A Semantic Architecture for Theory Construction in the Residential Energy Consumption Domain

Abstract

Theory construction remains an open challenge in the artificial intelligence (AI) community. A theory construction module in an AI system for knowledge discovery must provide support for generating, developing and appraising theories. In this paper we explore a semantic architecture for a theory construction module in an AI system. The component incorporates an ontology for analogical reasoning that is used to generate new theories, a Bayesian network for theory development and cluster analysis for theory appraisal. The module is designed and evaluated in the energy domain, on a use case application for understanding the factors that influence electricity consumption behaviour of residential households in South Africa.

Introduction

Scientific knowledge discovery is a non-trivial process that is usually performed by humans. It consists of tasks that are difficult to formalise and automate. This process can however benefit greatly from the application of artificial intelligence (AI) tools and techniques incorporated in an AI system for scientific knowledge discovery (Krenn et al. 2022; Langley 2022; Moodley and Seebregts 2023). AI-driven knowledge discovery systems will be required to have the capacity to pursue scientific research, collect measurements, find regularities, form hypotheses, and gather additional data to test them (Langley 2022; Moodley and Seebregts 2023).

Haig proposed a theory of scientific method which is referred to as the abductive theory of method (ATOM) (Haig 2018). ATOM is considered to be broader than the commonly applied methods i.e. the hypothetico-deductive method and the inductive method. It consist of two main processes i.e. phenomenon detection and theory construction. Phenomenon detection deals with determining empirical regularities from data while theory construction is concerned with abductively inferring the existence of an underlying causal mechanism to explain the detected phenomena. Theory construction, according to Haig consists of three major sub-processes i.e. theory generation, development and appraisal. Haig suggests that theories can be generated abductively and developed using analogical modelling. These theories can then be appraised using the inference to the best explanation or theory of explanatory coherence.

Wanyana and Moodley used ATOM to design an intelligent agent architecture for knowledge discovery and evolution in physical data driven systems (Wanyana and Moodley 2021). The system was designed and evaluated for understanding electricity consumption behaviour in South African households using an application use case approach. A hybrid AI approach was used and it was demonstrated how different AI techniques i.e. unsupervised machine learning, ontologies and Bayesian networks can be leveraged in an AI system for scientific knowledge discovery. The theory construction module of the architecture however, was limited in functionality and they acknowledged that it required further exploration. It used a weak form of analogical reasoning using cluster analysis. In this paper we extend this work focusing on and delving deeper into the design of a theory construction module. This module incorporates a domain ontology that supports analogical reasoning based on existing scientific knowledge. We use the same application, i.e. understanding household consumption behaviour.

The main contribution of this paper is an approach for integrating different AI tools in the theory construction workflow. Our focus is on the residential electricity consumption domain. We propose an ontology that captures aspects of scientific knowledge about domestic electricity consumption in such a way that theories can be drawn from it using analogical reasoning. We demonstrate how the ontology can be used to construct Bayesian causal models that are used to develop the generated theories as well as how a theory can be appraised on a given data set.

Background and Related Work

Analogical reasoning

Analogical reasoning is important in science and it lies at the heart of analogical modelling (Haig 2018). It is a core cognition process that produces inferences and new ideas by applying previous knowledge and experience (Han et al. 2018). It uses relational similarity across different contexts and it is regarded as a fundamental component of scientific discovery (Gentner and Smith 2013; Han et al. 2018). The conventional analogical form is generally described in a likeness relation of $A : B :: C : D$. This indicates that C is related to

D in the target domain as how A is related to B in the source domain (Casakin and Goldschmidt 1999). This analogical form transforms into the problem $A : B :: C : X$ where X is required to be established. According to Feldbacher-Escamilla and Gebharter, analogical reasoning focuses on trying to find "systems *S'* that are analogous or similar enough to systems *S* about which *H* claims this and that, and formulating a corresponding hypothesis *H'* for these similar enough systems *S'* " (Feldbacher-Escamilla and Gebharter 2020). This argument structure takes the format below:
Similarity Premise: Generally, case *S'* is similar to case *S*
Base Premise: Q' holds in *S'*
Conclusion: Q, which is similar to *Q'*, holds in *S*.

ATOM deals with the generation of explanatory theories and so, Haig uses the term analogical abduction to refer to the explanatory analogical modelling involved. The general argument schema for analogical reasoning according to Haig is as follows:
Hypothesis H about property Q was correct in situation S1.*
Situation S1 is like the situation S2 in relevant respects.
Therefore, an analogue of H might be appropriate in situation S2.*
Haig does not explicitly define what a situation is but gives an example from Darwin's theory i.e.

"*The hypothesis of evolution by artificial selection was correct in cases of selective domestic breeding*.

Cases of selective domestic breeding are like cases of the natural evolution of species with respect to the selection process.

Therefore, by analogy with the hypothesis of artificial selection, the hypothesis of natural selection might be appropriate in situations where variants are not deliberately selected for".

Domestic Electricity Consumption

Understanding daily electricity consumption behaviour of residential customers is essential to predict long term demand and planning for national energy infrastructure (Toussaint and Moodley 2020). Electricity use in domestic buildings results from occupants' need for energy services, such as light, comfort, cooking and entertainment. Understanding and predicting energy consumption is complex and is influenced by multiple interlinked and interacting socio-economic, dwelling and appliance related factors (Jones, Fuertes, and Lomas 2015).

