Application of COCOA Schema to ABP

Project no.:	2745	**Date:**	22/05/2024
Author(s):	Cristina Munilla, Tom Nelson, Alan Spey	**Email:**	Cristina.Munilla@artesia-consulting.co.uk Francesca.cecinati@artesia-consulting.co.uk
Reviewed by:	Francesca Cecinati	**Reference:**	AR1601

1 Context

Project Discovery[1] developed new approaches to segment non-household (NHH) customers with the aim of benchmarking water consumption, improving demand forecasting, assisting in responses to water stress, and to help identify water efficiency opportunities. This includes understanding what is driving demand in NHH properties and exploring how granular consumption data could be used to augment the segmentation and targeting.

Through Project Discovery, the Commercial Consumption Analysis (COCOA) Schema was developed. **The COCOA Schema** consists of the following components:

- **The COCOA Classification**, which aims at classifies NHH properties based on water usage behaviours. It is built on a functional water use classification and a data-driven classification, developed by studying consumption profiles from smart metering data.

- **The COCOA Benchmarking**, an estimate of expected consumption for properties in each of the COCOA Classification groups, for each calendar month.

To develop the COCOA Schema, an extensive literature and data review was carried out. This review also investigated the existing schemas that could be used to base the COCOA Schema on. Among others, the **Standard Industry Classification (SIC)** schema was identified as the most widely used, detailed, and open-access schema, and was chosen to develop COCOA. SIC was preferred to other schemas, including the **Address Base Premium (ABP)** Classification from the Ordnance Survey (OS), which is not freely accessible and is less detailed.

However, through the development of Project Discovery, it was observed that the SIC classification for NHH properties in the Central Market Operating System (CMOS) database is often incomplete. Only approximately 30% of NHH properties have an associate SIC classification.

On the other hand, toward the end of the project, it emerged that ABP is widely available across wholesalers, possibly with better matching rates. For this reason, MOSL identified the need to re-map COCOA to the ABP Classification, to complement the existing version.

[1] Artesia Consulting Ltd, 2024, *Project Discovery – Segmentation and benchmarking of non-household properties*, AR1560, Project 2579

This technical note describes the process of mapping the ABP Classification to the existing COCOA Schema.

2 Understanding of the ABP Classification

The ABP Classification Codes from OS are part of a comprehensive data product called AddressBase Premium. This product provides detailed and accurate address and property information across Great Britain. AddressBase Premium includes various classification codes to describe the type and usage of each address. Hence, codes indicate the type and usage of properties, such as residential, commercial, or mixed-use.

The codes follow a hierarchical structure, starting with broad categories at the primary level and becoming more specific through subcategories, down to the quaternary level. The available classification of the primary level is shown in Table 1.

Table 1: ABP Classification codes – Primary code

Primary Code	Primary Code description
R	Residential
C	Commercial
M	Military
L	Land
O	Other (Ordnance Survey Only)
P	Parent Shell
U	Unclassified
X	Dual use
Z	Object of interest

Table 2 illustrates how the subcategories of the code provide additional detailed information. It is worth mentioning here that there are only certain classification codes that extend to the quaternary classification level.

Table 2: ABP Classification codes examples

Primary Code	Primary Code description	Secondary Code	Secondary code description	Tertiary Code	Tertiary Code description	Quaternary Code	Quaternary code description
C	Commercial						
C	Commercial	A	Agricultural				
C	Commercial	A	Agricultural	1	Farm / Non-Residential Associated Building		
C	Commercial	A	Agricultural	2	Fishery		
C	Commercial	A	Agricultural	2	Fishery	FF	Fish Farming
C	Commercial	A	Agricultural	2	Fishery	FH	Fish Hatchery
C	Commercial	A	Agricultural	2	Fishery	FP	Fish Processing
C	Commercial	A	Agricultural	2	Fishery	OY	Oyster / Mussel Bed

The most recent ABP AddressBase Version 2.0 (release date April/2023) was obtained from the OS website, a description accompanies the excel classification download[2].

The review of the structure of the ABP classification, provided in the COCOA Excel tool for reference, revealed that there were 579 unique classification codes. Table 3 illustrates the distribution of these codes across different levels of detail and subcategories.

Table 3: ABP Classification code number of categories

Total	Primary Code	Secondary Code	Tertiary Code	Quaternary Code
579	9	62	228	280

It is worth it mentioning here that ABP codes focus on classifying properties based on their geographic and usage characteristics, making them valuable for spatial and urban planning. while SIC codes classifies industries based on economic activities, providing a framework for economic analysis and industry classification.

Another consideration is that the different ABP levels are often used simultaneously. For example, some records may be classified using tertiary codes, while others use quaternary codes. In contrast, SIC levels are typically used independently.

This differentiation is going to have effects and certain limitations in the process of mapping to the COCOA schema and these will be detailed in the relevant sections.

[2] Ordnance Survey, 2023, *AddressBase, AddressBase Plus & AddressBase Premium – Classification Scheme*, Ver 2.0

3 Methodological approach

The COCOA Schema relies on SIC Codes to classify non-households based on their water use. The final schema was stratified in two different layers:

- **Functional classification:** this layer aimed to group and capture the functional use of water in the businesses. Here, `functional use` is understood as how, when, and why water is used. This classification was based on expert judgment and our understanding of how businesses use water. This layer is made up of 32 clusters based on the nature of the business and the way water is consumed.
- **Data-driven classification:** this layer enhances the functional classification by integrating insights from a data-driven clustering exercise that is free from any pre-existing biases or assumptions. The data-driven clustering was based on annual consumption profiles and resulted in the identification of 9 different clusters. These clusters allow for a better understanding of how NHH consumption varies across seasons for different types of businesses.

The final version of the COCOA schema relies on SIC codes. However, given the findings of the low availability of this information for NHHs (30% of the properties in the analysis we carried out) paired with the increased availability of ABP information, it was recognised as necessary to adapt the COCOA Schema to include the ABP classification, enhancing the current version.

To keep consistency across both SIC and ABP classification, the same classification approach was followed for ABP that is:

- **Map ABP classification to the COCOA functional classification**. This step required a manual mapping based on expert judgement and evaluation of functional use for each of the 579 ABP codes.
- **Map ABP classification to the COCOA data-driven classification**. This step required to match the available granular data to ABP codes and compare the granular data profiles to the 9 existing clusters. Then each ABP code was assigned to a cluster based on the majority of properties in it, exactly like it was done for SIC in Project Discovery. Note that we do not aim at re-defining clusters, so that the same Schema will be consultable through SIC or ABP alike.

3.1 Map ABP classification to the COCOA functional classification

Under this task the structure and organisation of the ABP Classification was evaluated against the functional mapping designed for COCOA. Based on this review, all available ABP classification codes (primary, secondary, tertiary, and quaternary levels where available) were associated to the functional groups as per the COCOA schema.

The resulting association between the COCOA functional categories and the ABP Classification's primary and secondary layers is illustrated in Figure 1 and Figure 2 respectively, where the COCOA functional categories are shown on the left hand side and the ABP classification on the right hand side of the plot. Note that these two charts include primary and secondary levels of the ABP codes respectively and exclusively, for visualisation purposes and clarity, but tertiary and quaternary have been mapped too.

