

RESEARCH

Open Access



Explaining and predicting annual electricity demand of enterprises – a case study from Switzerland

Carlo Stingl¹, Konstantin Hopf^{1*} and Thorsten Staake^{1,2}

From The 7th DACH+ Conference on Energy Informatics
Oldenburg, Germany. 11–12 October 2018

*Correspondence:

konstantin.hopf@uni-bamberg.de

¹Information Systems and Energy Efficient Systems Group, University of Bamberg, Kapuzinerstraße 16, 96047 Bamberg, Germany

Full list of author information is available at the end of the article

Abstract

In an attempt to channel sales activities, companies often focus on ‘high value targets’ that offer attractive prospective returns. In liberalized electricity markets, commercial customers with high electricity demand constitute such high value targets. The problem when acquiring new customers, however, is that the electricity consumption is not known to the sales organization in advance. This hinders the possibility to prioritize sales targets and thus increases the acquisition cost, reduces the competitiveness within the market and ultimately leads to higher cost for electricity customers. In this study, we investigate the annual electricity consumption of enterprises by means of a dataset with 1810 company addresses in a typical town in Switzerland. We use the industry branch of the enterprises together with open big data (geographic information, online-content, social media data and governmental statistical data) to explain and predict the electricity consumption of such. Our linear regression analysis shows that information on the economic branches of the enterprises, basal area of buildings, number of opening hours and social media data can explain up to 19% of variance in electricity consumption. Economic trends (e.g., in labor market and turnover statistics) reflect changes in the electricity consumption in the investigated years 2010–2014 for several economic branches.

We show, that the electricity consumption can be predicted better than a random predictor, however with a high uncertainty. Nevertheless, the open data sources can be used to identify a relevant group of companies with high consumption (more than 100,000 kWh per year) with good accuracy.

Keywords: Enterprise electricity consumption, Open big data, Load prediction, Random forest, Economic development, High consumption customers

Background

The electricity consumption of enterprises and their development over time is a relevant information for utility companies. This holds for large and small enterprises alike. Those enterprises with an electricity consumption of more than 100,000 kWh are relevant, because the energy retail market in this segment is competitive. In Germany, for instance, the churn rates among such firms, likely to switch their electricity

supplier, are constantly at a high level of 10% (Bundesnetzagentur 2017, p. 205) and any of the four largest energy utilities¹ offer special tariffs for them. Besides that, utility companies have different effort handling supply and invoicing of electricity customers above this level (StromNZV 2005, §12). In Switzerland, the electricity retail market is only liberalized above this consumption level. Identifying these large electricity consumers is therefore relevant for utilities' sales departments, but not so much the short-term load-forecasting that has attracted much attention from researchers in the past.

Another relevant customer segment are Small and Medium Enterprises (SMEs), because they account for at least one third of the global energy demand in industry and service (International Energy Agency 2015). In some countries, the estimated share of energy consumption of SMEs is even higher, with contributions of over 60% of the industrial sector in Italy (Trianni and Cagno 2011) and 50% in the manufacturing sector in the U.S. (Trombley 2014). Numerous new enterprises are created constantly. In Switzerland, up to 40,000 new ones are founded every year (Swiss Federal Statistical Office 2017). For those newly created firms, no data on electricity consumption is available, but could be beneficial for load planning and grid operation. For example, when a new enterprise is founded or a new industrial area is designated in a city, it is relevant for local utility companies to estimate the upcoming load. Available synthetic standard load profiles for typical businesses usually cover only a limited number of consumer types ('general business', 'shop', 'bakery') and focus on the daily load distribution, but do not help to predict the overall annual consumption of enterprises.

Besides that, information on typical electricity consumption in economic branches is interesting for enterprises themselves, given that they can compare their consumption to branch standards and take actions when competitors have lower energy demand. Likewise, Simpson et al. (Simpson et al. 2004) state that implementing environmentally friendly practices to gain higher energy efficiency can lead to a competitive advantage for SMEs. In order to obtain the desired electricity consumption of enterprises, it seems appropriate to make use of public available data, such as the *economic branch* of the enterprise and *open big data* from online sources. The meaningful use of open big data sources can lead to value-adding applications (LaValle et al. 2011; Davenport 2014), in particular in the energy retail industry (Hopf 2018), and the use of big data analytics is becoming increasingly important to firms (Constantiou and Kallinikos 2015). The challenge in analyzing big data sources lays not only in the amount of data, but also in the characteristics variety, velocity and veracity. Thus, the analysis and sensemaking from raw data is necessary.

By means of a dataset with 1810 enterprise addresses in a typical town in Switzerland, we investigate the annual electricity consumption of such and use the industry branch together with open big data (geographic data, online-content, social media data and governmental statistical data) to *explain* the electricity consumption. We also evaluate to what extent a statistical model based on the public available data sources can be used to *predict* the electricity consumption of enterprise customers of this utility company.

The results of this study help to better understand the enterprise's electricity consumption per se, it allows utility companies to better plan upcoming loads or changes through economic developments, and helps companies to identify the group of high consumption customers.

Moreover, the results of this study may also help to improve modeling the electricity load in the grid. To the best of our knowledge, there is no similar study that investigates the electricity consumption of enterprises in the context of open data.

The remainder of the paper is structured as follows: First, we formulate the three research questions we will answer. Thereafter, we provide an overview of existing works and show that this is the first study investigating annual electricity consumption of enterprises together with economic branch information and open big data. Thereafter, we explain the predictor variables in detail and answer the three research questions. We discuss the findings, their implications, and future research in the concluding section.

Research goal

We formulate three Research Questions (RQs) guiding through our paper, where the first is:

RQ 1 *To what extent can the electricity consumption of enterprises be explained with the base area of the enterprise building, economic branch affiliation, opening hours and online user-reviews?*

Besides that, we analyze the development of the electricity over 5 years, identify trends that are reflected in the data and compare the trends with governmental statistical data. The correlations between trends in electricity consumption may indicate a decoupling of economic development and electricity demand, give further insights on the electricity consumption of enterprises, may lead to the identification of further influencing factors for prediction models and helps to assess the reliability of the presented models. Thus, the second RQ is:

RQ 2 *Are economic trends (e.g., in turnover statistics or job opportunities) reflected in the electricity consumption of enterprises in different industries?*

Finally, we investigate to what extend the available data can be used to predict the average annual electricity consumption of enterprises and whether this can be used as an alternative to prediction models with historical consumption data. This raises the third and last RQ:

RQ 3 *How well can the annual power consumption of enterprises be predicted by using the given data sources?*

Related work

A large body of research investigates the modeling and forecasting of energy demand with various purposes (Jebaraj and Iniyar 2006). However, to the best of our knowledge, our study is the first empirical work trying to explain annual electricity consumption of enterprises on an individual level with open big data that is available online to the public.

Existing studies often have a macroeconomic and long-term focus, explaining or predicting electricity consumption for whole countries (Wolde-Rufael 2006; Al-Bajjali and

Shamayleh 2018; Bianco et al. 2009; Mohamed and Bodger 2005), sectors (Al-Ghandoor and Samhoury 2009) or cities (Farahat 2004).