We focus on an application use case on consumption behaviour for households in South Africa, where economic volatility, income inequality, geographic and social diversity contribute to increased variability of daily household consumption behaviour. We draw from an existing study that combined both expert domain knowledge and cluster analysis to determine patterns of daily electricity consumption by South African households (Toussaint and Moodley 2020). The South African Domestic Electrical Load (DEL) data sets (Toussaint 2020) used in this use case were collected during the National Rationalised Specification Load Research Programme from 1994 to 2014. The data sets contain metered household electricity consumption data, and socio-demographic survey data for a diverse sample population spanning urban, informal and rural environments, five climatic zones, a large spectrum of income groups, newly to long-term electrified households, and different dwelling structures in South Africa and Namibia.

Ontologies for electricity consumption behaviour

Some ontologies have been designed in the domestic electricity consumption domain. However, these focus on energy saving and management and not on achieving analogical reasoning towards the construction of theories as well as representing known theories. For instance, Kott and Kott created an ontology to capture knowledge about electrical devices, energy sources and energy reservoirs found in households with the focus on the possibility of creating models for effective micro-grid management (Kott and Kott 2019). The taxonomy in this ontology captures home appliances like appliances for cleaning, garden, air conditioning, kitchen, body care, home office, entertainment, heating, home utility systems and transport ignoring other household attributes.

Shah et al. introduce an electrical appliances domain ontology that inherits concepts and relationships described in SUMO ontology (Shah et al. 2011). Focus is placed on electrical appliances and therefore other household attributes like socio-economic attributes and dwelling factors as well as contextual aspects like weather are not considered.

Saba et al. developed an ontology for intelligent energy management in order to save energy (Saba et al. 2019). The solution is built using OWL and SWRL. The ontology is designed in such a way that the concepts and rules are towards actions for saving energy.

The objective of the ThinkHome ontology (Reinisch et al. 2011) is to realise energy efficient, intelligent control mechanisms. To design the ontology, seven categories of information are brought together. These are: 1) building information e.g. walls, materials, lay out spaces, 2) actor information e.g. schedules, preferences and context, 3) process information e.g. system process and user activities, 4) exterior influences e.g weather and climate, 5) comfort information e.g thermal comfort and visual comfort, 6) energy information e.g. environmental impact and energy providers, 7) resource information e.g. building automation services.

Putra, Hong, and Andrews describe the development of an ontology of occupant behavior. The ontology extends the drivers, needs, actions and system (DNAS) framework (Putra, Hong, and Andrews 2021). Although the drivers, needs, actions and system remain unchanged, the extension includes more detailed characteristics of an occupant e.g. socio-economic characteristics, geographical location, subjective values, occupant activities, individual and collective adaptive action. This ontology falls short as far as the objective of our proposed ontology is concerned since their focus is only on occupant characteristics and behavior and not appliances or detailed dwelling factors.

Ontologies for analogical reasoning

Some work has been done on analogical reasoning and ontologies. Forbus, Mostek, and Ferguson proposed an analogy ontology that represents key entities and relationships in analogical processing (Forbus, Mostek, and Ferguson 2002).

Their ontology is based on Gentner's structure-mapping theory, a psychological account of analogy and similarity. Raad and Evermann used analogical reasoning to support ontology alignment (Raad and Evermann 2015). Basing on an ontology and embracing the aspect of analogical reasoning, an idea generation tool called the retriever was developed (Han et al. 2018). Retriver is designed to solve proportional analogical problems of the form: $A : B :: C : X$. Retriever inputs a known term C and returns it along with two re-representations and three unknown Xs retrieved by the tool through the association" or "related to" functions. Alsubait, Parsia, and Sattler present an approach for generating analogies in multiple choice questions of the form: "A is to B as C is to D" for students assessment (Alsubait, Parsia, and Sattler 2012). The analogies are mined from existing description logic ontologies.

Theory construction

We use the residential electricity consumption domain as an application use case for designing the theory construction module and focus our use case on understanding consumption behaviour of households in South Africa. Current studies in this domain attempt to identify and understand the complex interplay between socio-economic factors in different countries. While some studies in South Africa have already identified certain factors that influence usage behavior, these are not exhaustive. Our goal is to determine other possible factors that we do not know yet that might influence electricity usage behaviour in South African households.

We therefore perform analogical reasoning to identify other factors that could possibly explain certain usage behaviours. These possible factors can be developed further into theories by first constructing and situating the factor in a causal model and then appraising it on a given data set. Figure 1 shows the components and sub processes of our proposed theory construction module. The module incorporates an ontology, Bayesian network and cluster analysis for the different steps of the theory construction process.

Theory generation

The knowledge discovery task is to discover new knowledge and understand electricity consumption behaviour of households. Drawing largely from Haig's argument schema presented in the background and related work, we formulate an argument schema for analogical reasoning for discovering knowledge pertaining to residential electricity consumption behaviour in a particular country. Haig's argument schema draws analogies from looking at *how situations are similar in relevant aspects*. We refer to these relevant aspects as the context. The notion of similarity in this case involves the consumption behaviour of households in different countries. For example we would expect a rural household in South Africa to have a similar context, and thus similar consumption behaviour to a rural household in Botswana, a neighbouring country, compared to a rural household in Germany.

The property in this case is the electricity consumption behavior. A theory is considered to be an explanation for some electricity consumption behavior in terms of the factors that influence electricity consumption. The goal is therefore to obtain a theory that is correct in one country and then consider the same theory for another country with similar context and establish if it holds.The former country is considered as the source country while the latter is considered as the target country.