Figure 1: Sankey chart showing the associations between ABP primary codes (right hand side) and the COCOA functional classification (left hand side)

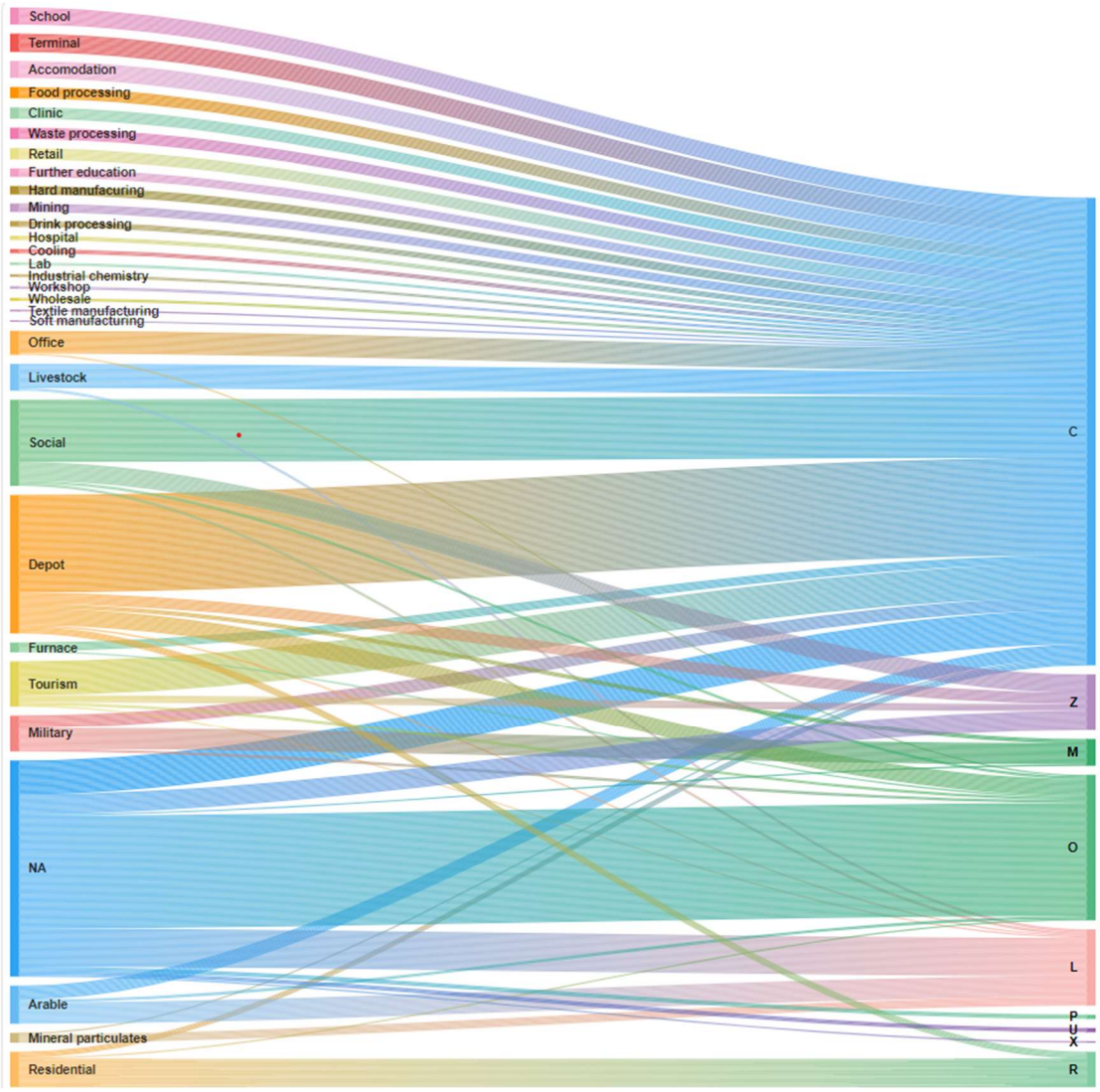
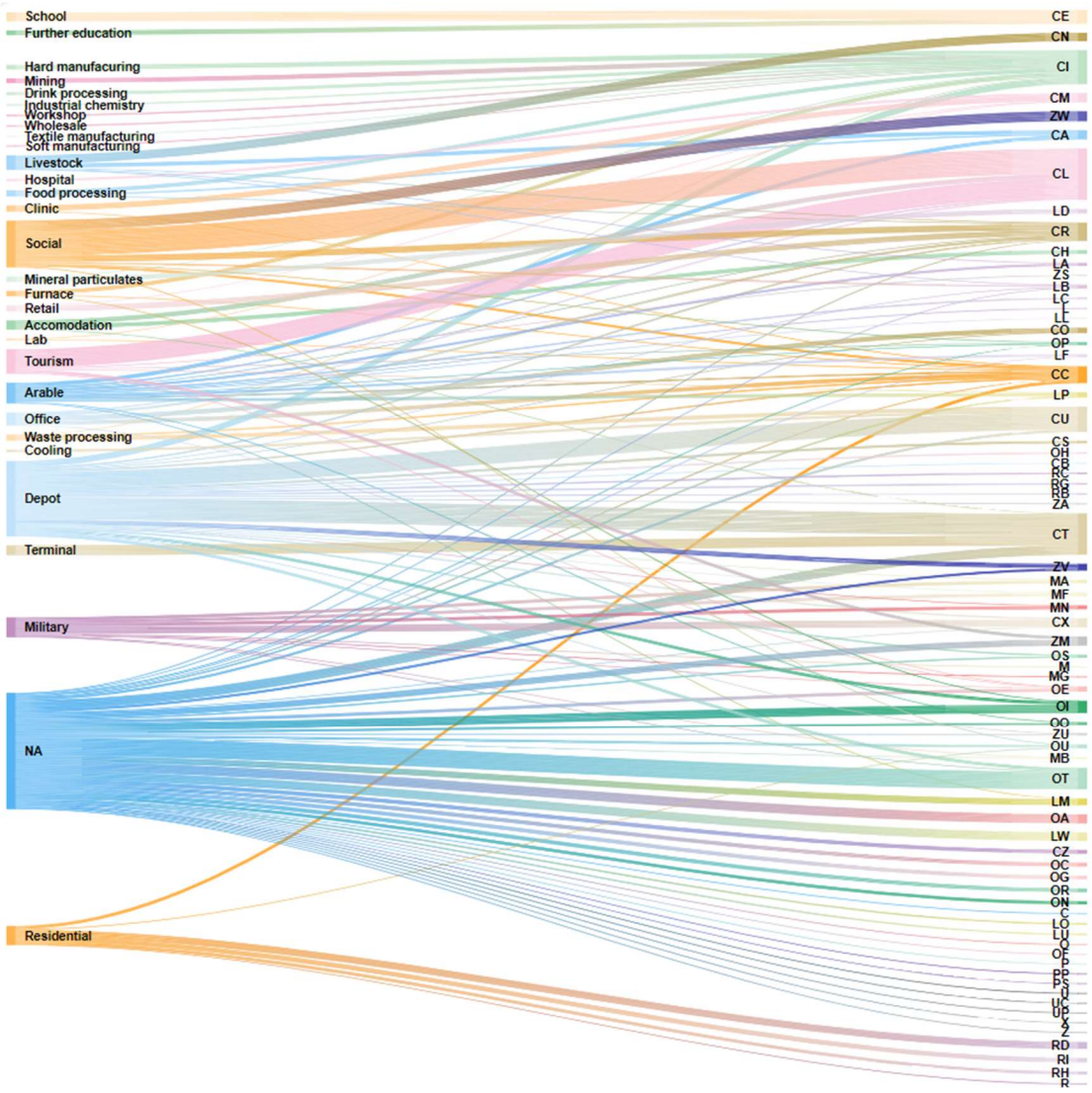


Figure 2: Sankey chart showing the associations between ABP secondary codes (right hand side) and the COCOA functional classification (left hand side)

