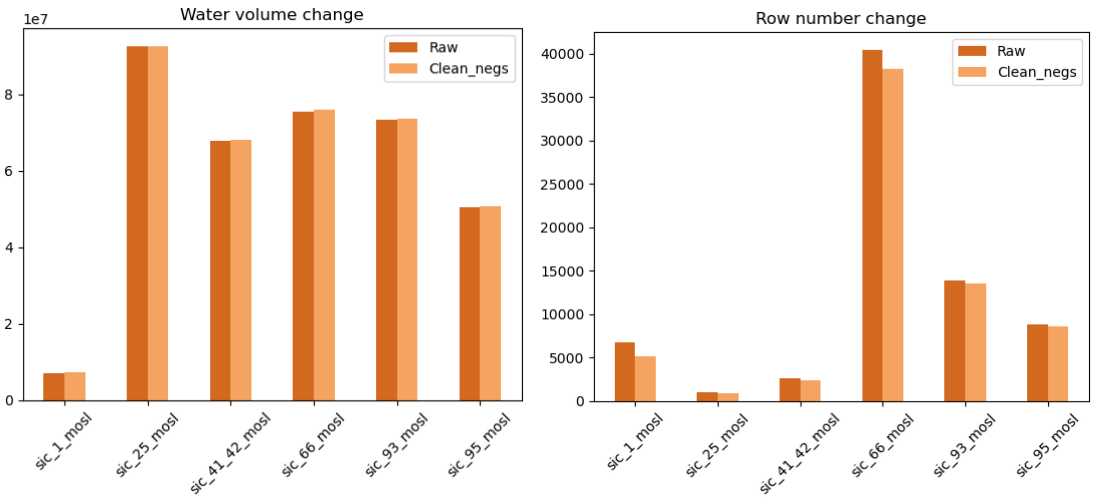

Figure 4. Water volume after cleaning of negative and nil water use.

3.2.4. Large water usage outlier removal

Very large water records are present in the data due two possibilities: recording errors, and due to users, that require a large amount of water. Businesses that require such large amounts of water are likely already aware of their usage or will have been intervened, hence they are not the target population for this work. For each SIC, we remove large water usage outliers by stablishing average usage and then attempting outlier detection.

Three methods were explored for detection of large outliers: z scores for data normalisation, median to mean ration, and interquartile range method. Here we describe the data reduction caused by the IQR method as it offered the cleanest data set. More details on which methods were explored, their theoretical basis and results are available in Appendix 1.

The total decrease in water volume after outlier removal was 77% while percentage of rows (customers) dropped was 11%. These outliers are likely to be both wrong waters use values alongside the larger water users, which are driving the total consumption mean up. Removing these very large water users allows for a finer look at businesses of all sizes, and these users are likely aware of their usage.

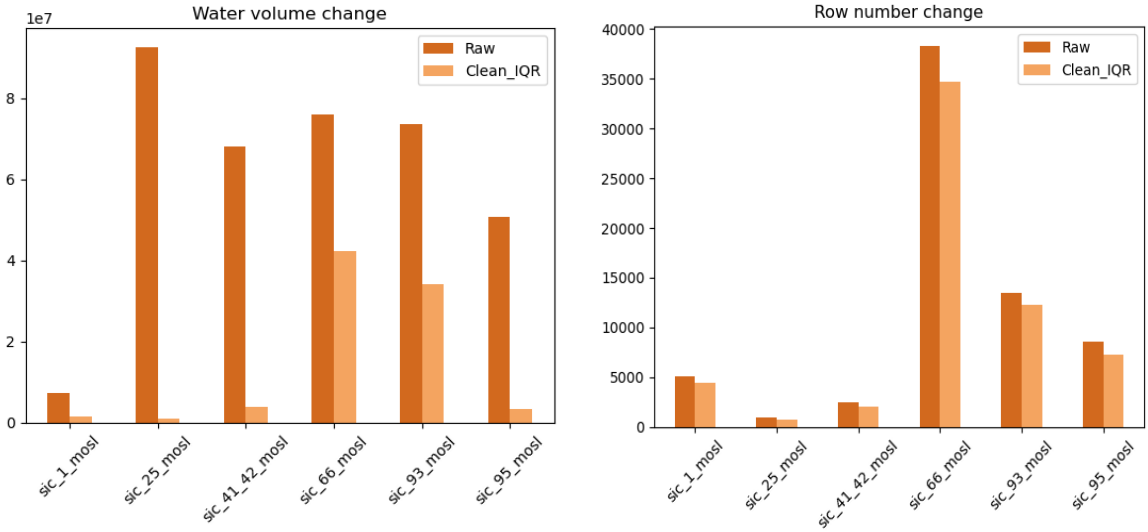

Figure 5. Water volume and rows (customers) pre and post large water use outlier removal

3.2.4. Data Linkage

We explored a number of public datasets to link with the MOSL dataset. However, it was not possible to identify any sufficiently comprehensive publicly available datasets containing UPRNs. Although the ONS UPRN is promoted as a national standard metric by the government, uptake since its launch appears very limited to date, and the UPRN field is not currently consistently used within any open datasets at national level.

We identified aggregate counts for employment and turnover per business activity available at different geographic levels. Whilst we could not link customers directly to external data sets at an individual ('atomic') level (e.g., using UPRN), we were able o place customers into small geographic areas based on their postcode, and hence link to zonal business statistics for the small geographic area. Thus, data linkage was possible using ONS boundary lookups matching companies' postcode to their corresponding 2021 Census MSOA, calculating average water use per SIC by MSOA, and also the average business size by MSOA, as employment and turnover per SIC.

For the education sector, linkage of education size data to water usage data using UPRN was not possible as the DfE uses a unique identifier system for all schools. Instead, education data was also linked to CMOS data using school postcodes. This approach requires a very clean education dataset as multiple premises can share a postcode; however, it is very unlikely than two or more schools will share a postcode. For this reason, this approach was carried out in a very clean set of education business activity data from CMOS.

Health datasets are not UPRN identified and use a specific NHS identifier on all available datasets we could find. Linkage to health data using postcode was not successful, as NHS data is segmented across many datasets that need to be linked together. We were not able to carry out this linkage.

3.3. Comparing Water usage distributions

The initial scope of this project included creating distributions of size and water usage for each business activity to support creating a dashboard that would allow a business to understand their water usage compared to similar business of the same size/activity rate. Smart meter data offers higher granularity and accuracy relative to the CMOS data, hence is the preferred data set for benchmarking. This does assume that the retailer has collected accompanying business activity data on smart metered customers that allows the customer activity to be described (e.g., SIC class, floorspace, number of employees, etc). We do not know what activity data is available alongside smart meter records, and the degree to which this data is standardised (which would support compilation of smart meter records from multiple retailers).

As we were not able to analyses smart meter data, we attempted to compare water use distributions and activity/size distribution using the CMOS data. Essentially, here we are making the assumption that volumetric water uses, and business activity rates are positively correlated. If so, we can then, albeit coarsely, identify customers withing a given SIC, that have above average water use, but are of below average size/activity – these customers are likely potential outliers on which water conservation interventions can be targeted.

