https://doi.org/10.20378/irbo-50522

Predictive Customer Data Analytics – The Value of Public Statistical Data and the Geographic Model Transferability

Completed Research Paper

Konstantin Hopf
University of Bamberg
Kapuzinerstraße 16, 96047 Bamberg
konstantin.hopf@uni-bamberg.de

Sascha Riechel
University of Bamberg
Kapuzinerstraße 16, 96047 Bamberg
sascha-riechel@gmx.de

Mariya Sodenkamp
University of Bamberg
Kapuzinerstraße 16, 96047 Bamberg
mariya.sodenkamp@uni-bamberg.de

Thorsten Staake
University of Bamberg
Kapuzinerstraße 16, 96047 Bamberg
thorsten.staake@uni-bamberg.de

Abstract

Companies pay high prices for detailed customer information (e.g., income, household type) for gaining insights and conducting targeted marketing campaigns. We argue that companies can utilize predictive analytics artifacts to derive such information from existing customer data in combination with freely available data sources, such as open government data. In this study, we use a machine learning artifact for a specific yet highly relevant case from the utility industry, trained on data of 7,504 energy customers and investigate two important aspects for predictive business analytics: First, we identified the sparsely available open government statistics and found that even that limited amount of open data can increase our artifact's performance. Second, we applied the predictive models, trained with a regional customer dataset, on households in other geographic regions with acceptable performance loss. The results support the development of systems aiding managerial decision-making, predictive marketing and showcase the value of open data.

Keywords: Predictive Analytics, Transferability, Supervised Machine Learning, Open Government Data, Public Sector Information, Data Integration

Introduction

Information on customers, beyond the customer core data that is usually gathered for order fulfillment and invoicing, attracts high interest among companies. Such data serves as source for targeted marketing campaigns, including up- and cross-selling, sales forecasts, or companies' strategic alignment.

Energy retailer particularly rely on such data to realize customer loyalty and energy-efficiency campaigns. Such campaigns for retaining existing or attracting new customers became necessary in consequence of increased market pressure due to the ongoing energy retail market liberalization in many countries. Customer churn rates in that industry are at substantially high levels: for example, 6.4% in Germany 2015 (BNetzA 2016 p. 184), 10.3% in Sweden 2014 and 12.7% in Norway 2014 (NordREG 2017). Another reason is the fact that utility companies need to carry out cost-effective energy saving campaigns among residential customers, triggered by a growing number of energy-efficiency mandates. Such mandates range from incentives for efficient behavior to subsidies for energy-efficient construction – for example in Germany (BMUB 2014) – to mechanisms that decouple the amount of energy sold from the profit of utilities in some states of the U.S. (Eto et al. 1997).

The knowledge on individual residential customer details (e.g., whether the household belongs to the group of families with children vs. retirees, electric vs. oil/gas heating, or the age of household members) can be utilized for the above-mentioned purposes in marketing and energy efficiency. For instance, utilities can tailor energy saving advices to specific customer groups or give consumption feedback that includes references to similar dwellings addressees, which has been proven to be more effective than unspecific feedback (Ayres et al. 2013; Fischer 2008; Tiefenbeck et al. 2013; Vassileva et al. 2012). From the retailing perspective, the shift from mass-marketing campaigns with high scattering loss towards individualized customer communication and one-to-one marketing campaigns, that treat each customer uniquely according to its living situation and preferences, has been shown to create competitive advantage (Bose 2002; Xu and Walton 2005).

For large-scale implementations of such targeted campaigns, companies need a reliable base of household data. It is common practice that companies purchase such information from data providers and spend a considerable amount of their marketing budget for such information. Together with our partners from the utility industry, we asked address trading companies for enriching a customer core dataset from a utility company. The prices ranged from about € 0.10 per address for general statistical data (e.g., population density, buying power, mean household size) to a range of € 0.10 – € 0.30 for one or more micro-census criteria per address (e.g., estimated number of residents and their most frequent age in one household, the housing type and age, and the interests of the residents). Some data providers offer even more precise customer data, such as "well-off middle and upper class singles", but for higher costs per address. Considering that the cost of a personalized mailing lies at approximately € 1 to € 2, the cost component of such an address purchase amounts to 5-30%.

Alternatively, companies conduct customer surveys themselves, motivate customers to use their online self-service portal (and ask for information during registration and use), or introduce customer-loyalty programs to grasp data on their customer base. Typical problems here are data sparsity (only a small portion of customers return questionnaires or register on online platforms) and self-selection bias (e.g., elderly people are less often willed to participate in online surveys).

Predictive business analytics methods provide the ability to make use of company-owned data and provide in this way *business value,* as highlighted recently by several information systems (IS) scholars (Lycett 2013; Mithas et al. 2013; Sharma et al. 2014). In the same vein, several data analytics case studies (Habryn et al. 2012; Hopf et al. 2016; Jank and Shrivastava 2015; Sodenkamp et al. 2015) were able to show how companies can make use of their own data (e.g., customer address information and their buying history, and contents in their customer relationship management systems) and predict household characteristics from such company-owned data. However, there appears to be a gap between the abstractly stated "value of business analytics" and the concretely shown business advantages in single case studies.

Our work helps to bridge the gap between abstract and concrete value statements about predictive IS. *By means of a relevant case of the utility industry – the prediction of household characteristics based on publicly available and company-owned data – we answer two research questions described below* with the help of a comprehensive dataset on 7,504 customers from seven utility companies in Switzerland and Germany. Our paper follows Müller et al.'s (2016) guidelines for presenting data analytics results for IS research. After the presentation of our concrete research questions and a brief summary of related literature, we highlight the theoretical contribution that we seek to make with our study. Thereafter, we describe our predictive IS artifact together with the underlying data, provide the results, offer interpretations on the findings, and summarize our answers to the research questions. We include proper links to all online available datasets and open-source software used in this study. The replication of our analysis should be possible with the provided references and the research presented in this paper.

Research Questions

Our first aspect in this paper is to ***investigate the value of government statistical data as an addition to company-owned data for the use in customer data analytics*** by means of a specific but relevant case study. In the recent years, governments have put large effort in making government data available to the public – most notably, the U.S. and the U.K. administrations have started open data initiatives (Immonen et al. 2014). Also the EU decided to publish public sector data (EU 2003, 2013a) and started open data initiatives: In 2007, the INSPIRE project (EU 2007) was launched that aims to build a geographic and environmental data infrastructure with public data of its member countries and in 2013, the Copernicus project (EU 2013b) that has the goal to make data public that stems from the European Space Agency on Earth Monitoring.

Open government data (also known as "public sector information") is intended to be accessible by everyone, stored using open formats and under open licenses, and attracts multiple volunteer initiatives creating applications and adjacent data collection projects by the crowd. Thereby, the public sector, private companies, civil society and citizens are involved (Jetzek, Avital, and Bjørn-Andersen 2013). Examples for created artifacts are visualizations, apps and derived new insights to the existing data (Kuk and Davies 2011). The open government data must thereby be seen in the broader context of open data in general. According to the Open Knowledge Foundation (2017), open data is defined as data or "knowledge [that] is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness". However, the focus of this paper is to investigate open *government* data and its use for predictive analytics.

The total economic value of published open data in Europe is estimated to be between € 27 billion (Dekkers et al. 2006) and € 140 billion (Vickery 2011). These potential values are high, but the actual value for companies is questionable, since open data in itself creates only value when it is used further (Immonen et al. 2014; Janssen et al. 2012). The first research questions that we aim to answer in this study is therefore:

RQ 1: To what extent can open government statistical data improve the predictive power of machine learning IS artifacts in the specific yet important case of identifying household characteristics of utility customers?

The second aspect of our research is to ***test the transferability of predictive analytics*** by means of the present case from the utility industry. This aspect is important for companies operating multi-national and want to overcome the creation of predictive machine learning models for each regional division. Furthermore, models for business *and* academia must be general to a large extent, being able to universalize findings using business analytics and machine learning. Building such models is sophisticated (Shmueli and Koppius 2011), and multiple activities are associated with the data collection (i.e., through customer surveys), insight generation from the data (including data preparation and cleaning, model generation and tuning), transforming the insights to decisions and to create business value through correct decisions (Sharma et al. 2014). Especially the step of rising data is problematic: customer surveys suffer from low response rates, the nonresponse bias (Groves 2006) and fraudulent answers (Bohannon 2016; Kuriakose and Robbins 2015). Moreover, Müller et al. (2016) state that models in (big) data analytics are often built for one specific dataset and provide therefore "little theoretical contribution, but are also vulnerable to changes and anomalies in the underlying data". Therefore, the aspect of model transferability relevant, but only sparsely addressed by related studies.

Therefore, we aim to answer the following second research question:

RQ 2: To what extent can a predictive IS artifact that has been trained on data from customers located in one geographic region be used to classify customers located in another geographic region?

