

# Exogenous Data for Load Forecasting: A Review

Ramón Christen<sup>1</sup>, Luca Mazzola<sup>1</sup><sup>a</sup>, Alexander Denzler<sup>1</sup> and Edy Portmann<sup>2</sup><sup>b</sup>

<sup>1</sup>Information Technology, HSLU - Lucerne University of Applied Sciences and Arts, Suurstoffi 1, 6343 Rotkreuz, Switzerland

<sup>2</sup>Human-IST Institute, University of Fribourg, Bd de Pérolles 90, 1700 Fribourg, Switzerland

**Keywords:** Smart Grid, Energy Prediction, STLF, Feature Selection, Exogenous Data Analysis.

**Abstract:** Electrical power load forecasting defines strategies for utilities, power producers and individuals that participate in a smart grid. While it is well established in planning processes for production and utilities, the importance of accurate forecasting increases for individuals. The ongoing deregulation of the electricity market enables energy trading by individuals, requiring an accurate estimation of the production and consumption. Research on forecast for aggregated demand shows that including features for the forecast from sources, called exogenous, additional to the purely historical consumption data allows to obtain higher accuracy. In fact, their usage demonstrated to be able to explain the large variability observed in the power demand, taking into account the individual influences. Anyway, the influence of exogenous data is hardly investigated for individual forecasting, due to the minor prevalence of this analysis to date. This review shows the benefit of exogenous data usage and the necessity of detailed research on the input features and their influence on detailed, individual level, forecasts of power demand. Eventually, this contribution is concluded by the presentation of open issues and research directions for electric smart communities that the authors would like to address.


## 1 INTRODUCTION


Electrical power is hardly storable. In fact, the energy production depends directly on the demand. The higher the demand, the more power must be produced. That means, any change in the demand has an immediate impact on the grid stability. Therefore, ensuring a stable grid requires a regulation of the production. On a producer level, power plants and grid operators count on a serious planning in order to guarantee a permanent power demand coverage. They expect to know the future demand for maintaining short-term grid balances and for planning grid extensions based on long-term power demand.

In contrast, on consumer level, the ongoing deregulation of the energy market allows direct access to the electricity market. Consumers with own power production, so called prosumers, become able to share their own productions with peers and trade energy on local markets (Mazzola et al., 2020). This drives the development of smart grids with the intention, to use the power as close as possible to the source. Electrical energy should be used more efficiently, close to production and with less grid usage, thus incurring in less

infrastructure charges. However, a change to locally produced sustainable energy increases the volatility in the power grid and makes it hard to control with respect to stability. Regarding this issue, it is again essential for power plants to have an accurate estimation of the power demand in advance, in order to react on the volatile and fast changing demand. For prosumers, on the other side, it is equally worthwhile to estimate the expected power consumption and self-production capabilities. In fact, this information allows balancing two conflicting objectives: optimal price control on network level with maximised self-sufficiency for prosumer.

The end-user's consumption itself is affected by many external factors. Among others, weather conditions and consumer's behaviour and activities, define the circumstances for the consumption. On the other side, sustainable power production almost completely depends on weather conditions, such as wind and sun irradiation. Many studies evidence this dependency of the power consumption and production on external factors; the dependency on so called exogenous variables. Therefore, load forecasting often include data from exogenous variables in addition to historical load data. This additional data seems to provide promising information for increasing forecast accuracy. Furthermore, the enhancement of everyday ob-

<sup>a</sup>  <https://orcid.org/0000-0002-6747-1021>

<sup>b</sup>  <https://orcid.org/0000-0001-6448-1139>

jects by electrification, the consequent growing diffusion of consumer grade Internet of Things (IoT) devices and the stream of information through social networking services open up new possible sources for exogenous data.

Load forecasting is an activity with differing challenges. Parameters such as the forecast time window or the aggregation level can completely change the focus and imply different challenges. That also applies for exogenous data as this information is used to increase the forecast accuracy. Its value for contributing on better results strongly depends on the forecast focus. Therefore, it is of paramount importance to precisely know the level of support provided by each variable for load forecasting. Although the influence of several well studied exogenous variables on the power consumption is known, the real increase for forecast quality and certainty is broadly unknown.

