Hopf *Energy Informatics* (2018) 1:4
https://doi.org/10.1186/s42162-018-0009-3

Energy Informatics

RESEARCH Open Access

Mining volunteered geographic information for predictive energy data analytics

Konstantin Hopf

Correspondence:
konstantin.hopf@uni-bamberg.de
Information Systems and Energy Efficient Systems Group, University of Bamberg, Kapuzinerstraße 16, 96047 Bamberg, DE, Germany

Abstract

Background: Users create serious amounts of Volunteered Geographic Information (VGI) in Online platforms like OpenStreetMap or in real estate portals. Harvesting such data with the help of business analytics and machine learning methods yield promising opportunities for firms to create additional business value through mining their internal and external data sources. Energy retailers can benefit from these achievements in particular, because they need to establish richer customer relations, but their customer insights are currently limited. Extending this knowledge, these established companies can develop customer-specific offerings and promote them effectively.

Methods: This paper gives an overview to VGI data sources and presents first results from a comprehensive review of these crowd-sourced data pools. Besides that, the value of two exemplary VGI data sources (OpenStreetMap and real estate portals) for predictive analytics in energy retail is investigated by using them in a household classification algorithm that recognizes specific household characteristics (e.g., living alone, having large dwellings or electric heating).

Results: The empirical study with data from 3,905 household electricity customers located in Switzerland shows that VGI data can support the recognition of the 13 considered household classes significantly, and that such details can be retrieved based on VGI data alone.

Conclusion: The results demonstrate that the classification of customers in relevant classes is possible based on data that is present to the companies and that VGI data can help to improve the quality of predictive algorithms in the energy sector.

Keywords: Volunteered geographic information, Energy data analytics, Energy retail, Predictive analytics, Machine learning, Household classification

Background

Predictive data analytics becomes increasingly important for organizations and enable them to remain competitive and agile in continuously changing markets (Constantiou and Kallinikos 2015; Gillon et al. 2012; Mithas et al. 2013; Sharma et al. 2014). This holds especially for energy utilities that are exposed to a groundbreaking market transition, affecting them from the production and sales perspective. On the *energy production* side, the current fossil-nuclear energy supply is currently being replaced by renewable energy sources (Dangerman and Schellnhuber 2013), which forces utility companies to high

Springer Open

investments into their fixed assets. On the *energy retail* side, their market becomes more competitive due to market liberalizations in many countries and the fact that private spendings for energy are not much increasing, since the share of wallet for housing, water and energy remains nearly constant during the last 20 years in Europe at 20–25% (Eurostat 2017). Business model innovations in the utility industry are therefore crucial (Markard and Truffer 2006).

Big data analytics in the energy industry

Utility companies hold, however, millions of data points on their customers that can be developed to a valuable business asset. This available data will further increase in the future due to the roll-out of smart meter and Internet of things infrastructures. Techniques of big data analytics, especially predictive information systems will help utility companies to make sense of that data (Sharma et al. 2014; Zhou et al. 2016) and can enable them to create innovative products and services. Even though, first steps in development of algorithms and tools for predictive energy data analytics are done and basic knowledge on how to create insights from big data exists, but the currently existing models are often not reliable enough to be used in practice. Therefore, further research on better predictor variables, more suitable data processing and knowledge modeling techniques is necessary. *This work contributes to this research gap and investigates the use of freely available geographic data for predictive energy data analytics in the specific but relevant example of household classification.*

Predictive analytics in energy retail: the use of household classification

A possible reaction to the harsher market requirements for energy retailer is to improve the customer communication and to offer product or services that better fit to the needs of their customers (Gebauer et al. 2014), establish targeted customer communication or one-to-one marketing. Examples for such products are tariffs for specific customer groups, energy consulting, and energy- or infrastructure-related cross-sales (e.g., heating and cooling appliances, photovoltaic installations, Internet access). For the development and commercialization of such products and services, detailed knowledge about customers is necessary. Typical variables needed in marketing campaigns are household characteristics (e.g., families with children, household type and size) or knowledge on the customers' attitudes and intentions (e.g., towards sustainability, purchase intentions for products). Recent studies have shown, that various household characteristics (such as household type, type of the heating system, number of residents, etc.) can be identified using 30-min electricity smart meter data (Beckel et al. 2014). Based on 15-min smart meter and hourly weather data, household characteristics related to energy-efficiency can be identified (Sodenkamp et al. 2017; Hopf et al. 2018). Other studies show that households with old heating systems can be identified to be addressed in an energy-efficiency campaign (Kozlovskiy et al. 2016), or customers using an electric vehicle (Verma et al. 2015), which is also an interesting insight for energy utilities. In all studies on household classification, several attempts have been made to improve the quality of predictive artifacts. Multiple data preparation and machine learning algorithms have been tested so far, but the models still need further improvements and lowered error rates to become more reliable and ready for real applications in the industry. I argue that *additional predictor variables need to be taken into account to increase the quality of household classification* *models, since the information that is possible to extract from electricity consumption data is limited*. In this paper, I therefore present VGI as a promising source of additional data that is freely available online and has emerged in the past years.

Volunteered geographic information as a data source for predictive analytics

A significant number of web portals arose during the past years, due to the fact that users collaborate and create crowd-sourced data that often has a geographic reference associated. Goodchild (2007) coined the term *VGI* for this phenomenon of user-generated geospatial data. His work attracted a considerable amount of research in the past ten years and lead also to attention in other disciplines than geography. Unique features of this data source are the free availability and the variety of subjects – even for some subjects, data was never-collected before (Sester et al. 2014).

In this work, *all user-generated geospatial data is considered as VGI as long as it was collected with the (explicit or implicit) consent of the user*. This definition does not include data provided or published in a non-voluntary way (e.g., accidentally published or revealed by hackers).

This definition is necessary here, since there is no clear consensus in related works on the concept of VGI. On the one hand, there is the understanding of VGI in a strict sense, restricting it to user contributions that are made "with the intent of providing information about the geographic world" (Elwood et al. 2012). Examples for this interpretation of VGI are the co-creation of maps (e.g., OpenStreetMap, Wikimapia) or the collection of environmental data (e.g., Christmas Bird Count). On the other hand, user contributions without the intrinsic motivation to publish geographic data exist, since users create for tweets, posts, likes, reviews, photos or other content that has a spatial datum. Stefanidis et al. (2013) name such information Ambient Geospatial Information (AGI), but they are seen as VGI in a broader sense in this paper. Harvey (2013) suggests to delimit the term VGI to data that was collected with a clear *opt-in* of users. He suggests the term Contributed Geographic Information (CGI) for all geographic data that was collected without that clear consent, but having an *opt-out* possibility for users as. Harvey names "cell phone tracking and RFID-equipped transport cards" as examples, but the use of such data, outside the objective it was collected for, stays obviously often in conflict with privacy or data protection regulations. Therefore, the above given definition of VGI in contrast to non-voluntary generated data is preferred here.

A vast number of VGI initiatives are already documented in the literature: (Elwood et al. 2012) mention 99 initiatives that they found in 2009 and categorize them according to geographic extend, the date initiated, the sponsoring entity and the primary purpose of the initiative. Ballatore et al. (2013), Sester et al. (2014), such as Rinner and Fast (2015) document further platforms and suggest different categorization schemes for initiatives, by focusing on the usage of VGI data, the data types and formats used. Due the dependence on continuous contributions from users, a large turnover of newly emerging, changing and perishing projects is visible. Based on the existing collections (Elwood et al. 2012; See et al. 2016), I collected 116 VGI initiatives that have been documented in the literature and analyzed them together with students. The list was further extended with a number of real estate advertisement portals where users can search for hiring, renting, selling or buying houses and dwellings. In total, we identified a sample of 127 initiatives. This collection is not intended to be an extensive list of VGI initiatives, but to give some idea of the types and categories of such data sources. From the 116 VGI initiatives that stem from the literature, only 64 (53.7%) still have an active web site or could be found via the Internet[1]. However, some initiatives have become very popular and attracted many active users (such as OSM, or Geocaching).