For each of the 6 SIC sectors analysed, we converted water usage and customer size estimate variables to z scores to allow comparison on common scales (Figure 6). It is important to recognise that there is no direct (individual level) linkage between company size and the water usage data, therefore we cannot know with certainty if a larger business has a higher water consumption. Therefore, we next further explored these data (and the consumption-size correlation assumption) through a further analysis using geographical (MSOA) data linkage.

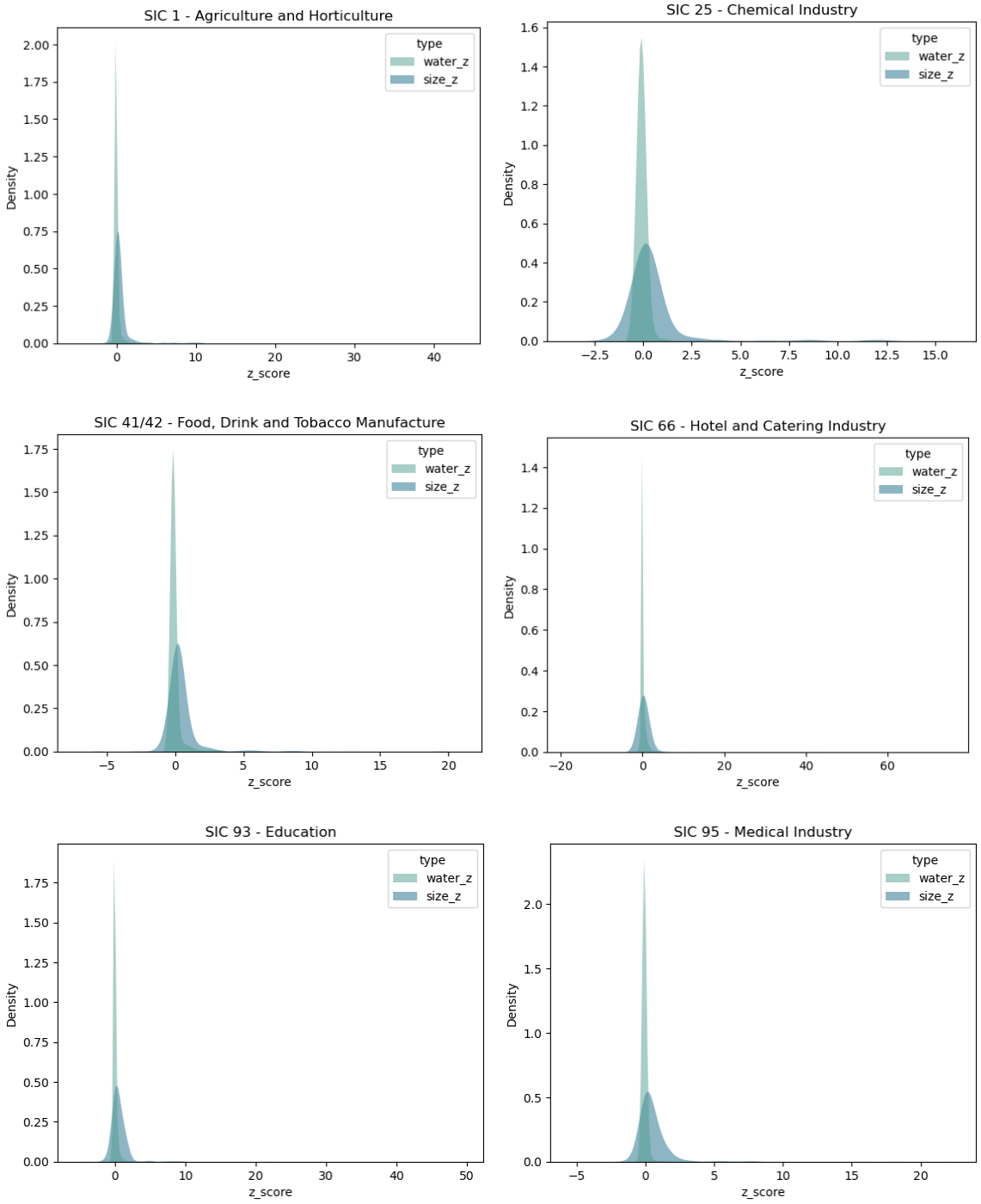

Figure 6. Distribution of Z scores for water usage and business size by SIC group

3.4. Exploring the relationship between water usage and size data

Using the data linkage methods described above, we calculated the average business size per MSOA using both employment size bands and turnover size bands for each business activity division. For each MSOA, we divided water usage by number of businesses to get an average water usage in the MSOA and liked the water usage estimates and the size estimates by their MSOA for each business activity. Usage using Pearson's correlation, we explored the relationship between business size and

water. In order to be able to estimate appropriate water usage for a business of a given size, there must be a relationship between water usage and size. We assumed this relationship would be linear.

The correlations were calculated between water usage and three different size estimates: average employment size, average turnover size, and a computed size estimate which was a product of employment and turnover average size. None of the business activities we have used showed any correlation between any of the size estimates and water usage at MSOA level.

Table 2.MSOA level correlation between water use and activity size metrics by SIC

SIC Division		Employment		Turnover		Size estimate	
		Corr	p	Corr	p	Corr	p
1	Agriculture and Horticulture	n/a	n/a	-0.06	0.49	n/a	n/a
10	Chemical Industry	-0.02	0.73	0.02	0.71	-0.02	0.73
41/42	Food, beverage, & tobacco	0.00	0.91	-0.01	0.71	-0.02	0.68
66	Hotel and catering industry	0.05	0.01	0.05	0.01	0.02	0.22
93	Education	0.01	0.57	-0.01	0.63	-0.00	0.86
95	Healthcare	0.09	0.68	-0.02	0.47	-0.02	0.39

The lack of correlation between water usage and any of these size metrics could be due to several reasons. First, we judge that the CMOS data lacks accuracy relative to smart meter data, due for example, to issues with water use estimation. Therefore, further work on benchmarking and dashboard construction should focus on smart meter analysis. Second, we judge that during our sub-setting of the CMOS data to select only customers from the six main business activities, biases were introduced. The cleaning was done by keeping those rows which were true for conditions related to how they are labelled for both their SIC division and master division, but we saw that this excluded a large number of rows which belonged to that division due to mislabelling. The format of SIC data and poor SIC categorising accuracy has caused the clean data to miss some subcategories for each division. As such, we weren't able to do a more detailed analysis of the relationship between size and water usage for each business activity group. We estimate that if size and water usage is correlated for one subgroup withing a business activity but not for another, this signal might be lessened by the presence of the other groups in the data.

The geographical matching, we have carried out is also not as accurate as a one-to-one business matching that the UPRN data, if available, would enable. Instead, it should be considered a workaround due to the lack of UPRN usage in open access datasets. IDBR data are business count estimates which will also cause noise withing the data and lead to less accurate results.