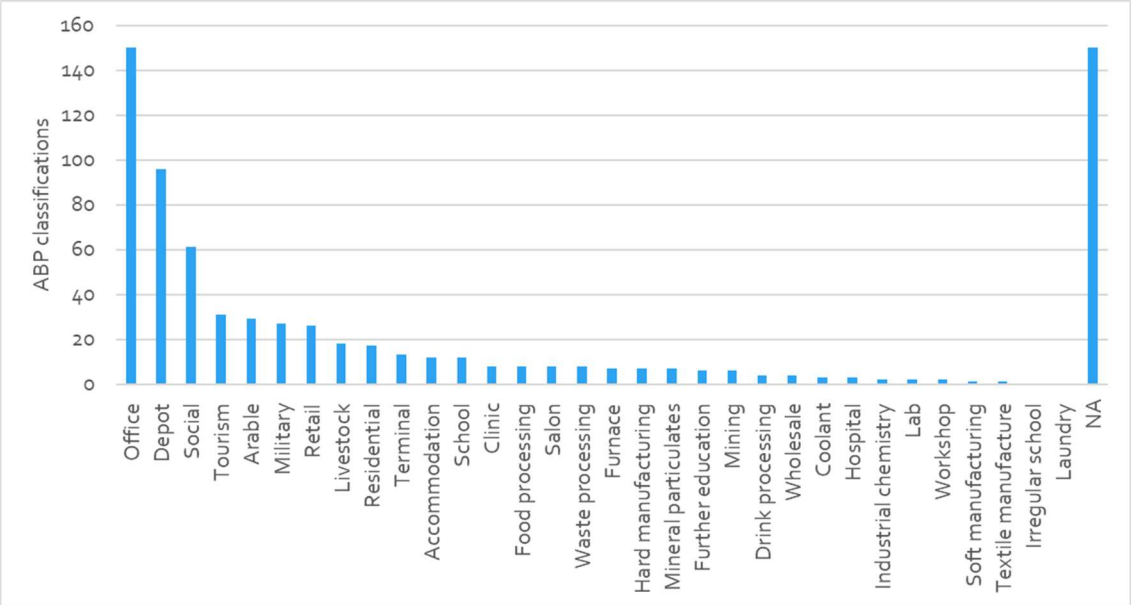


The association of the ABP codes to the COCOA functional categories posed certain challenges, as listed below:

1. **Existence of ABP codes that could not be associated to a particular COCOA functional group:** This is due to the OS classification system typically focusing on landmarks rather than their industry usage. As a result, certain landmarks, such as telephone boxes and car parks could not be classified. This can be observed in Figure 1 and Figure 2 above, where within the COCOA functional categories we observed a "NA" group. A total of 150 codes were affected (see also Figure 3, below), most of them (79) belonging to O -Other (Ordnance Survey Only)-, followed by L -Land- and C -Commercials. In total these account for 26% of ABP classifications, although we anticipate that most of these will not have a water supply anyway.

2. **Lack of representation of some COCOA functional groups:** COCOA functional classification is comprised of 32 levels, however, following the mapping of ABP codes to these levels, the ABP Classification categories lacked representation in two of these groups, with a further two having only a single corresponding ABP category.

Figure 3: ABP matches to COCOA Classification categories



3.2 Map ABP classification to the COCOA data-driven classification

The data driven ABP mapping approach depended heavily on the data available, hence, it was important to have a full understanding of the ABP code representation in the sample; therefore, this section is structured as follows:

- **Data available:** provides a summary of the data provided for this task, and the main quality assurance checks carried out.
- **Exploratory data analysis** provides a summary of the ABP code coverage in the sample.
- **ABP data-driven mapping approach** details the process followed to map the ABP codes to the data clusters.
- **Evaluation of data-driven classification against ABP codes** presents a summary of the results achieved through this process.

3.2.1 Data available

For this piece of work we have used Thames Water (TW) logged data available at hourly resolution, which covered around 4,000 properties. This selected sample also contained relevant and required information to be able to carry out the mapping. That is, all available SPIDs contained SIC code, ABP Classification code, consumption in terms of litres per day and building area. The sample selection and quality analysis (QA) of this dataset are explained

in detail in the Stage 2 report produced for Project Discovery[3]. Therefore, no additional details are provided here.

Aside from TW, Anglian Water (AW) logger data with information on business area, consumption in terms of litres per day, and SIC ABP code was also available from Project Discovery. As with TW, all the relevant information regarding the QA of this dataset is available in the relevant report.

TW data, which was selected for its greater temporal availability, was used to perform the mapping of individual properties to clusters (it should be noted that the data-driven exercise required one complete year of logger data. This requirement was met by the TW data but not by the AW data as detailed in Stage 2 report of Project Discovery). A sample of 3,903 TW properties had the required combination of hourly data for a year and correct ABP codes and was used to map the ABP codes to data-driven clusters.

Data from TW and AW was combined to analyse the effectiveness of these clusters. A sample of 3,516 AW properties and 4,997 TW properties containing SIC code, ABP code, property area in m^2, and average consumption in litres per day was created. This was used to compare the differences in ABP and SIC predictions.

3.2.2 Exploratory data analysis (EDA)

The complete list of ABP classifications contains 579 unique references, 429 if we exclude those codes that were identified as not having associated supply (section 2). The sample available to us contains 175 unique codes, 155 if we exclude codes without identified supply. Table 4 provides more details of the representation of ABP codes in the sample, considering the different ABP subcategories.

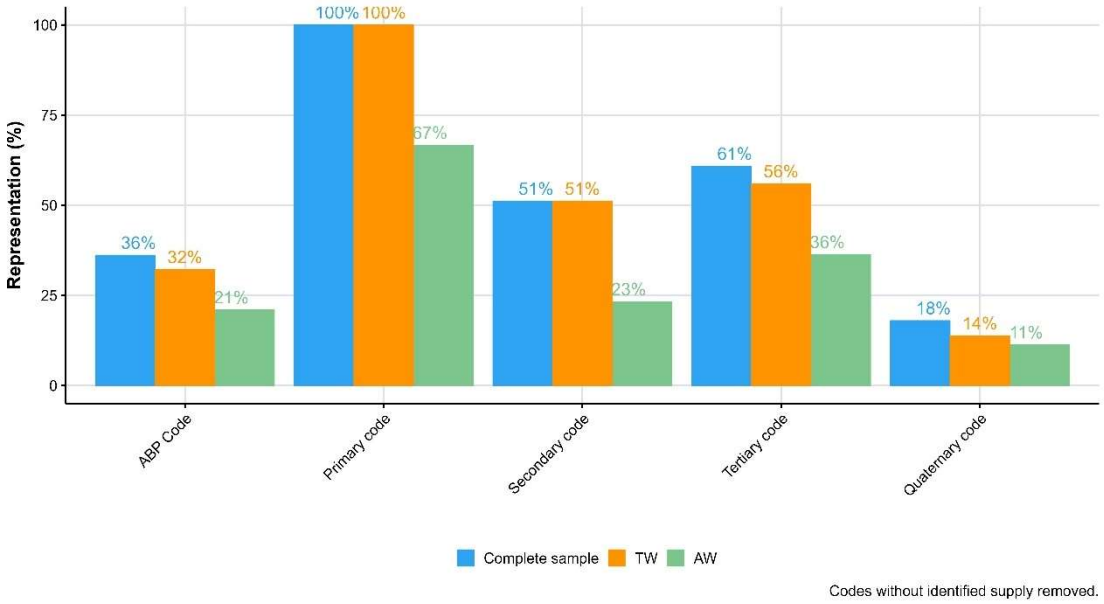
[3] Artesia, 2023, *Project Discovery – Stage 2 Report*, AR1551, Project 2579.

Table 4: Sample representation of the ABP codes (number of unique code references per group)

Dataset	Subset	Codes with no identified supply removed	ABP Code	Primary Code	Secondary Code	Tertiary Code	Quaternary Code
ABP full list of codes		FALSE	579	9	62	228	280
		TRUE	429	6	43	143	239
ABP code representation in the sample	**Full sample**	FALSE	175	9	27	97	44
	Full sample	TRUE	155	6	22	87	43
	AW	TRUE	90	4	10	52	27
	TW	TRUE	138	6	22	80	33

Focusing on those codes with identified supply, the representation in percentage terms is illustrated in Figure 4. It can be observed here that overall, the representation of ABP codes in the sample ranges is 21% for AW, 32% for TW, and 36% for the complete sample. All primary levels are represented, while secondary levels are represented at 51%. Tertiary levels are represented at 56% for TW, and quaternary levels are represented at 14%.