In a comprehensive study, Schlomann et al. (Schlomann et al. 2013) describe the main electricity consumption and structural data of companies in the German trade, commerce and services sector and provide an extrapolation for final energy consumption by energy source.

Besides that, several works aim at modeling the electric grid, with various focuses and research goals. Those include improving grid stability (Kinney et al. 2005), integrating renewable energy on a large scale (Pruckner et al. 2012) and advancing communication in smart grid systems (Godfrey et al. 2010).

Further related works focus on the micro-level energy demand of different consumer groups. Besides the electricity consumption of residential customers (Kavousian et al. 2013; Apadula et al. 2012), also the consumption of enterprises was investigated so far. For the short-term predictions of electricity consumption of enterprises, Gundin et al. (Gundin et al. 2002) investigate three industrial electricity consumers and use variables such as historic demand, the number of production days, capacity utilization, size and sector of the enterprises to predict the weekly power consumption of individual companies with a Relative Root Mean Squared Error (RRMSE) of 12–18%. On the level of individual enterprises, Braun et al. (Braun et al. 2014) predict the energy consumption of a supermarket with linear regression models using weather and consumption data with an Root Mean Squared Error (RMSE) of less than 4%.

For SMEs² specifically, research focused on improving energy efficiency (Trianni and Cagno 2011; Bradford and Fraser 2008; Thollander and Dotzauer 2010). Lee et al. (Lee et al. 2014) estimate the weekly electricity profile of SMEs based on the mean daily consumption and operational hours of an enterprise in combination with clusters obtained from smart meter data of 196 known SMEs. However, no further studies on modeling the electricity consumption of enterprises on a micro-level could be identified that include a decent number of companies.

In summary, numerous research on modeling and forecasting of energy demand, either aggregated or on the level of individual consumers exist. However, we could not identify studies that explain or predict the annual electricity demand of individual enterprises with data present at utility companies and open big data.

Predictors for SME electricity consumption and modeling

As a first step of our research, we identified online data sources that are publicly available and may serve as predictors for enterprise electricity demand. We identified the free geographic data from OpenStreetMap (OSM) as the first data source that we use to obtain the building basal area of the main company building, the economic branch and opening hours that can be retrieved from the companies website or from business directories, and user ratings from social media platforms. We underline that the investigated data sources comply with the characteristics of big data (LaValle et al. 2011), known as the four V's (volume, variety, velocity, veracity). Even when the investigated data is not 'big' in terms of volume, the other characteristics are fulfilled: online content is mostly unstructured or semi-structured and changes over time, different data types are considered, user-generated content may contain errors or wrong information and the amount of data increases by the number of companies investigated.

Figure 1 illustrates the identified predictors and the relationships between the variables that are investigated in this study. We justify the relations between the investigated factors and the electricity demand of enterprises below.

Building size and energy consumption

The size of the companies' building(s) has a significant influence on the electricity consumption. For instance, the annual electricity consumption per square meter in company buildings in Germany is estimated to lie between 155–183 kWh/m^2 (Schlomann et al. 2013). Accordingly, in residential buildings, the size of houses is one of the most important factor influencing the electricity consumption (Kavousian et al. 2013).

As a proxy for the actual building size, we consider the basal area of the building next to the company address, as mapped in OSM. We select OSM as the geographic information data source, because it is the currently largest free mapping website and the data quality is high (Jokar et al. 2015). There is, indeed, the possibility to store the number of building floors in the OSM database, which would enable to obtain the actual floor area of the whole enterprise building, but this functionality is only rarely used³.

Economic branch

As a second influencing factor, we consider the economic branch a company belongs to, given that the electricity demand strongly depends on the kind of business conducted. We adopt the "General Classification of Economic Activities" scheme from the Swiss Federal Statistical Office (Swiss Federal Statistical Office 2008). This allows us to compare the energy consumption development in different years also to compare with several economic trends that we investigate later in this study. The different branches are listed in Table 1.

Opening hours

We assume that longer opening hours lead to higher electricity consumption. This information can be retrieved using the Google Places API⁴. The information from this service contains opening and closing times for each day of the week. Based on this information, the amount of open hours per week can be calculated.

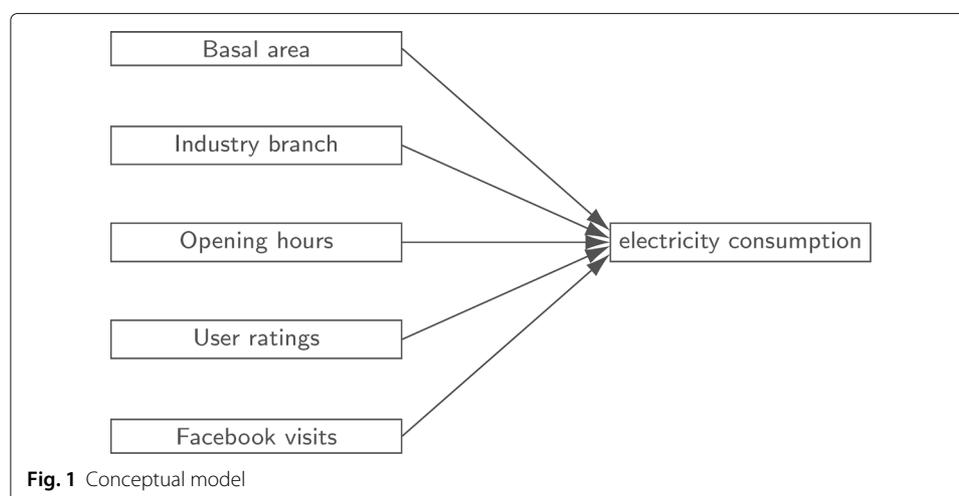


Fig. 1 Conceptual model

Table 1 Economic branch classification and number of companies in the dataset with the different open big data variables available

Sec.	Economic branch	Total	Company location with data for			Terms used for mapping
			Opening hours	Facebook visits	Online ratings	
C	Manufacturing	29	17	6	6	bäckerei, konditorei, bakery
D	Electricity, gas, steam and air conditioning supply	41	12	0	0	energie, gas, strom
E	Water supply; sewerage, waste management and remediation activities	3	0	0	0	umwelt, müll
F	Construction	409	152	7	3	bau, handwerker, maler, zimmerei, schreiner, gips, fenster, sanit
G	Wholesale and retail trade; Repair of motor vehicles and motorcycles	195	92	47	41	groceries, grocery, obst, gemüse, lebensmittel, getränk, lidl, aldi, coop, auto, car, motorrad, brillen, optiker, mode, kleidung, schuhe, fashion
H	Transportation and storage	17	2	0	0	transport, logistik, mobil
I	Accommodation and food service	180	117	70	76	hotel, hostel, restaurant, imbiss
J	Information and communication	95	51	3	3	software, it-, tele, computer, edv, informatik, medien, video, radio, zeitung, druckerei, buch
K	Financial and insurance activities	129	46	12	9	versicherung, vorsorge, bank, anlage, credit, fnanz, invest, trading, vermögen
L	Real state activities	307	59	3	3	immobilien, immo, estate, wohn
M	Professional, scientific and technical activities	70	23	4	1	archite, design, ingenieur, werbe, übersetz
N	Administrative and support service activities	3	3	1	0	travel, reise
O	Public administration and defence; compulsory social security	168	42	5	3	amt, asyl, stadt, museum
Q	Human health and social work activities	102	61	27	28	apotheke, arzt, praxis, medizin, ortho, zahn, physio
R	Arts, entertainment and recreation	3	1	0	0	fitness, gym, spa
S	Other service activities	85	28	18	16	coiffeur, friseur, haar, frisör
-	no mapping possible	446	156	19	15	
		2282	862	222	204	

Online user ratings

As a fourth influencing factor of the electricity consumption of enterprises, we take user ratings on companies' social media websites into account.