3.5 Education

For commercial business activities, finding business characteristics data which has UPRN data has not been possible, however education business activities are state funded, and their data is managed by the DfE, which hosts many open access datasets including school's identifier data. We found data for primary, secondary, and post-16 schools which contained number of registered students, number of students by gender, type of school, number of staff, and number of full-time equivalent staff. This data contained unique reference number (URN) identifiers which are specific to each school but not linkable to UPRN.

3.5.1. Bespoke data cleaning and data matching

Education business activity water usage data in the CMOS Enriched dataset is not very clean due to SIC codes being misclassified. Creating different datasets for each of the different groups of educational business activities is then a manual task. Education data was separated from the rest of the data in the CMOS enriched datasets using key term searches within the data. Key terms used are 'school', 'academy', 'schools', 'college', 'university', 'primary', 'secondary', 'education', 'learning', 'childcare', 'acad', 'learners', 'junior', 'academies'. Note that this process itself is prone to error too. Some reasons why businesses names didn't match the activity include incomplete or misspelled names such as: *inglewood junior* instead of *Inglewood junior school* and *st philip howard catholic voluntary acad* instead of *academy.*

After key term matching, we have 5569 rows (educational establishments). We have different size datasets for primary, secondary and colleges; further education; and higher education businesses. Data matching for the first groups (primary, secondary and colleges) is then carried out using postcode data due to no matchable identifiers being present. Since SIC code data is not reliable, we separate primary, secondary and college school data from the rest of the educational business activity data by matching to the school size data by postcode. After matching, we are left with 3554 rows containing water usage data and pupil numbers as well as girls to boys split.

3.5.2. Correlation between water usage and school characteristics

We found no significant relationship between number of pupils or staff numbers and water usage for the aggregated data set of primary schools, secondary schools and colleges (corr. 0.08).

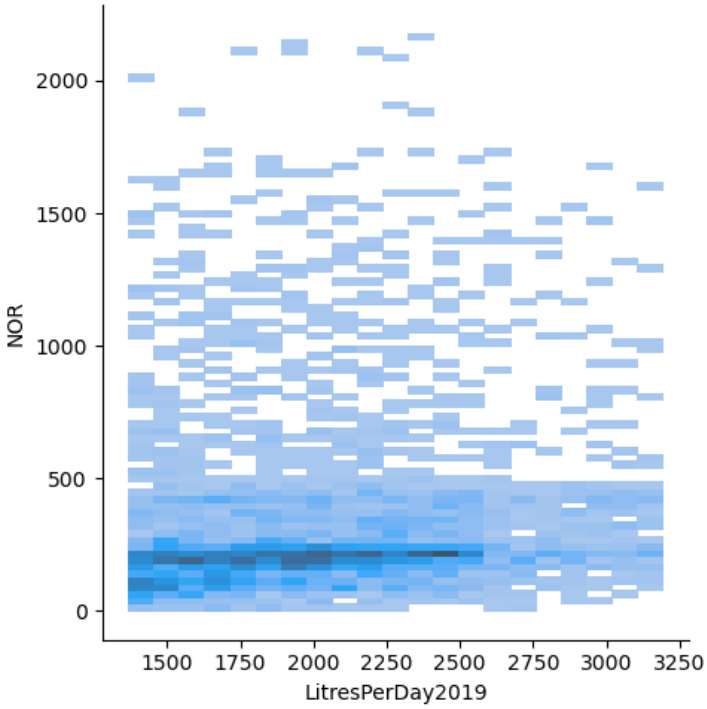

Figure 7. Scatterplot of number of registered pupils v water use in 2019: Primary schools

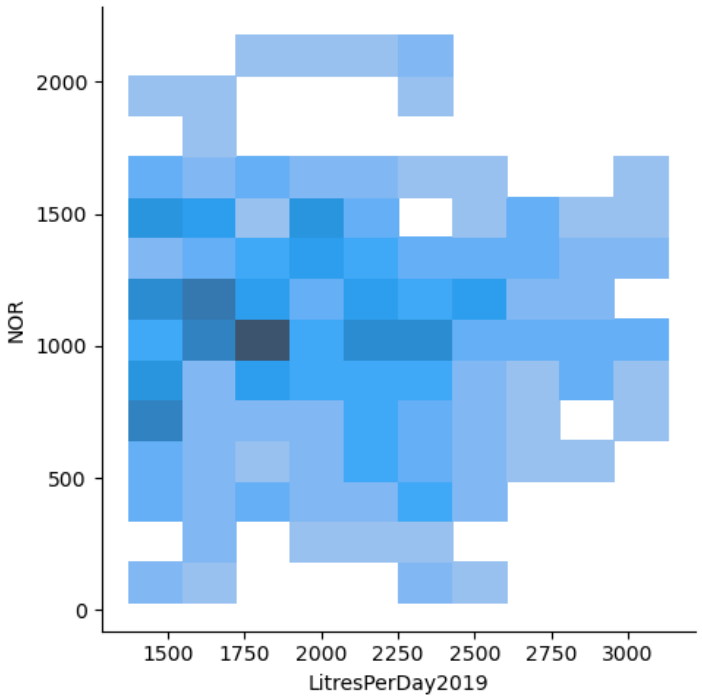

Figure 8. Scatterplot of number of registered pupils v water use in 2019: Secondary schools

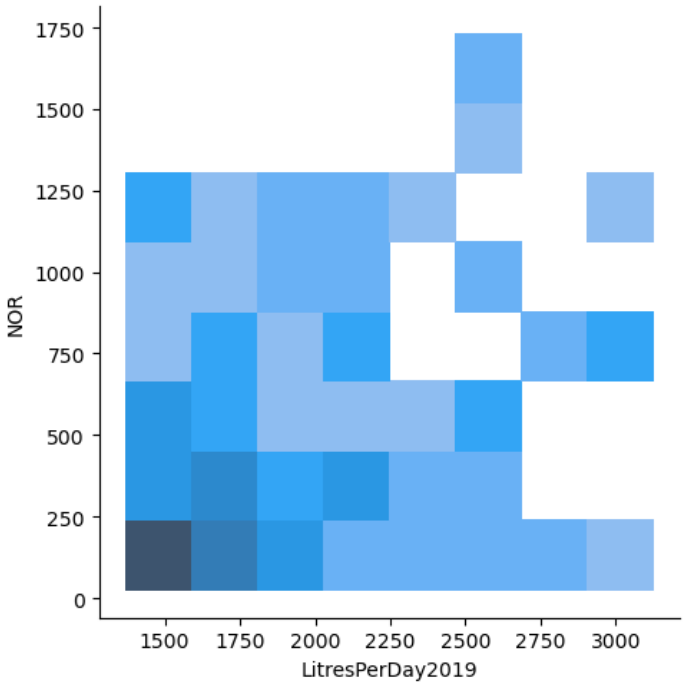

Figure 9. Scatterplot of number of registered pupils v water use in 2019: Boarding schools

We also found no significant linear relationship between number of pupils and water usage, when disaggregated by education establishment type (Figures 8, 9). We segmented the data into the different school types: primary schools (corr. 0.19); secondary schools (corr. 0.01); post-16 education (n=4) and all years (corr. 0.11). There was insufficient college data for a reliable correlation to be calculated. Number of pupils in primary schools had the highest correlation to water usage. This correlation could be a real signal; however, these results should be validated using smart meter data due to the CMOS data lack of accuracy in recording water usage.