Figure 4: Representation of the ABP codes with identified supply in the sample. ABP codes refers to the overall representation

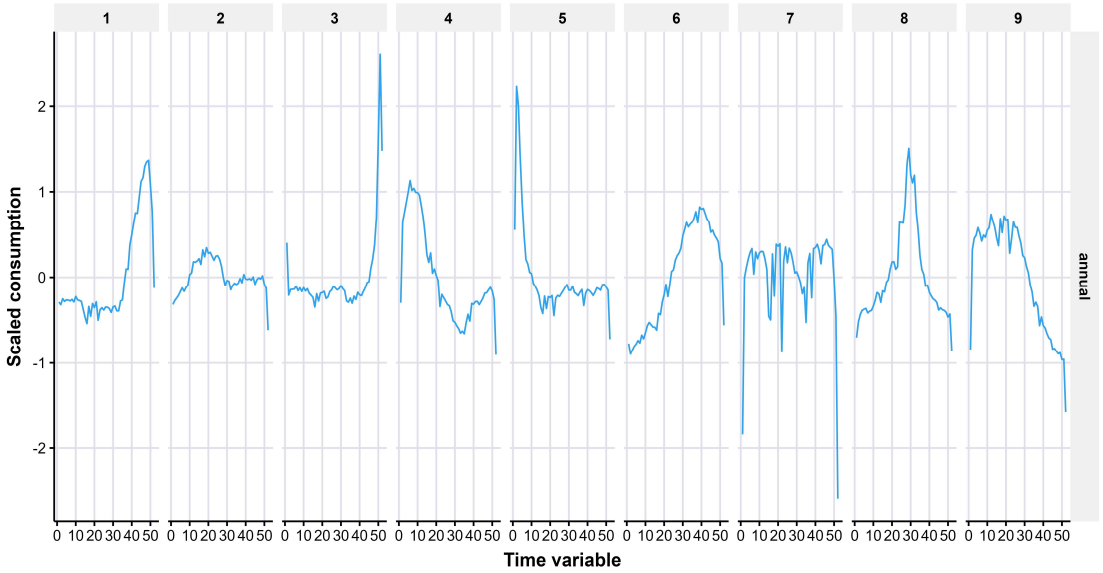


These findings serve as a basis for the decisions taken during the data-driven mapping exercise.

3.2.3 ABP data-driven mapping approach

In Project Discovery, the available logger data was used to identify typical consumption profiles (clusters) in NHHs. As a result of this exercise, 9 clusters were identified. For reference these are shown in Figure 5.

Figure 5: Annual profiles at weekly resolution



As a result of the clustering exercise, each individual property had been assigned to a cluster out of the 9 above. Since this was done at property level, properties with the same SIC code may have been spread across multiple clusters. However, the ultimate objective of this analysis was to associate SIC Codes to specific patterns of consumption. Therefore, a majority-based approach was used then to the determine the most suitable cluster for each SIC code. Where this majority-based approach was not feasible to apply, a set of individual rules was then applied to generate the final assignation (all this process and the details of the set of rules can be found in the Stage 2 report of Project Discovery).

This piece of work concerns the mapping of the ABP codes to these clusters. To preserve consistency, the approach to map these codes to the above-mentioned clusters has remained consistent with the SIC code clustering. That is, we first assigned a cluster to each property, then the assignation of codes to clusters was initially based on the majority, and for those codes where this approach was not feasible (i.e., not enough properties in a given ABP code, no representation of a given ABP code, etc.) a set of rules were then defined and tailored to the specific characteristics and data availability of the ABP classification. The specific details of this process are explained next.

One of the aspects we had to address during this analysis was deciding on the optimal level of granularity to use for ABP codes (i.e., primary, secondary, tertiary and quaternary levels, as explained in section 2).

Using the full quaternary ABP class code and filtering those code for which we had at least 5 properties in the sample, 64 codes had enough properties, covering 11% of possible codes. However, this represented 95% of properties within the sample.

Using the tertiary code increases this coverage to 17% of all available codes, and 96% of all properties, but decreases granularity. For example, tertiary code CI1 (Factory/Manufacturing) covers 24 ABP codes, including boat building, breweries, and brick works. These are differentiated by the COCOA functional grouping as hard manufacturing, drink processing, and furnace respectively. When limiting the code to tertiary level, these same properties are

all classified as Factory/Manufacturing. Based on the observed loss of granularity, we decided to use the full ABP code for mapping.

For codes with more than 5 properties, and following the majority-based approach, the most common cluster for that code was selected.

For those ABP codes where no majority could be identified or with 5 or fewer properties in the sample, the data-driven cluster was assigned using the standard cluster for the given default/functional group. The standard cluster for each functional group was determined using a majority-based approach whenever the number of properties in that functional group was equal or greater than 10 properties.

For example, ABP code C103GA is described as a "Servicing Garage" and the functional group for this ABP as per the COCOA schema was determined to be "Workshop". There were three properties in the sample with this classification, hence, less than the minimum sample requirement. Therefore, the standard data-driven cluster for "Workshop" was used for this ABP code.

In cases where the number of properties in the functional group was lower than 10, or when a majority could not be identified, the standard profile for functional groups derived during Project Discovery (therefore based on SIC codes) was used.

For example, the functional group "Military" contained only 5 properties in the ABP sample. Each of these five properties were assigned to a different data driven cluster, so no data driven cluster could be found for this group. This group was then assigned a data driven cluster of 2, using the assignation from the SIC schema.

It is worth it mentioning here that due to the sample size in the ABP data, and the fact that the majority of properties were offices and retail, some of the less common codes were not found in our dataset. This led to some of the default/functional and data-driven classifications found in the SIC code work being unavailable in the ABP code work.

3.2.4 Evaluation of data-driven classification against ABP codes

As a result of this mapping exercise, it was observed that, out of the 9 possible data-driven clusters, 8 contained at least one ABP code using this sample. Cluster 7 contained 36% of all ABP codes, and 71% of all properties, including the majority of offices and retail spaces. Cluster 2 was the second largest, containing 33% of all codes and 18% of all properties in the sample.

Figure 6 shows the matching between the functional groups and data-driven clusters assigned to individual ABP codes, where data-driven clusters could be mapped. Most functional groups showed strong links to only one data-driven cluster.

Some functional groups mapped on to multiple data-driven clusters, implying there are different usage types within that functional grouping. For example, the functional group "Social" includes ABP codes for pubs, village halls, and outdoor leisure centres, which can then be separated by their data-driven clusters (2, 4, and 8 respectively).

Retail is perhaps the most split group, containing 6 data-driven clusters in relatively similar proportions. Again, the data-driven clusters serve to unpick differences in usage, including showrooms, supermarkets, and petrol filling stations.

Figure 6: Relationship between data-driven clusters (right) and Functional groups (left) using ABP classification, where data-driven clusters are available