This question is further operationalized by the following additional two subordinate research questions:

RQ 2.1: How does the classification quality change, when a predictive model that was trained with customers from one utility company is applied to customers of other utility companies?

RQ 2.2: To what extent can open government statistical data improve the transferability of classification models between companies?

Related Works and Contribution to Theory

This section gives a brief summary of the related research regarding the prediction of household characteristics for energy customers based on data that is available to energy utilities (i.e., customer core data, energy consumption and other data sources), the research related to open government data and the geographical transferability of predictive models.

Several studies exist that *infer household characteristics* (e.g., size and type of the residency, number of persons in the household, heating type and age) from electricity consumption data that is collected by utilities for order processing and invoicing. This data stems either from conventional meters that are typically read out once a year, or from connected smart meters that send the readings to the utility in up to 15-min intervals. To the best of our knowledge, the first study on household classification was done by Beckel et al. (2013, 2014) who showed the principal feasibility to obtain household characteristics from 30-min smart meter data using predictive machine learning artifacts. In recent studies, we suggested several improvements to this work (Hopf et al. 2014) and adapted the approach to 15-min smart meter data and showed that even purchase interest in a product (solar installations) is possible (Sodenkamp et al. 2017). Since the dissemination of smart meters is currently stumbled in many countries and due to privacy regulations, energy retailers have not always access to fine-grained energy consumption data. Therefore, the development of such artifacts based on annual energy consumption data still remains a relevant task. In a recent work, we developed such a predictive artifact on the base of annual electricity consumption data in combination with free available geographic data (Hopf et al. 2016) and tested it with a dataset from a utility company in Switzerland with 3.986 customers. This artifact in an extended version is used to answer our research question in this paper.

Multiple studies from the IS field conceptualize the *open data* phenomenon in general and investigate its characteristics (Chatfield et al. 2015; Janssen et al. 2012; Jetzek, Avital, and Bøjrn-Andersen 2013; Kalampokis et al. 2011; Marton et al. 2013; Ponte 2015). Other studies focus on the diffusion and adoption of open data in organizations (Alanazi and Chatfield 2012; Gil-Garcia et al. 2007; Ham et al. 2015; Immonen et al. 2014; Maccani et al. 2015, 2017; Oliveira and Santos 2016). Besides that, multiple case studies have been published that describe the application of open data, for instance in recommender systems for products (Heitmann and Hayes 2010; Pham and Jung 2014), for sustainable mobility projects (Yadav et al. 2017), or the forecasting of electrical load (Vercamer et al. 2016). However, the only quantitative study about the value of open *government* data that we found is that of Jetzek, Avital, and Bøjrn-Andersen (2013), who investigated how open government data can stimulate value generation (in terms of social welfare) based on macro-economic data from 61 countries. The actual value of open government data for companies (e.g., by using the data in business analytics) remains, however, open and is therefore subject of the research presented in this paper.

Transferability of predictive models have been investigated in studies from the field of geography (e.g., Vanreusel et al. 2007; Wenger and Olden 2012), recreation demand (Loomis et al. 1995), forecasting of travel demand for transportation planning (Everett 2009; Sikder et al. 2013) or accident prediction (Sawalha and Sayed 2006). An evaluation of the geographic transferability of predictive models in customer data analytics was – to the best of our knowledge – not subject to existing research.

With this study, we ultimately aim to contribute three distinct features to the body of knowledge related to predictive analytics IS. First, this research empirically evaluates the quality and validity of an existing predictive IS artifact and helps thereby to lay the grounding for future development of design principles for such systems. Second, this research provides descriptive insights on predictive analytics that are already considered as an important toolset for IS research (Shmueli and Koppius 2011). Third, we see our research as a relevant step to develop a theory on the value of different data sources (here: open government data) in business analytics applications by using predictive IS artifacts.

Predictive Customer Data Analytics in the Utility Industry

The subject of our study is a predictive customer data analytics (machine learning) IS artifact that is used to identify household characteristics (e.g., size and type of the dwelling, number of residents, type of heating) from private customers for targeted marketing or energy efficiency campaigns. The artifact is schematically illustrated in Figure 1. The technical implementation of the artifact used in this work, stems from a previous study (Hopf et al. 2016) and was extended to answer the stated research questions.

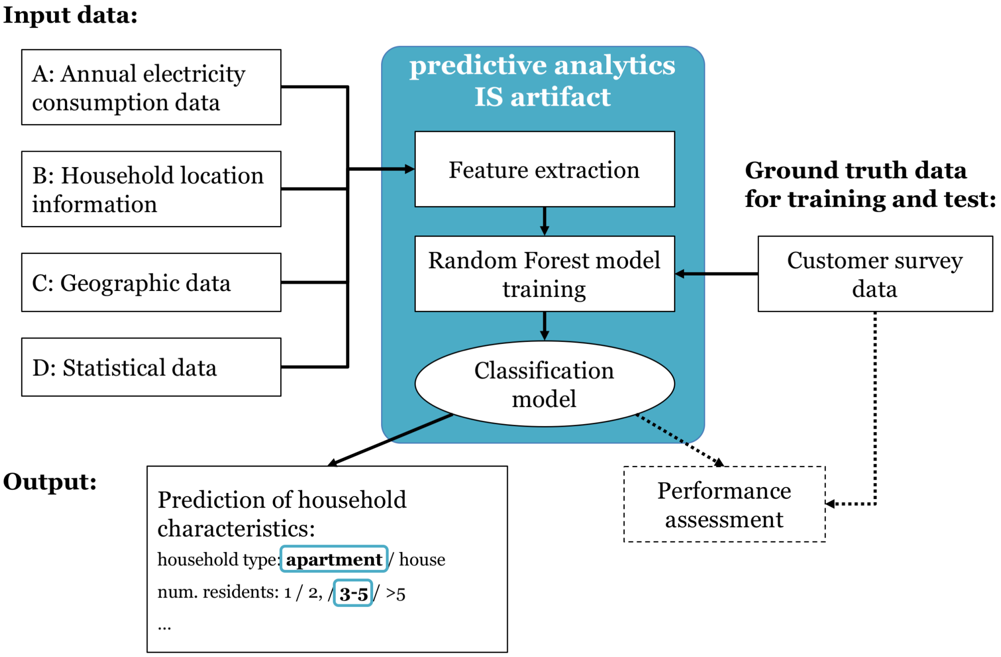
Input data:
A: Annual electricity consumption data
B: Household location information
C: Geographic data
D: Statistical data
predictive analytics IS artifact
Feature extraction
Random Forest model training
Classification model
Ground truth data for training and test:
Customer survey data
Performance assessment
Output:
Prediction of household characteristics:
household type: apartment / house
num. residents: 1 / 2, / 3-5 / >5
…

Figure 1: Predictive customer data analytics IS artifact that predicts household characteristics

As *input*, four data sources have been considered: A) Annual electricity consumption data and B) household location information (i.e., the address of the household); both items of information are known to utility companies due to order fulfillment and billing purpose. The geographic data (C) and statistical data (D) can be obtained by the company through public data sources. The resulting *outputs* of the IS artifact are predicted household classes: the type of the household (apartment or house), the number of residents (1, 2, 3-5, or more than 5), the type of heating system, etc. For *model building* (training) and performance evaluation (test), ground truth data is needed that is usually obtained using customer surveys.

For the implementation of the predictive artifact, we used the statistical programming environment R and the open-source packages “randomforest” (Liaw and Wiener 2002) for classification and “caret” (Wing et al. 2015) for sampling and cross-validation. Implementation details on the features from statistical data are provided later in this paper.

Dataset

For our analysis, a dataset on 7,504 residential customers from seven utility companies in Switzerland and Germany (see Table 1) was provided by our research partner BEN Energy AG, Zurich (Switzerland).

For each customer, the annual electricity consumption and pseudonymized address data is known from the companies' core database. Additionally, we have access to survey results on household characteristics (number of persons in the household, living area, heating type, etc.) that were obtained via an online customer portal.

Data subset name	CH	DE
Location	Switzerland	Germany
Number of customers with survey responses	5,446	2,058
Number of customers that are usable for analysis *Reasons for exclusions are:* *1) address-conversion to geo-coordinates not possible* *2) less than 100 consumption days or less than three years*	5,020	1,843
Number of utility companies	6	1
Timespan of the electricity consumption	2009 – 2014	2009 – 2016

Table 1: Characteristics of the dataset: regional distribution and excluded data

Dependent Variables: Household Properties

Household properties are the dependent variables in the considered prediction model. The information on household properties for all customers have been obtained through surveys on an online customer engagement portal that aims to motivate energy-efficient behavior (Graml et al. 2011). The portals are provided by our research partner as software-as-a-service solutions for utility companies and have been individually branded for each utility. Therefore, the survey data is comparable for all utility companies. For the property 'living area' and 'number of residents', customers can insert integer values. We split the classes for these properties according to previous research (Hopf et al. 2016). For the properties 'household type', 'space heating type' and 'water heating type', the class labels have been taken directly from the survey results. The household properties and classes are listed in Table 2, together with the absolute and relative distribution in both countries.