We further explored the education data using a several other activity metrics. We found no significant difference in water usage based on type of funding, boarding school status, gender or year the school opened. We found a positive correlation between water usage and number of pupils in boarding schools (Figure 9), but we only had 95 schools in this sample. Number of staff or full-time equivalent staff variables also did not correlate strongly to water usage (corr. 0.05).

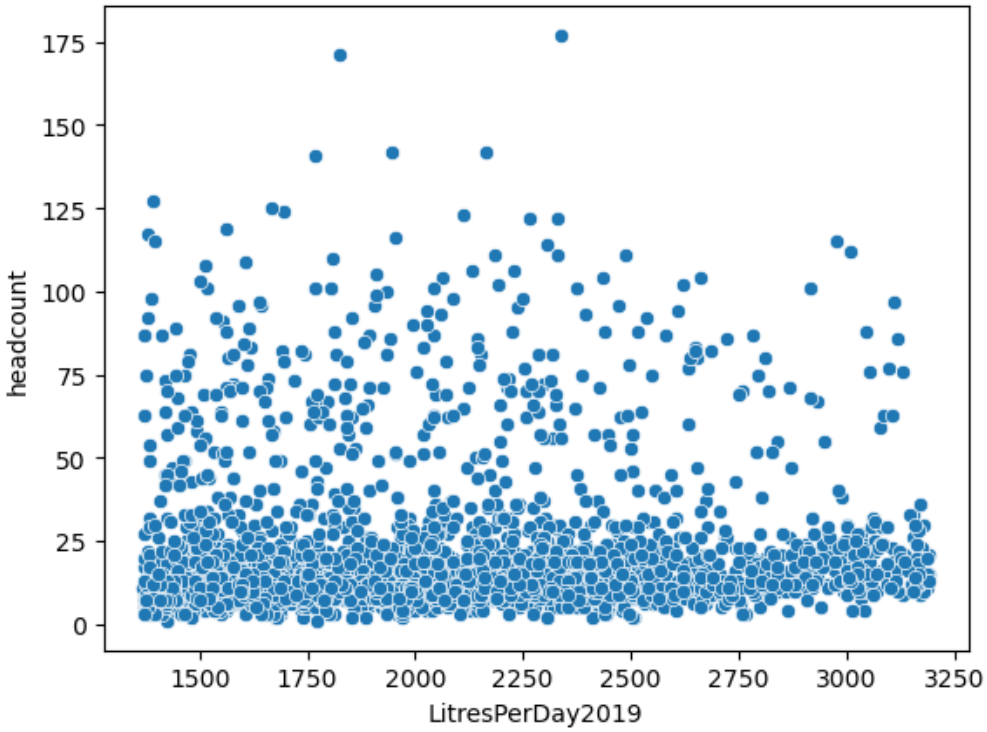

Figure 10. Scatterplot of number of employees and water usage per school

We segmented the sample by type of education as we did with number of students to see whether there was a strongest relationship in primary schools than other school types. We found that primary schools had the strongest relationship between headcount and water usage (corr. 0.18) and secondary schools had no correlations (corr. 0.00). The college dataset had only three college hence no relationship could be established.

Primary school headcounts show a slight positive effect on water usage. We note that the primary school sample is much larger than that of the secondary school. We suspect that the key term data cleaning might be more effective in extracting primary than secondary schools. A study of secondary school key terms might be useful in obtaining a larger secondary school dataset.

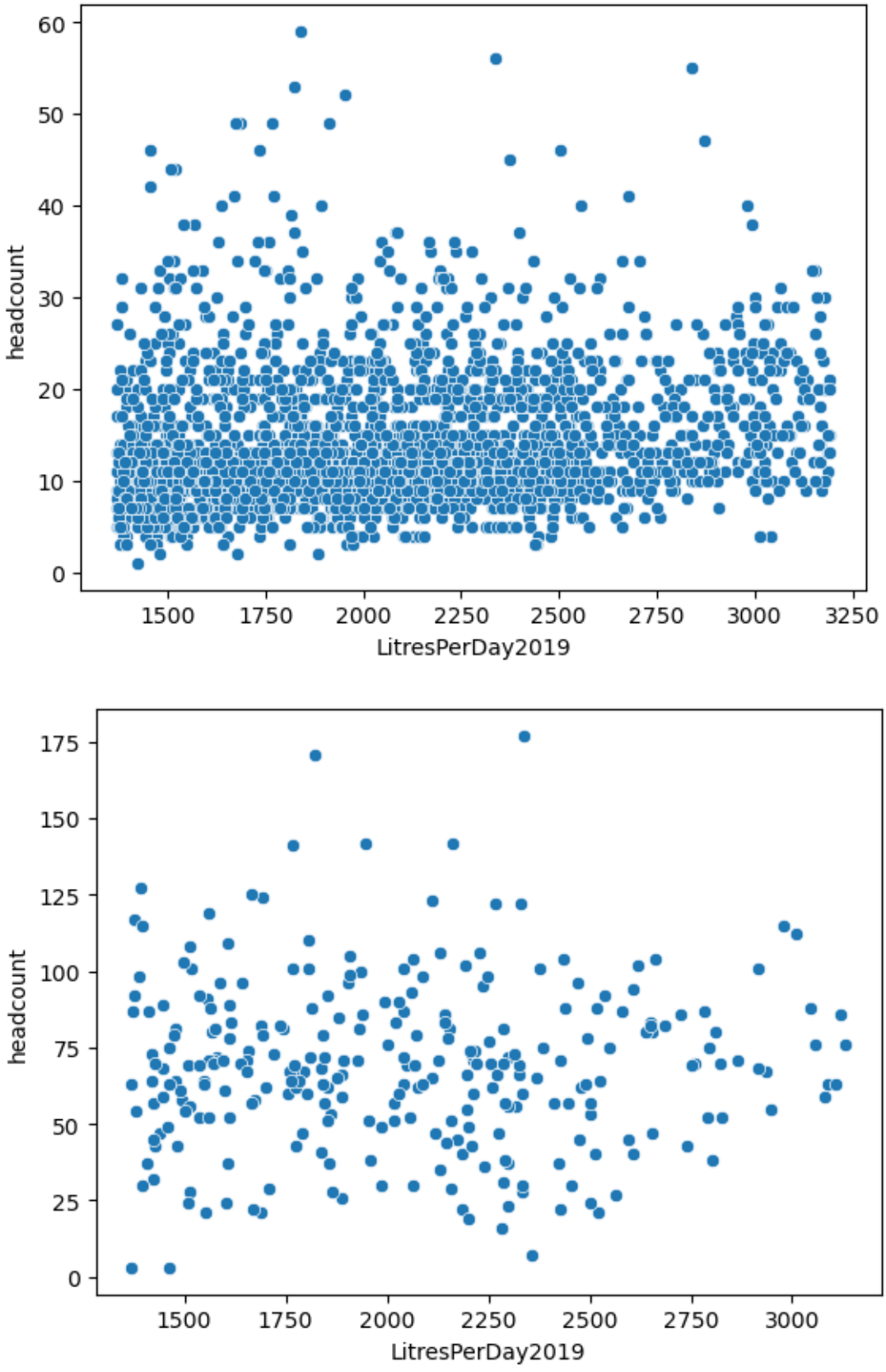

Figure 8. Water usage and staff headcount in primary schools (top) & secondary schools (bottom):