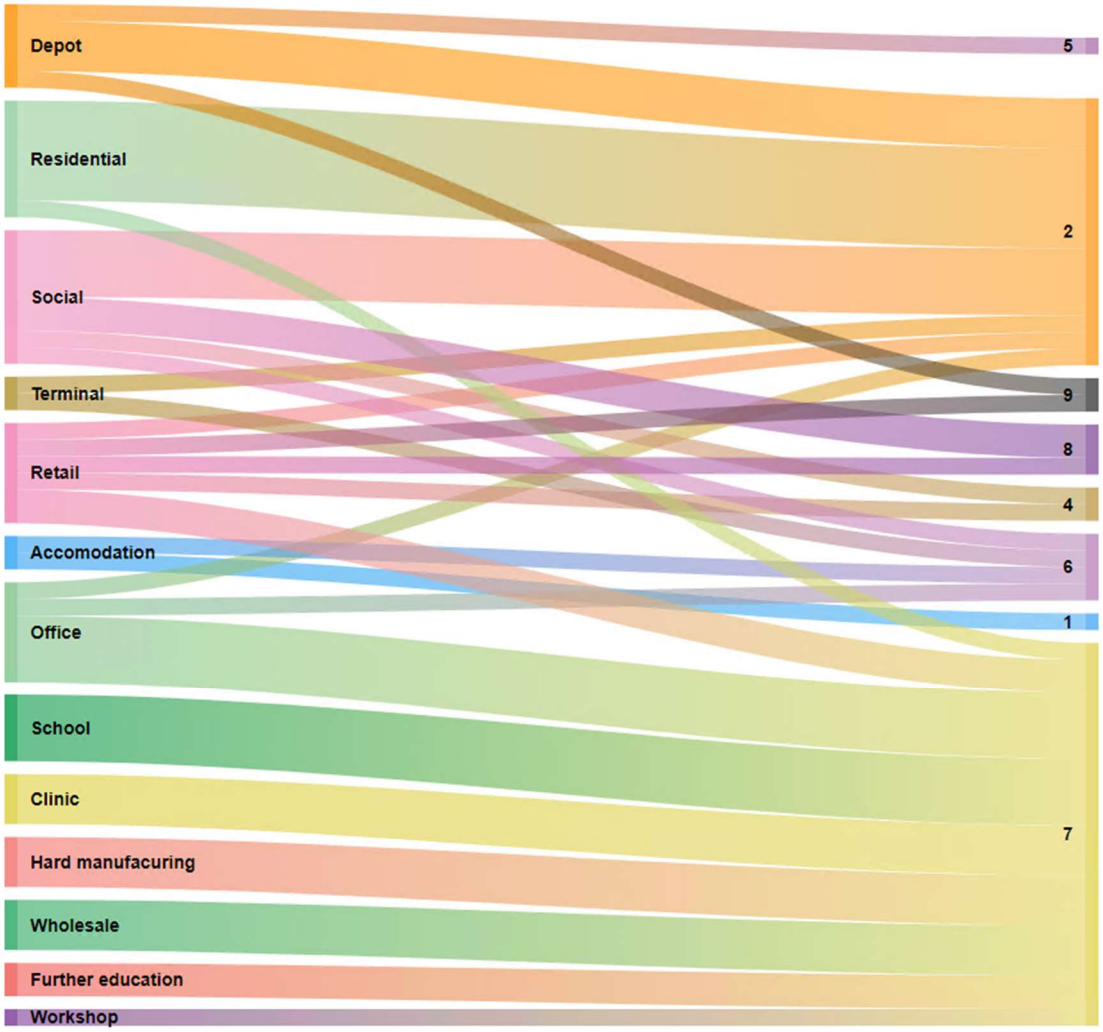


Table 5 shows the assignation of ABP codes to default/functional and data-driven classifications in the COCOA schema. One of the observations here is the existence of new strata in the schema, that is, combinations of functional groups and data-driven groups not previously found in the COCOA schema based on SIC codes. These strata are highlighted in green on the table. In the same way, rows in orange show those strata that were present in the COCOA Schema based on SIC codes but not using ABP. And finally, rows with a data-driven classification of NA could not be matched to a data-driven cluster using the rules above.

Table 5: COCOA classification: relationship between functional classification, data-driven classification, and ABP categories. New groups are highlighted in green. Groups present in SIC mapping but not in ABP mapping are orange

Default/Functional classification	Data-driven classification	ABP code
Accommodation	1	CH01

	6	CH CH01YH CH02 CH03 CL02 CL02CG CL02CV CL02HA CL02HO CL02YC CT06
	8	
	9	
Arable	1	
	2	
	8	CA CA03 CA03SH CA03VY CA03WB CC06 CC06CY CC06MC CL06PF CL06RG CR08GC L LA LA01 LA02 LA02OC LC LC01 LF LF02 LF02AU LL LM03 LP LP01 LP02 LP04 OO04 OP02
Clinic	2	
	7	CC06CR CM CM01 CM02 CM02HC CM02HL CM05 CM05ZS
Cooling	NA	CU03 CU03ED CU03EP
	1	
Depot	1	
	2	CB CC06CB CI04CS CI04PL CI04SO CI04TS CI05SF CI05TD CI08 CO01FM CR02EV CS CS01 CS02 CT01AF CT01AI CT01HS CT01HT CT03 CT03PK CT03PP CT03PU CT03VP CT04 CT04AE CT04CF CT04RH CT04RT CT05 CT07 CT09 CT10 CT10BG CT10BU CT11 CT11WG CT12 CT13 CT13NB CT13NF CT13TK CU CU01 CU03WF CU04 CU04WC CU04WD CU04WM CU04WS CU04WW CU06TE CU06TX CU08 CU08GG CU08GH CU08OT CU09 CU09CQ CU09OV CU09RA CU09SE CU12 CX CX07 LB MA99AG MF99UG MN99VG OH OH01 OI OI02 OI05 OI07 OS02 OS03 OT OT17 OT18 OT19 RB RC RC01 RG RG02 ZA ZM05 ZU ZV ZV01 ZV02 ZV02MI ZV02OI ZV02QI
	3	
	4	
	5	CT
	6	
	7	
	8	
	9	CU06
Drink processing	NA	CI01BW CI01CD CI01DY CI01WN
	5	

	8	
Food processing	NA	CA02FP CA04 CI01DA CI01FL CI01FO CI01OH CI01SR CR04FK
	1	
	2	
	4	
	6	
	7	
	8	
	9	
Furnace	NA	CC06CN CI01BR CI01GW CI01PG CI01SW CI07 OI09
	4	
	7	
	8	
	9	
Further education	7	CE01 CE01FE CE01HE CE05 CE06 CE07
Hard manufacturing	1	
	2	
	3	
	4	
	5	
	6	
	7	CI CI01 CI01AW CI01BB CI01MG CI01TL CI01YD
	8	
	9	
Hospital	4	CM03 CM03HI CM03HP
Industrial chemistry	NA	CI01CM CI01OR
	1	
	2	
	4	
	7	
	9	
Irregular school	7	

	9	
Lab	2	CC06MY CM04
	6	
	8	
Laundry	6	
Livestock	NA	CA01 CA02 CA02FF CA02FH CA02OY CN CN01 CN02 CN02AX CN03 CN03HB CN03SB CN04 CN05 CN05AN CN05MR CR04LV LB99AV
	2	
	4	
	5	
Military	NA	CX01 CX01PT CX02 CX02FT CX03 CX03AA CX04 CX05 CX06 M MA MA99AR MA99AS MA99AT MB MB99TG MF MF99UR MF99US MF99UT MG MN MN99VR MN99VS MN99VT OE OT06
	2	
	4	
Mineral particulates	NA	CI01CW LD LD01 LD01CC LD01CO LD01RN LD01TC
	1	
	3	
	6	
	7	
	8	
Mining	NA	CI02 CI02MA CI02MD CI02MP CI02OA CI02QA
	7	
	8	
Office	1	
	2	CC08
	3	
	4	
	5	
	6	CR09BS

	7	CC CC02 CC05 CC12 CL03 CL03RR CO CO01 CO01EM CO01GV CO01LG CO02 CR01 CR02 OE01
	8	
	9	
Residential	2	CC03 CC03HD CC03PR CC03SC OU05 OU08 RD RD01 RD02 RD03 RD04 RD06 RD07 RD08 RD10 RH RH01 RH02 RH03 RI RI01 RI02 RI02NC RI02RC RI03
	4	
	7	R
	9	
Retail	1	
	2	CR08CS
	4	CM06
	7	CR CR02PO CR04 CR08
	8	CR08SM
	9	CR05
Salon	2	
School	1	
	7	CE CE02 CE03 CE03FS CE03IS CE03JS CE03MS CE03NP CE03PS CE04 CE04NS CE04SS
Social	2	CC04YR CC07 CL06AH CL06BF CL06CK CL06CU CL06DS CL06EQ CL06FB CL06FI CL06GF CL06GL CL06GR CL06HF CL06HR CL06LS CL06ME CL06MF CL06QS CL06RF CL06SI CL06SK CL06SX CL06TB CL06TN CL06WA CL06WP CL06WY CL06YF CL07 CL07CI CL07EN CL10 CL10RE CR06 CR06BA CR06NC CR06PH CR09OL CR10 LB99PI LB99SV OP OU04 ZW ZW99AB ZW99CA ZW99CH ZW99CP ZW99GU ZW99KH ZW99LG ZW99MQ ZW99MT ZW99SU ZW99SY ZW99TP
	4	CC04
	5	
	6	CR09
	8	CL06 CR07
	9	

Soft manufacturing	2	
	4	
	5	
	7	
	8	CI01PW
	9	
Terminal	2	CT01 CT01AP CT01AY CT08 CT08BC CT08VH CT09CL CT09CX CT09MO CT13FR CT13SP CT13VF
	4	
	6	CT08RS
	8	
Textile manufacturing	NA	CI01PM
	2	
Tourism	1	CL CL01 CL01LP CL04 CL04AC CL04AM CL04HG CL04IM CL04MM CL04NM CL04SM CL04TM CL06HV CL07EX CL07TH CL08 CL08AK CL08AQ CL08MX CL08WZ CL09 CL11 CL11SD CL11SJ LP03PD OE05 OP03 ZM04 ZM05CE ZM05WI ZS
	4	
	6	
Waste processing	7	
	8	CC09 CC10 CI06 CU02 CU07 CU07WR CU07WT CU10
Wholesale	1	
	2	
	3	
	4	
	6	
	7	C CI04 CI05 CR04FV
	8	
Workshop	2	
	6	
	7	CI03 CI03GA
	8	
	9	

To finalise the schema, those strata that were not present in the original COCOA schema based on SIC codes and those classified as "NA" were reassigned to the standard data-driven cluster for each functional group (how the standard profiles for each functional group were selected is detailed in Stage 2 report of Project Discovery). For instance, for "Accommodation", the standard data-driven profile is 6, hence, ABP codes classified as 1 in Table 5 were reclassified to cluster 6 as presented in Table 6. This way, the final schema based on ABP codes is illustrated in this same table.