	Property	Class	Class sizes (absolute and relative)			
			CH		DE	
1	Living area	≤ 95 m²	1,059	21,10%	623	33,80%
		≤ 145 m²	1,852	36,91%	748	40,59%
		> 145 m²	2,107	41,99%	472	25,61%
2	Number of residents	1	612	12,26%	354	19,23%
		2	2,004	40,14%	653	35,47%
		3-5	2,251	45,09%	808	43,89%
		> 5	125	2,50%	26	1,41%
3	Household type	house	2,228	44,38%	863	46,88%
		apartment	2,792	55,62%	978	53,12%
4	Space heating type	electric	781	15,65%	39	2,12%
		not electric	4,210	84,35%	1,801	97,88%
5	Water heating type	electric	2,431	50,03%	187	10,49%
		not electric	2,428	49,97%	1,596	89,51%

Table 2: Household properties with class labels and the distribution in both countries

Explanatory Variables (Features)

The input data is reduced to a small set of expressive features. Since the focus of this work is the evaluation of the transferability of predictive models and the test of statistical data to improve the model performance, we use empirically defined features on electricity consumption and geographic information from previous research on similar problems (Hopf et al. 2016; Sodenkamp et al. 2015), where detailed motivation on these features is given.

Three features are used to describe *electricity consumption and household location data*:

- Logarithmic annual consumption, normalized by the days in which the energy was consumed
- Consumption trend as the relative change between the consumption of different years, obtained with a linear regression model that uses the normalized consumption of each year as dependent variable and the number of the respective year as the independent variable.
- Neighborhood comparison as the Z-score of the household's logarithmized normalized consumption deviation from its neighborhood (in the postal code region).

Using the address information of one household, we obtained 66 features from the *free geographic database OpenStreetMap*. The calculated values can be subsumed under four categories:

- Features on the topology: describing the structure of and relations between one household and spatial neighbors (e.g. lon./lat., frequency objects in the surroundings, distance to city center)
- Features on landmarks and points of interests: describe the meaning of an object within the spatial context it appears (frequency, distance, and other measures to sights, shops, cafes, etc.)
- Features about buildings (e.g. mean/variance of the surface area, the distance to buildings, and the type of buildings in the surrounding)
- Features about land use (land use type embracing the household, area distribution in different land use types, etc.)

Random Forest Classification Model

The core of the studied predictive IS artifact is a Random Forest (Breiman 2001) classifier. The working principle of this supervised learning algorithm is that it generates several (in our instance 500) uncorrelated decision trees based on the training data and forms a predictive model based on the best decision trees.

In fact, there are hundreds of supervised machine learning algorithms that could also be used for the problem, but the Random Forest algorithm showed good performance in previous studies from the energy data analytics field (Hopf et al. 2016; Sodenkamp et al. 2017). Besides that, in a comprehensive study, Fernández-Delgato et al. (2014) tested 179 classifiers from 17 algorithm families and come to the conclusion that random forest or its variants are most likely to be the bests currently known classifier.

One can argue that such 'black box' classification algorithms are largely incomprehensible (Martens and Provost 2014) and so, the interpretation of the functional principle cannot fully be done (Müller et al. 2016), but our goal in this paper is to draw conclusions on the value of statistical data for exactly such classification models and the regional transferability of such algorithms. From a theoretical point of view, it seems to be better, to use more explanatory mathematical models (such as logistic regression), but it is questionable whether the results would also be valid for state of the art classification algorithms. Furthermore, we explicitly do not test multiple classification algorithms or tune the classifier's parameters, since we do not aim to find the "best model that works for our case", but we want to derive statements on transferability and the value of statistical data.

Performance Assessment Using Matthews Correlation Coefficient (MCC)

We assess the classification performance by comparing the predictions made by the classification model with the ground truth data and calculate the classification performance. There are multiple classification performance metrics, which all have advantages and drawbacks. The widely used *accuracy* (measuring the percentage of correct classified examples), for example, is easy to interpret, but it is known that accuracy is influenced by the relative class sizes (Baldi et al. 2000) and therefore unsuitable for our case. Other measures, such as *precision* and *recall*, or measures associated with the *Receiver Operating Characteristic (ROC)* curve (Fawcett 2006) are only applicable for binary classification problems. Therefore, we opt to assess the classification quality using *Matthews Correlation Coefficient (MCC)*. This correlation coefficient quantifies the association between the observed X and predicted Y class memberships of examples. In the case of binary classification problem, it is equal with the phi statistic (Cramer 1946). We use MCC definition for multiclass problems (Gorodkin 2004; Jurman et al. 2012) relying on the covariance (cov):

$$MCC = \begin{cases} \sqrt{\phi^2} & for\ two\ class\ problems \\ \dfrac{cov(X,Y)}{\sqrt{cov(X,X) * cov(Y,Y)}} & for\ n\ class\ problems \end{cases}$$

MCC can take values between -1 and 1, where 0 represents random classification, 1 indicates the ideal classification, and -1 is the total disagreement between the predictions and real observations.

For a methodologically correct estimation of the model performance (Hastie et al. 2009, Chap. 7; Shmueli and Koppius 2011), we use 10-fold cross-validation, whenever possible, to calculate the MCC results and avoid selection bias in splitting the data randomly into datasets for training and test. With this approach, the data is split into 10 parts of equal size (folds) and classification MCC estimation is replicated 10 times, where each time another fold is used as test data.

Value of Government Statistical Data for the Use in Predictive Customer Data Analytics

In this part of our work, we evaluate government statistical data for its use in predictive customer data analytics. First, we give an overview to government statistical in Germany, Switzerland and the EU, identify statistics that are household-related, define features from the statistical data, and present an approach to connect the statistical features with customer data in organizations. Second, we assess the overall value of the connected statistical data using our predictive analytics IS artifact and give the answer to RQ 1.

Government Statistical Datasets for Predictive Customer Analytics

For the identification of government statistical datasets in our study area covering Germany and Switzerland, we thoroughly studied the web-catalogs of all existing statistical offices. Each officially published statistical dataset contains figures for one specific *statistical geographic region (SGR)* that is identified by a unique number. In *Germany*, federal and state statistical offices exist. The Federal Statistical Office (FSO) offers data for SGRs on the level of German states, which is too coarse for business analytics on individual households. Besides that, statistical data is offered by the joint database of German state statistical offices[1] that contains data for SGRs on the level of down to single municipalities. In *Switzerland*, statistical data is provided by the FSO of Switzerland[2]. The statistical office of the *EU*[3], Eurostat, aggregates data from the EU administration, it's member states and associated states, such as Switzerland. Most of the data has a national level of detail, but some datasets also have the level of smaller SGRs.

All data catalogs are structured differently, so that an easy mapping of statistical datasets was not possible. According to our goal to *find datasets that are usable for predictive analytics in the utility industry*, we defined four groups of statistics and categorized the datasets as follows:

Category A: Statistics we identified to be helpful for predictive customer data analytics:

- Housing, number of rooms, usage of real estate
- Population statistics including the population density, age, sex, migration
- Number of businesses (small, medium, large) and economic development (GDP in the region)

[1] http://www.statistikportal.de/Statistik-Portal/en/, last access 28.11.2016

[2] https://www.bfs.admin.ch/bfs/en/home.html, last access 28.11.2016

[3] http://ec.europa.eu/eurostat/home, last access 28.11.2016

Category B: Statistics that are relevant, but have a low geographical granularity:

- Tourism (e.g., arrivals and nights spent at tourist accommodation)
- Health (e.g., health personnel and beds, discharges and hospital days of in-patients)
- Education (e.g., pupils and students by enrolment)
- Research and development (e.g., R&D expenditures, employment in high-tech sector)

Category C: Statistics containing information not immediately related to household customers

- Territory (e.g., sizes of territories and number of municipalities by county)
- Civil register (e.g., marriages, births and deaths in an area, adoption)
- Agricultural economics (e.g., plantings, agricultural businesses, forestry, wood harvest)
- Economy and markets (e.g., bankruptcy, paid insurance output, money growth)
- Employment (e.g., employees in the public sector)
- Energy (e.g., national energy import and export, coal plants and nuclear plants)
- Traffic (e.g., accidents, agricultural vehicles, cars and trucks)

Category D: Statistics that are not available in all countries or the content of the statistics is different in the respective countries so that the data is not comparable

- Politics (e.g., elections and referendums)
- Ethics and religion
- State finances and taxes

Only the statistics in Category A are applicable for predictive analytics in the utility industry. We list them in Table 4, highlight the data source and define three types of features from the statistical datasets: relative frequencies (rf.*), absolute frequencies (num.*) and averages (mean.*). If the data was given in several intervals (e.g., age of houses), the mean was calculated as the weighted average using the mean of each interval.