3.5.3. Linear Regression Models

Following these results, we judged that the relationship between water usage and the different school size metrics might be more complex than suggested by a single size metric, and so we developed a linear regression model with independent variables of number of pupils, number of girls, number of boys and staff headcount. We found no significant relationship between these metrics and water usage for either primary (r squared = 0.04) or secondary (r squared = -0.00) schools.

3.5.4. Litters per pupils

The primary aim for this objective was to support creation of a dashboard customers could use to inform themselves about their water usage in relation to other businesses in their SIC. Initially this work would be done under the assumption that the obtained size metrics and water usage presented a linear relationship. However, without being able to match size and water usage data, it was impossible to stablish an average water usage for a particular size group. However, through matching the data geographically, we can not only be able to stablish the average water usage for each business activity, but also to validate the assumptions around the relationship between size and water usage.

After testing these assumptions, we have seen a lack of linearity in the relationship between the two. However, this is not necessarily to mean these data can't be used to inform customers. We find that by dividing schools into three size groups by number of pupils (small, medium, large). These sizes were defined by the quartiles of number of pupils per school: small was from minimum school size to first quartile included, medium was from first quartile to third quartile included, and large was stablished as schools with a number of pupils larger than the third quartile. We observe significant differences between the means of the three groups (Figure 12) and therefore could use this metric to inform customers about their water usage. This can then be used to compare a school's water usage to the average for the group. We suggest that this approach should be validated using the smart meter data.

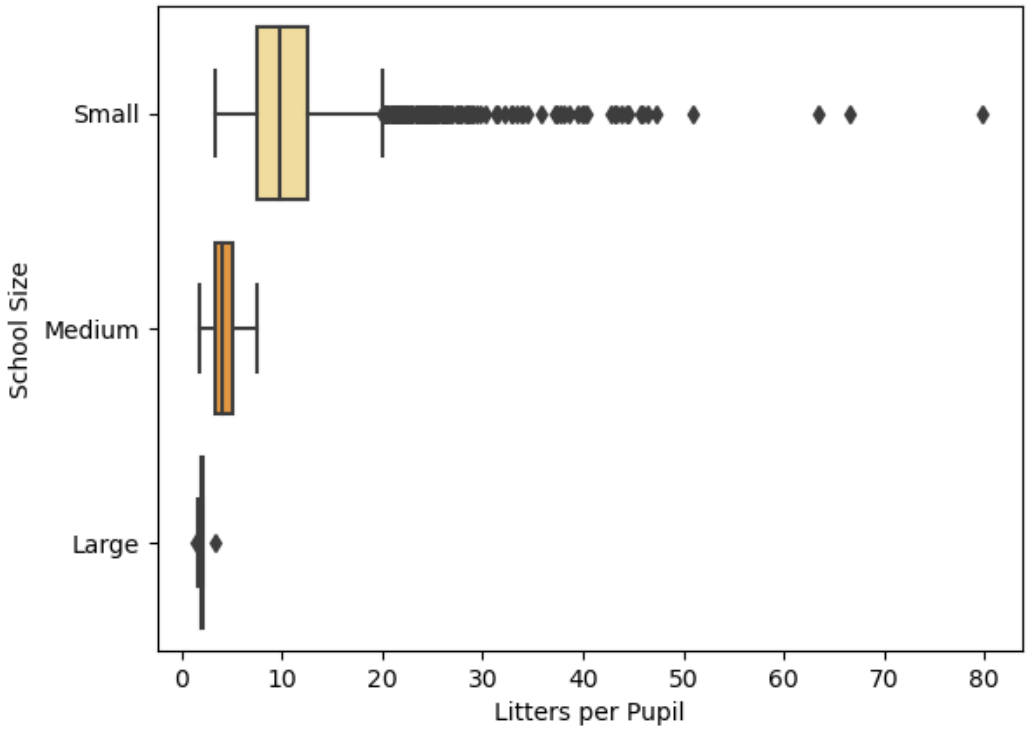

Figure 9. Boxplots showing the distribution of water usage by each school size group

4. Leakage Analysis

Using smart meter data, we had planned to identify SICs that were more prone to leakage than others. For example, we might hypothesis that schools experience relatively high rates of leakage due to limited funds for new infrastructure and maintenance. Using the leak detection algorithm from our prior work (Hulse et al., 2022) we had sought to estimate leakage rates for different activities by SIC, and so add a high water use risk factor to the analysis conducted above. If an SIC appears to have an above average rate of leakage, this adds a further measure to the analysis above, when considering audit/intervention (i.e., if a customer is smaller than average but an above average consumer ***and*** is in a SIC group that overall appears to have above average leakage, there is a further rationale to investigate that customer.

Continuous flow urinals are an already a known source of leakage. We suggest that incorporating type of toilet in premisses (if possible) or gender split might be useful in identifying leaks. Alternatively, it could be explored whether our leak algorithm is able to identify patterns typical of continuous flow urinals so recommendations can be made towards their replacement if necessary.

We also sought, as a secondary objective, to analyse the smart meter data for the Hadfield DMA, where all properties are on a smart meter. This would then allow us to match DMA leakage to leakage per property, estimated using different leak algorithm parameters. This would then enable us to recommend a set of leak parameters (flow duration, and flow above minimum night flow) indicative of a 'standard leak.' This standard definition could then be used in further work that needs to estimate and compare leakage in SM networks.

Unfortunately, due to the very late arrival of the SM data, we were not able to carry out this work.

5. Conclusions and recommendations

5.1. CMOS Data

1. The analysis found no significant relationships between water use and customer activity descriptors. The reasons for this are uncertain but are likely attributable to: (1) an inability to link available public open-source data for the market to individual customers due to lack of suitable activity descriptors, use of a linking field (e.g. UPRN) and business data confidentiality; (2) signal noise (i.e. high rates of leakage swamp any observable activity rate-consumption relationship by SIC; and (3) lack of robustness / confidence in the CMOS data.