This paper gives a review of previous work on load forecasting with focus on the use of exogenous variables for predicting power demand. The review is based on literature searches run between March, 16th and April, 3rd 2020 on both IEEE Xplore and Google Scholar libraries. While IEEE Xplore provides publications in computer science and engineering, Google Scholar has been considered for covering a wide area of scientific publications. The literature research comprises the terms: *Short-Term Load Forecast (STLF)*, *power prediction*, *load forecasting*, *exogenous variables / features / data*, *feature selection* and *social media*. It considers more than 50 publications, with a focus on recent trends. In fact, about 85% appeared in 2010 or later (half of which in the last 3 years). The analysis shows findings in: a) the correlation of exogenous data to the power demand, b) considered data in load forecasting and c) feature selection approaches for accuracy improvements. In addition, this paper gives an overview of the most common applied forecasting methods and points out the demand of an in-depth analysis on the use of exogenous data for increasing the prediction accuracy.

The remainder of the paper is organised as follows: Section 2 presents the use of exogenous data, by disclosing the impact and advantage of considering it in load forecasting. Following, Section 3 reveals the broad variation of used exogenous data in the analysed approaches. A grouping and mapping of the various variables with respect to the application incidence directs the focus of the research. Section 4 highlights, for various parameters, their information value in power utilisation forecasting. Additionally, Section 5 and Section 6 review common feature selection practices. For covering all characteristics of these variables in load estimation, this part also in-

clude a discussion of methodologies and issues existing. Based on this review, Section 7 discloses a revealed knowledge gap about the true information value of exogenous data in load forecasting, before our conclusion (Section 8) terminate this contribution.

## 2 IMPACT OF EXOGENOUS DATA IN LOAD FORECASTING

This section answers the question about *why exogenous data is used in load forecasting*. Predicting the energy consumption is an old endeavor (Gross and Galiana, 1987). Whether for long-term decisions in respect to grid assets or short-term load balance estimations in modern smart grid approaches, the prediction always needs to be as accurate as possible with respect to its energy consumption. Balancing the production and demand of power or trading energy on a deregulated market among others, require a high agility of power producers and storage assets to guarantee at the same time grid stability. Because of this requirement, both may profit from an accurate and reliable prediction. In an initial effort, experts purely estimated the power demand for the next few hours, days or weeks based on the historical behaviour. This simple estimation features a high uncertainty as it assesses the prospective behaviour on history only, without considering any context to it. Certainly, it suffices for rough decisions but not for issues with a finer granularity or a higher complexity, such as balancing significant and quick load changes.

The power demand as well as the production from sustainable sources depend on higher-level circumstances. Power generation units such as Photo Voltaic (PV), for instance, produce energy depending on the intensity of solar irradiation. Similarly, the power demand depends on the operation of electrical loads which is driven by external factors. In fact, users define the consumption profile by turning on an off their devices, under the needs determined by higher-level circumstances. Accordingly, researchers try to use information of this external factors, so called exogenous data, to improve the load forecasting quality. The correlation between the climate and the power consumption have already been discussed half a century ago (Heinemann et al., 1966). And the results of many dependency studies demonstrate a positive correlation with the proposed forecasting approaches, when using exogenous data in load forecasting.

Using exogenous data as additional source for input data in load forecasting allows for the extraction of context related features. This features provide information from higher-level circumstances that di-

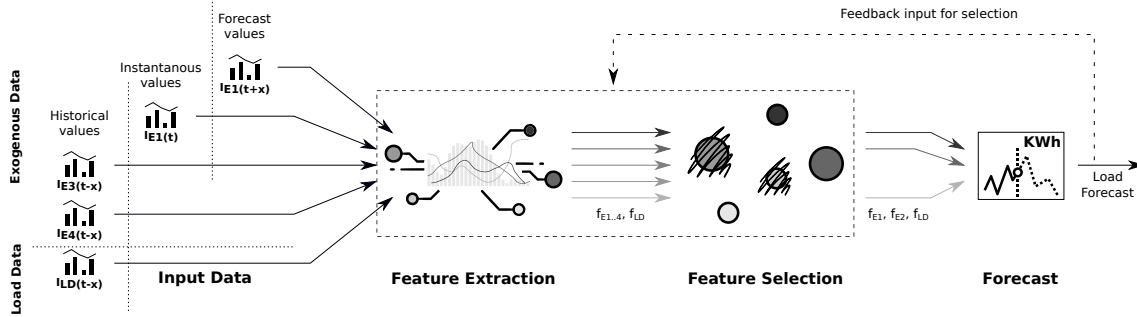


Figure 1: Load forecast feature sources.

rectly influence the power time series. As depicted in Figure 1 the information from exogenous data can originate from recordings, current measurements or forecasts. The extracted features from the exogenous data can be used together with the information from power load data for building a pool of features which describe the target power time series. Alternatively, exogenous data can also be used decoupled and therefore without load data as shown in (Kandil et al., 2006) where they evaluated an approach for missing historic load data. In a typical application, the extracted features pass a selection procedure that pick a few features with high relevance which are subsequently forwarded to the input of the forecast algorithm. For a refinement, the parametrisation of the feature extraction and selection methods can also consider the output of the load forecast in a feedback loop.