Several popular online services offer built-in rating functionalities that make statements about the quality or price level of companies possible. These evaluations, which were originally intended as a recommendation for other users, represent the popularity of places and might therefore serve as explanatory variables for the electricity consumption. We assume that companies with numerous ratings and activity on social media are more popular and have more customers than comparable companies lacking such an online presence. Consequently, comparable companies with more customers should also exhibit a higher electricity demand.

Such user ratings also served as predictors in other studies. Ye et al. (Ye et al. 2011), for example, show that user ratings and the number of reviews have a positive impact

on online hotel bookings. Facebook activity can be used to predict attendance of football matches (Egebjerg et al. 2017), user-generated content related to music albums has a positive correlation with sales (Dhar and Chang 2009) and movie ticket sales can also be predicted using online ratings (Duan et al. 2008). Social media content was also used in other areas including the prediction of election results or macroeconomic developments (Yu and Kak 2012).

We select the platforms Facebook, Yelp and Google as sources for user-generated content in this work.

Analysis

In this section, we describe the available datasets, our data preparation steps and present our analysis. We use explanatory linear regression models to answer the first RQ, correlation analysis to answer the second RQ, and evaluate predictive models to answer the third RQ.

Experimental data and data preparation

For our study, a dataset with 2282 names and addresses of enterprise locations together with annual electricity consumption in the years 2010–2014 was available. This dataset is a typical data base that is present to any energy retailing company having enterprises as customers.

All enterprises are located in an exemplary city in Switzerland⁵. We converted the address into a geographic coordinates using a geocoding service, being able to further retrieve online location data.

The electricity consumption per year was normalized by the number of consumption days, giving us the Consumption per Day (CPD). This CPD ($M = 284.58 \text{ kWh}$, $SD=1379.07 \text{ kWh}$) is suspected to contain a number of extremely high values. Initially, we transformed the consumption with the natural logarithm, resulting in an approximately normal distribution. Following Tukey (Tukey 1977), we replaced the consumption in 38 cases, where the log-transformed consumption was 1.5 times the inter-quartile-range higher than the median, with the value of the 95% percentile (1091.46 kWh). This replacement was performed to remove extreme values that might distort the linear models and leads us to an adjusted CPD of $M = 171.66 \text{ kWh}$ ($SD = 371.07 \text{ kWh}$).

We obtained the branch membership for each company location by collecting a number of words describing the business activity from three data sources. First, we used the words in the company name. Second, a business directory⁶ was used to obtain descriptions of each company. Third, keywords from the Google Places API⁷ were retrieved.

Considering the collection of all words, describing the business activities of the companies, we associated them with the respective economic branch when the textual description contained a certain keyword (see Table 1). In some cases, the branch was manually attributed. This mapping enabled us to associate economic branches for 1810 of the 2282 company locations.

We exclude all branches from our analysis with less than 25 company locations, because of low statistical validity of the findings. To get an impression of the data, we show descriptive statistics for all variables, the correlation between the variable and the logarithmized electricity consumption in Table 2. Following Cohen (Cohen 1988), all variables

Table 2 Open big data variables with presence for the company locations, descriptive statistics and the correlation with normalized electricity consumption (log)

Variable	Presence	Mean	(SD)	Correlation	Significance
Base area	1810	1120.94	(2283.62)	$\rho = 0.15$	$p < 0.01$
Economic branch	1810	-	-	$\eta^2 = 0.09$	-
Open hours per week	700	53.13	(28.04)	$\rho = 0.18$	$p < 0.01$
facebookVisits	202	141.92	(341.39)	$\rho = 0.13$	$p < 0.01$
Number of reviews	189	13.64	(23.61)	$\rho = 0.12$	$p < 0.01$

show a weak positive correlation with the electricity consumption, which suggests a further examination of the relationship using linear regression models.

We have no information on the size of the enterprises (turnover or number of employees), but we assume that a large portion are SMEs and we find evidence in two descriptive facts on the data. First, we found 1467 unique enterprise names enabling us to group the addresses to enterprises. Each enterprise has $M = 1.65$ ($SD = 3.79$) locations, but the majority (80%) of enterprises have only one address. The grouping of addresses was just a descriptive analysis and we use the company locations independently from their affiliation to an enterprise in the remaining analysis of the paper. Second, the median of the base area of all enterprises is $476.28m^2$ (e.g., a square with a side length of $22m$).

Explanatory models of the electricity consumption

In this first analysis, we use linear regression models with ordinary least squares estimation⁸ and answer RQ 1 based on the data. The regression models are described in Eq. 1 in a general form. For each observation i , we consider the mean CPD_i for all years as the dependent variable and transform the values with the natural logarithm, given that the distribution of this variable is approximately log-normal. In different models, we use n explanatory variables $x_j, j \in \{1, \dots, n\}$ to investigate combinations of them. While β_0 represents the intercept, $\beta_j, j \in \{1, \dots, n\}$ are regression coefficients that describe the size of the effect of the variables x_j .

$$\log(CPD_i) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_n x_{ni} + \epsilon_i \quad (1)$$

The explanatory variables basal area, opening hours, user ratings and Facebook visits are numeric and are used as we obtained the values from the open data sources. The industry branch is a categorical variable which we represented as a binary dummy variables for all branches, whereas the economic branch “S” (other service activities) serves as default and is encoded in case all dummy variables are zero. ϵ_i denote the error terms in the regression model. We estimate separate models for the different influencing factors first (Model 1 – 5) to see the direct effect of the variables on the electricity consumption and the amount of explained variance (R^2). Model 6 and 7 combine the different variables.