The data available for both countries is sparse, since most datasets are available either for Germany or for Switzerland. In fact, only the features *rf.male, num.residents, mean.residentialAge, num.migration, mean.HouseAge, rf.newHouses* and *mean.NumRooms* are available for both countries.

The final step that needs to be taken to make government statistical data usable for data analytics in companies, is to *connect the statistical data to the customer data*. The statistical data is associated to specific SGRs and the definition of the SGRs differs for each statistical office, but mostly the political borders of municipalities, districts or states are used.

One way to integrate statistical data into the customer database is to simply use tables that are prepared and published by the statistical offices containing a matching criterion for customer data (e.g., the city name, postal code, or the geo-reference of one point in this SGR). We identified three major problems of this matching approach: 1) using the postal code or the city name as matching criterion, customers cannot be uniquely assigned to the SGR, since the spatial extend of postal code regions or cities do not fully overlap with the SGR borders defined by statistical offices; 2) spelling errors in city names cause serious problems; 3) when a specific geographical point for each SGR is provided by the statistical office (e.g., geo-coordinates of the administrative center), the household cannot be assigned to the SGR by using shortest-distance to this point, since the location of the given point is arbitrary and must not be the geometric center of the statistical unit.

Category	EU	DE	CH	Description	Feature Name
building-statistics	X[a]		X	Frequency of single family homes	rf.singleFamilyHome
	X[a]		X	Frequency of multiple family homes	rf.multipleFamilyHome
			X	Frequency of residential homes with ancillary use	rf.residentialHome WithAncillaryUse
			X	Frequency of houses with partial residential use	rf.housePartlyResidential
		X	X	Average house age	mean.HouseAge
		X	X	Share of new houses	rf.NewHouses
			X	Number of residents which hold a share or are owner of the building they live in in the region	rf.homeOwners
		X	X	Average number of rooms per apartment	mean.NumRooms
	X[a]			Amount of buildings with one apartment	rf.oneDwellingBuildings
	X[a]			Amount of buildings with two apartments	rf.twoDwellingBuildings
	X[a]			Amount of buildings with three or more apartments	rf.threeOrMoreDwelling-Buildings
	X[a]			Amount of buildings that are not used for residential purposes	rf.nonResidential-Buildings
socio-demographics	X	X	X	Amount of male population in %	rf.male
			X	Amount of permanent residents	rf.permResidentials
	X	X	X	Average age of permanent residents	mean.ResidentialAge
		X	X	Quantity of permanent residents	num.permResidents
	X	X	X	Difference of immigration and emigration	num.migration
	X	X	X	Quantity of total residents (permanent and non-permanent)	num.residents
economic	X[b]	X		Amount of businesses that have 0 employees	rf.zeroEmployee-Businesses
	X[b]	X	X	Amount of businesses with 0-9 employees	rf.smallBusinesses
		X	X	Amount of businesses with 10-250 employees	rf.mediumBusinesses
		X	X	Amount of businesses with 250 and more employees	rf.bigBusinesses
	X[b]	X		Quantity of GDP in Euro per citizen	num.gdpEuroPerCitizen
			X	Relative frequency of investments in buildings by the public sector	rf.publicInvest
			X	Relative frequency of investments in new buildings instead of renovation	rf.newInvest

Table 4: Statistics that have been identified as meaningful for predictive customer analytics (a = no or incomplete data for Germany, b = no or incomplete data for Switzerland)

Therefore, we developed a *methodology for geo-referenced data integration of customer core data and statistical datasets*, that we illustrate in Figure 2. The first step in this approach is the conversion of customer addresses into geo-coordinates. This can be done involving an online service provider. In the second step, the SGR for each customer location is determined. For that, we loaded the digital maps published by respective governmental institutions in standardized formats (i.e., Swiss federal statistical office[4], German federal agency for cartography and geodesy[5] and Eurostat[6]) and obtained the identifier for

[4] https://www.bfs.admin.ch/bfs/de/home/dienstleistungen/geostat.html, last access 28.11.2016

[5] http://www.geodatenzentrum.de/geodaten/gdz_rahmen.gdz_div, last access 28.11.2016

[6] http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units, last access 28.11.2016

that region encompassing the customer location. Technically, the SGRs are represented as polygons in the digital maps that can be used together with open source packages for geographical positioning in R. In detail, we used the packages *"rgdal" (Bivand, Keitt, et al. 2017)*, *"sp" (Pebesma et al. 2017)*, *"rgeos" (Bivand, Rundel, et al. 2017)* and *"photon" (Giraud and Viry 2015)* for the geographical computations. Finally, the obtained geographic identifier can be used to match the available statistical data.

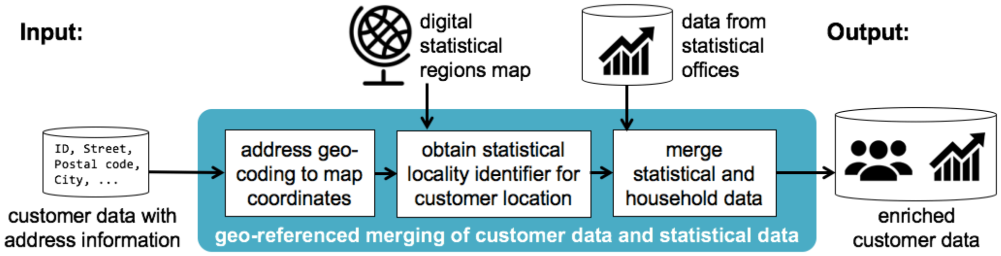
Input:
digital statistical regions map
data from statistical offices
Output:
ID, Street, Postal code, City, ...
customer data with address information
address geo-coding to map coordinates
obtain statistical locality identifier for customer location
merge statistical and household data
geo-referenced merging of customer data and statistical data
enriched customer data

Figure 2: Methodology for geo-referenced data integration of customer data and statistical datasets

Classification Performance Improvement Through Statistical Data

To evaluate the performance of statistical data, we operationalized the predictive IS artefact described in the previous Section, and show the results in Figure 3 Different feature sets have been used: 1) we used solely consumption features as the base case as red bars, 2) consumption and geographic features as yellow bars, 3) consumption and all available statistical features as dark blue bars, and 4) the consumption and geographic features together with different kind of statistical features (all, only features from Swiss or German statistical offices, and only features from Eurostat) highlighted in different shades of green. The MCC results in Figure 3 are obtained in 10-fold cross-validation. We included also the standard deviations of MCC values depicted as 'error-bars' in the graphs.

The classification of all properties is – except the heating types in Germany – better than random, since MCC is larger than 0. The result for the low performance in German heating type prediction is the uneven distribution in the class sizes (2% 'electric' vs. 98% 'not electric' for 'space heating type', and 10% 'electric' vs 90% 'not electric' in 'water heating type'). Further model tuning and data preparation techniques such as oversampling can be used to improve the recognition of such household properties, but this is out of the scope in this work.

From the results, we respond to our research question RQ 1 (*To what extent can open government statistical data improve the predictive power of machine learning IS artifacts in the specific yet important case of identifying household characteristics of utility customers?*):

Electricity consumption and statistical data: In Switzerland, statistical data improves the classification based on solely consumption data (red vs. blue bars in Figure 3a) slightly, but statistically significant (paired t-test, $t = -9.1478$, $df = 4$, p-value < 0.001).

For *Germany, there is no statistically significant improvement of statistical data* (paired t-test, $t = -0.8224$, $df = 4$, p-value $= 0.2285$). One reason for the lower effect in Germany might be that the size of SGRs in Germany is with 33.32 km^2 on average 95% larger than in Switzerland, where the average area of one SGR is 17.12 km^2 (this difference is statistically significant with p-value < 0.0001 in a t-test with $t = 11.405$).

Electricity consumption, geographic information, and statistical data: By including also geographic information in the models, the added value from open statistical data is *lowered for some properties and disappeared completely for others* (orange vs. green bars).

We assume that the geographic data describes the regional differences between the households quite well and the level of detail in the statistical data is too low to add further information to the prediction model.

Different sources of statistical data: From the results, we can see that using all statistical features (from national and EU agencies) achieved the highest classification performance, but using only one source of statistical features – either national or from EU – does only lead to a small, but significant loss in classification performance (paired t-tests, t > 2.3, df = 4, p-value < 0.05).