By looking at the distribution of active initiatives according to data types (Rinner and Fast 2015) and categories (Sester et al. 2014) in Fig. 1, one can see that the majority of projects deal with location data, mapping, or environmental and social media topics. The category "business" is also frequent, but this figure is misleading, since ten German-speaking real estate portals were assigned to this category that might over-represent this category in comparison to others. Generally, VGI initiatives contain a considerable amount of data that is related to an energy context and is therefore relevant for the utility industry, given that knowledge from the freely available data sources can be derived. However, the mining of such data sources is associated with large effort to access the data. For example, only 26 of 75 active platforms have an explicit API for data access.

So far, the use of VGI data in the energy domain is sparse, but could bring value to energy retailers and service providers in the utility industry. This paper aims therefore to motivate the use of VGI, illustrates the usage of such data and the possible contribution for predictive data analytics in the field of energy retail considering two prominent examples of VGI initiatives.

Research Objective and Methodology

Research presented in this article aims to provide an answer the *research question: To what extend can volunteered geographic information improve the quality of predictive models in energy data analytics, in the specific case of household classification?*

The approach to answer this research question implies two steps of data analysis: In a first step, the VGI data is be made accessible for the further analytical procedure. This implies to use techniques from cartography, data integration, data fusion and data generalization (Sester et al. 2014), because the user-generated data is very heterogeneous in terms of existing data structures, producers and its quality. In a second step, the usefulness of VGI data is assessed by using the data for household classification and calculating

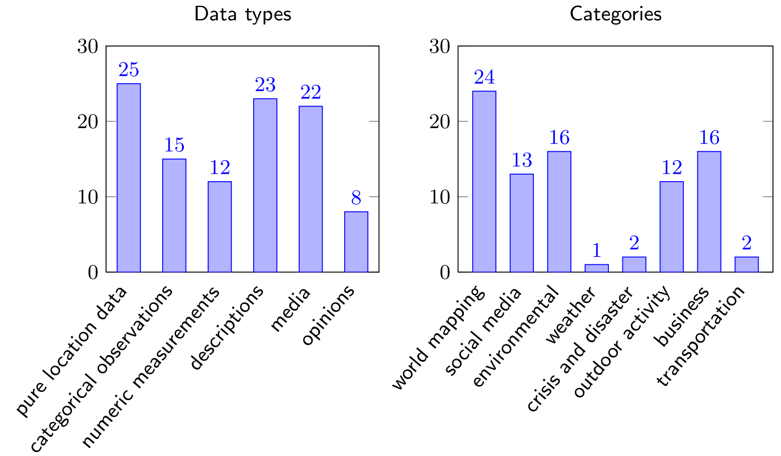

Fig. 1 Frequencies of $n = 75$ active VGI initiatives 2017, according to data types (Rinner and Fast 2015) and categories (Sester et al. 2014); some portals have multiple data types and categories associated

performance gains. This goes along with the suggestion of Mondzech and Sester (2011) to assess the quality of VGI by means of application needs.

Starting with the mining of OSM as a first data source for energy data analytics, that contains geographic information and Points of Interest (POIs) data, online portals for real estate advertisements (containing information on energy-efficiency, rental prices, living quality, etc.) are evaluated.

Related Work

To the best of my knowledge, the first application of VGI in energy data analytics was mentioned in our recent paper (Hopf et al. 2016). Likewise, Zhou et al. (2016) mention several data sources in their literature review on big energy data analytics, but do not mention VGI data. However, the use of VGI data was documented for many applications and it was shown that this data source is valuable. Examples are the eHealth-domain (Eysenbach 2008; Mooney et al. 2013), or disaster management (Haworth and Bruce 2015) where volunteers brought quick help into crisis areas. A prominent example is the Haitian Earthquake (Zook et al. 2010), where a large number of volunteers worldwide provided remarkable aid and created very detailed maps for emergency forces within days. The traditional creation of such maps by professional map-makers would have taken much more time (Anhorn et al. 2016; Horita et al. 2013).

In the field of energy data analytics, the identification of household characteristics based on customer data from utility companies was investigated in several studies. The literature can be grouped into two categories based on the applied methods: *unsupervised machine learning* (also known as clustering or segmentation) using energy consumption from utility customers and other data sources (Beckel et al. 2012; Chicco 2012; Kwac et al. 2013; McLoughlin 2013). Because of their nature, these analyses give descriptive insights on groups of customers, but do only provide cluster-membership information on the level of single households. Thus, further interpretations of experts are necessary to use the results in practice.

Supervised machine learning methods are used to identify household properties based on energy consumption data (this field is also known as ’household classification’). The approach was first mentioned by Beckel et al. (2013, 2014) who show that prediction of individual household characteristics is possible using 30-min electricity smart meter data. This approach was improved (Hopf et al. 2016) and adapted it for 15-min smart meter data to identify energy-efficiency related household properties (Sodenkamp et al. 2017; Hopf et al. 2018). Similar approaches using smart meter data have been used to detect heat pumps (Fei et al. 2013), predicting occupancy in residential homes (Albert and Rajagopal 2013), the enrollment in energy-efficiency campaigns (Zeifman 2014), or the existence of electric vehicles in households (Verma et al. 2015). However, these studies investigate only single characteristics of households, but not multiple as in the household classification studies.

The usage of smart meter data for such analysis in often limited due to technical (e.g., low number of smart meters installed), organizational or legal reasons (e.g., sales departments may not have access to high-resolution data). It is therefore reasonable to investigate to what extend household classification is feasible using annual electricity consumption data together with the household location, since both is available in energy retail for billing purposes. Recently published results of household classification studies with annual electricity consumption and location data show promising results (Hopf et al. 2016; Hopf et al. 2017), but the currently achieved prediction accuracy needs further improvements for practical applications. The investigation of further freely available data sources, such as VGI is therefore reasonable.

Household classification based on annual electricity consumption and location information

The household classification methodology used in this work is illustrated in Fig. 2. It consists of two dimensionality reduction and data preparation steps (empirical feature extraction and automatic feature selection), followed by the training and test of suitable supervised machine learning algorithms. For the supervised machine learning part, multiple algorithms (e.g., k-Nearest-Neighbor, Support Vector Machines, Naïve Bayes, Random Forest, Artificial Neural Networks) are candidates for the artifact development and the most suitable algorithm is then chosen. Typically, Breiman's (2001) Random Forest algorithm yield good results for the investigated classification problems (Hopf et al. 2016; Hopf et al. 2017), but also in other studies (Fernández-Delgado et al. 2014). Finally, the classification results are evaluated using well-recognized performance metrics (Hastie et al. 2009, chap. 7) such as Accuracy, Precision, Recall or Area Under ROC Curve (AUC). In the remaining section, a detailed description of the empirically defined electricity and geographic features is given as an example of the necessary data transformation in energy (big) data analytics and the use of VGI data.