Table 3 shows the estimated coefficients for linear regression models for the variables base area, opening hours, number of visitors on Facebook and the combined number of reviews on Yelp, Google and Facebook independently. All variables have a statistically significant effect in the individual models. The estimated effects can be interpreted as follows: Per m^2 basal area, the electricity consumption increases by $e^{0.239} = 1.269979$ kWh, per additional opening hour, the consumption increases by 1.0% ($e^{0.009937} = 1.009987$). Per additional online rating, the consumption increases by

Table 3 Linear regression models explaining logarithmized CPD with each influencing factor separately

	Model 1	Model 2	Model 3	Model 4
(Intercept)	2.36*** (0.23)	3.59*** (0.13)	3.84*** (0.04)	4.28*** (0.12)
log(area + 1)	0.24*** (0.04)			
opening hours per week		0.01*** (0.00)		
combined number of ratings			0.02*** (0.00)	
number of facebook visits				0.00*** (0.00)
R ²	0.02	0.03	0.02	0.08
Adj. R ²	0.02	0.03	0.02	0.08
Num. obs.	1810	700	1810	202
RMSE	1.67	1.67	1.67	1.57

****p* < 0.001

2.5% ($e^{0.02429} = 1.024587$). The increase in consumption per Facebook per additional visit is small with 0.14% ($e^{0.001366} = 1.001367$) and only estimated based on a smaller sample, but the effect is statistically significant.

According to the low estimates of the coefficients in the models, the explained variance (*R*²) of the logarithmized CPD is quite low, ranging from 2% to 8%. The *R*² for Model 4 is slightly higher than for Model 1–3, even though the effect of Facebook visits is small. We assume that this is a result of the different numbers of observations (202 instead of 1810) that are available, given that only those companies offered a Facebook page.

The influence of the economic branches is included in Model 5 (Table 4).

In this model, the branch membership has a significant influence on the electricity consumption and the explained variance is higher than in the Models 1–4.

Table 4 Linear regression models explaining logarithmized CPD with the branch information and combined models with multiple influencing factor

	Model 5	Model 6	Model 7
(Intercept)	2.65 (0.18)***	1.77 (0.46)***	1.82 (0.59)**
branche C	2.85 (0.35)***	3.05 (0.49)***	
branche D	1.25 (0.31)***	2.42 (0.54)***	
branche F	1.24 (0.19)***	1.13 (0.32)***	
branche G	1.56 (0.21)***	1.27 (0.34)***	1.26 (0.33)***
branche I	2.17 (0.21)***	1.94 (0.35)***	1.83 (0.34)***
branche J	1.08 (0.24)***	1.19 (0.37)**	
branche K	1.04 (0.23)***	1.26 (0.38)***	
branche L	1.15 (0.20)***	1.46 (0.36)***	
branche M	0.65 (0.26)*	0.85 (0.44)	
branche O	0.88 (0.21)***	0.39 (0.39)	
branche Q	0.90 (0.24)***	0.99 (0.36)***	0.98 (0.34)**
opening hours per week		0.00 (0.00)	0.01 (0.00)*
combined number of ratings		0.01 (0.01)*	0.01 (0.00)*
log(area + 1)		0.13 (0.05)*	0.09 (0.08)
R ²	0.09	0.15	0.19
Adj. R ²	0.08	0.13	0.18
Num. obs.	1810	700	298
RMSE	1.61	1.57	1.51

****p* < 0.001, ***p* < 0.01, **p* < 0.05

Model 6 and 7 in Table 4 show the estimates for multinomial regression including also variables from online data sources. By adding the number of opening hours, Facebook visits and the basal area to the model, the estimates for branches M and O are not anymore significant, but the explained variance increased (adjusted $R^2 = 0.13$).

In Model 7, we consider only service-oriented enterprises with direct customer contact, because these companies have also a sufficient number of online ratings and social media data present. Interestingly, the opening hours have a slightly higher influence in this model and the explained variance could be further increased (adjusted $R^2 = 0.18$). One reason for that can also be that the companies in these branches are more homogeneous. We conclude that we can explain electricity consumption of enterprises to some extent and thereby answer our first RQ.

Reflection of economic trends in electricity consumption of enterprises

In the available dataset, the annual electricity consumption for the years 2010–2014 is available. In this analysis, we want to see whether economic trends are reflected in the energy consumption of typical enterprises in different economic branches and thus answer RQ 2.

For data on economic trends, the Swiss Federal Statistical Office offers numerous official statistics. For the years 2010–2014, datasets on employment, turnover and electricity consumption were retrieved, where the same branch classification as in Table 1 was used⁹. All statistics are aggregations on the level of the local canton of the city, except for energy consumption, where the data for whole Switzerland was used. We answer our second RQ for each of the considered statistic data below.

Labor market statistics No significant correlation between labour market statistics and the electricity consumption exists in most branches. However, in the construction branch a strong and significant correlation ($p < 0.1$) is present.

Turnover statistics Turnover statistics are available for the secondary sector (manufacturing, industry, crafts, energy and construction) in Switzerland. Sales for each quarter were reported as indices (annual average 2010 corresponds to 100%). The annual average was calculated for these quarterly figures, which in turn was used to calculate the correlation with electricity consumption. The results are shown in Fig. 2. No significant correlations ($p < 0.1$) could be found for the sectors C (manufacturing industry / manufacture of goods) and D (energy supply). However, there is a strong linear correlation for the construction industry (F).

Nationwide electricity consumption The majority of economic branches (12 of 16) show a positive correlation, of which D, F and M have a very strong and significant correlation with $\rho > 0.7$. The relationship between nationwide consumption and that of enterprises in our dataset can give a perception of how representative they are for all of Switzerland. While a positive correlation leads to the assumption that findings from those branches have more general importance, this assumption can not be made for branches with a strong negative correlation (K and S).

Sec	Economic branch	Examples	Correlation of electricity consumption and branch statistics				
			open jobs	turnover	employees	places of employment	electricity consumption
C	Manufacturing	31	-	cor:0.283 p=0.644	cor:-0.187 p=0.814	cor:0.029 p=0.971	cor: -0.15 p= 0.809
D	Electricity, gas, steam and air conditioning supply	43	-	cor:-0.658	cor:-0.122 p=0.879	cor:-0.644 p= 0.356	cor:0.911 p= 0.031
E	Water supply; sewerage, waste management and remediation activities	3	-	-	cor:-0.909 p= 0.091	cor:-0.930 p= 0.070	cor:0.244 p=0.692
F	Construction	431	cor:-0.871 p= 0.054	cor:0.841 p= 0.075	cor:0.779 p=0.222	cor:-0.985 p= 0.015	cor:0.897 p= 0.039
G	Wholesale and retail trade; Repair of motor vehicles and motorcycles	203	cor:0.457 p=0.439	-	cor:0.960 p= 0.04	cor:0.908 p= 0.092	cor:0.702 p=0.186
H	Transportation and storage	19	cor:0.319 p=0.6	-	cor:0.178 p=0.822	cor:-0.813 p=0.188	cor:0.627 p=0.257
I	Accommodation and food service	195	cor:-0.039 p=0.951	-	cor:-0.688 p=0.312	cor:0.807 p=0.194	cor:0.186 p=0.764
J	Information and communication	100	cor:0.248 p=0.688	-	cor:0.841 p=0.159	cor:0.735 p=0.265	cor:0.712 p=0.178
K	Financial and insurance activities	132	cor:0.115 p=0.854	-	cor:-0.974 p= 0.026	cor:0.455 p=0.545	cor:-0.725 p=0.166
L	Real state activities	336	cor:0.014 p=0.986	-	cor:0.196 p=0.805	cor:-0.872 p=0.128	cor:0.337 p=0.663
M	Professional, scientific and technical activities	81	-	-	cor:-0.985 p= 0.015	cor:-0.398 p=0.602	cor:0.863 p= 0.06
N	Administrative and support service activities	3	cor:-0.718 p=0.172	-	cor:-0.725 p=0.275	cor:-0.995 p= 0.005	cor:0.373 p=0.536
O	Public administration and defence; compulsory social security	204	cor:-0.699 p=0.189	-	cor: 0.224 p=0.776	cor: 0.629 p=0.371	cor:-0.183 p=0.768
Q	Human health and social work activities	106	cor: 0.296 p=0.629	-	cor: -0.635 p=0.365	cor: -0.327 p=0.673	cor:0.458 p=0.438
R	Arts, entertainment and recreation	3	cor: -0.32 p=0.599	-	cor: -0.345 p=0.655	cor: -0.957 p= 0.044	cor:0.252 p= 0.682
S	Other service activities	91	-	-	cor: -0.949 p= 0.051	cor: 0.156 p=0.844	cor:-0.740 p=0.153