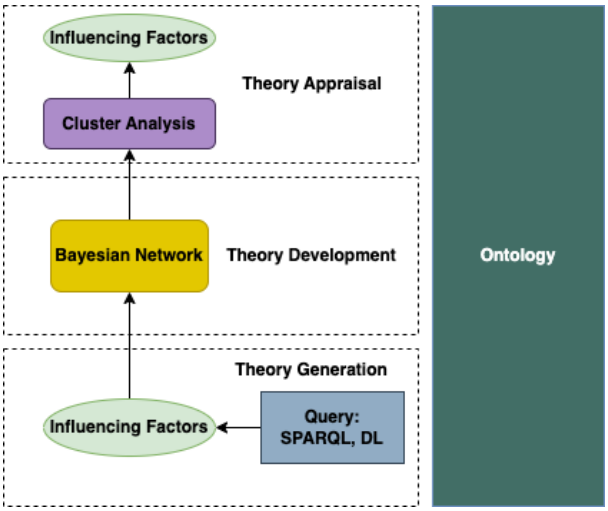

Figure 1: The sub processes of theory construction and the respective components applied.

Therefore, the general argument schema for analogical reasoning used is as follows:
Theory τ about property P was correct for households in source country S.
Source country S is like target country T in relevant aspects.
Therefore, theory τ might be appropriate for households in target country T.

A simple example of this form of analogical reasoning would be:

Geyser ownership positively affects domestic electricity consumption in households in Botswana.

Botswana is like South Africa with respect to the fact that they are both Sub-Saharan African countries.

Therefore, by analogy, geyser ownership may positively affects domestic electricity consumption in households in South Africa

Analogical reasoning with the help of an ontology is applied in the process of theory generation. A concrete representation of analogical reasoning using an ontology is shown in algorithm 1 lines 1-4. The purpose of the ontology is to articulates the concepts and properties required for analogical modelling in theory construction from known theories in domestic electricity consumption. It specifies factors that influence domestic electricity consumption supported by published studies. The algorithm shows the procedure applied in using the ontology to obtain lines of inquiry. The input to the algorithm is the ontology and a target country. The process starts by determining the context of the target country on which to base the similarity (Algorithm 1, line 1).

After the context is determined, the next step is to retrieve all available studies conducted in countries with same con-

Algorithm 1: Theory construction in domestic electricity consumption

Input: ontology O, target country T
Output: potential theory τ_p

1: Determine the context C of T to base the analogy on.
2: Retrieve studies conducted in countries with the same context C
3: Remove influencing factors that are known for T from the studies
4: τ = Influencing factors that are not known for T.
5: Select on of the factors, τ_i, to follow up
6: Retrieve direct super classes of τ_i from O
7: Add τ_i to Bayesian network
8: Query O to determine if variables in the Bayesian network are features in the data set
9: Confirm if τ_i may hold on the data set
10: $\tau_p \leftarrow \tau_i$
11: **return** τ_p

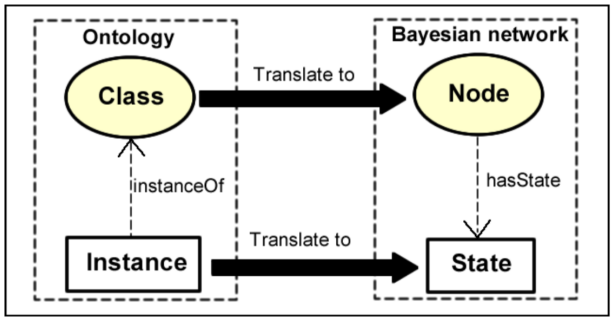

Figure 2: Transforming ontology classes and instances to nodes and states of the corresponding Bayesian network (Ogundele et al. 2017).

text as the selected context. From the retrieved studies and their respective countries, influencing factors that are known for the target country T are removed leaving a set of influencing factors that are not known to apply to T (Algorithm 1, line 3-4). These form the lines of inquiry to be followed up in T. Details are presented using an example use case from the domestic electricity consumption domain.

Theory development

A Bayesian network is used to situate a possible factor in a causal model, to further explore and explain how the factor can influence energy consumption behaviour. Using the approach proposed by Ogundele et al. (Ogundele et al. 2017), we generate a Bayesian network using the ontology.

Algorithm 1, line 4 returns instances of influencing factors that form theories. Before constructing the Bayesian network to develop a theory for a given influencing factor, it is important to first establish the respective corresponding direct super classes for these instances in the ontology (Algorithm 1, line 6). A DL query is used for this.

In Algorithm 1, line 7, a Bayesian network is generated to develop the theory. To generate the Bayesian network, as

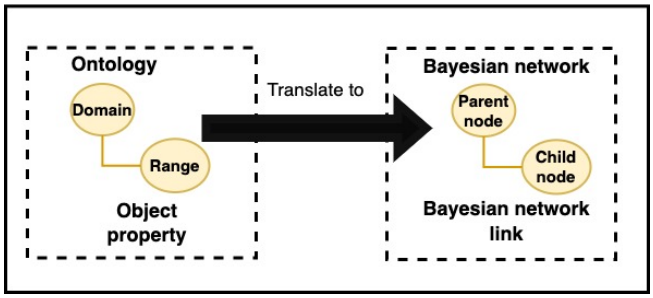

Figure 3: Transforming object properties and the respective classes into links and nodes respectively in the corresponding Bayesian network.

suggested by Ogundele et al. (Ogundele et al. 2017), the corresponding direct super class of a given influencing factor of choice is converted into a node and its instances into states. Figure 2 shows the mapping of the classes and instances in the ontology to nodes and states in the Bayesian network respectively. To generate the links and obtain the rest of the nodes that should be included in the network, all object properties for which the chosen influencing factor participates as a *domain* or *range* in the ontology are retrieved. These are added to the Bayesian network as links and the respective participating nodes are added as the other nodes in the network (see figure.3). If the influencing factor is a domain, then the link is an outgoing one and the respective node is a parent node. On the other hand, if the influencing factor is a range, then the link is an incoming link and the respective node is a child node.