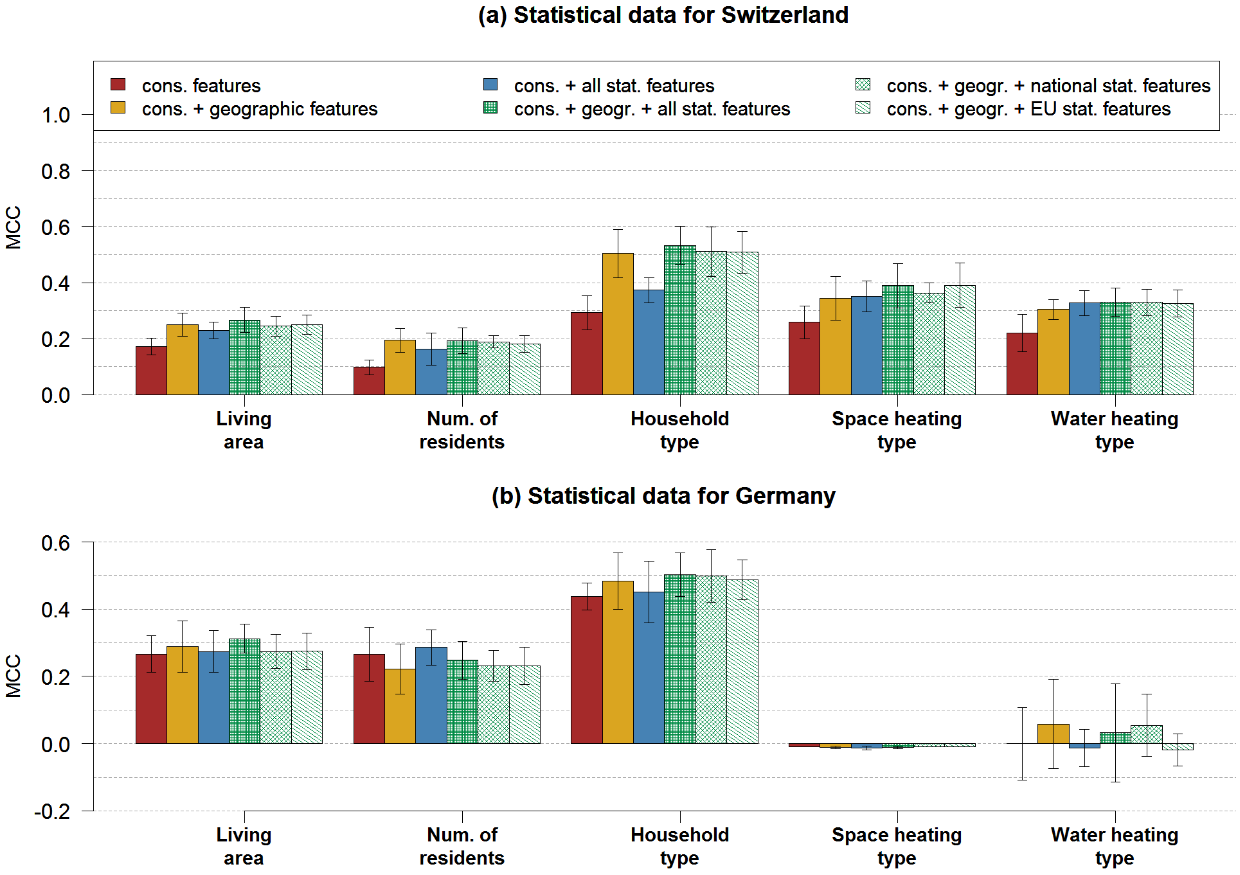

Figure 3: MCC results for classification with statistical data in Switzerland (a) and Germany(b); the statistical data can improve the classification performance slightly in Switzerland, but not in Germany

Transferability of Machine Learning IS Artifacts

In the second part of the paper, we test the geographical transferability of prediction models between the two countries Germany and Switzerland. For that, we operationalized our IS artifact again, and tested the impact of additional customer data from other utility companies such as statistical data to improve the classification performance for transferred models. In this analysis, we used all available features for both regions. The available statistical features for Germany and Switzerland can be seen in Table 4.

It might appear questionable at first glance to selected two countries in central Europe with mostly German-speaking customers to study transferability of predictive models. But indeed, both countries differ to a large extend among some major differences. The main distinction between the countries regarding to our study is the energy market that has been liberalized in Germany 1998 and is planned to be liberalized in Switzerland not earlier than 2018 (Swiss Federal Council 2014), but for our study time-frame (2009-2014), the Swiss market has regional monopoles. A second difference is the higher prosperity in Switzerland that can be seen in the per capita income 2016 that was USD 48.730 in Germany and USD 62.882 in Switzerland (World Bank 2017). Apart from that, the energy price deviates in the other direction: € 0.29/kWh in Germany and € 0.18/kWh in Switzerland that results mainly due to regulatory differences in taxation. Therefore, we argue that findings presented in this paper cannot provide evidence for the international transferability of predictive models, but our analysis allows a good first estimate to what extent predictive models can be applied in different countries. To the best of our knowledge, our study is the first investigation of predictive analytics artifacts transferability with a considerable dataset.

Transferability Between Different Geographic Regions

We tested the geographic transferability in four setups, as illustrated in Figure 4. The MCC for training and test with data from the same country was obtained using 10-fold cross-validation. For the training with data from one country and the classification of households in the other country, the entire data set from one country was used for training and all available data from the other country was used for testing.

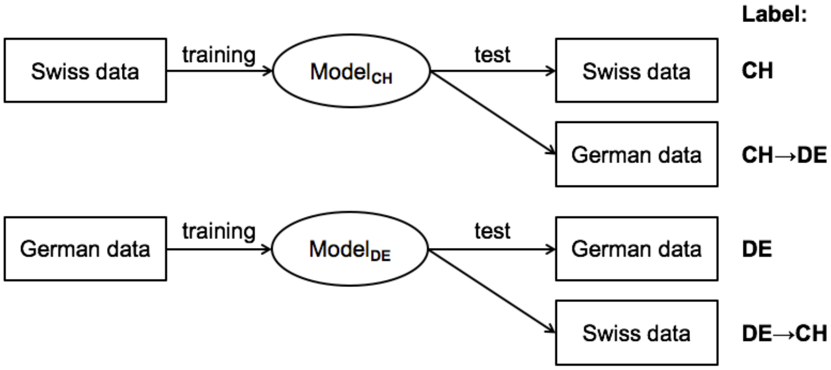

Figure 4: Illustration of the four considered cases to test the geographic transferability of machine learning models

The results for each household property are given in Figure 5. We can answer our RQ 2 (*Can a predictive IS artifact that has been trained on data from customers located in one geographic region be used to classify customers located in another geographic region?*) positively: classification models that are trained with customer data from one geographic region can be applied to customers located in another geographic region, yet with a lower classification quality.

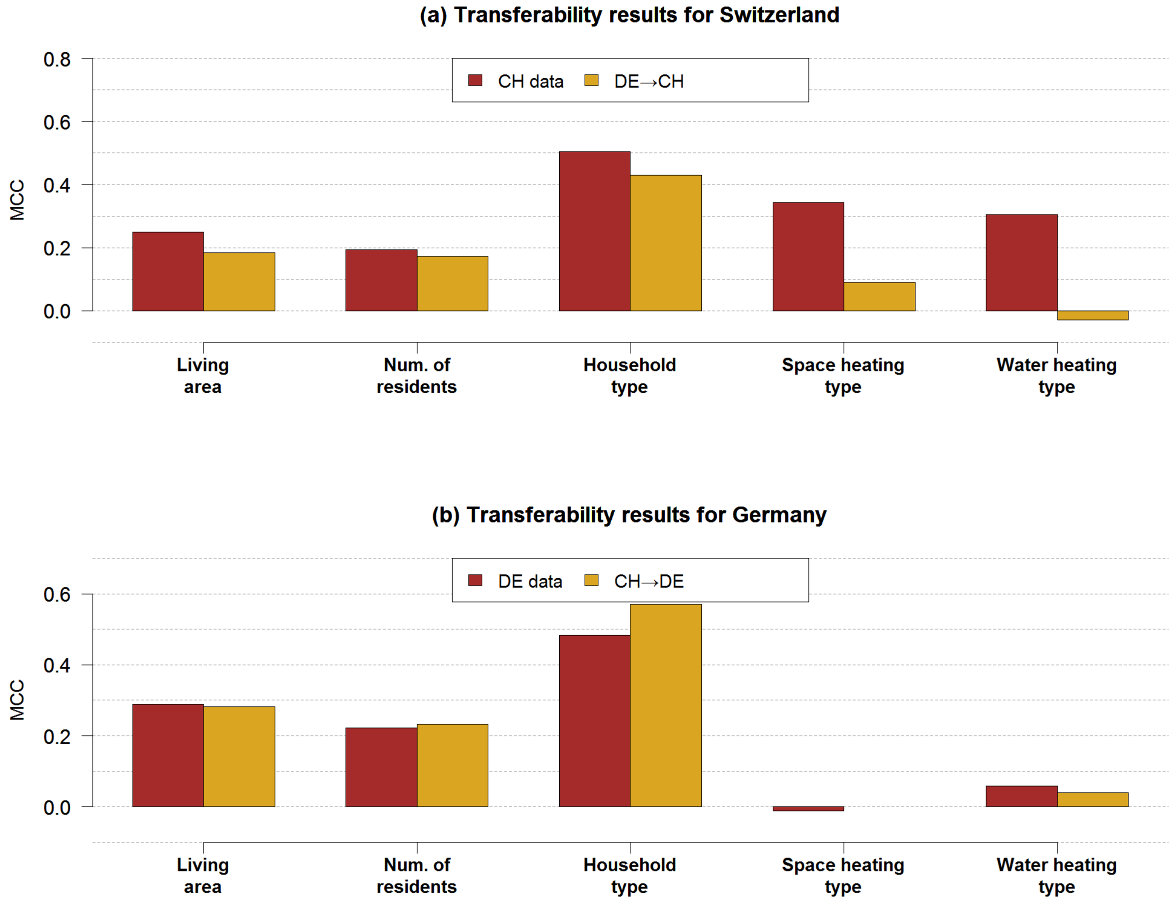

Figure 5: MCC results for classification of households based on different datasets; reading example: 'CH': training and test with Swiss data, 'DE→CH': training with German data and classification for Swiss data