Exogenous data can be seen as meta-data of the power load. They provide information about higher-level circumstances that affect the consumption as well as the production. As in (López et al., 2017) and (Janicki, 2017), the literature shows that considering this additional information can improve the load forecast quality. However, this also shows the high complexity of the dependency of power load on various influencing variables. There is no single but also not a fixed set of variables that completely describe the consumption nor the production. The broad variation of exogenous data and their impact on load forecasting is discussed in the following sections.

### 3 EXOGENOUS DATA IN LOAD FORECASTING: CHARACTERIZATION

In this section, an *exploration on which exogenous data are used in load forecasting* is provided. Energy consumption forecast mainly extracts key information from recordings of the target variable to pre-

dict the future demand. These recordings are scanned for describing key values under the assumption that time series similarly continue as in history. The extracted characteristics that accurately define history time series such as frequencies, wavelet components or patterns that follow a typical structure, serve as input variables for the forecast as in (Chen et al., 2008; Zheng et al., 2017; Silva et al., 2017; Jiang et al., 2017). Some approaches, such as Rana and Koprinska (Rana and Koprinska, 2012b; Rana and Koprinska, 2013), try to fully include this source for better forecast results. In this work, they claimed to achieve better forecast results by means of a shift invariant transformation of the frequency components in the exogenous data. However, despite the high informative value of the history data, there is a possible valuable improvement in forecasting accuracy of the power load by extending the input variables with exogenous data.

The literature review revealed a wide range of additionally included exogenous data. More than 50 different variables could be identified from 48 analysed publications. By clustering them into typology-based general categories, we were able to identify the following groups:

**Weather data** (*humidity, precipitation, temperature, wind-chill, etc.*)

**Calendar data** (*date, events, moving holidays, summer break, etc.*)

**Day information** (*before/after holiday, (non-) working day, weekday, etc.*)

**Socio-economic** (*economic trends, gdp, # of employments, etc.*)

**Demographic information** (*birth rate, dwelling count, population, etc.*)

**Others** (*no. of sensors, occupants, devices, etc.*)

Anyway, our analysis discovered a high variation in the usage of these factors. Counting the occurrence

of all variables in the analysed forecast approaches reveals large differences in the consideration received by the higher-level categories. With a significant gap, weather, calendar and day information are clearly the default choices. 50% and more of the approaches include data from these groups. In contrast, only about 10% or less use socio-economic data, demographic information or other data to enrich the input variables. Table 1 shows the inclusion of exogenous data regarding the different higher-level groups in a descending order. It emphasises the large gap of favoured additional data sources to the minor ones.

Table 1: Inclusion of exo. data in load forecast.

Weather data	63%
Calendar data	55%
Day information	53%
Socio-economic data	8%
Demographic information	5%
Other data	5%

A breakdown of the variables in weather data presents four main clusters: (i) temperature related data with a usage proportion of approx. 37%. It appears to be the most relevant weather variable and includes any temperature data such as the air temperature, dry- and wet-bulb temperatures or the wind-chill index. Followed by (ii) humidity and (iii) wind-speed information that equally share about 20%. The sky coverage (iv) has still a proportion of 12%. The remaining 30% comprises rarely used variables such as precipitation or air pressure information.

A significant parameter in load forecast is the forecast time window. Due to different problem representations depending on the time window, the literature separate the forecast time windows predominantly in four time section: Very-Short-Term Load Forecast (VSTLF), STLF, Medium-Term Load Forecast (MTLF) and Long-Term Load Forecast (LTLF). However, the time span definition for each of them is differently designated in literature, as shown in Table 2. This review integrates the VSTLF into the group STLF because of the minimal differences shown in terms of exogenous data usage between those two categories.

Table 3 to 5 provide an overview of the distribution of three forecast key values and the inclusion of exogenous data accordingly. Table 3 divides the forecast time window in STLF, MTLF and LTLF. The forecast resolution and the aggregation level are compared in table 4 and table 5 respectively. In all tables, the first column represents the proportion of the variables of all evaluated approaches. The columns after the double line separation represent the distribu-

tion of included exogenous information for each variable. The review focus on the use of exogenous data in power demand forecasting. Approaches for other variable forecasts or without exogenous data are out of this scope. Hence, a change of the focus may yield different results in statistics.