In a case where there is no direct relationship between a factor in the theory and the property under study i.e. consumption behavior, then the intermediate factor is also included in the network.

Theory appraisal

The ontology is designed to capture the data set details including the features it contains. The ontology can therefore be queried to check if the features in question are present in the data set that is to be used to appraise the given theory (Algorithm 1, line 8). The theory is appraised on a respective data set using an applicable machine learning technique, in this case, cluster analysis to either confirm whether it may hold or refute it in the given target location (Algorithm 1, line 9). The human however has to check the theory in order to fully confirm it.

An ontology for theory construction in the Domestic Electricity Consumption Domain

This section presents a conceptual model and ontology for theory construction in the domestic electricity consumption domain.

Design and approach

The approach applied in the design of the ontology is a slight variation of the Unified Process ONtology (UPON) building

methodology (De Nicola, Missikoff, and Navigli 2009) proposed by Ogundele et al. (Ogundele et al. 2015, 2016). In this approach, the ontology is used to consolidate and represent categorical knowledge obtained from an unstructured data source. The ontology is further used to construct decision networks. This approach consists of the following steps: Definition of design and purpose, knowledge acquisition, model design, model analysis, model formalisation and ontology evaluation. The involves identifying, structuring and categorising factors that have either a positive or negative influence on a given property of the domain. This approach has been used effectively for systematically acquiring and structuring scientific knowledge for interactive decision making in two different domains, i.e. the health domain (Ogundele et al. 2015, 2016, 2017) and the finance domain (Drake and Moodley 2021). We have also found that the approach aligns well with the structure of knowledge in the electricity consumption domain.

Conceptual model

A literature review was conducted to identify and structure the factors that influence domestic electricity consumption in the residential sector as well as their categories. To design the conceptual model, we use the categorisation of the factors that influence domestic electricity consumption that is presented by (Jones, Fuertes, and Lomas 2015) in their review of the field. This categorisation encompasses the socio-demographic factors presented in other categorisations e.g. the categorisation according to (Frederiks, Stenner, and Hobman 2015) and the categorisation according to (Mutumbi, Thondhlana, and Ruwanza 2022) which categorise the factors into socio-demographic factors and psychological factors. The categorisation of (Jones, Fuertes, and Lomas 2015) also states whether a given factor has a significant positive influence, a significant negative influence or a neural effect on domestic electricity consumption based on the studies used in their review.

Based on this categorisation we place factors into three major categories: 1) Socio-economic factors which encompasses the demographic factors like age, employment status, family composition among others 2) physical dwelling factors e.g. dwelling floor area and the number of room and 3) appliance factors including appliance ownership e.g entertainment, cooling, water heating and cooking appliances as well as appliance usage.

Context is relevant when establishing the similarity between the source and target locations during analogical modelling. Therefore, the context of the studies has to be captured. The contextual similarity applied during the process of analogical reasoning in this study is drawn from the features of the study from which the influencing factors are obtained. The features of the studies that are captured by Jones, Fuertes, and Lomas that do not constitute influencing factors are: - the Country that was studied e.g Belgium, India, Norway, South Africa etc, - the location of the study in terms of city e.g. Antioch, England, Hong Kong etc, - the type of community that was studied i.e. rural vs urban, - the geographic region of the study i.e. Europe/UK, the rest of the world, and - the kind of weather in that area where the study was conducted e.g. "...the data was collected in the winter of 2008", "...summer and winter of 1999 to 2000", "hot climatic zones" etc. Although all these features of the studies are captured, they are not all regarded as context. As far as context is concerned, only community, geographic region and weather are considered to form the study context. This implies that the contextual information as well as the notion of similarity can be drawn from whether the source and target country or city are in the same region, have similar weather or whether the communities under consideration are both either rural or urban.

Although Jones, Fuertes, and Lomas divide the geographical regions into only two categories i.e. Europe/UK and the rest of the world, we apply the world bank classification of countries[1]. This is because the categorisation of countries' geographical regions by Jones, Fuertes, and Lomas provides very little contextual information on the studies.

The world bank classifies countries basing on the income level, geographic region and operational lending for the presentation of key statistics. This is also done to allow users to aggregate, group and compare statistical data as needed. In this study, interest is placed on the classification by geographical region. The world bank classifies countries into the following seven geographical regions: East Asia and Pacific, Europe and Central Asia, Latin America and the Caribbean, Middle East and North Africa, North America, South Asia and Sub-Saharan Africa. The world bank geographical regions are considered because some work from world bank on household electricity consumption specific to given geographical regions e.g. Sub-Saharan Africa (Blimpo, Postepska, and Xu 2020) has been done and we can draw from it. These geographical regions along with the rest of the contextual information i.e. community and weather as well as the influencing factors are formalised using the domestic electricity consumption ontology.

The DEConto ontology

This section presents an ontology that formalises the conceptual model presented in the previous section.

The factors that influence domestic electricity consumption are formalised as the Domestic Electricity Consumption ontology (DEConto) using OWL and Protégé. Figure 4 shows the different classes or concepts represented in the ontology.