Value of Additional Customer Data from Other Utilities

To answer RQ 2.1, we tested whether the model trained with data from one utility company can be applied to customers from another utility company. A distribution of the data points per utility is shown in Table 5. For this analysis, we considered the data from company U3 (having the largest database for model training). This model was then used to predict household classes for customers of U1, U4, U5 and U6 and the performance were assessed with MCC. The results are shown in Figure 6. We do not use the German utility U7 for this analysis, since the transferability of models between countries was subject to our previous RQ 2 and the lower number of common statistical features available for Germany and Switzerland would influence the results. As a benchmark, we consider the average classification results for Switzerland (as calculated for answering RQ 1).

Utility	U1	U2	U3	U4	U5	U6	U7
Country	CH						DE
Num. of customers	975	26	3,593	130	106	190	1,843

Table 5: Distribution of customers in our dataset to the seven different utility companies

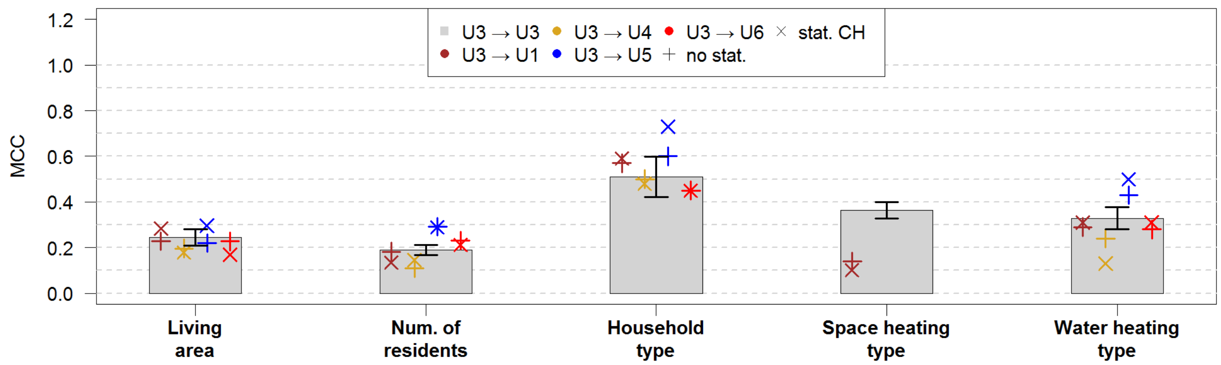

Figure 6: MCC results for classification of households based on internal and additional customer data, the benchmark (training and classification with the same company data) is illustrated with grey bars and standard deviations obtained with cross-validation

From the results, we can conclude that – in our exemplary scenario – it is possible to apply the model of one utility company to other and there is no significant deviation in MCC classification performance over all household properties and for the considered utility company datasets (paired t-test, t = 0.84726, df = 4, p-value > 0.4). We can therefore respond to our RQ 2.1 (*How does the classification quality change, when a predictive model that was trained with customers from one utility company is applied to customers of other utility companies?*), that the classification performance is neither negatively nor positively influenced on average, but the variance in classification performance is higher than the expected variance using models for classification that were trained using data stemming from the same utility company. The transferability of models between the considered utilities is therefore possible.

Impact of Statistical Data for Model Transferability

Finally, we analyzed whether the statistical data influences the performance of transferred classification models between different companies. For that, we compared the average performances of classification results per household property with or without the use of statistical data (in this case all statistical data available for Switzerland was used). According to the depicted results in Figure 6, we find no significant improvement or decrease of classification performance by adding statistical features in the classification (two-sided paired t-test, t = -0.0095453, df = 4, p-value > 0.99). Our research question RQ 2.2 *(To what extend can open government statistical data improve the transferability of classification models between companies?)* must therefore be answered negatively.

Summary and Implications

In this paper, we used a machine learning IS artifact that identifies household characteristics from company-owned and public available data – which is a specific, but relevant case from the utility industry – to answer two important questions for the field of predictive analytics: First, we tested whether open governmental statistics can help to considerably increase the predictive quality of a machine learning tool and second, to what extent the tool that was trained on one geographic region can be transferred to another geographic region. With that work, we empirically evaluated the quality and validity of a previously developed predictive analytics IS artifact (Hopf et al. 2016), lay the base for future development of design principles for predictive analytics IS artifacts and the development of a methodology to assess the value of different data sources in an application-based manner.

To *investigate the value of open government statistical data for the customer data analytics in organizations,* we reviewed the available official government statistics published from Germany, Switzerland and the EU, identified statistics that are household-related, defined features from the statistical data, and showed how the open government data can be connected to the customer core data in organizations. From this work, we conclude that a large portion of the statistical data is hardly usable by organizations in our investigated case. The handful of features that are useable in our case brought only notable performance improvements in Switzerland, not in Germany. As reasons for that, we identified the low geographical granularity (data is only available on the level of municipalities or above) and the low number of statistical datasets that are available for different countries, even in the EU. Studies on open data in Europe estimate its value to be € 27 – € 140 billion (Dekkers et al. 2006; Vickery 2011). Our impression is, that it will be challenging, to realize parts of this value propositions as profit in organizations using open data in customer-related data analytics, since the contribution of open data for predictive analytics in the presented case (which is a veritably well suited case for the utilization of company-external data), does not reflect the stated value. To convert open government data into actual business value, the published datasets need to be more detailed: regarding to the geographical level of detail and with respect to the variety of variables in different topics. However, the data within utility companies itself provides already insightful information that helps to perform targeted energy efficiency or marketing campaigns and can be enriched with free available geographic data. For example, the number of residents in a household (based on the Switzerland data) can be recognized correctly in 64% of the single-households, 48% of the two-person-households and 56% of households with 3-5 persons. By randomly distributing the customers, one would achieve 12%, 40%, 45% correctness respectively (considering the relative frequencies of household members in Table 2).

The second part of our paper was dedicated to the que*stion* of *whether the considered predictive analytics models are geographically transferrable*, which means that they can be trained using customer data from one region and can be applied to customer data in another region. Our answer to this question – based on data from two countries in central Europe that have a different energy market – is clearly positive: a statistical model in the case of the utility industry can be learned with data from one region and can be applied to data from another region. The statistical data could not improve the results in this case. We finally tested whether data from another company can improve the predictive model of the own company and found that, in the considered case, no performance improvement was possible.

Nevertheless, we see this study as a starting point for further research, especially regarding the transferability of business analytics models to countries from much different cultural regions. This aspect could just be investigated considering Germany and Switzerland as study areas in our work so far, since we had no access to utility data outside these regions. The validity check of our results in further countries is therefore desirable and should be subject to future research. Besides that, future works must identify potential influence factors (economic, social, environment, etc.) that lower statistical model transferability. Additional research might also consider data from data providers in comparison with results from predictive customer data analytics as presented in this study. This would give a further baseline to assess the value of predictive customer data analytics artifacts.

As a bottom line, utility companies may be better off than others when they start to use predictive customer data analytics and transfer existing knowledge (such as their family status, household characteristics, etc.) on a limited number of customers to all customers. Therefore, they can enrich their customer data with knowledge on a small portion of customers that they already hold, instead of purchasing customer characteristics from data provider for relatively high prices with sometimes unknown data quality. The integration of government statistical data can improve models to only a small extend, since the available data is too sparse. In addition, even companies that operate on a large geographic (i.e., cross-border) sales area can use their predictive IS artifacts for all customers and are not limited to geographic regions.

Acknowledgements

We thank our research partner BEN Energy AG, Switzerland, who provided data for this study and gave feedback on earlier versions of this manuscript. The research presented in this paper was financially supported by Eureka member countries and European Union (EUROSTARS Grant number E!9859 - BENgine II).