Table 6: COCOA Classification: relationship between final data-driven classification from Stage 2, functional classification from Stage 1, and SIC codes and classes

Functional Classification	Data Driven Classification	ABP Classification
Accommodation	6	CH01, CL02CG, CL02CV, CL02, CL02HA, CL02HO, CH02, CH, CH03, CT06, CH01YH, CL02YC
Arable	8	CA, LA, LL, LF02AU, LC, CC06CY, CC06, LF02, LF, OP02, CR08GC, LA01, LC01, CA03, L, LM03, OO04, CC06MC, LA02OC, LP, LA02, CL06PF, LP04, LP02, LP01, CL06RG, CA03SH, CA03VY, CA03WB
Clinic	7	CM05ZS, CC06CR, CM01, CM02, CM02HL, CM02HC, CM, CM05
Coolant	1	CU03ED, CU03EP, CU03
Depot	2	MF99UG, CT04AE, CT01AI, CT01AF, RC01, CB, LB, RB, ZA, MA99AG, CT10BG, CS02, CT10BU, CU09CQ, OI02, CT03, RC, ZV01, CC06CB, CT04CF, OI05, CI04CS, CU12, ZV02, CR02EV, CU01, CX, CO01FM, RG, CU08, CU08GG, CU08GH, CS01, CT04, CT13, CT01HS, CT01HT, OH01, OH, OI07, OI, CX07, RG02, CI08, CT05, ZV02MI, ZV02QI, CT13NF, MN99VG, CT13NB, CU09OV, ZV02OI, CU08OT, ZM05, ZV, CU09, CT12, CI04PL, CT03PP, CT03PU, CT03VP, CT03PK, CU04, OS02, CU09RA, CT04RT, OT19, OT18, CT07, OT17, CT04RH, CU09SE, CI05SF, CI04SO, CS, CT13TK, CU06TE, CU06TX, OS03, CI05TD, CI04TS, CT11, OT, CT09, ZU, CU, CT10, CU04WW, CU04WC, CU04WD, CU04WM, CU04WS, CT11WG, CU03WF
Depot	5	CT
Depot	9	CU06
Drink processing	8	CI01BW, CI01CD, CI01DY, CI01WN
Food processing	7	CI01DA, CR04FK, CA02FP, CI01FL, CI01FO, CI01OH, CA04, CI01SR
Furnace	4	CI01BR, CC06CN, CI01GW, CI07, OI09, CI01PG, CI01SW
Further education	7	CE01, CE01FE, CE01HE, CE07, CE06, CE05
Hard manufacturing	7	CI01AW, CI01BB, CI01, CI, CI01MG, CI01YD, CI01TL
Hospital	4	CM03HI, CM03HP, CM03
Industrial chemistry	2	CI01CM, CI01OR
Industrial chemistry	4	

Lab	8	CM04, CC06MY
Livestock	2	CN05, CN, CN02AX, CN05AN, CN02, LB99AV, CN01, CN03SB, CN03, CA01, CA02FF, CA02FH, CA02, CN03HB, CR04LV, CN05MR, CA02OY, CN04
Military	2	MF, MF99UR, MF99UT, MF99US, CX03AA, CX03, MB, MA, MA99AR, MA99AT, MA99AS, CX05, OT06, MG, OE, CX02FT, CX02, CX04, M, MB99TG, CX06, MN99VR, MN99VT, MN99VS, MN, CX01, CX01PT
Mineral particulates	6	CI01CW, LD01CC, LD01CO, LD, LD01, LD01RN, LD01TC
Mining	7	CI02, CI02MD, CI02MA, CI02MP, CI02QA, CI02OA
Mining	8	
Office	2	CC08
Office	6	CR09BS
Office	7	CR01, OE01, CO02, CO01GV, CC, CO01EM, CC12, CC02, CL03, CO01LG, CO, CO01, CC05, CL03RR, CR02
Residential	4	RD01, RI01, RI02, RD02, RD, CC03HD, CC03PR, RH02, RH03, RH01, RD07, RH, RI02NC, CC03, RD10, RI02RC, R, RI03, RI, CC03SC, RD06, RD03, OU08, RD08, RD04, OU05
Retail	2	CR08CS
Retail	4	CM06
Retail	7	CR04, CR02PO, CR, CR08
Retail	8	CR08SM
Retail	9	CR05
School	7	CE02, CE, CE03FS, CE03IS, CE03JS, CE03MS, CE03NP, CE04NS, CE03, CE03PS, CE04, CE04SS
Social	2	ZW99AB, CL06LS, CL06AH, CR06BA, CL07, CL06BF, ZW99CA, ZW99CP, ZW99CH, CC07, CL07CI, CL06SX, CL06CK, CL06CU, CL06YF, CL06DS, CL07EN, CL06EQ, CR10, CL06FI, CL06FB, CL06GL, CL06GF, CL06GR, ZW99GU, CL06HF, CL06HR, ZW99KH, CL10, ZW99LG, ZW99MT, CL06ME, ZW99MQ, CL06MF, CR06NC, CR09OL, LB99PI, OU04, ZW, CR06PH, CR06, CL06TN, CL06QS, CL10RE, CL06RF, CL06SI, CL06SK, OP, LB99SV, ZW99SU, ZW99SY, ZW99TP, CL06TB, CL06WA, CL06WY, CL06WP, CC04YR
Social	4	CC04
Social	6	CR09
Social	8	CL06, CR07
Soft manufacturing	7	CI01PW
Soft manufacturing	9	
Terminal	2	CT01AY, CT01, CT01AP, CT08BC, CT09CX, CT09CL, CT09MO, CT13FR, CT13SP, CT08, CT13VF, CT08VH
Terminal	4	CT08RS
Textile manufacture	2	CI01PM
Tourism	4	CL08AK, CL01, CL08AQ, CL11, CL04AC, CL04AM, CL09, ZM05CE, ZM04, CL07EX, CL04HG, CL06HV,

		CL04IM, CL, CL01LP, OE05, CL04NM, CL04MM, OP03, CL08MX, CL04, LP03PD, CL04SM, CL11SJ, CL11SD, ZS, CL07TH, CL04TM, CL08WZ, ZM05WI, CL08
Waste processing	7	CU02, CC09, CI06, CC10, CU10, CU07WR, CU07, CU07WT
Wholesale	7	C, CR04FV, CI04, CI05
Workshop	7	CI03GA, CI03

Using the ABP code classification, 16% of the properties in the sample had the same strata assignation as using SIC codes.

4 Benchmarking consumption models evaluation

Once the functional and data-driven mapping of ABP codes was complete, the next step was to evaluate if the benchmarking models developed in Project Discovery could be applied to the ABP mapping and/or whether any amendments or considerations had to be made. Equally important, was to evaluate the performance of the models on the ABP code mapping compared to the SIC code mapping.