Strength of correlation ρ : >0.7 0.7 to 0.3 0.3 to 0.1 0.1 to 0.1 -0.1 to -0.3 -0.3 to -0.7 <-0.7
all significant values with $p < 0.1$ are written in bold

Fig. 2 Correlation of electricity consumption with governmental statistical data in the years 2010–2014

In summary, some interesting points have emerged from the study of the links between the electricity consumption and other statistical surveys. In some sectors, for example, there are strong and significant correlations between electricity consumption and various labour market statistics. However, there is no uniform picture of the nature of the interrelationships: whereas there is a strongly positive correlation in the retail sector, the correlations in the other sectors are usually negative. A further investigation of these interrelationships and the causalities behind them can be a goal of further research.

In addition, there is a positive correlation for most industries between the development of electricity consumption of enterprises in our dataset and the development of consumption throughout Switzerland.

Prediction of annual power consumption

In this final analysis, we answer RQ 3 and test, how well our presented models can be used to predict the electricity consumption of an enterprise for which no electricity consumption data is known.

For prediction, we consider the linear regression model 5 and 6 (see Table 4). In previous studies, linear regression models showed a good prediction performance, even in comparison with neural network and decision tree machine learning algorithms (Al-Ghandour

and Samhoury 2009; Tso and Yau 2007). However, we compare the prediction performance of the linear regression model with a Random Forest (Breiman 2001) regression model, trained with the same data as model 6.

To measure the prediction error, we use the actual electricity consumption per day y_i and compare it to the predicted consumption \hat{y}_i for every company $i \in \{1, \dots, n\}$. We can then compute the Mean Absolute Percentage Error (MAPE):

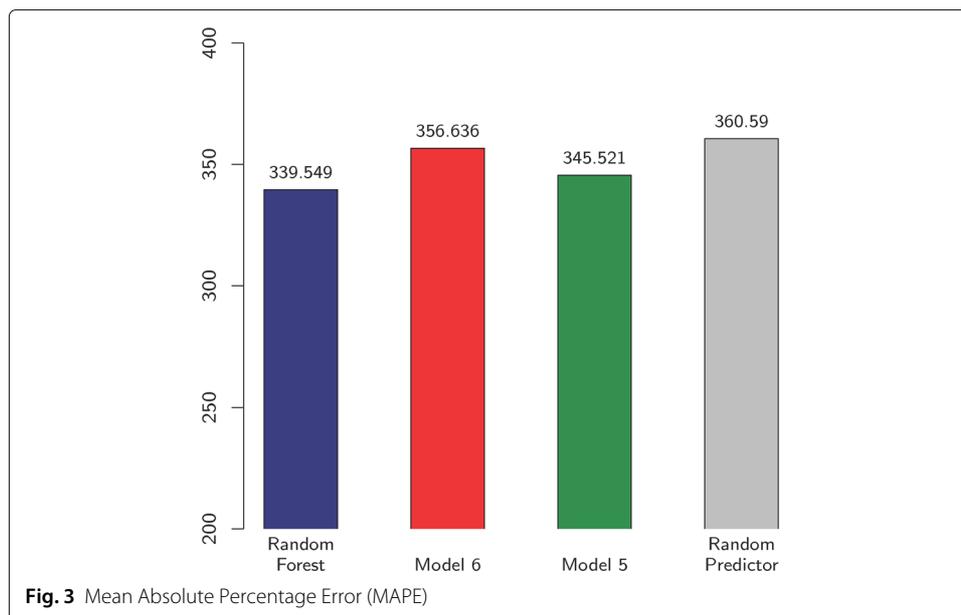
$$MAPE = \frac{100}{n} * \sum_{i=1}^n \left(\left| \frac{y_i - \hat{y}_i}{y_i} \right| \right) \tag{2}$$

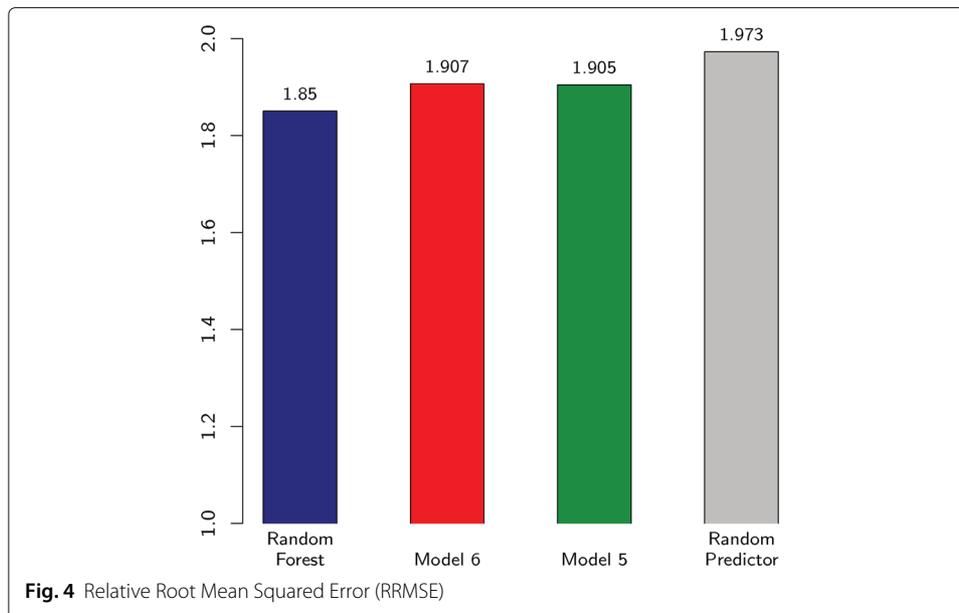
To get an impression to what extent the prediction deviates from the average electricity consumption \bar{y} , we consider the RRMSE:

$$RRMSE = \frac{\sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}}{\bar{y}} \tag{3}$$

For an unbiased estimation of the errors, we use 10-fold cross-validation¹⁰. As a benchmark measure, we consider a random predictor taking the average electricity consumption of all company locations.