Empirical and automatic dimensionality reduction

Model building is one of the core techniques used in big data analytics (Abbasi et al. 2016), since knowledge, insights and business value does not automatically come from simply applying analytic tools to data. It must be leveraged by analysts and managers into strategic and operational decisions to generate value (Müller et al. 2016; Sharma et al. 2014). One of the major challenges in sense-making from big data is to identify the relevant features and encounter the "curse of dimensionality" (Keogh and Mueen 2011) lowering the quality of predictive systems. This problem is serious: Whereas in the last decade,

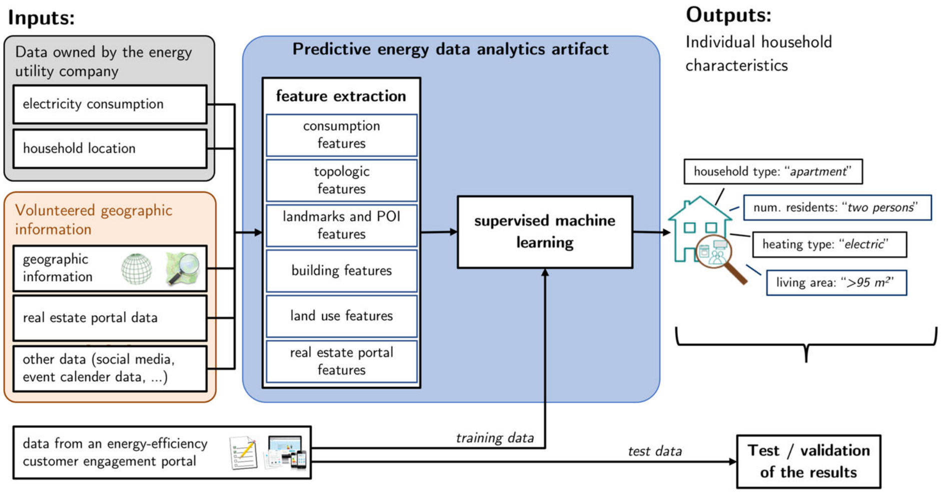
Inputs:
Data owned by the energy utility company
electricity consumption
household location
Volunteered geographic information
geographic information
real estate portal data
other data (social media, event calender data, ...)
Predictive energy data analytics artifact
feature extraction
consumption features
topologic features
landmarks and POI features
building features
land use features
real estate portal features
supervised machine learning
Outputs:
Individual household characteristics
household type: "*apartment*"
num. residents: "*two persons*"
heating type: "*electric*"
living area: "*>95 m²*"
data from an energy-efficiency customer engagement portal
training data
test data
Test / validation of the results

Fig. 2 Household classification and evaluation methodology

a number of 50-100 features was called a “large” feature set (Kudo and Sklansky 2000), today we are confronted with hundreds or even thousands (Hua et al. 2009) of features.

Much research has been conducted on automatic dimensionality reduction – also known as *feature selection*. The interested reader is referred to the general introduction of Guyon and Elisseeff (2003) or the comprehensive literature reviews made by Chandrashekar and Sahin (2014), Liu and Motoda (2008) or Saeys et al. (2007). However, the capability of such automatic feature selection algorithms is limited, at least for the investigated class of machine learning problems in energy retail. *Empirical definition of relevant features* is therefore a cornerstone in the development of such tools, since it enables the modeling of human knowledge into the data. By defining features empirically, also expert knowledge or commonsense can be coded into the data.

Electricity consumption features

In this study, annual electricity consumption is considered that is available to all utility companies for billing purpose. In previous works (Hopf et al. 2016; Hopf et al. 2017), three features for such data have been suggested and used that are also applied in this study:

1. Logarithmic annual consumption, normalized by the number of days in which the energy was consumed
2. Consumption trend as the relative change between the consumption of different years (obtained with a linear regression model)
3. Neighborhood comparison as the Z-score of the household’s logarithmized normalized consumption deviation from its neighborhood (in the postal code region)

Geographic features from OpenStreetMap data

The empirical definition of features for VGI map data is done using the currently largest VGI initiative OSM (Jokar Arsanjani et al. 2015) as the study subject. The data in this project is – typically for geographic map data – characterized by multiple spatial dimensions and contextual data that cannot be easily used in data analytics methods. In detail, the data in OSM consists of points, poly-lines and spatial relations that are annotated with tags (containing plain-text data) to give semantic meaning to the objects. Therefore, feature extraction a necessary step making the data accessible in energy data analytics.

The feature extraction is done using problem specific knowledge applied to the data to derive meaningful variables (Guyon and Elisseeff 2006): Some features are directly derived from the *raw data* (e.g., frequencies of objects in an area). Besides that, a literature survey in geographic information science was conducted and *spatial landscape metrics* (e.g., describing the spatial envelope or the structure of a landscape) were identified as domain-specific features. The identified features were grouped into *topologic features* describing the structure of and relations between one household and spatial neighbors (e.g. longitude, latitude, frequency objects in the surroundings, distance to city center), and *semantic features* considering the meaning of an object within the spatial context it appears. In the group of semantic features, I differentiate between *point-based* features and *polygon-based* features (i.e. *building* and *land use*). In the remaining section, the categories of features are described.

Topologic features — In geodesy, topology focuses on the *structure of and relations between spatial objects*, e.g., whether two objects overlap or whether one object is contained in another (Becker 2012). In contrast to the semantic relation of a geographical object to another, the topology covers only shape and arrangement of objects. Example topologic features are *geographic location* (longitude, latitude, altitude) that representing the local environment, such as climate conditions or the number of sun hours per day. The *number of objects* in the surrounding of a household is a proxy for the population density and *distance to city center* is a distinction criterion for urban or rural area

Semantic features — A charted point or polygon in a map can have different meaning, depending on the size of the object, in which region it appears and which objects are nearby. The semantics focuses on the *meaning of an object within the context it appears*. The available source for semantic geographical data are tag-annotations associated to objects in OSM.

It is obviously difficult to express the proximity of two or more objects by means of a semantic distance. Suggested similarity measures, as used for example in information retrieval (Schwering 2008; Janowicz et al. 2011; Ballatore et al. 2012), can not be easily adapted in the context of this study, since they mainly express the similarity of two objects, but not the similarity of areas on a map as needed for the characterization of households. Therefore, I defined semantic features from the OSM internal taxonomy for tags of geographic objects that is known as *OSM map features* (the double meaning of this term cannot be avoided here) that is maintained by the OSM community.

The community defines 26 *primary map features* that are recommended to be used (others may be used, but are mostly infrequent), but this tag-taxonomy is not well suited for the use in energy data analytics:

- map features have no direct meaning for the field of application (such as *aeroway*, *barrier* or *emergency*)
- the wide range of some map features must be diversified (such as *amenity*, *natural* or *man_made*)
- map features cover an overlapping field of objects (for example *office*, *craft*, *shop*, *amenity* and *building* can describe the existence of company locations)
- some of the tags are not included in the set of primary map features, but are interesting for the field of application, because they are frequently used and cover important aspects of the surrounding of a house

Geographic Object Categories (GOCs) for semantic feature definition — To overcome with these ambiguities and the incompleteness of the map feature definition, a set of GOC was defined that describes the relevant geographic information in the context of energy data analytics. All tags that belong to the primary map features (including the keys and the values) are associated to one or more GOCs. All GOC are then used to define semantic features for classification. These definition of GOCs and the derivation of features for later classification was done in three steps:

1. *Top-down:* all primary map features and their tags[2] have been investigated and meaningful groups of objects have been derived.
2. *Bottom-up:* actual tag frequencies in the surrounding of ca. 4,000 households were downloaded from OSM, analyzed and used for group definition.
3. *Feature derivation:* For each GOC, features for the classification are defined.

The result of this categorization is shown Table 1. Nine point-based GOCs (marked with ★) and two polygon-based categories are defined. The polygon-based categories are buildings (marked with ⧫) and land use (marked with ♠). Map features that belong to object categories are marked with X. Consecutively, the calculation of features based on the categorization is described.