2. Cleaning the current CMOS dataset will be a laborious and manual task. We suggest some of this work might be automized by employing fuzzy matching algorithms, which an approximate string-matching machine learning tool to find similarities within string variables. This might be able to, to some extent, group businesses with their activity by using name cues.

3. Overall, we suggest (see below) that any further work to develop water use benchmarks is focussed on smart meter data where customer level activity descriptors can be quantified and matched at the level of individual customers.

4. While the primary objectives weren't delivered due to difficulties with administration of data sharing, we believe that a reshaping of the project's objectives might anyway be necessary to producing an industry level data product. A focus on the SICs that account for 50% of NHH water demand, may be inappropriate as:

- Very high-water consumers in manufacturing probably have a poor relationship to size/activity descriptors and need to be identified and separately targeted (if not already) for bespoke water conservation measures; this should be feasible given the relatively small number of very large users.

- Water use in agricultural and horticultural businesses is likely more sensitive to geography / weather than other sectors, making it harder to identify water usage – activity relationships. Benchmarks here may thus require data on crops, yields etc. While such specificity might be possible at SIC class level, other methods of targeting water conservation might be more useful than benchmarking, give the data constraints.

- Health related activities currently tend to defy water benchmarking analysis due to the complex and varied data administration systems governing the NHS.

5. The retail industry is more likely to see a benefit to being included into this benchmarking exercise. They are low water users but will make use of toilet facilities which are believed to be the more wasteful facilities. For the remaining business activities, we suggest a dedicated effort to compile a very clean set of linked data.

5.2. Smart Meter Data

6. SM data is currently a sample for which manual cleaning of ambiguous rows might be more accessible than for the CMOS data. Key term matching has been shown to be quite effective in selecting certain business types but has the limitation that it might cause a bottleneck that excludes certain business groups. Bringing all SIC codes in line with more up-to date classifications would be desirable towards an easier handling of the data which is more readily available for linkage and analysis.

7. The data samples in some of these business groups seem however to be too small, reducing the power of the analysis. Statistical power analysis might be helpful towards establishing the best sample size for each business activity towards a more dedicated data finding exercise, as it will describe the likelihood of an estimate to predict an effect given a certain sample size. We suggest that other literature findings might have collected such data, and this might in turn be open access and available for linkage. Another approach towards collecting business characteristics data might be using data scraping algorithms that allow the scientist to collect data which is available on the web.

8. Although we have not worked with the business characteristics data that TW has provided MOSL, it is presumably a detailed dataset including variables beyond those we consider above, that can better explain water usage. Cleaning and linking this dataset to the SM data for the companies involved would allow a more granular analysis of water usage by business activity. That is, rather than attempting a whole industry analysis of coarse data sets, or data collected by retailers to different reporting formats, work on a sub-set of high quality smart meter data, with effort expended to generate the associated customer level activity data might prove a more productive approach (i.e. for key SICs, build small sets of SM and activity level data). Thus we recommend that a small, powerful enough sample of business SM data is collected and linked to available size metrics.

9. NHH SMs are not yet universal, and so we suggest attention is also paid to ensuring individual customer's data is collected in a consistent standardised format by retailers in order to enhance industry level data amalgamation and analysis.

5.3. Existing Dashboards

10. A number of dashboards are available online that attempt a to offer a similar service such as the MOSL School Benchmark dashboard and the TW Business water saving calculator. In exploring these dashboards and in light of the lack of a linear relationship between water usage and business size metrics, we decided to seek the expertise of the team that developed the TW business water saving calculator. We met with Andrew Tucker, the water demand reduction manager in TW and Sally Bremner, the Non-Household Demand Reduction Manager, in order to enquire about the TW dashboard development process.

11. The TW Business water savings calculator was a large undertaking involving Ricardo contractors who underwent the data collection, data analysis and dashboard development. Water usage data was provided by TW whereas business characteristics data was either already owned by Ricardo or obtained during a bespoke data collection exercise. This dashboard covers a slightly different range of business activities, including offices and retail but is missing the agricultural and horticultural, and the health business activities.

12. Some issues arise with the TW dashboard, such as nil returns due to small sample sizes. However, TW has made available to MOSL their water usage data and their business characteristics data which was obtained during this process. Although we did not get to use this, we believe it to be a valuable resource in going forward.

13. The main challenge in completing this project is the lack of data reliability and the dimensions of the data cleaning task required to be able to carry out a much smoother data linkage process. A cleaner water usage data will allow data linkage, even when just on geographic boundaries, to be carried out in a more reliable manner, decreasing study limitations and need for assumptions.

6. References

DfE (2023) *Schools, pupils and their characteristics, Academic year 2021/22*. (2022, June 9). https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics/2021-22 Department for Education.

Hulse J, Mitchell G and Newing A (2022) *An Evaluation of Granular Consumption Data for the Non-Household Water Market*. LIDA Data Science Programme with Market Operator Services Limited and Yorkshire Water Services Ltd. 59pp

NHS (2023) Hospital Admitted Patient Care Activity, 2021-22 NHS Digital. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-admitted-patient-care-activity/2021-22#resources National Health Service.

ONS (2023). *UK business activity, size and location: methodology - Office for National Statistics*. https://www.ons.gov.uk/businessindustryandtrade/business/activitysizeandlocation/methodologies/ukbusinessactivitysizeandlocationmethodology

7. Appendices

7.1 Large water usage outlier removal

7.1.1. Z-Scores

Z-scores are a common method used for outlier detection. A Z-score calculates how many standard deviations a value is from the mean value, allowing outlier distance from the mean to be calculate din a standardised manner. Z score identification works best with normally distributed data, however our current data doesn't follow this distribution (it's highly skewed), hence Z-scores aren't the best metric for this dataset.