The `Influencing_factors` class captures the factors that influence domestic electricity consumption. These are categorised into three main categories which form three of the sub classes of the `Influencing_factors` class i.e. `Appliance_factors`, `Dwelling_factors` and `socio-economic_factors`.

The ontology represents the `Study` class which captures the details of the study that asserts that a given factor indeed affects domestic electricity consumption. This relationship is represented using the object property `assertsInfluencingFactor`. The `assertsInfluencingFactor` object property has

[1]https://datahelpdesk.worldbank.org/knowledgebase/articles/378834-how-does-the-world-bank-

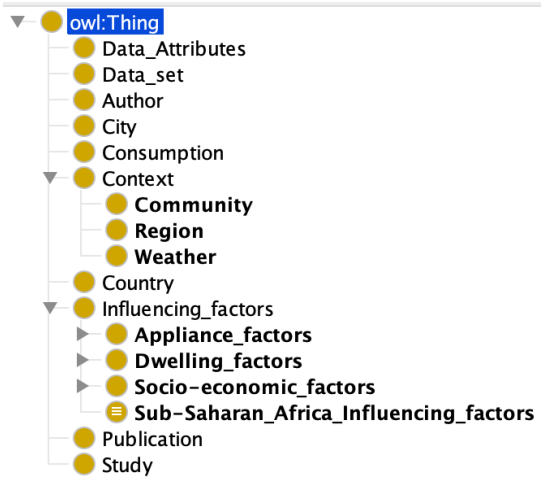

Figure 4: The classes in the domestic electricity consumption ontology.

three sub properties i.e. `assertNegativeInfFactor`, `assertPositiveInfFactor` and `assertNeutralInfFactor`. These factors capture the direction of influence that a factor has on consumption i.e. the factor leads to reduction in consumption, increase in consumption or it has no effect on the consumption respectively. A study is published as a publication which is represented as the `Publication` class. A publication refers to a published or publishable document that contains evidence that asserts an influencing factor. It indicates the source of the theory. The relationship between *study* and `publication` is captured using the `isPublishedas` object property. The ontology also captures the city and country in which the study was conducted represented using the classes `City` and `Country` respectively. The relationship between the study and the country in which it was conducted is captured using the object property `hasCountryofStudy`. The year in which the study was carried out is also recorded using the data property `hasYearofStudy`.

A publication has authors, represented using the `Author` class. This relationship is captured using the `hasAuthor` object property. The publication also has some other properties captured using data properties i.e. the digital object identifier (DOI), the uniform resource locator (URL) and the publication year.

The `Consumption` class represents the dependent variable consumption that is determined by the influencing factors. The `Consumption` class has four instances i.e. `very high` , `high` , `medium` and `low`. The relationship between the influencing factor and consumption is captured using the `leadsTo` object property.

The ontology captures the context to facilitate the understanding of some auxiliary aspects of the studies represented in the ontology. Context is represented as the class `Context`. Context can be the community in which the study was conducted i.e. rural or urban. It can also be region e.g. Sub-Saharan Africa or weather e.g. summer or winter.

The ontology has a `Data_set` class. The country in which the data set was collected is also captured. This is represented using the object property: `wasCollectedin`. The attributes contained in a given data set are captured using the `Data_Attributes` class and the relationship between a data set and its attributes is represented using the `hasFeatures` object property.

Evaluation

Theory generation using Analogical reasoning

Let us consider South Africa as our target country and the regional context of Sub-Saharan Africa in which it belongs. We need to retrieve the studies conducted in Sub-Saharan Africa along with their respective locations e.g. country. This can be achieved from the ontology using the SPARQL query below.

```
SELECT ?Study ?Country
    WHERE {
         ?Study deconto:hasCountryofStudy
?Country.
         ?Country deconto:isinRegion
deconto:Sub-Saharan_Africa}
```

Among the studies and countries returned, a specific source country may be determined. The retrieval of the studies and their respective countries is therefore followed may be followed by selecting a source country and determining the factors that influence consumption as per the studies conducted in that country. For example, "in Botswana, the presence of geysers leads to higher electricity consumption" can be a source situation with Botswana as the source country that may need to be explored in another geographical location e.g. in South Africa to determine if it is still the case. A filter is used in the query to only retrieve influencing factors that are in the source country but are not in the target country. This is done in order to generate new lines of inquiry.

Considering the studies captured in the ontology, the filter SPARQL query below returns the influencing factors that are published in studies from Botswana but are not in studies published from South Africa. As far as the current state of the ontology is concerned, the query only returns one influencing factor i.e. *geyser_present*. By applying analogical reasoning, on a general note, this kind of query retrieves influencing factors that can be considered as good lines of inquiry in the target country, South Africa.

```
   SELECT ?Influencing_factors
     WHERE {
         ?StudyBT
deconto:hasCountryofStudy
deconto:Botswana.
         {{?StudyBT
deconto:assertNegativeInfFactor
?Influencing_factors} UNION
         {?StudyBT
deconto:assertPositiveInfFactor
?Influencing_factors} UNION
         {?StudyBT
deconto:assertNeutralInfFactor
```

Influencing_factors and (influencingfactorAssertedby some (Study and (hasCountryofStudy some (Country and(isinRegion value Sub-Saharan_Africa)))))

Figure 5: Studied influencing factors that are specific to Sub-Saharan African Countries.