We show the results in Figs. 3 and 4. The prediction error is high for all considered models. Expectably, the random predictor has the worst performance in all metrics, the Random Forest model shows the best performance, with both regression models in between. Interestingly, the inclusion of open big data (basal area and opening hours) in the regression model 6 leads to a higher predictive error than only using economic branches (model 5) as a predictor. However, this could also be a result of model overfitting. We could not achieve significant less prediction errors by considering only the companies with strong relations to consumers (those in economic branches I, G, Q or S).

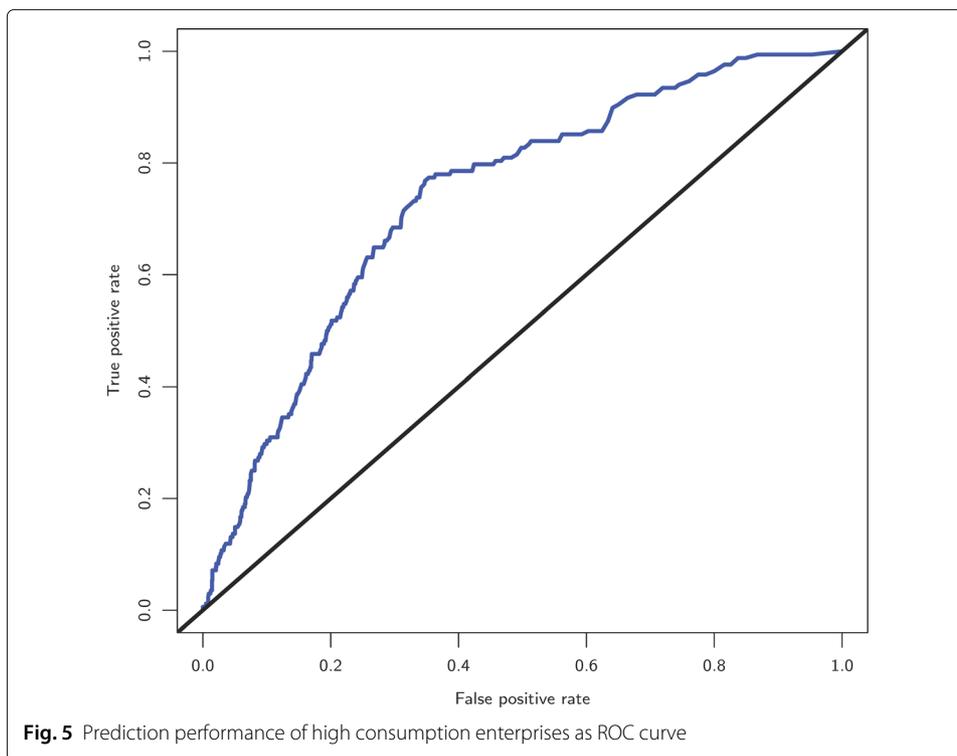




Previous literature achieved forecast errors for long-term power consumption in the industrial sector of approximately 2% (Farahat 2004) and suggests that for energy suppliers in long-term forecasts an error of up to 10% is acceptable, which is clearly exceeded here. In addition, Savka (Savka 2005, p. 52ff) shows that predicting electricity consumption for one year in advance in the industrial and commercial sector is possible values of 6% and 3%, respectively. Those accurate load forecasts have been enabled by time series data of past consumption, which was *not used for our predictions*. We conclude that the detailed prediction of the actual electricity consumption based on open big data is not reliable, but can give a first estimate when historic consumption of a potential customer is not available.

In some cases, the actual electricity consumption of enterprises is not necessary and it is sufficient to identify high energy consumers with annual electricity consumption of more than 100,000 kWh. We therefore train a Random Forest classification model with the branch information and open big data features and use the Receiver Operating Characteristic (ROC) curve for evaluation (see Fig. 5). This curve shows the performance of a binary classifier by plotting the true positive rate against the false positive rate of classification. The Area Under ROC Curve (AUC) is a well-known metric to evaluate classifier (Fawcett 2006) and is in our case $AUC = 0.74$. A random classification is considered as a diagonal line from (0,0) to (1,1) in the plot corresponding to an $AUC = 0.5$. For further information, we provide the feature importance scores of the Random Forest prediction model in Table 5.

In conclusion, we can answer RQ 3 as follows: The prediction of the annual power consumption of enterprises based on public available data is possible better than random, but still associated with a high prediction error. Nevertheless, the identification of companies with a high electricity consumption of more than 100,000 kWh annually is possible based on branch information and open big data.



Discussion and conclusion

In this paper, we investigated the annual electricity consumption of 1810 company addresses in an exemplary Swiss city together with information on the economic branch and open big data from various sources (geographic information, online content, social media data and governmental statistical data). In contrast to previous studies, we used only explanatory variables from public available online sources. Based on the data, we answered three research questions and can draw the following three conclusions from our research:

First, the electricity consumption of SMEs can be explained with open big data and information on the company branch using linear regression models. In detail, the size of the companys’ buildings increases the electricity consumption by 1.27kWh per additional m^2 , each online review increases the consumption by 2.5%, each opening hour by 1.0% and each Facebook visit by 0.14%, when using the variables as single predictors. Nevertheless, only a small part of the variance in electricity demand can be explained (from 2% to 8%) with the simple models using only one explanatory variable. By using all variables

Table 5 Random Forest feature importance scores for the prediction of high consumption enterprises

	Class low consumption	Class high consumption	Mean Decrease Accuracy	Mean Decrease Gini
brancheBFS	1.92	21.85	8.61	167.74
log(area+1)	10.44	33.86	21.35	309.43
number of facebook visits	2.74	4.48	4.26	25.56
opening hours per week	4.83	12.87	8.54	118.31
combined number of ratings	-4.99	10.27	-1.451	30.36

and adding the branch information to a combined model, our linear regression analysis shows that up to 19% of variance in electricity consumption can be explained among the service-oriented enterprises with direct customer contact, and up to 13% of variance considering all branches.

Second, economic trends in different industries (e.g., in turnover statistics or job opportunities) are reflected in the electricity consumption of SMEs to some extent, especially in the labor-intensive construction industry. The electricity consumption of enterprises in some economic branches developed alongside open statistical surveys (such as economic development or labour market statistics) over time with strong and significant correlation.

Third, the annual power consumption of enterprises can be predicted by using the considered public available data sources. The exact prediction of the electricity consumption using linear regression and Random Forest regression led, however, to a high average forecasting error of 340%. A random predictor, which always assumes the average as a prediction, has an error of 360%. Nevertheless, the identification of companies with a high energy consumption of more than 100,000 *kWh* is possible with an $AUC = 0.74$.

Implications and contribution

Our study contributes to the sparse literature on explaining and predicting the electricity consumption of enterprises by investigating new predictor variables for the electric load of such and investigating the topic with a comprehensive dataset of 1810 company addresses.