Point-based features on landmarks and POIs (★) — Most of GOCs describe point-information on a map regardless of whether the objects are represented as points or polygons in the OSM database, since there is no standard on how to map POIs. On the one hand, some building-polygons are directly tagged as restaurants, public institution, etc. On the other hand, POIs are mapped as separate points. The characterization of one household, however, may rely on the frequency of POIs or the distance to it. The size or shape of a restaurant or the next park is considered as unimportant in this context. Therefore, nine GOCs are defined as point-based categories and the following features are defined for each category (each calculated using a predefined map section, usually a rectangular box of $500 \times 500\,m$)

1. *Frequency*: the total number of objects in this GOC
2. *Average distance*: the arithmetic mean of all distances from the household location to the objects
3. *Minimal distance*: the distance to the nearest object in the GOC to the household location

Features on buildings (⧫) — The category of buildings constitute a separate GOC because they are of main concern in energy data analytics. This category is solely affected by map feature *building*. Other map features (e.g., *levels, building:levels, building:roof, roof:shape*) may contain additional information, but do not identify polygons as buildings. The following semantic features are defined for all buildings:

1. Size of the buildings (mean and variance)
2. Distance to buildings (mean and variance)
3. Type of the next building
4. Type of building that contains the household location on the map
5. Most frequent building type in the surrounding

Features about land use (♠) — Finally, land use information constitutes a separate category. Within this category, three land use types are distinguished: *residential, city*, and *countryside*. The following features are defined for land use information:

1. The land use type the household location lies in
2. The area of the land use polygon the household location lies in
3. The total area of all three land use types

Feature extraction from real estate portals

As a second source of VGI data, publicly available online real estate advertisements are considered. This data is user-generated and has – in most cases – a geographic location associated. In the context of household classification, the use of this data source seems natural, given that household and building characteristics are often similar in neighborhoods.

Table 1 Overview to the definition of GOCs based on primary and secondary map features in the OSM database

OSM map feature	# objects	Building ♦	Land use ♠	Public Institution ★	Business ★	Food ★	Transportation ★	Recreation ★	Culture ★	Sights ★	Countryside ★	Road system ★
Primary Map Features												
Aerialway	< 5						X					
Aeroway	< 5						X					
Amenity	7 942			X	X	X	X	X	X	X	X	
Barrier	1 387											
Boundary	1 454											
Building	141 420	X		X	X		X				X	
Craft	< 5				X							
Emergency	1 506			X								
Geological	< 5								X			
Highway	65 194											X
Historic	< 5								X			
Landuse	10 000		X									
Leisure	1 813		X					X				
Man_made	209								X			
Military	< 5											
Natural	4 036		X					X		X	X	
Office	< 5			X	X							
Places	647											
Power	18 277											
Public_transport	3 199						X					
Railway	5 393						X					
Route	10 904						X					X
Shop	< 5				X	X						
Sport	657							X				
Tourism	< 5			X				X	X	X		
Waterway	1 874							X				

Table 1 Overview to the definition of GOCs based on primary and secondary map features in the OSM database *(Continued)*

OSM map feature	# objects	Building ♦	Land use ♠	Public Institution ★	Business ★	Food ★	Transportation ★	Recreation ★	Culture ★	Sights ★	Countryside ★	Road system ★
Secondary Map Features												
Bus	1 170						X					
Bicycle	3 128							X				
Cuisine	268					X						
Cycleway	591							X				X
Denomination	210			X					X			
Horse	410							X				
Parking	805											X
Recycling	584											
Religion	502			X					X			

The major issue regarding the use of real estate portal data is the sparsity of data (i.e., high number of missing values). There are two reasons for missing values in features from that data source. First, the number of real estate advertisements in the surrounding of a target address is strongly dependent on the location (there might be, for example, more advertisments in city centers than in suburbs) and on the point in time when the data is retrieved (e.g., more offers at the end of school years and less offers at the start of courses in university cities). Second, the amount of information provided in advertisments varies to a large extent, because users only insert information on household properties that exist (e.g., elevator in the house, pool in the garden) or do not want to provide some information.

For a sample of 4,521 household locations in Switzerland, 8,341 advertisements from a Swiss real estate portal with 133 different attributes have been downloaded with a webcrawler between March 23 and May 10, 2017. Thereby, all advertisments within a radius of 1,000 m around each household address were considered. Figure 3 illustrates the availability of advertisements for the considered locations. For the majority of advertisements, only basic attributes are present (e.g., living area, rent per month, rent per month, textual description). Table 2 shows the most frequent attributes that are present in more than 50% of the advertisments. Infrequent attributes, only avaliable in less than 1,000 advertisments, were discarded to not distort the analysis and being able to obtain generalizable results in this exploratve study. Future studies may try to impute values based on further data (Saar-Tsechansky and Provost 2007), or infer the information from textual descriptions using text mining.

The feature values from this data source are then calculated by aggregating the data from all advertisments within a radius of 1,000 m around each household. In the case of numeric values, the arithmetic mean, in the case of logical values (e.g., existence of cellar or parking space), the relative frequency, and in the case of categorical variables, the relative frequency for each category is calculated.

Data Analysis and Results

In this section, results from the application of two VGI portals (OSM and data from real estate advertisements) in household classification are presented. A dataset on utility customers was available for this study. The experimental data is described below. Thereafter, details on the implemented machine learning algorithm are given and results from the evaluation are presented. Finally, a brief discussion of the results is given.

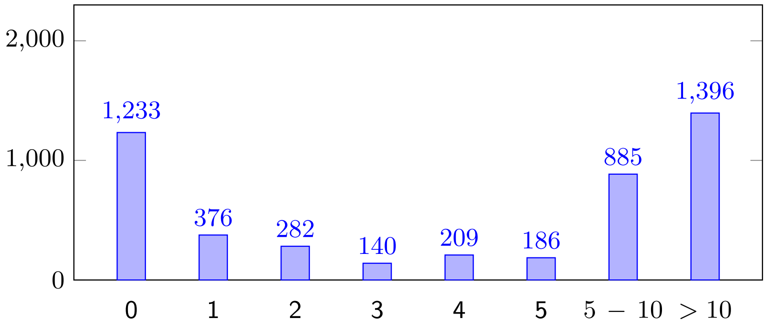

Fig. 3 Frequencies of real estate advertisements within a radius of 1000 m around $n = 4,521$ households in the available utility dataset

Table 2 Data in a Swiss real estate advertisement portal: list of variables with less than 50% missing values

Attribute	Data type	Missing values (in %)
Description	Text	1.39
Main dwelling category	Categorical (e.g., house, dwelling, business)	2.01
Type of dwelling	Categorical (e.g., studio, single-family home, shared flat)	2.01
Geographic location	Coordinates	22.35
Amount rooms	Number	24.13
Living area	Number	30.64
Floor	Categorical (e.g., 1st, 2nd, attic)	32.77
Rent per month	Number	36.14
Area living flat	Number	42.86
Additional costs per month	Number	49.23
Net rent per month	Number	49.27

Experimental data

To evaluate the proposed features from VGI data, a machine learning procedure is used. As data for training and test, information on residential customers from a Swiss utility company is available. The dataset ($\mathcal{A}$) encompasses annual electricity consumption between 2009 - 2012 and address data on 3,905 customers. For all customers in $\mathcal{A}$, information on household properties is known (see Fig. 3). The data was obtained with online-surveys on a customer engagement web portal to increase households energy-efficiency, maintained by our research partner BEN Energy AG, Zurich. All users gave their consent for data processing and analytics in an anonymized way.