4.1 Model implementation

Under Project Discovery, once the COCOA Classification was developed, granular data was used to develop models linking classification to consumption for each stratum (combination of functional and data-driven categories). Based on the findings from the EDA in Project Discovery, linear models based on building area and consumption were built across all groups in the COCOA levels and per month. Although the main objective was to create a model per group and month, due to certain limitations and data availability, some exceptions had to be implemented as follows:

1. Models were accepted for subgroups (considering both default/functional classification category and data-driven clusters) for which a sufficiently large sample size was available (over 10 SPIDs), and the final models demonstrated a significance level greater than 0.2. (**Strata models**).
2. In cases where subgroups did not meet the criteria in the first step, a second set of models was constructed, focusing on functional/default classification categories to increase the sample size. A more relaxed threshold for model significance was applied in this scenario, however, the requirement of at least a sample size of 10 remained in place. (**Functional/Default classification models**).
3. Finally, in instances where neither the models from step 1 nor step 2 were successful, a single universal model per month would be used. (**General model**).

The final models used in each case are summarised in Table 7.

Table 7: Models used per stratum

Default classification	Data driven classification	Selected model	Average adjusted R^2 value
Accommodation	6	Strata	0.377
	8	Strata	0.527
	9	Default Classification	0.403
Arable	1	Default Classification	0.205
	2	General	0.089
	8	General	0.089
Clinic	2	Default Classification	0.12
	7	Default Classification	0.12
Coolant	1	General	0.089
Depot	1	Default Classification	0.182
	2	Default Classification	0.182
	3	Default Classification	0.182
	4	Default Classification	0.182

	5	Default Classification	0.182
	6	Default Classification	0.182
	7	Default Classification	0.182
	8	Strata	0.531
	9	General	0.089
Drink processing	5	General	0.089
	8	Strata	0.735
Food processing	1	Strata	0.671
	2	Default Classification	0.728
	4	Default Classification	0.728
	6	General	0.089
	7	Strata	0.94
	8	Default Classification	0.728
	9	Default Classification	0.728
Furnace	4	General	0.089
	7	General	0.089
	8	General	0.089
	9	General	0.089
Further education	7	General	0.089
Hard manufacturing	1	Default Classification	0.622
	2	Default Classification	0.622
	3	Default Classification	0.622
	4	Default Classification	0.622
	5	Default Classification	0.622
	6	Default Classification	0.622
	7	Strata	0.708
	8	General	0.089
	9	Default Classification	0.622
Hospital	4	Strata	0.958
Industrial chemistry	1	Default Classification	0.281
	2	Strata	0.267
	4	Default Classification	0.281
	7	Default Classification	0.281
	9	Default Classification	0.281
Irregular school	7	Default Classification	0.047
	9	Default Classification	0.047
Lab	6	General	0.089
	8	General	0.089
Laundry	6	Strata	0.388
Livestock	2	General	0.089
	4	General	0.089
	5	General	0.089
Military	2	General	0.089
	4	General	0.089
	1	General	0.089

Mineral particulates	3	General	0.089
	6	General	0.089
	7	General	0.089
	8	General	0.089
Mining	7	General	0.089
	8	General	0.089
Office	1	Strata	0.473
	2	Default Classification	0.193
	3	Default Classification	0.193
	4	Default Classification	0.193
	5	Default Classification	0.193
	6	Strata	0.272
	7	Strata	0.2
	8	Strata	0.521
	9	Default Classification	0.193
Residential	4	Strata	0.485
	9	Default Classification	0.523
Retail	1	General	0.089
	2	Default Classification	0.156
	4	Default Classification	0.156
	7	Default Classification	0.156
	8	Strata	0.268
	9	Default Classification	0.156
Salon	2	Strata	0.431
School	1	Default Classification	0.53
	7	Strata	0.535
Social	2	Default Classification	0.102
	4	Default Classification	0.102
	5	Default Classification	0.102
	6	Default Classification	0.102
	7	Default Classification	0.102
	8	Default Classification	0.102
	9	Default Classification	0.102
Soft manufacturing	2	Default Classification	0.115
	4	Default Classification	0.115
	5	Default Classification	0.115
	7	Default Classification	0.115
	9	General	0.089
Terminal	2	Default Classification	0.117
	4	Strata	0.337
	8	Default Classification	0.117
Textile manufacture	2	General	0.089
Tourism	4	General	0.089
	6	General	0.089
Waste processing	7	General	0.089

Wholesale	1	Strata	0.419
	2	Strata	0.289
	3	Default Classification	0.299
	4	Strata	0.308
	6	Default Classification	0.299
	7	Strata	0.348
	8	Default Classification	0.299
Workshop	2	Default Classification	0.367
	6	Default Classification	0.367
	7	Strata	0.359
	8	Default Classification	0.367
	9	Strata	0.617

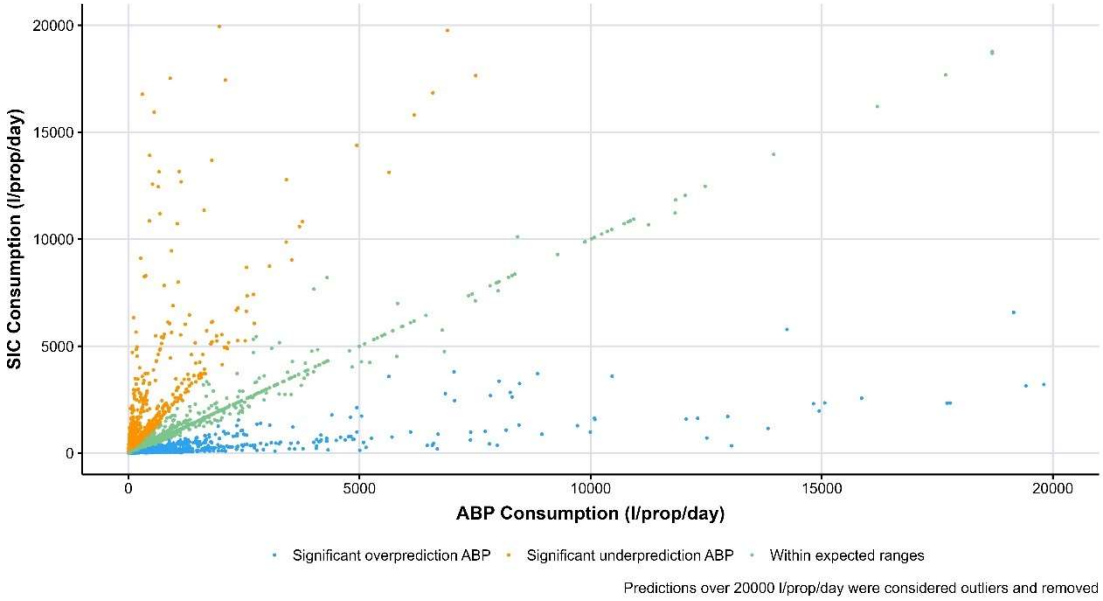
4.2 Model evaluation

The benchmarking models developed under Project Discovery, where applied to the sample. Results were then compared against predictions obtained using the COCOA schema based on SIC codes.

This is illustrated in **Error! Reference source not found.**, where the x-axis represents predictions using the ABP COCOA schema and the y-axis represent predictions using the SIC COCOA schema. Each individual point represents the averaged predictions for a single property. Points are color-coded to indicate the alignment of predictions. Green points show properties where the two sets of predictions are closely aligned. Blue and orange points represent properties where there is a significant difference, exceeding 50%, between the predictions.

We observe in this figure that for 31% of cases, predicted consumption for a property was identical using the ABP classifications as when using the SIC classifications. And for over 55% of properties alignment between both dataset is considered to be within expectation (green datapoints).