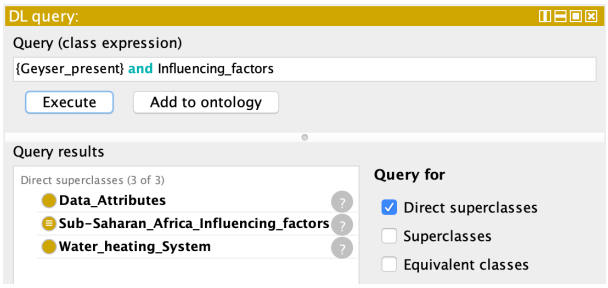

Figure 6: DL query to determine the direct superclasses of the returned influencing factors

```
?Influencing_factor}}
     FILTER NOT EXISTS
          {{ ?StudySA
deconto:hasCountryofStudy
deconto:South_Africa.
          {{?StudySA
deconto:assertNegativeInfFactor
?Influencing_factors} UNION
          {?StudySA
deconto:assertPositiveInfFactor
?Influencing_factors} UNION
          {?StudySA
deconto:assertNeutralInfFactor
?Influencing_factor}}.}} }
```

An inferred class in the ontology can also be used to obtain the required lines of inquiry. For example, the description logic (DL) query in figure 5 is used to create an inferred class that can be used to directly obtain influencing factors specific to only Sub-Saharan Africa. This class can be used to obtain lines of inquiry when dealing with a target location that is in the Sub-Saharan region. In this case, a user would have to choose from the instances provided by the ontology which of them they would want to follow up.

Theory development using a Bayesian network

Considering the infuencing factors, to clearly represent this and further develop the theory, a Bayesian network for the factor(s) that are under question is drawn from the ontology. The Bayesian network is designed following the mappings in figures 2 and 3.

The SPARQL filter query above returns instances of the influencing factors. Before designing the Bayesian network, it is important to first establish the respective corresponding direct superclasses for these instances. This is achieved using a DL query as shown in figure 6. Based on the output in the SPARQL filter query above, the corresponding direct superclass of interest for `Geyser_present` returned by the DL query in figure 6 is `Water_heating_System`.

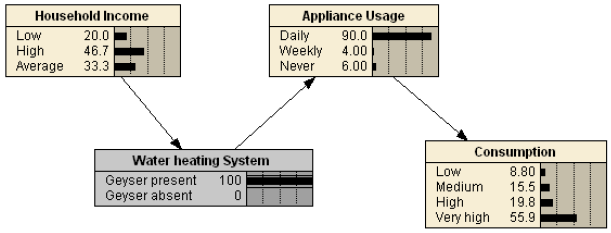

Figure 7: A sample Bayesian network for the influence of geyser ownership on electricity consumption.

The instances of `water_heating_System` i.e. `Geyser_absent` and `Geyser_present` are obtained from the ontology and made states in the `water_heating_System` node.

Figure 7 shows the Bayesian network obtained from the ontology to develop the the *geyser_present* influencing factor into a theory. As far as the geyser ownership Bayesian network is concerned, consumption behavior is the target variable. The ontology contains an object property: *affects* which captures the relationship between household income and all sub classes of *appliance_ownership* including *water_heating_System*. Also, `water heating systems` influences *Water heating appliance usage* which in turn influence electricity consumption. This may be explored to check whether it also holds in the South African households.

Theory Appraisal Using Machine Learning

The generated theory (see the Bayesian network in figure 6) i.e. household income affects appliance ownership, specifically `geyser_present` which affects its usage and in turn positively influences domestic electricity consumption is further explored on the South African domestic electricity load data set to either confirm if it may hold or to refute it. The ontology is first queried to determine if the variables in the obtained Bayesian network are present as features in the data set at hand. This can be achieved using a sample query shown below.

```
SELECT ?Data_Attributes
     WHERE { deconto:Domestic_electricity
_load_monitoring_data
deconto:has_features ?Data_Attributes.}
```

K means clustering, which has been shown to be effective for cluster analysis on this data set (Toussaint and Moodley 2019), can now be used to perform cluster analysis with respect to income since, from the Bayesian network, income affects ownership of water heating systems. We performed cluster analysis on a subset of the data set, on households in Cape Town, a major city in South Africa. The optimal number of clusters was determined using the elbow method which involves plotting the within-cluster sum of squares (WCSS) or inertia versus the number of clusters as shown in appendix 1, figure 11. The elbow method seeks to identify the point where the rate of decrease in WCSS begins to slow down beyond which only results in marginal improve-

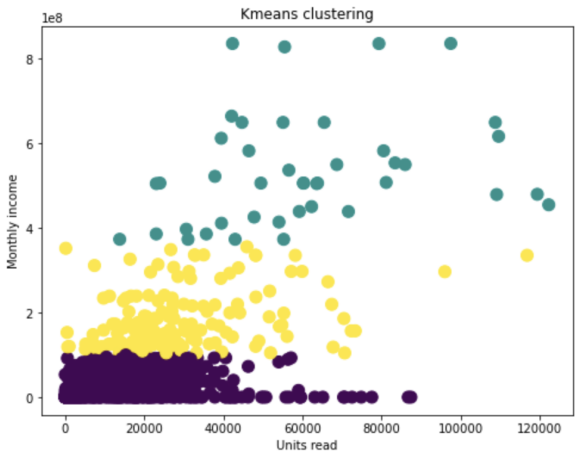

Figure 8: K means clustering of electricity consumption of households in Cape Town