Our results have implications for grid planning, load forecasting and energy modeling in utility companies. Competitors may use the public available data for benchmarking, as we show that the explanation and prediction of enterprise energy consumption can be supported by open big data, as firms or researchers can include the estimated influence of basal area, industry branch, opening hours, number of user ratings and Facebook visits into their energy models. Besides that, we showed how companies with a high energy consumption ($> 100,000 \text{ kWh}$) can be identified, which is a beneficial insight for electricity retailers.

We underline, that all data for the considered predictor variables stems from public available online sources and is available to researchers and practitioners for future works.

Our results extend findings from the most comprehensive study investigating the electricity demand of enterprises (Lee et al. 2014) that uses data from 196 Irish SMEs). We find support in our data that operational hours of enterprises are valid predictors of the electricity demand, but find evidence to the obvious fact that the economic branches of an enterprise affects the electricity demand to a large extent (which Lee et al. (Lee et al. 2014) found no evidence for).

Limitations and future research

With an explained variance of up to 19%, the identified factors do not provide a full explanation of the electricity consumption of companies and further factors should be considered for a complete picture. Possible ones include the annual revenue, number and size of production equipment or the number of employees. We motivate further research to investigate such factors.

Given that a large portion of companies in our dataset are SMEs, the results presented are especially valid for SMEs and can explain the energy consumption for the companies that account for a large proportion of overall electricity consumption.

A subject of future research can be the extension of our analysis on enterprises to a broader geographic scope. So far, only companies from a single municipality from Switzerland have been considered in our case study. To lower the forecasting error of our prediction of enterprise energy consumption, further advanced prediction models (such as artificial neural networks or recurrent neural networks) could be tested. For the analysis of the reflection of economic trends in energy consumption of enterprises we used a correlation analysis. However, a panel data analysis using regression models with a time-dimension would be helpful to further verify the findings and could be subject of future work.

Furthermore, more open big data sources could be examined as influencing factors of enterprise electricity consumption. This research could be inspired by previous work on analyzing household electricity consumption with open geographic data. Hopf et al. (Hopf et al. 2016) for example used features derived from OSM to a much greater extent than this paper, including topological features, land use and landmarks in their analysis of household consumption.

Endnotes

¹ The “four strongest companies” in the German electricity retail markets are, according to the Bundesnetzagentur (Bundesnetzagentur 2017): RWE, E.ON, EnBW and Vattenfall.

² The European Commission (European Commission 2015) defines SMEs as enterprises with less than 250 employees and either annual turnover of less than 50 Mio. EUR or a balance sheet total of less than 43 Mio. EUR. The reviewed literature has either followed this definition (Trianni and Cagno 2011) or used comparable ones only focusing on the fact that the number of employees is less than 250 employees (Thollander and Dotzauer 2010).

³ The respective tag ‘floor’ or ‘addr:floor’ are just used 239 times in Switzerland (<http://taginfo.openstreetmap.ch/search?q=floor>, last accessed on March 22, 2018).

⁴ <https://developers.google.com/places/web-service/details>, last accessed on March 26, 2018.

⁵ The municipality is comparably large with approximately 44,000 inhabitants in 2015, the average municipality in Switzerland in the same year had $M=3638$ ($SD=12,016$) inhabitants (Swiss Federal Statistical Office 2018).

⁶ <http://www.tel.search.ch>, last accessed on March 22, 2018.

⁷ <https://developers.google.com/places/web-service/details>, last accessed on March 26, 2018.

⁸ using the “lm”-function in R version 3.4.3.

⁹ All statistics are openly available at STAT-TAB, <https://www.pxweb.bfs.admin.ch/pxweb/en/>, last accessed on March 26, 2018.

¹⁰ The cross folds are created using stratified random sampling on economic branch, using the package ‘caret’ in R (Kuhn 2015).

¹¹ <http://cran.r-project.org/>, last accessed on March 22, 2018.

Abbreviations

API: Application programming interface; AUC: Area under ROC curve; CPD: Consumption per day; OSM: OpenStreetMap; MAPE: Mean absolute percentage error; POI: Point of interest; POIs: Points of interest; ROC: Receiver operating characteristic; RQ: Research question; RMSE: Root mean squared error; RRMSE: Relative root mean squared error; SME:

Small and medium enterprise; VGI: Volunteered geographic information; WGS84: World geodetic system 84; XML: Extensible markup language

Acknowledgments

We kindly thank BEN Energy AG (Zurich, Switzerland) for their support, expertise and valuable feedback during the study.

Funding

The financial support from Eureka member countries and European Union (EUROSTARS Grant number E19859 - BENGINE II) is gratefully acknowledged. Publication costs for this article were sponsored by the Smart Energy Showcases - Digital Agenda for the Energy Transition (SINTEG) programme.

Availability of data and material

Due to its nature, open data is available to the public and can be retrieved from the respective source. All computational methods used are open source and available via the Comprehensive R Archive Network¹¹. 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).

About this supplement

This article has been published as part of *Energy Informatics* Volume 1 Supplement 1, 2018: Proceedings of the 7th DACH+ Conference on Energy Informatics. The full contents of the supplement are available online at <https://energyinformatics.springeropen.com/articles/supplements/volume-1-supplement-1>.

Author's contributions

CS conducted the data collection and the statistical analysis. KH and CS wrote the manuscript. TS provided critical review and wrote parts of abstract and introduction. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Information Systems and Energy Efficient Systems Group, University of Bamberg, Kapuzinerstraße 16, 96047 Bamberg, Germany. ²Department of Management, Technology and Economics, ETH Zurich, Weinbergstrasse 5, 8092 Zurich, Switzerland.

Published: 10 October 2018

References

- Al-Bajjali SK, Shamayleh AY (2018) Estimating the determinants of electricity consumption in Jordan. *Energy* 147:1311–1320
- Al-Ghandoor A, Samhouri M (2009) Electricity Consumption in the Industrial Sector of Jordan: Application of Multivariate Linear Regression and Adaptive Neuro-Fuzzy Techniques. *JJMIE Jordan J Mech Ind Eng* 08:3
- Apadula F, Bassini A, Elli A, Scapin S (2012) Relationships between meteorological variables and monthly electricity demand. *Appl Energy* 98:346–356
- Bianco V, Manca O, Nardini S (2009) Electricity consumption forecasting in Italy using linear regression models. *Energy* 34(9):1413–1421
- Bradford J, Fraser ED (2008) Local authorities, climate change and small and medium enterprises: identifying effective policy instruments to reduce energy use and carbon emissions. *Corp Soc Responsib Environ Manag* 15(3):156–172
- Braun MR, Altan H, Beck SBM (2014) Using regression analysis to predict the future energy consumption of a supermarket in the UK. *Appl Energy* 130:305–313
- Breiman L (2001) Random forests Vol. 45. pp 5–32
- Bundesnetzagentur (2017) Monitoring report 2017. <https://www.bundesnetzagentur.de/SharedDocs/Downloads/EN/Areas/ElectricityGas/CollectionCompanySpecificData/Monitoring/MonitoringReport2017.pdf>. Accessed 22 Aug 2018
- Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences. In: Revised edition ed. Routledge
- Constantiou ID, Kallinikos J (2015) New games, new rules: big data and the changing context of strategy. *J Inf Technol* 30(1):44–57
- Davenport T (2014) Big data at work: dispelling the myths, uncovering the opportunities. Harvard Business Review Press, Boston
- Dhar V, Chang EA (2009) Does Chatter Matter? The Impact of User-Generated Content on Music Sales. In: *Journal of Interactive Marketing*. vol. 23. pp 300–307
- Duan W, Gu B, Whinston A (2008) Do online reviews matter? An empirical investigation of panel data. *Decis Support Syst* 11;45(4):1007–1016
- Egebjerg NH, Hedegaard N, Kuum G, Mukkamala RR, Vatrapu R (2017) Big Social Data Analytics in Football: Predicting Spectators and TV Ratings from Facebook Data. In: 2017 IEEE International Congress on Big Data (BigData Congress). pp 81–88
- European Commission (2015) User guide to the SME Definition. <http://ec.europa.eu/DocsRoom/documents/15582/attachments/1/translations>. Accessed 22 Aug 2018
- Farahat MA (2004) Long-term industrial load forecasting and planning using neural networks technique and fuzzy inference method. In: 39th International Universities Power Engineering Conference, 2004. UPEC 2004. vol. 1. pp 368–372
- Fawcett T (2006) An introduction to ROC analysis. *Pattern Recogn Lett* 27(8):861–874
- Godfrey T, Mullen S, Griffith DW, Golmie N, Dugan RC, Rodine C (2010) Modeling Smart Grid Applications with Co-Simulation. In: 2010 First IEEE International Conference on Smart Grid Communications. pp 291–296