Figure 7: Predicted consumption from SIC and ABP models per property



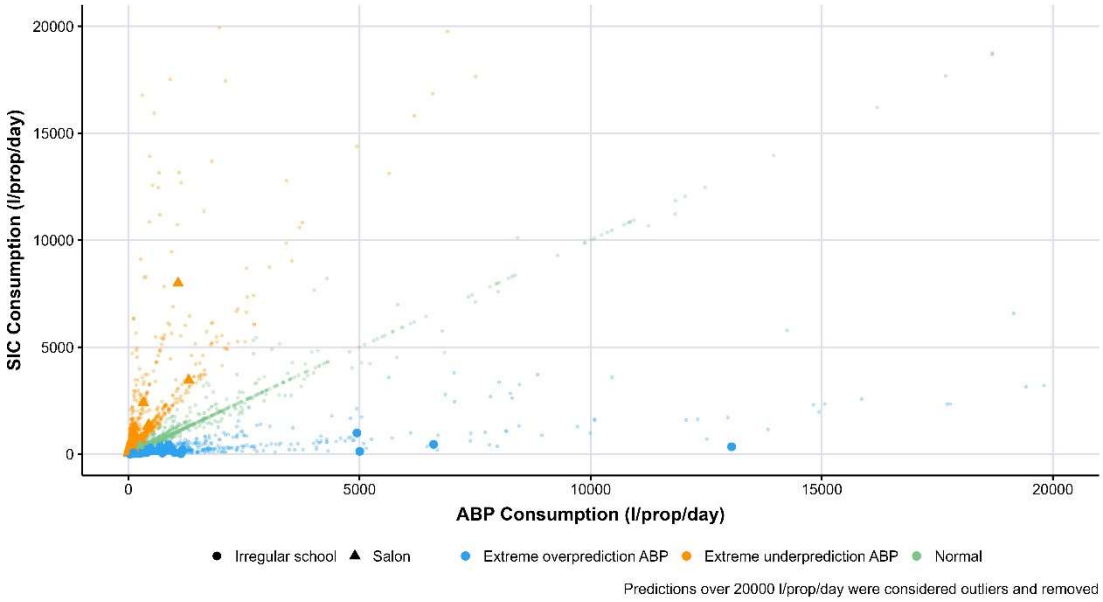
There are, however, properties where predictions using SICs and ABPs differ significantly. This is not fully unexpected. As explained in section 2, ABP codes are fundamentally less geared toward the understanding of demand than SIC codes because the focus is more on the building and land type rather than business activities. This had significant implications in the mapping of the ABP codes to functional groups and the subsequent application of benchmarking models.

For example, one of the functional groups that stands out is "Salon". Based on the work carried out under Project Discovery, "Salon" presents one of the highest consumptions per square meter compared to other functional groups. However, when using ABP, this functional group could not be identified, resulting in the reclassification of these properties to other groups, in most cases "Retail" or "Social" which had significant lower consumption per square meter as per the benchmarking models. It is not surprising then that this functional group is significantly underpredicted (compared to the prediction using SIC codes) when using ABP.

An example of the opposite scenario (extreme overprediction using ABP) would be "Irregular school". "Irregular schools" presents one of the lowest consumptions per square meter compared to other functional groups. However, when using ABP, this functional group could not be identified, resulting in the reclassification of these properties to other functional groups including, but not limited to, "Residential", "Wholesale", or "Workshop". As a result, most of the properties in this functional group (as per the SIC COCOA schema), are significantly overpredicted when using ABPs.

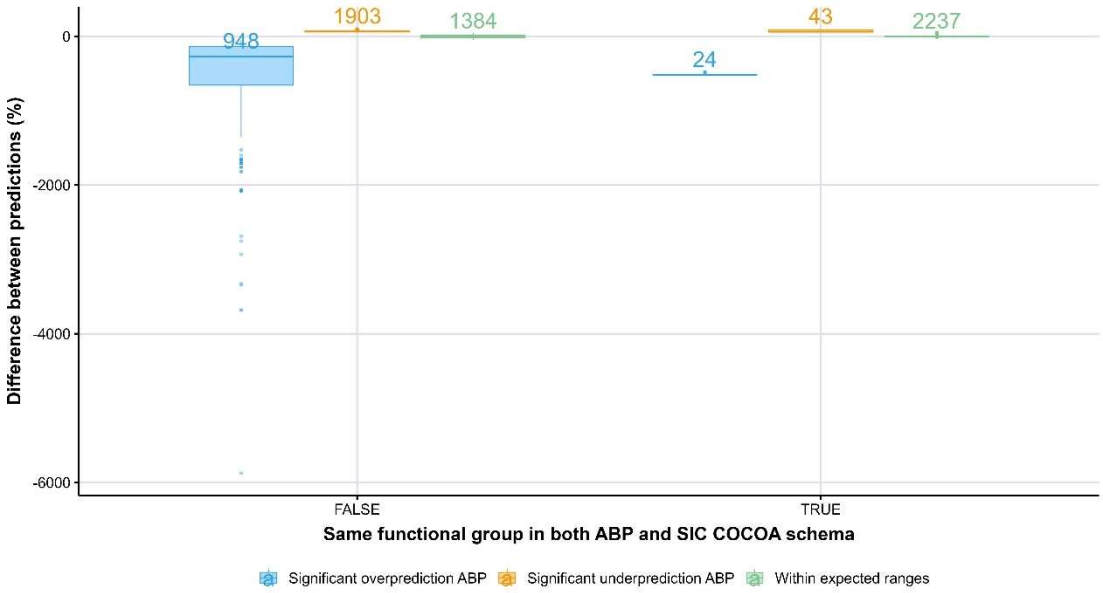
These examples are illustrated in Figure 8, where "Salons" identified by a triangular shape, and "Irregular schools" a circular shape.

Figure 8: Predicted consumption from SIC and ABP models per property. "Irregular schools" and "Salons" highlighted.



As a summary, we can evaluate the differences between predictions based on the functional group properties were allocated to. This is shown in Figure 9. It can be observed here, that for those properties were the functional group using ABP and SICs coincided (TRUE), most predictions remain within expected ranges. On the contrary (FALSE), the assignation of different functional groups leads to over 50% of predictions in properties to be out of expected ranges.

Figure 9: Difference in predicted consumption (%) using ABP and SIC codes, split by same or different assignation of functional group, and color coded by prediction status (within expected ranges, significant underprediction of ABP, significant overprediction of ABP). The numbers on the plot indicate the total number of properties per group.



Since the tool and methodology was developed specifically for SIC code which are much more granular and geared toward business classification, hence, it is fair to assume that results using SIC codes will be more reliable than those generated for ABPs. Based on this analysis, it would be beneficial to consider revisiting and tailoring the methodology for ABPs if deemed advantageous. It is important to remind here that the primary focus of this work was mapping ABPs to the existing schema, not reviewing or tailoring the methodology.

5 Limitations and recommendations

Aside from the recommendations and limitations stated in Project Discovery, this section highlights the main limitations and any recommendations for the purpose of the ABP code classification mapping.

- ABP data-driven clustering was carried out with a small sample of ABP codes, many of which were duplicates. Only 11% of possible ABP codes appeared more than 5 times in the sample, so could be used in the data-driven approach. As a result, the remaining ABP codes were mapped using the Default/Functional group.
- ABP codes are fundamentally less geared toward the understanding of demand than SIC codes because the focus is more on the building and land type rather than business activities. This resulted in the re-assignation of functional groups in the sample, and ultimately resulted in significant differences in the consumption predictions in the tested sample using the COCOA tool for a number of properties (section 4.2).
- Since the tool and methodology was developed specifically for SIC code, considering the particularities of this datasets, it is fair to assume that results using SIC codes will be more reliable than those generated for ABPs. Based on this analysis, it would be beneficial to consider revisiting and tailoring the methodology for ABPs if deemed advantageous.

6 Conclusions

Under Project Discovery, the COCOA schema was developed, and this project culminated with the generation of the COCOA tool. Within this tool, NHH customers were segmented in different strata with the aim of benchmarking water consumption. The development and usage of this tool was fully based on Standard Industry Classification (SIC) code classification of each non-household (NHH) property.

Following conversations and feedback provided, it was identified as an opportunity for development the adaption of the COCOA tool to incorporate AddressBase Premium (ABP) Classification Codes from Ordnance Survey (OS). This document has outlined the process and outcomes of updating the tool by mapping the ABP codes to the existing strata in the COCOA schema is presented in Table 5.

The review of the ABP classification and subsequent comparison to SIC classification revealed fundamental differences. The most relevant one being ABP primarily considering geographic and usage characteristics, while SIC focusing on economic activities. This contrast presented challenges during the mapping exercise, shedding light on areas for future development.