Using the address information, geographic features from OSM were calculated using the offered Application Programming Interface (API) following the methodology described in Geographic features from OpenStreetMap data section. Features from real estate advertisments were computed as described in Feature extraction from real estate portals section. For the analysis in this paper, 34 features from the real estate portal data were present in more than 1000 advertisments and therefore used (the features include for example: category and type of the dwelling, rent per month, net rent per square meter, number of rooms, living area, year of construction, distances to next bus station / highway / kindergarden / school / shopping, existence of balcony, cable TV). Missing values have been encoded with -1 being able to use the data in the further analysis.

Only for half of the customers ($n = 1{,}718$), a sufficient number of five real estate advertisements was available (see Fig 3). This sample of customer data with a sufficient number of real estate advertisements $\mathcal{B} \subset \mathcal{A}$ is used for the remaining analysis.

I assume that the sample of customers in $\mathcal{B}$ is representative for utility customers and find evidence for that in the statistics presented in Table 3. The distribution of household properties are quite equal in both datasets, except the size and type of the residency: sample $\mathcal{B}$ contains less larger dwellings and more apartments. By looking at the electricity consumption, there is a difference in the means of both datasets: $M = 12.24$ kWh ($SD = 6.85$) in $\mathcal{B}$ and $M = 13.03$ kWh ($SD = 7.02$) in $\mathcal{A}$. The difference mainly results from the fact that smaller homes and less houses are present in $\mathcal{B}$. The electricity consumption of large homes with living ares above 145 m^2 show for example no significant difference ($t(282) = -1.347, p > 0.10$). Therefore, one can assume that the sample is representative.

Table 3 Household properties used to evaluate the household classification methodology; the relative class sizes in the complete sample and the finally used part, such as the mean and standard deviation of Accuracy and F1-scores for the Random Forest model based on all data sources are given

Property	Class labels	Class size	Class size	Accuracy (%)		F1-Score (%)	
		$\mathcal{A}$	$\mathcal{B}$	Mean	(SD)	Mean	(SD)
Household type	apartment	52.03%	69.23%	77.94	(5.30)	83.28	(3.01)
	house	47.97%	30.77%			66.36	(13.18)
Living area	≤ 95	34.07%	34.95%	59.78	(2.50)	65.00	(4.28)
	$96 - 145$	33.23%	40.61%			58.17	(4.10)
	> 145	32.70%	24.43%			56.40	(5.24)
Number of	1 person	13.07%	17.31%	58.08	(5.07)	50.39	(11.97)
Residents	2 persons	40.32%	43.01%			59.76	(5.03)
	$3 - 5$ persons	43.92%	38.12%			59.07	(6.09)
	> 5 persons	2.69%	1.47%			0	(0.00)
Heating type	electric	12.86%	9.73%	90.05	(0.52)	68.08	(3.59)
	other	87.14%	90.27%			71.68	(4.30)
Water heating	electric	50.49%	44.97%	70.03	(3.87)	68.08	(3.59)
type	other	49.51%	55.03%			71.68	(4.30)

Classification results

The household classification methodology, depicted in Fig. 2, is applied using the Random Forest classifier with its implementation in the package 'randomForest' (Liaw and Wiener 2015) for the statistical programming environment GNU R[3]. No automatic feature selection is done, since the Random Forest algorithm has a good internal ability to weight the importance of certain features. The aim of this study is to investigate, how well household characteristics can be recognized from different data sources. Following this explorative scope, explicitly neither a tuning of the Random Forest algorithm parameters nor a model selection (including test of other classifiers) is performed. Initially, a number of 500 trees is used (this is the standard value of this main parameter to control the algorithm in the used implementation) and a sensitivity analysis on how different values of this parameter affect the results is conducted.

The classification performance is evaluated using AUC as a well-known metric assessing the performance of classification models (Han et al. 2012). The metric quantifies the classification performance from 0 (lowest) to 1 (best), where 0.5 is considered as random classification. So, all values that are clearly higher than 0.5 or lower values are acceptable. For comparison with other studies and for better interpretation of the results, Accuracy and the harmonic mean of Precision and Recall, called F1-score (Hastie et al. 2009, chap. 7), is included in Table 3. To obtain a solid estimation of the classification performance, 10-fold cross-validation is applied. In this statistical technique, the data is randomly split into $k = 10$ independent parts (so-called *folds*). Training and test is performed k times while in each iteration, one fold is used for test and the remaining folds for training. It is generally acknowledged that this technique gives a proper estimate of the classification performance, even with small datasets. In fact, the procedure even leads to a conservative estimate of the performance, given the fact that cross-validation slightly overestimates the variance of prediction results (Arlot and Celisse 2010). The classification performance results are presented in Fig. 4 together with 95%-confidence interval obtained using the Student's t-Distribution ($df = 9$).

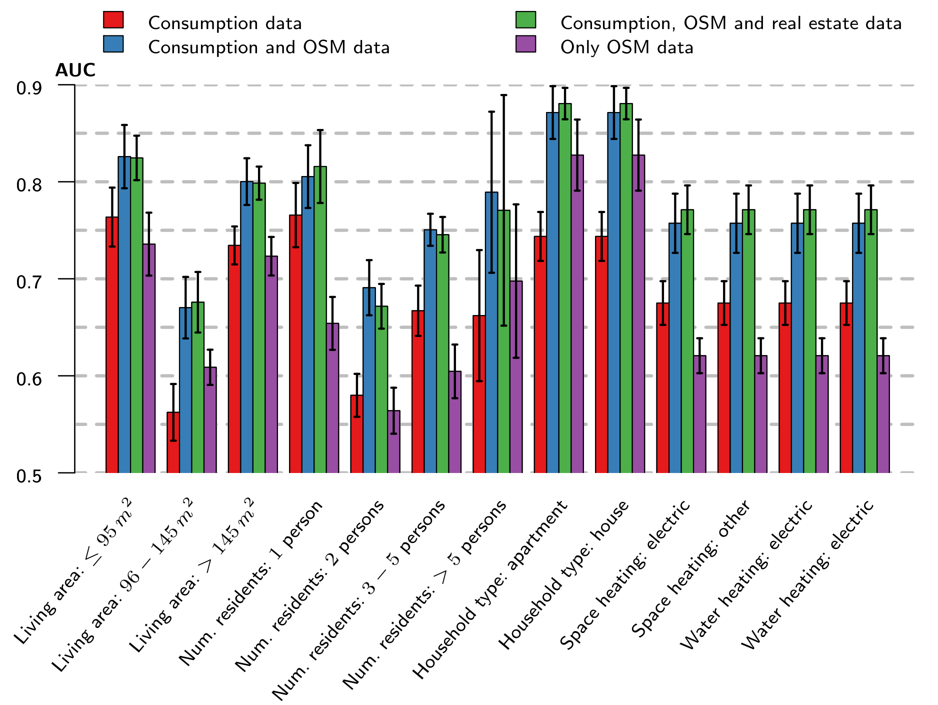

Fig. 4 Classification results (AUC) for different data sources with 95% confidence interval obtained using 10-fold cross-validation

One can see that all predictions made by the presented classification model are significantly better than random ($AUC > 0.5$). Besides that, OSM data improves the classification performance significantly. This result replicates the findings from previous studies (Hopf et al. 2016; Hopf et al. 2017). The additional use of real estate portal data shows small improvements for the properties space and water heating, living area and the class >5 persons. Especially the small class can be better recognized using the real estate portals.

On average, the use of both geographic data sources improved the classification performance in AUC by 12.8% (the improvements range between 8% for single households and 22% for dwellings $> 145\,m^2$), but especially the household type could be improved by 16.6% in both classes.

A very interesting result is that classification only based on OSM data can also lead to good results. The household type, for example, can be predicted better than using electricity consumption features. However, the prediction of household classes solely with OSM features leads to 14.19% lower AUC on average for all properties.