- Gundin D, Garca C, Gomez-Sanchez E, Dimitriadis Y, Vega-gorgojo G (2002) Short-Term Load Forecasting For Industrial Customers Using Fasart And Fasback Neuro-Fuzzy Systems. In: Proceedings of the 14th Power Systems Computation Conference. PSCC
- Hopf K (2018) Mining volunteered geographic information for predictive energy data analytics. *Energy Inform* 1(1):4
- Hopf K, Sodenkamp M, Kozlovskiy I (2016) Energy Data Analytics for Improved Residential Service Quality and Energy Efficiency. In: ECIS 2016 Proceedings. AIS electronic library, Istanbul
- International Energy Agency (2015) Accelerating Energy Efficiency in Small and Medium-sized Enterprises. https://www.iea.org/publications/freepublications/publication/SME_2015.pdf. Accessed 22 Aug 2018
- Jebaraj S, Iniyar S (2006) A review of energy models. *Renew Sust Energ Rev* 10(4):281–311
- Jokar AJ, Zipf A, Mooney P, Helbich M (2015) OpenStreetMap in GIScience. In: Lecture Notes in Geoinformation and Cartography. Cham: Springer International Publishing
- 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
- Kinney R, Crucitti P, Albert R, Latora V (2005) Modeling cascading failures in the North American power grid. *Eur Phys J B - Condens Matter Complex Sys* 46(1):101–107
- Kuhn M (2015) Classification and Regression Training. In: R Documentation. <https://www.rdocumentation.org/packages/caret/versions/6.0-78>. Accessed 22 Aug 2018
- LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N (2011) Big data, analytics and the path from insights to value. *MIT Sloan Manag Rev* 52(2):21
- Lee TE, Haben SA, Grindrod P (2014) Modelling the Electricity Consumption of Small to Medium Enterprises. In: Russo G, Capasso V, Nicosia G, Romano V (eds). *Progress in Industrial Mathematics at ECMI 2014*. Cham: Springer International Publishing. pp 341–349
- Mohamed Z, Bodger P (2005) Forecasting electricity consumption in New Zealand using economic and demographic variables. In: *Energy*. vol. 30. pp 1833–1843
- Pruckner M, Bazan P, German R (2012) Towards a simulation model of the Bavarian electrical energy system. In: *GI-Jahrestagung*. pp 597–612
- Savka D (2005) Evaluation of errors in national energy forecasts. Rochester Institute of Technology
- Schlomann B, Kleeberger H, Pich A, Gruber E, Mai M, Gerspacher A, et al. (2013) Energieverbrauch des Sektors Gewerbe, Handel, Dienstleistungen (GHD) in Deutschland für die Jahre 2007 bis 2010. In: Fraunhofer-Institut für System- und Innovationsforschung
- Simpson M, Taylor N, Barker K (2004) Environmental responsibility in SMEs: does it deliver competitive advantage? *Business strategy and the environment* 13(3):156–171
- StromNZV (2005) Verordnung über den Zugang zu Elektrizitätsversorgungsnetzen (Stromnetzzugangsverordnung - StromNZV). 2005. *Bundesgesetzblatt* 46:2243–2251. <https://www.gesetze-im-internet.de/stromnzhv/BJNR22430000.html>. Accessed 4 Aug 2018
- Swiss Federal Statistical Office (2017) Neu gegründete Unternehmen nach Kanton und Wirtschaftssektor. <https://www.bfs.admin.ch/bfs/de/home/statistiken/industrie-dienstleistungen/unternehmen-beschaeftigte/unternehmensdemografie.html>. Accessed on 22 Aug 2018
- Swiss Federal Statistical Office (2008) NOGA 2008: General Classification of Economic Activities. Swiss Federal Statistical Office. <https://www.bfs.admin.ch/bfs/de/home/statistiken/industrie-dienstleistungen/nomenklaturen/noga/publikationen-noga-2008.assetdetail.344611.html>. Accessed on 22 Aug 2018
- Swiss Federal Statistical Office (2018) Sustainable Development, Regional and International Disparities / Statistical Basis and Overviews. <https://www.bfs.admin.ch/bfs/en/home/statistics/regional-statistics/regional-portraits-key-figures/communes.assetdetail.2422865.html>. Accessed 22 Aug 2018
- Thollander P, Dotzauer E (2010) An energy efficiency program for Swedish industrial small- and medium-sized enterprises. *J Clean Prod* 18(13):1339–1346
- Trianni A, Cagno E (2011) Energy Efficiency Barriers in Industrial Operations: Evidence from the Italian SMEs Manufacturing Industry. In: ACEEE's Summer Study on Energy Efficiency in Industry
- Trombley D (2014) One small step for energy efficiency: Targeting small and medium-sized manufacturers. In: American Council for an Energy Efficient Economy
- Tso G, Yau K (2007) Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy* 9;32(9):1761–1768
- Tukey JW (1977) *Exploratory data analysis*. vol 2. In: Reading, Mass
- Wolde-Rufael Y (2006) Electricity consumption and economic growth: a time series experience for 17 African countries. *Energy Policy* 34(10):1106–1114
- Ye Q, Law R, Gu B, Chen W (2011) The Influence of User-Generated Content on Traveler Behavior: An Empirical Investigation on the Effects of E-Word-of-Mouth to Hotel Online Bookings. *Comput Hum Behav* 27:634–639
- Yu S, Kak SC (2012) A Survey of Prediction Using Social Media. In: CoRR, dblp computer science bibliography. <https://arxiv.org/abs/1203.1647>. Accessed 22 Aug 2018