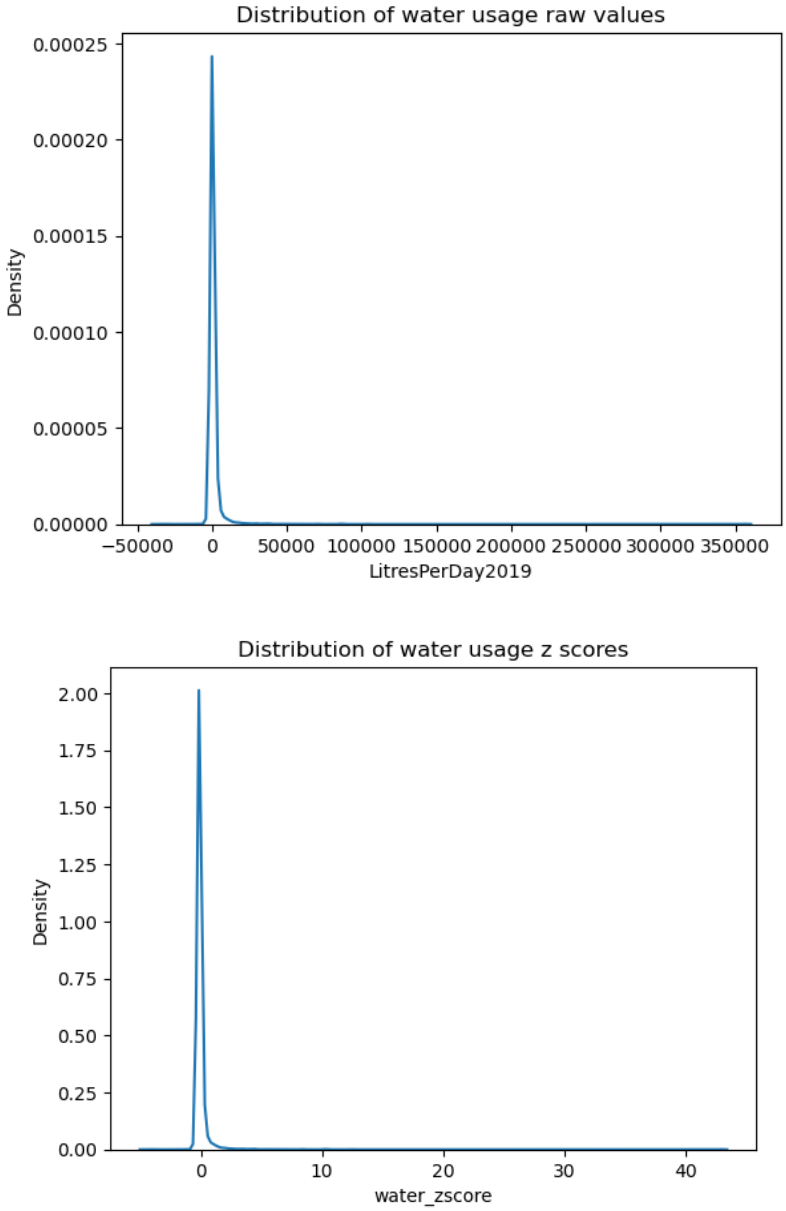

Figure 13. *Distribution of average monthly water usage for raw values and for z score statistics*

7.1.2. IQR Outlier Detection

The interquartile method is a statistical outlier detection method based on the first and third quartile. The range is calculated as follows:

$$Upper\ fence = Q3 - 1.5IQR$$

Where upper fence references outliers to the right of the curve, Q3 refers to the thirst quartile and IQR stands for Interquartile Range. We only need to use this formula for outliers at the right of the curve as we use different data cleaning methods for outliers at the left (negative and 0 water usage cleaning).

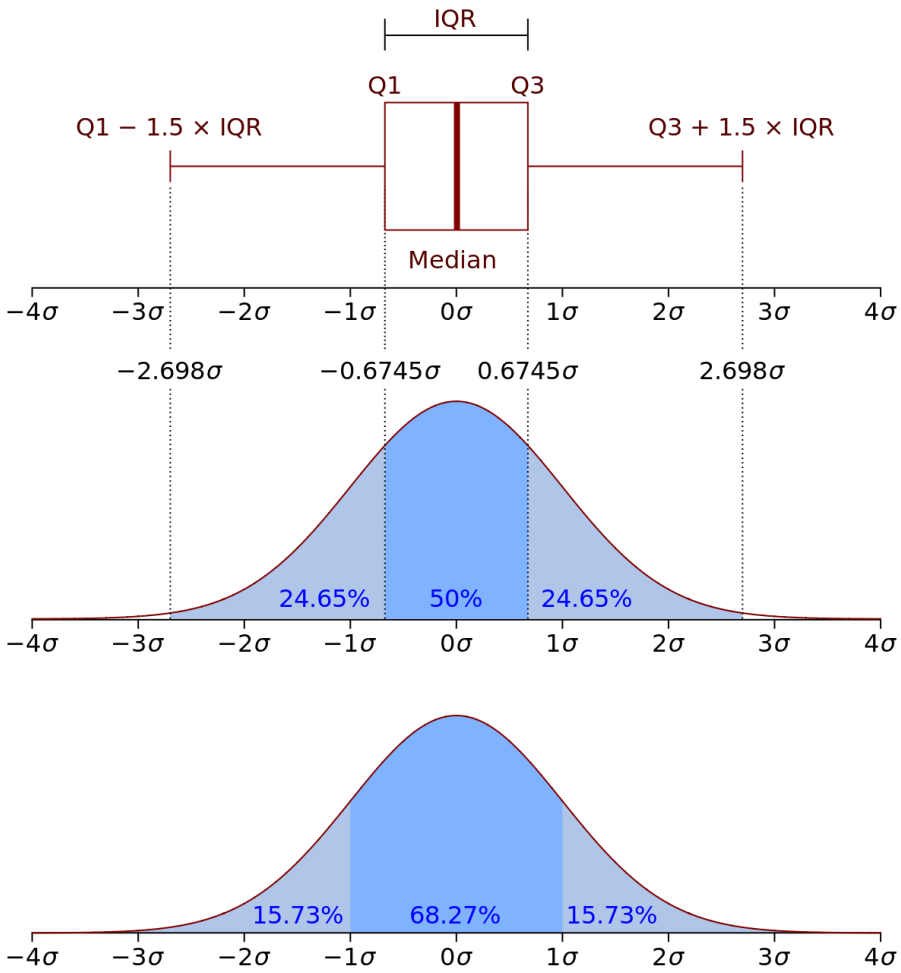

Figure 14. *Illustrated example of using IQR outlier detection method on normally distributed data*.

7.1.3. MMR Outlier detection methods

The MMR (mean median ratio) is a measure of central tendency that quantifies skewedness. A M:M ratio of 1 suggest that mean = median which in turn suggests a perfect normal distribution (top in Figure 15). For this method, we calculated the mean to median ratio of the raw data, and then subtracted 4, 3.5, 3, 2.5, and 2 standard deviations from the mean to improve the central tendency and compare the distributions of each subset.

Figure 16 shows results of this analysis, and that there is little difference from four to two standard deviations, but that there are significant difference in distribution. Percentage change in volume at 4sd is 59%, at 3.5sd 60%, at 3sd 61, at 2.5 sd 62%, at 2 sd 64% and percentage change in number of rows at 4sd 1.2%, at 3.5sd 1.4%, at 3 sd 1.6%, at 2.5 sd 1.9%, at 2 sd 2.5%.