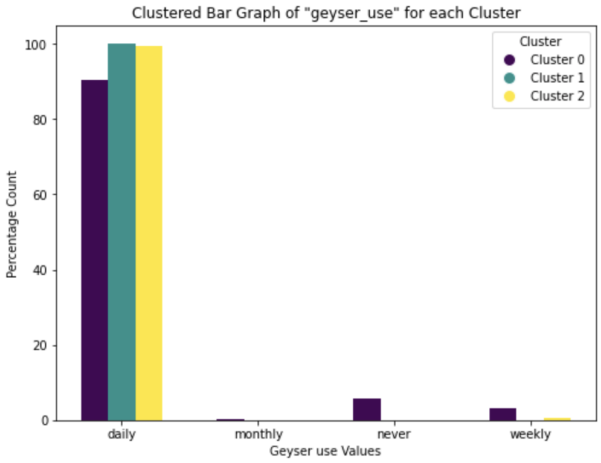

Figure 10: Geyser usage patterns for each cluster.

ments in performance while making the model more complex. The optimal number of clusters k obtained was three clusters (see figure 8). Cluster 1 and 2 have households that have the highest electricity consumption in the city of Cape Town and these households own the highest number of geysers i.e. 2 to 5 geysers (see figure 9). Also, as shown in figure 10, households in clusters 1 and 2 report a slightly higher percentage of daily geyser usage compared to households in cluster 0. This supports the theory being investigated that the ownership of water heating appliances, specifically geysers, influences the domestic electricity consumption patterns.

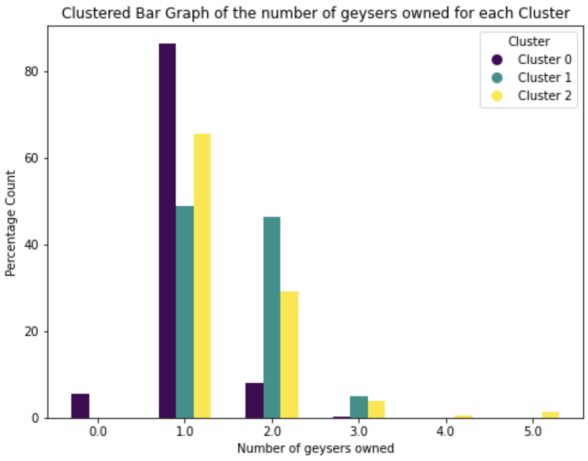

Figure 9: Geyser ownership patterns for each cluster

Discussion and Conclusion

In ATOM, theory construction is triggered by a detected phenomenon and it is done with the goal of explaining the said phenomenon. However, in this paper, theories are generally generated from existing knowledge using analogical modelling with a focus of obtaining new lines of inquiry that can be developed and appraised on respective data sets. Existing domain knowledge is structured and maintained using an ontology. The ontology is a formalisation of a conceptual model that is based on a categorisation of factors that influence domestic electricity consumption by Jones, Fuertes, and Lomas. This study does not disregard the effect of psychological factors on domestic electricity consumption presented in other categorisations. However, focus is placed solely on socio-demographic factors and not the psychological factors. The reason for this is two-fold. Firstly, although psychological factors may influence changes in the consumption over time, consumption in the residential sector strongly relates to socio-demographic factors (Abrahamse and Steg 2009; Frederiks, Stenner, and Hobman 2015). Secondly, most of the studies and data sets have mainly focused on a range of socio-demographic factors. Unsurprisingly, the South African data set at hand does not contain psychological factors. It is on this note that we apply the categorisation presented in (Jones, Fuertes, and Lomas 2015) in designing the conceptual model.

In conclusion, considering the availability of a data set from a specific location, the theory construction workflow conducted with the help of ontologies, Bayesian networks and machine learning as specified in Algorithm 1 can facilitate the generation, development and appraisal of theories. The abstract algorithm presented in this paper has the potential to be applied in other areas. However, more work needs to be done to check the extent to which it is generalisable. Future work entails consolidating the entire theory construction workflow into an AI system that fully automates scientific knowledge discovery.

Acknowledgements

References

Abrahamse, W.; and Steg, L. 2009. How do socio-demographic and psychological factors relate to households' direct and indirect energy use and savings? *Journal of economic psychology*, 30(5): 711–720.

Alsubait, T.; Parsia, B.; and Sattler, U. 2012. Automatic generation of analogy questions for student assessment: an Ontology-based approach. *Research in Learning Technology*, 20.

Blimpo, M. P.; Postepska, A.; and Xu, Y. 2020. Why is household electricity uptake low in Sub-Saharan Africa? *World Development*, 133: 105002.

Casakin, H.; and Goldschmidt, G. 1999. Expertise and the use of visual analogy: implications for design education. *Design studies*, 20(2): 153–175.

De Nicola, A.; Missikoff, M.; and Navigli, R. 2009. A software engineering approach to ontology building. *Information systems*, 34(2): 258–275.

Drake, R.; and Moodley, D. 2021. INVEST: Ontology Driven Bayesian Networks for Investment Decision Making on the JSE. 252–273.

Feldbacher-Escamilla, C. J.; and Gebharter, A. 2020. Confirmation based on analogical inference: Bayes meets jeffrey. *Canadian Journal of Philosophy*, 50(2): 174–194.

Forbus, K. D.; Mostek, T.; and Ferguson, R. 2002. An analogy ontology for integrating analogical processing and first-principles reasoning. *AAAI/IAAI*, 2002: 878–885.

Frederiks, E. R.; Stenner, K.; and Hobman, E. V. 2015. The socio-demographic and psychological predictors of residential energy consumption: A comprehensive review. *Energies*, 8(1): 573–609.