References

Alanazi, J., and Chatfield, A. 2012. "Sharing Government-Owned Data with the Public: A Cross-Country Analysis of Open Data Practice in the Middle East," in *AMCIS 2012 Proceedings*, AIS electronic library, July 29. (http://aisel.aisnet.org/amcis2012/proceedings/EGovernment/16).

Ayres, I., Raseman, S., and Shih, A. 2013. "Evidence from Two Large Field Experiments That Peer Comparison Feedback Can Reduce Residential Energy Usage," *Journal of Law, Economics, and Organization* (29:5), pp. 992–1022.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. 2000. "Assessing the Accuracy of Prediction Algorithms for Classification: An Overview," *Bioinformatics* (16:5), pp. 412–424. (https://doi.org/10.1093/bioinformatics/16.5.412).

Beckel, C., Sadamori, L., and Santini, S. 2013. "Automatic Socio-Economic Classification of Households Using Electricity Consumption Data," in *Proceedings of the Fourth International Conference on Future Energy Systems*, D. Culler and C. Rosenberg (eds.), Berkeley and California and USA: ACM, pp. 75–86.

Beckel, C., Sadamori, L., Staake, T., and Santini, S. 2014. "Revealing Household Characteristics from Smart Meter Data," *Energy* (78), pp. 397–410.

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., Hijmans, R., and Rouault, E. 2017. *Rgdal: Bindings for the Geospatial Data Abstraction Library*. (https://cran.r-project.org/web/packages/rgdal/index.html).

Bivand, R., Rundel, C., Pebesma, E., Stuetz, R., and Hufthammer, K. O. 2017. *Rgeos: Interface to Geometry Engine - Open Source (GEOS)*. (https://cran.r-project.org/web/packages/rgeos/index.html).

BMUB. 2014. "Aktionsprogramm Klimaschutz 2020," Kabinettsbeschluss, Kabinettsbeschluss, Berlin, Germany: Bundesministerium für Umwelt, Naturschutz, Bau und Reaktorsicherheit, March 12. (http://www.bmub.bund.de/service/publikationen/downloads/details/artikel/aktionsprogramm-klimaschutz-2020/).

BNetzA. 2016. "Monitoring Report 2016," Bonn, Germany: Federal Network Agency and Federal Cartel Office, November 30. (https://www.bundesnetzagentur.de/EN/Areas/Energy/Companies/DataCollection_Monitoring/MonitoringBenchmarkReport2016/Monitoring_Benchmark_Report_2016_node.html), in German.

Bohannon, J. 2016. "Many Surveys, about One in Five, May Contain Fraudulent Data," *Science*, February 24. (https://doi.org/10.1126/science.aaf4104).

Bose, R. 2002. "Customer Relationship Management: Key Components for IT Success," *Industrial Management & Data Systems* (102:2), pp. 89–97. (https://doi.org/10.1108/02635570210419636).

Breiman, L. 2001. "Random Forests," *Machine Learning* (45:1), pp. 5–32.

Chatfield, A., Reddick, C., and Al-Zubaidi, W. 2015. "Capability Challenges in Transforming Government through Open and Big Data: Tales of Two Cities," in *ICIS 2015 Proceedings*, Fort Worth, USA: AIS electronic library, December 13. (http://aisel.aisnet.org/icis2015/proceedings/eBizeGov/20).

Cramer, H. 1946. *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

Dekkers, M., Polman, F., te Velde, R., and de Vries, M. 2006. "Measuring European Public Sector Information Resources: Final Report of Study on Exploitation of Public Sector Information – Benchmarking of EU Framework Conditions," European Commission, June. (http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=1198).

Eto, J., Stoft, S., and Belden, T. 1997. "The Theory and Practice of Decoupling Utility Revenues from Sales," *Utilities Policy* (6:1), pp. 43–55.

EU. 2003. *Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the Re-Use of Public Sector Information*.

EU. 2007. *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*.

EU. 2013a. *Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 Amending Directive 2003/98/EC on the Re-Use of Public Sector Information Text with EEA Relevance*.

EU. 2013b. *Commission Delegated Regulation (EU) No 1159/2013 of 12 July 2013 Supplementing Regulation (EU) No 911/2010 of the European Parliament and of the Council on the European Earth Monitoring Programme (GMES) by Establishing Registration and Licensing Conditions for GMES Users and Defining Criteria for Restricting Access to GMES Dedicated Data and GMES Service Information Text with EEA Relevance*.

Everett, J. D. 2009. "An Investigation of the Transferability of Trip Generation Models and the Utilization of a Spatial Context Variable," Dissertation, Dissertation, Knoxville: University of Tennessee. (http://trace.tennessee.edu/utk_graddiss/47).

Fawcett, T. 2006. "An Introduction to ROC Analysis," *Pattern Recognition Letters* (27:8), pp. 861–874.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. 2014. "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?," *The Journal of Machine Learning Research* (15:1), pp. 3133–3181.

Fischer, C. 2008. "Feedback on Household Electricity Consumption: A Tool for Saving Energy?," *Energy Efficiency* (1:1), pp. 79–104.

Gil-Garcia, J. R., Chengalur-Smith, I., and Duchessi, P. 2007. "Collaborative E-Government: Impediments and Benefits of Information-Sharing Projects in the Public Sector," *European Journal of Information Systems* (16:2), pp. 121–133. (https://doi.org/10.1057/palgrave.ejis.3000673).

Giraud, T., and Viry, M. 2015. *R Interface to the Photon API (Version 1.0)*. (https://github.com/rCarto/photon).

Gorodkin, J. 2004. "Comparing Two K-Category Assignments by a K-Category Correlation Coefficient," *Computational Biology and Chemistry* (28:5), pp. 367–374.

Graml, T., Loock, C.-M., Baeriswyl, M., and Staake, T. 2011. "Improving Residential Energy Consumption at Large Using Persuasive Systems," in *ECIS 2011 Proceedings*, Helsinki, Finland: AIS electronic library, June.

Groves, R. M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys," *The Public Opinion Quarterly* (70:5), pp. 646–675.

Habryn, F., Bischhoffshause, J. K. von, and Satzger, G. 2012. "A Business Intelligence Solution for Assessing Customer Interaction, Cross-Selling, and Customization in a Customer Intimacy Context," in *ECIS 2012 Proceedings*, AIS electronic library, May 15. (http://aisel.aisnet.org/ecis2012/206).

Ham, J., Lee, J.-N., Kim, D., and Choi, B. 2015. "Open Innovation Maturity Model for the Government: An Open System Perspective," in *ICIS 2015 Proceedings*, Fort Worth, USA: AIS electronic library, December 13. (http://aisel.aisnet.org/icis2015/proceedings/eBizeGov/15).

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning*, Springer Series in Statistics, New York, NY: Springer New York. (http://link.springer.com/10.1007/978-0-387-84858-7).

Heitmann, B., and Hayes, C. 2010. "Using Linked Data to Build Open, Collaborative Recommender Systems.," in *AAAI Spring Symposium Series*, pp. 76–81. (https://www.aaai.org/ocs/index.php/SSS/SSS10/paper/view/1067).

Hopf, K., Sodenkamp, M., and Kozlovskiy, I. 2016. "Energy Data Analytics for Improved Residential Service Quality and Energy Efficiency," in *ECIS 2016 Proceedings*, Istanbul, Turkey: AIS electronic library, June.

Hopf, K., Sodenkamp, M., Kozlovskiy, I., and Staake, T. 2014. "Feature Extraction and Filtering for Household Classification Based on Smart Electricity Meter Data," in *Computer Science-Research and Development* (Vol. (31) 3), Zürich: Springer Berlin Heidelberg, pp. 141–148. (https://doi.org/10.1007/s00450-014-0294-4).

Immonen, A., Palviainen, M., and Ovaska, E. 2014. "Towards Open Data Based Business: Survey on Usage of Open Data in Digital Services," *International Journal of Research in Business and Technology* (4:1), pp. 286–295.

Jank, W., and Shrivastava, U. 2015. "A Data Driven Framework for Early Prediction of Customer Response to Promotions," in *AMCIS 2015 Proceedings*, Puerto Rico: AIS electronic library, June 26. (http://aisel.aisnet.org/amcis2015/BizAnalytics/GeneralPresentations/18).

Janssen, M., Charalabidis, Y., and Zuiderwijk, A. 2012. "Benefits, Adoption Barriers and Myths of Open Data and Open Government.," *Information Systems Management* (29:4), pp. 258–268.

Jetzek, T., Avital, M., and Bjørn-Andersen, N. 2013. "Generating Value from Open Government Data," in *ICIS 2013 Proceedings*, AIS electronic library, December 16. (http://aisel.aisnet.org/icis2013/proceedings/GeneralISTopics/5).

Jetzek, T., Avital, M., and Bøjrn-Andersen, N. 2013. "The Generative Mechanisms Of Open Government Data," in *ECIS 2013 Completed Research*, AIS electronic library, July 1. (http://aisel.aisnet.org/ecis2013_cr/156).