The results presented in Fig. 4 are obtained using 500 trees in the Random Forest algorithm. This is a typical parameter value and the standard in the used implementation. In an empirical study involving 29 datasets, Oshiro et al. (2012) document that larger numbers of trees do not significantly improve the classification performance, otherwise lower numbers of trees decreease the classification quality. To assess the stabilit of the models in this study and get an impression on possible improvements to the presented approach, a sensitivity analysis for this important parameter is conducted using $n = \{10, 100, 250, 500, 1000\}$ trees for the classification using all data available in this study. The results are shown in Fig. 5 as point estimates with a 95% confidence interval obtained using the t-Distribution from 10-fold cross-validation. For all binary properties (household and heating types) only one class is shown, because AUC results for the other classes are equal in each property. It is clear, that a low number of 10 trees leads to considerably lower classification performance. However, the model cannot be significantly improved by changing the number of trees between the values $n = \{100, 250, 500, 1000\}$. Thus, the standard value of 500 trees is quite a good choice in this presented case.

Discussion

The results from this study show that it is possible to considerably improve household classification models based on annual electricity consumption data by using freely available VGI data. Thus, *the stated research question can be answered positively*.

The results are in line with previous studies on factors that influence residential electric load (Kavousian et al. 2013; Heiple and Sailor 2008), which identify geographic data and household details as good explanatory variables for the energy consumption in households. So it is reasonable that geographic information – together with annual electricity consumption or alone – can serve as predictor for relevant details about residential energy customers, as shown in this paper. The real estate portal data has, despite the large amount of building related information contained, not helped much to increase the predictive quality.

Besides that, the presented results confirm also findings from a previous study with a similar household classification approach that was trained and validated with data from Switzerland and Germany, where also the transferrability of trained models (i.e., trainig of a model in one region and then applying this model to customers located in another country) was positively tested (Hopf et al. 2017).

Conclusion and Implications

The phenomenon of VGI, first documented by Goodchild (2007), has created a considerable amount of data in the recent years that is freely available to everybody. Although this data can lead to valuable insights, it's application in energy data analytics is sparse.

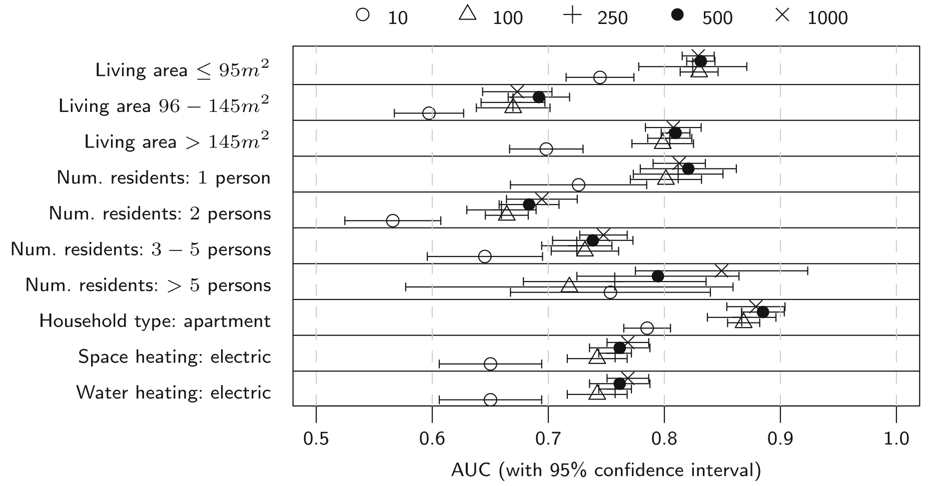

Fig. 5 Classification results (AUC) for different number of trees in the Random Forest algorihm

In this paper, an overview to VGI data from an energy industry viewpoint is presented. Besides that, it was shown, how the heterogeneous data can be transformed and finally be used in predictive models for the specific case of household classification. Such methods are able to predict characteristics for individual private households, such as household type and size, heating type and number of residents. This information can be used for value-added services, for improving the customer communication, tailoring energy-feedback for a more energy-aware behavior, or perform targeted one-to-one marketing campaigns (Tiefenbeck 2017).

The results presented in this paper highlight that VGI data sources are of value for predictive data analytics in the field of energy retail. So, VGI data is able to improve the classification quality in the presented case significantly by 12.8% on average over all considered household classes. All investigated household properties can also be predicted by just using geographic data without any further electricity consumption information, albeit with a loss in AUC of 14.19% compared to the model using all features in this study. Thus, the application of available crowd-sourced VGI data together with state of the art machine learning algorithms can serve as a blueprint for further application in business intelligence and analytics.

Even with annual electricity consumption data and location information (both is available to energy utilities for billing purposes), energy retailers and energy service providers can start right now to predict detailed information on their customers using machine learning and existing VGI data. So, they can enhance their customer communication with personalized data at low cost.

Limitations

Findings in this paper are drawn from a dataset on utility customers from Switzerland. Thus, the concrete classification performance results may vary in other studies, but the performance gain from VGI should be replicable. In a recent study it could be, for instance, shown that household classification models can be trained with data from one country and applied to another one (Hopf et al. 2017). Due to the sparse number of real estate advertisements, not the complete number of households could be used in the study. The collection of further real estate advertisements, and also data from additional VGI data sources, would give a more comprehensive picture of the topic.

A possible limitation to the work is the aspect of privacy preservation and user acceptance of predictive information systems. Krishnamurti et al. (2012) and Mah et al. (2012) have studied the acceptance of end users for smart grid applications empirically and came to the conclusion that private customers have a positive attitude towards smart grid applications (consumption feedback, dynamic prices, etc.). However, it is necessary that utility companies educate their customers on the added value of smart grid technology and the transparency of data processing (Gangale et al. 2013) in order to place smart grid services well in the market. An important driver of customer acceptance are also financial incentives such as dynamic pricing of energy that are made possible by smart meters (Motsch 2012).

Future work

The phenomenon of VGI should be subject to future research. For instance, the persistence and reliability of VGI initiatives can be investigated considering knwoledge on the motivational factors and contributor behaviors in free/libre open source software projects (Stewart and Ammeter 2002; Crowston et al. 2007; Crowston et al. 2008), as a starting point. Besides that, other VGI data sources may be tested for their contribution in energy data analytics applications, especially data sources that contain event and environmental data (adding a time-dimension), such as social networks.

Considering the data quality of VGI data, the empirical definition of new features can help to overcome issues of data sparsity and variety. This is a critical issue and became visible in this paper with the large number of missing values in real estate advertisments. Possible methods to overcome the problem without discarding instances or feature vectors include the imputation of values (Saar-Tsechansky and Provost 2007), the inference of values from textual descriptions using text mining, or by treating all missing values in a vector as a separate class in model trainig and prediction. The application and comparison of such approaches is motivated to be a subject of future research.

Regarding the household classification artifact, the prediction models can be further enhanced by testing other parameter values for the Random Forest algorithm, benchmark additional algorithms with different parameter combinations, and automatic feature selection.

Finally, it would be interesting to investigate whether VGI data sources have also value for other topics in the field of Energy Informatics, such as load forecasting or studies regarding consumer behavior in energy markets.

Endnotes

[1] The information has been collected between November 2016 and April 2017.