Gentner, D.; and Smith, L. A. 2013. Analogical learning and reasoning.

Haig, B. D. 2018. An abductive theory of scientific method. In *Method matters in psychology*, 35–64. Springer.

Han, J.; Shi, F.; Chen, L.; and Childs, P. R. 2018. A computational tool for creative idea generation based on analogical reasoning and ontology. *AI EDAM*, 32(4): 462–477.

Jones, R. V.; Fuertes, A.; and Lomas, K. J. 2015. The socio-economic, dwelling and appliance related factors affecting electricity consumption in domestic buildings. *Renewable and Sustainable Energy Reviews*, 43: 901–917.

Kott, J.; and Kott, M. 2019. Generic ontology of energy consumption households. *Energies*, 12(19): 3712.

Krenn, M.; Pollice, R.; Guo, S. Y.; Aldeghi, M.; Cervera-Lierta, A.; Friederich, P.; dos Passos Gomes, G.; Häse, F.; Jinich, A.; Nigam, A.; et al. 2022. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12): 761–769.

Langley, P. 2022. Agents of exploration and discovery. *AI Magazine*, 42(4): 72–82.

Moodley, D.; and Seebregts, C. 2023. Re-imagining health and well-being in low resource African settings using an augmented AI system and a 3D digital twin. *arXiv preprint arXiv:2306.01772*.

Mutumbi, U.; Thondhlana, G.; and Ruwanza, S. 2022. The Status of Household Electricity Use Behaviour Research in South Africa between 2000 and 2022. *Energies*, 15(23): 9018.

Ogundele, O. A.; Moodley, D.; Pillay, A. W.; and Seebregts, C. J. 2016. An ontology for factors affecting tuberculosis treatment adherence behavior in sub-Saharan Africa. *Patient preference and adherence*, 669–681.

Ogundele, O. A.; Moodley, D.; Seebregts, C. J.; and Pillay, A. W. 2015. An ontology for tuberculosis treatment adherence behaviour. In *Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists*, 1–10.

Ogundele, O. A.; Moodley, D.; Seebregts, C. J.; and Pillay, A. W. 2017. Building Semantic Causal Models to Predict Treatment Adherence for Tuberculosis Patients in Sub-Saharan Africa. In *Software Engineering in Health Care: 4th International Symposium, FHIES 2014, and 6th International Workshop, SEHC 2014, Washington, DC, USA, July 17-18, 2014, Revised Selected Papers 4*, 81–95. Springer.

Putra, H. C.; Hong, T.; and Andrews, C. 2021. An ontology to represent synthetic building occupant characteristics and behavior. *Automation in construction*, 125: 103621.

Raad, E.; and Evermann, J. 2015. The role of analogy in ontology alignment: A study on LISA. *Cognitive Systems Research*, 33: 1–16.

Reinisch, C.; Kofler, M.; Iglesias, F.; and Kastner, W. 2011. Thinkhome energy efficiency in future smart homes. *EURASIP Journal on Embedded Systems*, 2011: 1–18.

Saba, D.; Sahli, Y.; Abanda, F. H.; Maouedj, R.; and Tidjar, B. 2019. Development of new ontological solution for an energy intelligent management in Adrar city. *Sustainable Computing: Informatics and Systems*, 21: 189–203.

Shah, N.; Chao, K.-M.; Zlamaniec, T.; and Matei, A. 2011. Ontology for home energy management domain. In *Digital Information and Communication Technology and Its Applications: International Conference, DICTAP 2011, Dijon, France, June 21-23, 2011, Proceedings, Part II*, 337–347. Springer.

Toussaint, W. 2020. Domestic Electrical Load Data Descriptor.

Toussaint, W.; and Moodley, D. 2019. Comparison of clustering techniques for residential load profiles in South Africa. In Davel, M. H.; and Barnard, E., eds., *Proceedings of the South African Forum for Artificial Intelligence Research Cape Town, South Africa, 4-6 December, 2019*, volume 2540 of *CEUR Workshop Proceedings*, 117–132. CEUR-WS.org.

Toussaint, W.; and Moodley, D. 2020. Clustering residential electricity consumption data to create archetypes that capture household behaviour in south africa. *South African Computer Journal*, 32(2): 1–34.

Wanyana, T.; and Moodley, D. 2021. An Agent Architecture for Knowledge Discovery and Evolution. In Edelkamp, S.; Möller, R.; and Rueckert, E., eds., *KI 2021: Advances in Artificial Intelligence*, 241–256. Cham: Springer International Publishing. ISBN 978-3-030-87626-5.

Appendices

Appendix 1

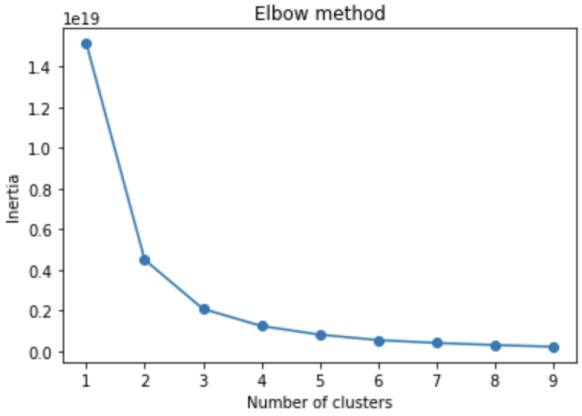

Figure 11: The elbow method to find the optimal number of clusters in the South African domestic electricity consumption data for the city of Cape Town