Jurman, G., Riccadonna, S., and Furlanello, C. 2012. "A Comparison of MCC and CEN Error Measures in Multi-Class Prediction," *PLoS ONE* (7:8). (https://doi.org/10.1371/journal.pone.0041882).

Kalampokis, E., Tambouris, E., and Tarabanis, K. 2011. "A Classification Scheme for Open Government Data: Towards Linking Decentralised Data," *International Journal of Web Engineering and Technology* (6:3), pp. 266–285.

Kuk, G., and Davies, T. 2011. "The Roles of Agency and Artifacts in Assembling Open Data Complementarities," in *ICIS 2011 Proceedings*, Shanghai, China: AIS electronic library.

Kuriakose, N., and Robbins, M. 2015. "Don't Get Duped: Fraud through Duplication in Public Opinion Surveys," SSRN Scholarly Paper No. ID 2580502, SSRN Scholarly Paper, Rochester, NY: Social Science Research Network, December 12. (http://papers.ssrn.com/abstract=2580502).

Liaw, A., and Wiener, M. 2002. "Classification and Regression by randomForest," *R News* (2:3), pp. 18–22.

Loomis, J., Roach, B., Ward, F., and Ready, R. 1995. "Testing Transferability of Recreation Demand Models Across Regions: A Study of Corps of Engineer Reservoirs," *Water Resources Research* (31:3), pp. 721–730. (https://doi.org/10.1029/94WR02895).

Lycett, M. 2013. "'Datafication': Making Sense of (Big) Data in a Complex World," *European Journal of Information Systems* (22:4), pp. 381–386. (https://doi.org/10.1057/ejis.2013.10).

Maccani, G., Donnellan, B., and Helfert, M. 2015. "Exploring the Factors That Influence the Diffusion of Open Data for New Service Development: An Interpretive Case Study," in *ECIS 2015 Completed Research Papers*, Münster, Germany: AIS electronic library, May 29. (https://doi.org/10.18151/7217419).

Maccani, G., Donnellan, B., and Helfert, M. 2017. "Adoption of Open Government Data for Commercial Service Innovation: An Inductive Case Study on Parking Open Data Services," in *AMCIS 2017 Proceedings*, AIS electronic library, August 10.

Martens, D., and Provost, F. 2014. "Explaining Data-Driven Document Classifications," *MIS Quarterly* (38:1), pp. 73–99.

Marton, A., Avital, M., and Jensen, T. B. 2013. "Reframing Open Big Data," in *ECIS 2013 Completed Research*, AIS electronic library, July 1. (http://aisel.aisnet.org/ecis2013_cr/146).

Mithas, S., Lee, M. R., Earley, S., Murugesan, S., and Djavanshir, R. 2013. "Leveraging Big Data and Business Analytics [Guest Editors' Introduction]," *IT Professional* (15:6), pp. 18–20.

Müller, O., Junglas, I., Brocke, J. vom, and Debortoli, S. 2016. "Utilizing Big Data Analytics for Information Systems Research: Challenges, Promises and Guidelines," *European Journal of Information Systems* (25:4), pp. 289–302. (https://doi.org/10.1057/ejis.2016.2).

NordREG. 2017. "Nordic Market Report 2015 - NMR Dataset," Helsinki, Finland: Energy Market Authority, January 13. (http://www.nordicenergyregulators.org/wp-content/uploads/2017/01/Dataset-NMR-2015.xlsx).

Oliveira, L., and Santos, C. 2016. "The Two Sides of the Innovation Coin," in *AMCIS 2016 Proceedings*, AIS electronic library, August 11. (http://aisel.aisnet.org/amcis2016/Adoption/Presentations/20).

Open Knowledge Foundation. 2017. "Open Definition 2.1 - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge," *Open Definition 2.1*, , August 14. (http://opendefinition.org/od/2.1/en/, accessed August 14, 2017).

Pebesma, E., Bivand, R., Rowlingson, B., Gomez-Rubio, V., Hijmans, R., Sumner, M., MacQueen, D., and Lemon, J. 2017. *Classes and Methods for Spatial Data [R Package Sp Version 1.2-5]*. (https://CRAN.R-project.org/package=sp).

Pham, X. H., and Jung, J. J. 2014. "Recommendation System Based on Multilingual Entity Matching on Linked Open Data," *Journal of Intelligent & Fuzzy Systems* (27:2).

Ponte, D. 2015. "Enabling an Open Data Ecosystem," in *ECIS 2015 Research-in-Progress Papers*, Münster, Germany: AIS electronic library, May 29. (http://aisel.aisnet.org/ecis2015_rip/55).

Sawalha, Z., and Sayed, T. 2006. "Transferability of Accident Prediction Models," *Safety Science* (44:3), pp. 209–219. (https://doi.org/10.1016/j.ssci.2005.09.001).

Sharma, R., Mithas, S., and Kankanhalli, A. 2014. "Transforming Decision-Making Processes: A Research Agenda for Understanding the Impact of Business Analytics on Organisations," *European Journal of Information Systems* (23:4), pp. 433–441.

Shmueli, G., and Koppius, O. R. 2011. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), pp. 553–572.

Sikder, S., Pinjari, A. R., Srinivasan, S., and Nowrouzian, R. 2013. "Spatial Transferability of Travel Forecasting Models: A Review and Synthesis," *International Journal of Advances in Engineering Sciences and Applied Mathematics* (5:2–3), pp. 104–128. (https://doi.org/10.1007/s12572-013-0090-6).

Sodenkamp, M., Kozlovskiy, I., Hopf, K., and Staake, T. 2017. "Smart Meter Data Analytics for Enhanced Energy Efficiency in the Residential Sector," in *Wirtschaftsinformatik 2017 Proceedings*, St. Gallen, Switzerland: AIS electronic library, January 23.

Sodenkamp, M., Kozlovskiy, I., and Staake, T. 2015. "Gaining IS Business Value through Big Data Analytics: A Case Study of the Energy Sector," in *ICIS 2015 Proceedings*, Fort Worth, USA: AIS electronic library, December.

Swiss Federal Council. 2014. "Bundesrat startet Vernehmlassung über die volle Strommarktöffnung," *Press release*, , August 10. (https://www.admin.ch/gov/de/start/dokumentation/ medienmitteilungen.msg-id-54746.html, accessed August 13, 2017).

Tiefenbeck, V., Tasic, V., Staake, T., and Fleisch, E. 2013. "Contrasting the Effects of Real-Time Feedback on Resource Consumption between Single- and Multi-Person Households," in *SSES Annual Meeting 2013*, Neuchatel, Switzerland.

Vanreusel, W., Maes, D., and Van Dyck, H. 2007. "Transferability of Species Distribution Models: A Functional Habitat Approach for Two Regionally Threatened Butterflies," *Conservation Biology* (21:1), pp. 201–212. (https://doi.org/10.1111/j.1523-1739.2006.00577.x).

Vassileva, I., Odlare, M., Wallin, F., and Dahlquist, E. 2012. "The Impact of Consumers' Feedback Preferences on Domestic Electricity Consumption," *Applied Energy* (93), pp. 575–582.

Vercamer, D., Steurtewagen, B., Van den Poel, D., and Vermeulen, F. 2016. "Predicting Consumer Load Profiles Using Commercial and Open Data," *IEEE Transactions on Power Systems* (31:5), pp. 3693–3701. (https://doi.org/10.1109/TPWRS.2015.2493083).

Vickery, G. 2011. "Review of Recent Studies on PSI Re-Use and Related Market Developments," *Information Economics, Paris*.

Wenger, S. J., and Olden, J. D. 2012. "Assessing Transferability of Ecological Models: An Underappreciated Aspect of Statistical Validation," *Methods in Ecology and Evolution* (3:2), pp. 260–267. (https://doi.org/10.1111/j.2041-210X.2011.00170.x).

Wing, M. K. C. from J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., The R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., and Scrucca, L. 2015. *Caret: Classification and Regression Training*. (http://CRAN.R-project.org/package=caret).

World Bank. 2017. "GDP per Capita, PPP (Current International $) | Data," *International Comparison Program Database*, , August 13. (http://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD? view=chart&year_high_desc=true, accessed August 13, 2017).

Xu, M., and Walton, J. 2005. "Gaining Customer Knowledge through Analytical CRM," *Industrial Management & Data Systems* (105:7), pp. 955–971. (https://doi.org/10.1108/02635570510616139).

Yadav, P., Hasan, S., Ojo, A., and Curry, E. 2017. "The Role of Open Data in Driving Sustainable Mobility in Nine Smart Cities," in *ECIS 2017 Research Papers*, Guinmarães, Portugal: AIS electronic library, June 10, pp. 1248–1263. (http://aisel.aisnet.org/ecis2017_rp/81).