[2] listed at http://wiki.openstreetmap.org/wiki/Map_Features, last accessed March 01, 2018

[3] R version: 3.4.2; 'randomForest' package version 4.6-12

[4] https://cran.r-project.org/, last accessed 13.01.2018

Abbreviations

AGI: Ambient geospatial information; API: Application programming interface; AUC: Area under ROC curve; CGI: Contributed geographic information; GOC: Geographic object category; GOC: Geographic object categories; OSM: OpenStreetMap; POI: Point of interest; POI: Points of interest; ROC: Receiver operating characteristic; RQ:Research question; SMD: Smart meter data; SVM: Support vector machine; VGI: Volunteered geographic information; WGS84: World geodetic system 84; XML: Extensible markup language

Acknowledgments

I thank Prof. Dr. Thorsten Staake and Prof. Dr. Hans-Arno Jacobsen, such as the faculty of 8th PhD Workshop "Energy Informatics" in Lugano 2017 for insightful comments and feedback to earlier versions of this manuscript. My thank also goes to the anonymous reviewers of this journal, for their valuable comments and suggestions. Besides that, I thank BEN Energy AG (Switzerland) for providing anonymized utility data for the analysis, Rolf Fakesch and Martin Bullin for their support in acquisition of real estate portal data, such as Philipp Stiber-Amend and André Fiedler for their help in collecting data on VGI initiatives.

Funding

Financial supported from Eureka member countries and European Union (EUROSTARS Grant number E!9859 - BENgine II) is greatfully acknowledged.

Availability of data and materials

Due to its nature, VGI data is available to the public and can be retrieved from the respective portals. All computational methods used are open source and available via the Comprehensive R Archive Network[4]. Other materials are referenced in this paper. The utility data used in this study cannot be published, because it contains confidential information (address data and electricity consumption).

Author's contribution

The author conducted the analysis and has written the manuscript. The author read and approved the final manuscript

Competing interests

The author declares that he has no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 15 January 2018 Accepted: 20 March 2018
Published online: 26 July 2018

References

Abbasi A, Sarker S, Chiang R (2016) Big Data Research in Information Systems: Toward an Inclusive Research Agenda. J Assoc Inf Syst 17(2):00026

Albert A, Rajagopal R (2013) Smart Meter Driven Segmentation: What Your Consumption Says About You. IEEE Trans Power Syst 28(4):4019–4030

Anhorn J, Herfort B, Albuquerque JPd (2016) Crowdsourced validation and updating of dynamic features in OpenStreetMap an analysis of shelter mapping after the 2015 Nepal, earthquake. In: Proceedings of the ISCRAM, 2016 Conference – Rio de Janeiro, Brazil, Rio de Janeiro. http://www.iscram2016.nce.ufrj.br/. Accessed 30 Apr 2016

Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Statist Surv 4:40–79. https://doi.org/10.1214/09-SS054

Ballatore A, Bertolotto M, Wilson D (2012) Geographic knowledge extraction and semantic similarity in OpenStreetMap. Knowledge and Information Systems 37(1):61–81

Ballatore A, Wilson DC, Bertolotto M (2013) A survey of volunteered open geo-knowledge bases in the semantic web. In: Quality issues in the management of web information, Springer. pp 93–120

Beckel C, Sadamori L, Santini S (2012) Towards automatic classification of private households using electricity consumption data. In: Pappas GJ (ed). Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings. ACM, Toronto and Ontario. pp 169–176

Beckel C, Sadamori L, Santini S (2013) Automatic socio-economic classification of households using electricity consumption data. In: Culler D, Rosenberg C (eds). Proceedings of the Fourth International Conference on Future Energy Systems. Berkeley and California, ACM. pp 75–86

Beckel C, Sadamori L, Staake T, Santini S (2014) Revealing household characteristics from smart meter data. Energy 78:397–410

Becker M (2012) Geodesy. In: Springer Handbook of Geographic Information. Springer, Berlin, Heidelberg. pp 95–117

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28. 00276

Chicco G (2012) Overview and performance assessment of the clustering methods for electrical load pattern grouping Vol. 42. pp 68–80

Constantiou ID, Kallinikos J (2015) New games, new rules: big data and the changing context of strategy. J Inf Technol 30(1):44–57

Crowston K, Li Q, Wei K, Eseryel UY, Howison J (2007) Self-organization of teams for free/libre open source software development. Inf Softw Technol 49(6):564–575. 00195

Crowston K, Wei K, Howison J, Wiggins A (2008) Free/Libre Open-source Software Development: What We Know and What We Do Not Know. ACM Comput Surv 44(2):7:1–7:35. 00330

Dangerman ATCJ, Schellnhuber HJ (2013) Energy systems transformation. Proc Natl Acad Sci 110(7):E549–E558

Elwood S, Goodchild MF, Sui DZ (2012) Researching Volunteered Geographic Information: Spatial, Data, Geographic Research, and New Social Practice. Ann Assoc Am Geogr 102(3):571–590

Eurostat (2017) Final consumption expenditure of households, by consumption purpose - Eurostat (Code: tsdpc520, Last update: 25/01/17). http://ec.europa.eu/eurostat/web/products-datasets/-/tsdpc520. Accessed 25 June 2017

Eysenbach G (2008) Medicine 2.0: Social Networking, Collaboration, Participation, Apomediation, and Openness. J Med Internet Res 10(3). https://doi.org/10.2196/jmir.1030

Fei H, Kim Y, Sahu S, Naphade M, Mamidipalli SK, Hutchinson J (2013) Heat Pump Detection from Coarse Grained Smart Meter Data with Positive and Unlabeled Learning. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13. ACM, New York. pp 1330–1338

Fernández-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15(1):3133–3181

Gangale F, Mengolini A, Onyeji I (2013) Consumer engagement: An insight from smart grid projects in Europe. Energy Policy 60:621–628. 00058

Gebauer H, Worch H, Truffer B (2014) Value Innovations in Electricity Utilities. In: Rønning R, Enquist B, Fuglsang L (eds). Framing Innovation in Public Service Sectors, Vol. 30. Routledge Studies in Innovation, Organization and Technology, Routledge. p 85ff

Gillon K, Brynjolfsson E, Mithas S, Griffin J, Gupta M (2012) Business Analytics: Radical Shift or Incremental Change? In: ICIS, 2012 Proceedings. AIS electronic library. ISBN: 978-0-615-71843-9. http://aisel.aisnet.org/icis2012/proceedings/Panels/4/

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Guyon I, Elisseeff A (2006) An Introduction to Feature Extraction. In: Guyon I, Nikravesh M, Gunn S, Zadeh L (eds). Feature Extraction, Vol. 207 of Studies in Fuzziness and Soft, Computing. Springer, Berlin, Heidelberg

Han J, Kamber M, Pei J (2012) Data mining: Concepts and techniques, The Morgan Kaufmann, series in data management systems, 3. edn. Elsevier, Amsterdam

Harvey F (2013) To Volunteer or to Contribute Locational Information? Towards Truth in Labeling for Crowdsourced Geographic, Information. In: Crowdsourcing Geographic Knowledge. Springer, Dordrecht. pp 31–42

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning, Springer Series in Statistics. Springer New York, New York

Haworth B, Bruce E (2015) A Review of Volunteered Geographic Information for Disaster Management. Geogr Compass 9(5):237–250

Heiple S, Sailor DJ (2008) Using building energy simulation and geospatial modeling techniques to determine high resolution building sector energy consumption profiles. Energy Build 40(8):1426–1436

Hopf K, Riechel S, Sodenkamp M, Staake T (2017) Predictive Customer Data Analytics – The Value of Public Statistical Data and the Geographic Model Transferability. In: Proceedings of the 38. International Conference on Information Systems (ICIS). AIS electronic library, Seoul

Hopf K, Sodenkamp M, Kozlovskiy I (2016) Energy data analytics for improved residential service quality and energy efficiency. In: Proceedings of the 24. European Conference on Information Systems (ECIS). AIS electronic library, Istanbul. http://aisel.aisnet.org/ecis2016_rip/73/