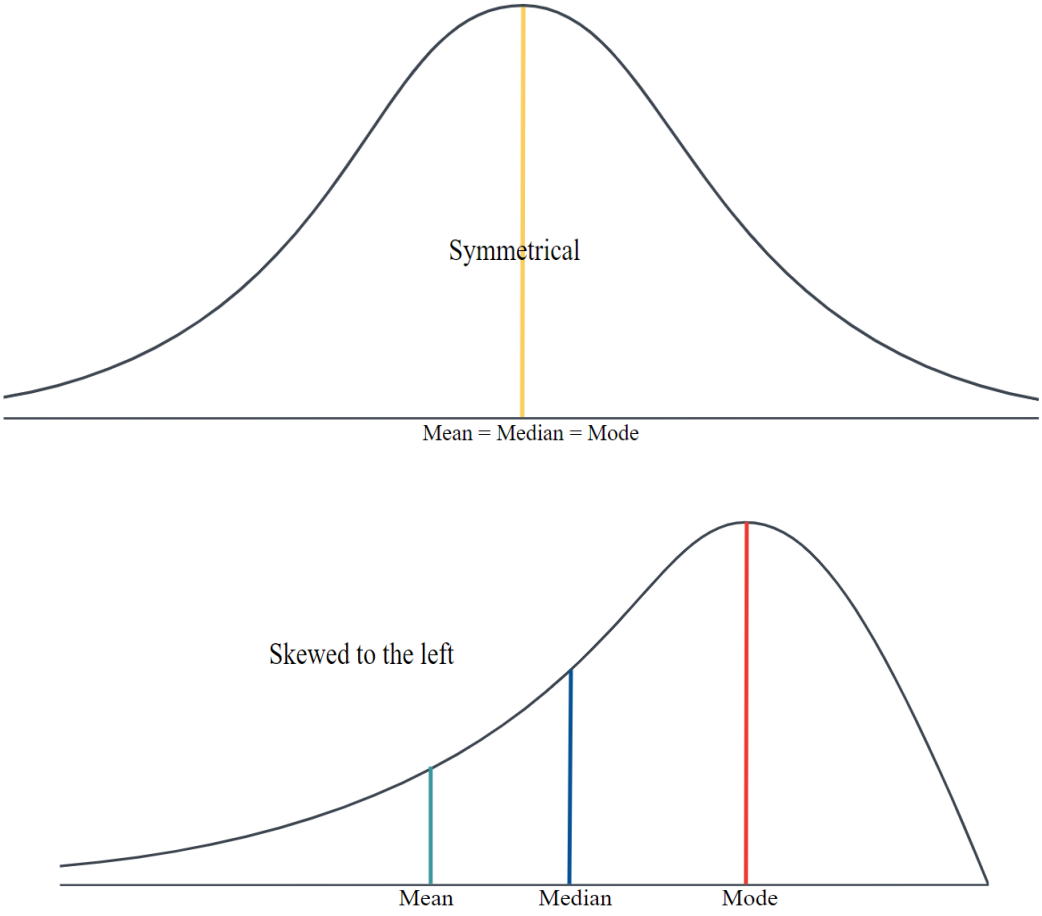

Figure 15. *Example of central tendency in normal and skewed distributions*

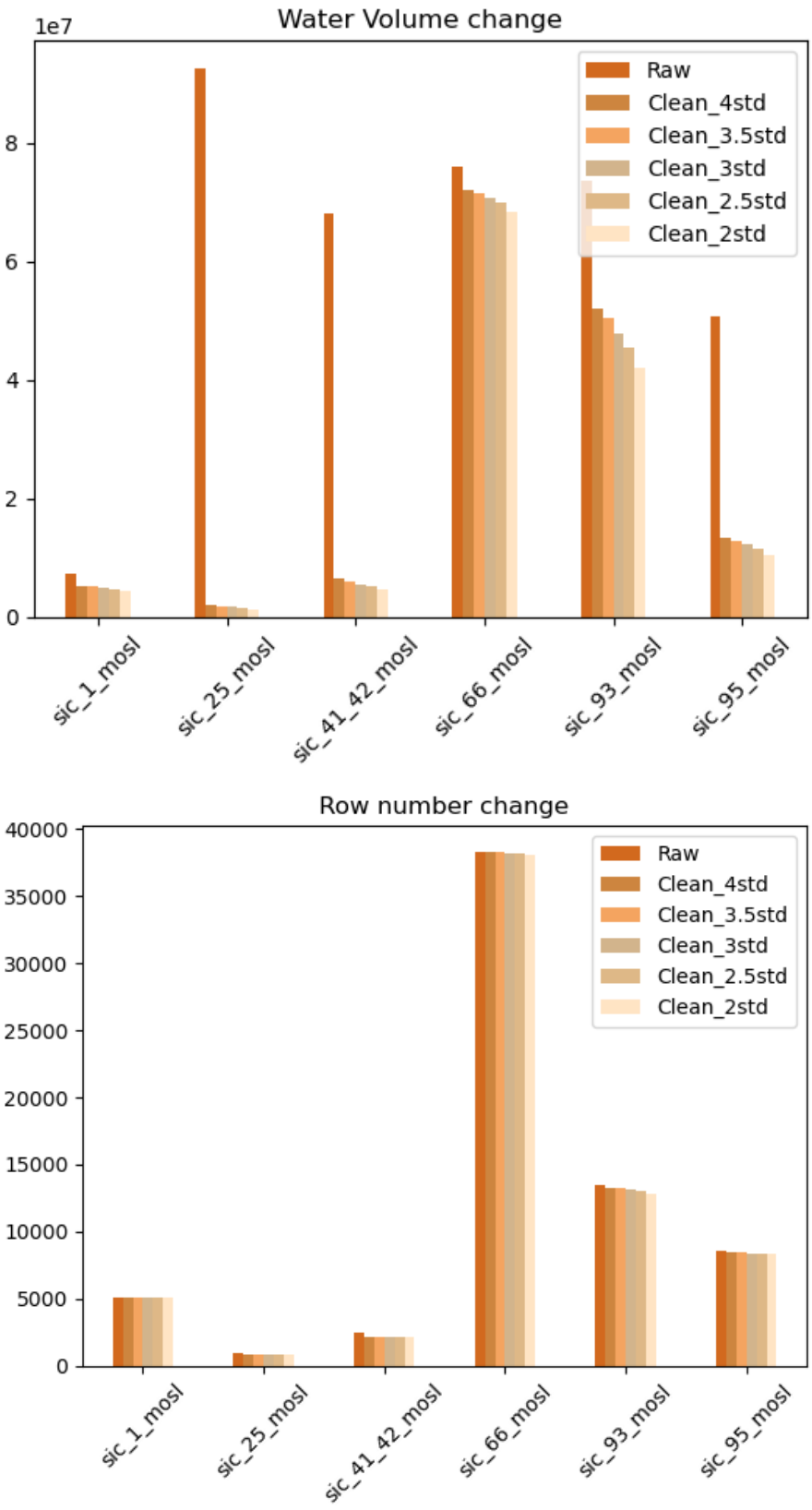

Figure 16. *Water volume and number of rows/customers – difference between raw and clean data after removal of large water use outliers using the MMR method for each standard deviation*

7.1.4. Comparing outlier detection methods

After comparing the distribution of water usage for each of the outlier detection methods, we decided to use the IQR method. This is stringent and leads to the highest data reduction, but also creates the more normal distribution, with the M:M method at 2 standard deviations still presenting large tails.

The target for this tool is average size business per size quartile, with businesses larger than this likely to require personalised methods to aid conservation. Therefore, we believe this method does the best job at capturing business with an average water usage that will best benefit from the tool.