Hopf K, Sodenkamp M, Kozlovskiy I, Staake T (2016) Feature extraction and filtering for household classification based on smart electricity meter data. In: Computer Science-Research and Development, Vol. (31) 3. Springer Berlin Heidelberg, Zürich. pp 141–148

Hopf K, Sodenkamp M, Staake T (2018) Enhancing energy efficiency in the residential sector with smart meter data analytics. forthcoming, https://doi.org/10.1007/s12525-018-0290-9

Horita FEA, Degrossi LC, de Assis LFG, Zipf A, de Albuquerque JP (2013) The use of volunteered geographic information (VGI) and crowdsourcing in disaster management: a systematic literature review. In: Proceedings of the 19. Americas Conference on Information Systems (AMCIS) 2013, Chicago, Illinois. AIS electronic library. https://aisel.aisnet.org/amcis2013/eGovernment/GeneralPresentations/4/

Hua J, Tembe WD, Dougherty ER (2009) Performance of feature-selection methods in the classification of high-dimension data. Pattern Recog 42(3):409–424

Janowicz K, Raubal M, Kuhn W (2011) The semantics of similarity in geographic information retrieval. J Spat Inf Sci 2011(2):29–57

(2015) OpenStreetMap in GIScience, Lecture Notes in Geoinformation and Cartography. In: Jokar Arsanjani J, Zipf A, Mooney P, Helbich M (eds). Springer International Publishing, Cham

Kavousian A, Rajagopal R, Fischer M (2013) Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. Energy 55:184–194

Keogh E, Mueen A (2011) Curse of Dimensionality. In: Sammut C, Webb GI (eds). Encyclopedia of Machine Learning. Springer, Boston. pp 257–258

Kozlovskiy I, Sodenkamp M, Hopf K, Staake T (2016) Energy informatics for environmental, economic and social sustainability: A case of the large-scale detection of households with old heating systems. In: Proceedings of the 24. European Conference on Information Systems (ECIS). AIS electronic library, Istanbul

Krishnamurti T, Schwartz D, Davis A, Fischhoff B, de Bruin WB, Lave L, Wang J (2012) Preparing for smart grid technologies: A behavioral decision research approach to understanding consumer expectations about smart meters. Energy Policy 41:790–797. 00084

Kudo M, Sklansky J (2000) Comparison of Algorithms that Select Features for Pattern Classifiers. Pattern Recogn 33(1):25–41. 00931

Kwac J, Tan C-W, Sintov N, Flora J, Rajagopal R (2013) Utility customer segmentation based on smart meter data: Empirical study. In: Smart Grid Communications (SmartGridComm) 2013 IEEE, International Conference on. IEEE, Vancouver. pp 720–725. https://doi.org/10.1109/SmartGridComm.2013.6688044

Liaw A, Wiener M (2015) randomForest: Breiman and Cutler's Random Forests for Classification and Regression. Fortran original by Leo Breiman and Adele Cutler. https://cran.r-project.org/web/packages/randomForest/index.html. Accessed 25 Oct 2017

Liu H, Motoda H (eds) (2008) Computational methods of feature selection, Chapman & Hall/CRC data mining and knowledge discovery series. Chapman & Hall/CRC, Boca Raton

Mah DN-y, van der Vleuten JM, Hills P, Tao J (2012) Consumer perceptions of smart grid development: Results of a Hong Kong survey and policy implications. Energy Policy 49:204–216. 00063

Markard J, Truffer B (2006) Innovation processes in large technical systems: Market, liberalization as a driver for radical change?. Research Policy 35(5):609–625. 00175

McLoughlin F (2013) Characterising Domestic Electricity Demand for Customer, Load Profile Segmentation, PhD thesis. Dublin Institute of Technology. http://arrow.dit.ie/engdoc/62

Mithas S, Lee MR, Earley S, Murugesan S, Djavanshir R (2013) Leveraging Big Data and Business Analytics [Guest editors' introduction]. IT Prof 15(6):18–20

Müller O, Junglas I, Brocke Jv, Debortoli S (2016) Utilizing big data analytics for information systems research: challenges, promises and guidelines. Eur J Inf Syst 25(4):289–302

Mondzech J, Sester M (2011) Quality Analysis of OpenStreetMap Data Based on Application, Needs. Cartographica Int J Geogr Inf Geovisualization 46(2):115–125

Mooney P, Corcoran P, Ciepluch B (2013) The potential for using volunteered geographic information in pervasive health computing applications. J Ambient Intell Humanized Comput 4(6):731–745

Motsch W (2012) Dynamische Tarife zur Kundeninteraktion mit einem Smart Grid. Vieweg+Teubner Verlag, Wiesbaden

Oshiro TM, Perez PS, Baranauskas JA (2012) How many trees in a random forest?. In: Perner P (ed). Machine Learning and Data Mining in Pattern Recognition. Springer Berlin, Heidelberg. pp 154–168

Rinner C, Fast V (2015) A Classification of User Contributions on the Participatory Geoweb. In: Harvey F, Leung Y (eds). Advances in Spatial Data Handling and Analysis, Advances in Geographic, Information Science. Springer International Publishing, Cham. pp 35–49

Saar-Tsechansky M, Provost F (2007) Handling missing values when applying classification models. J Mach Learn Res 8(Jul):1623–1657

Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517

Schwering A (2008) Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. Trans GIS 12(1):5–29

See L, Mooney P, Foody G, Bastin L, Comber A, Estima J, Fritz S, Kerle N, Jiang B, Laakso M, Liu H-Y, Milčinski G, Nikšič M, Painho M, Pődör A, Olteanu-Raimond A-M, Rutzinger M (2016) Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. ISPRS Int J Geo-Inf 5(5):55

Sester M, Arsanjani JJ, Klammer R, Burghardt D, Haunert J-H (2014) Integrating and Generalising Volunteered Geographic Information. In: Burghardt D, Duchêne C, Mackaness W (eds). Abstracting Geographic Information in a Data Rich, World, Lecture Notes in Geoinformation and Cartography. Springer International Publishing. pp 119–155

Sharma R, Mithas S, Kankanhalli A (2014) Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. Eur J Inf Syst 23(4):433–441

Sodenkamp M, Kozlovskiy I, Hopf K, Staake T (2017) Smart Meter Data Analytics for Enhanced Energy Efficiency in the Residential Sector. In: Wirtschaftsinformatik 2017 Proceedings. AIS electronic library, St. Gallen

Stefanidis A, Crooks A, Radzikowski J (2013) Harvesting ambient geospatial information from social media feeds. GeoJournal 78(2):319–338. 00212

Stewart K, Ammeter T (2002) An exploratory study of factors influencing the level of vitality and popularity of open source projects. In: ICIS 2002 Proceedings. AIS electronic library

Tiefenbeck V (2017) Bring behaviour into the digital transformation. Nat Energy 2:17085

Verma A, Asadi A, Yang K, Tyagi S (2015) A data-driven approach to identify households with plug-in electrical vehicles (PEVs). Appl Energy 160:71–79

Zeifman M (2014) Smart meter data analytics: Prediction of enrollment in residential energy efficiency programs. In: 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE. pp 413–416. 00007. https://doi.org/10.1109/TCE.2011.5735484. ISSN 0098-3063

Zhou K, Fu C, Yang S (2016) Big data driven smart energy management: From big data to big insights. Renewable and Sustainable Energy Reviews 56:215–225. 00052

Zook M, Graham M, Shelton T, Gorman S (2010) Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. In: SSRN Scholarly Paper ID 2216649. Social Science Research Network, Rochester. http://papers.ssrn.com/abstract=2216649