In Table 3 it is obvious that researcher pay more attention to STLF when considering exogenous data as supporting input data. Mid- and long-term forecast have only a proportion of 10% and 5% respectively. However, decoupled from the forecast time window, historic load data always seems to be a key input source for load forecast. For all time window, they have a prominent proportion of approx. 30% of the considered input data. In contrast, only long-term forecasts use socio-economic and demographic information seriously as additional input variables. Mid- and short-term forecast time window rather consider weather and day information whereas calendar data seems to provide valuable information for all time horizons.

According to the mainly provided forecast resolution in literature (see Table 4), it is separated in hourly, daily or yearly forecasts. The forecast resolution represents the time span of a single forecast value comprising either an instant demand or a peak value. Comparing Tables 3 and 4, the proportion of the forecast resolution shows a similar distribution as of the forecast time windows. This high correlation implies that forecasts with short time windows usually provide higher resolutions. The longer the forecast time window, the lower the forecast resolution. By considering the inclusion of exogenous data, it shows that only the proportion of other exogenous data differs from the relation to forecast window. This difference appears due to the evaluation of the study from Chen et al. that analyses the forecast behaviour by increasing the resolution (Chen and Cook, 2012).

The third key variable of the analysed studies in load forecasting, the node aggregation level presented in Table 5, points out a strong focus on utility level. More than 80% of all studies relate to forecasts on this level. That means, they consider the forecast of the power demand of numerous end users up to a summarised node on a utility level (i.e. of a grid branch, accommodation or production unit). The table also shows, that forecasts on utility and local levels include load data with a considerably higher proportion than individual forecasts. However, socio-economic and demographic data are only used on utility level. On the contrary, only the individual level includes other exogenous data in the inputs parameter list.

As evident from statistics, weather variables are the best scrutinised and most considered exogenous

Table 2: Definition of forecasting time horizons.

	(Hong, 2010)	(Ma and Ma, 2017)	(Matijaš et al., 2011)	(Mirowski et al., 2014)	(Raza and Khosravi, 2015)	(Mustapha et al., 2016)	(Zor et al., 2017)	(Hammad et al., 2020)
VSTLF	1d	<30'	1h	-	-	-	<1h	<1h
STLF	2W	30' - 6h	<30d	1h - 1W	1h - 1W	1h - 1W	1h - 2W	1h - 1W
MTLF	3Y	6h - 1d	<1Y	1W - 1Y	1M - 1Y	1W - 1Y	2W - 3Y	1W - 1Y
LTLF	30d	1d - 1W	>1Y	>1Y	1Y - 10Y	>1Y	>3Y	>1Y

Table 3: Forecast Time Window and Exogenous Data [%].

	%	Load	Weather	Calendar	Day	Socio-economic	Demographic	Others
STLF	85	29	25	21	22	1	0	2
MTLF	10	27	27	27	18	0	0	0
LTLF	5	29	0	14	0	29	29	0

Table 4: Forecast Resolution and Exogenous Data [%].

	%	Load	Weather	Calendar	Day	Socio-economic	Demographic	Others
Hourly	80	33	17	22	22	0	0	6
Daily	15	29	26	21	21	1	0	1
Yearly	5	20	10	20	10	20	20	0

data. Heinemann et al. studied the relationship between weather and power load data already about 55 years ago (Heinemann et al., 1966). The study describes a method to extract weather sensitive components from the total daily peak load. However, it only concerns the load during summer time. In the following years, several studies investigated the dependency of the power demand on the weather behaviour. And all agree on the existence of a notable correlation between these two domains. Furthermore, all emphasise possible improvements in load forecasts by including weather data such as in (Rahman and Hazim, 1993; Hernández et al., 2012; Sahay and Tripathi, 2014; Janicki, 2017).

As a result, numerous approaches tried to increase the forecast accuracy by extending the input variables with information from various weather variables (Rahman and Hazim, 1993; Mirasgedis et al., 2006; Howe, 2010; Chu et al., 2011). Recently, Silva et al. even defined weather variables and mainly temperature, humidity and wind speed as the most significant exogenous influences in STLF (Silva et al., 2019b).

In contrast to utility or local level, the forecast on individuals takes the direct environment more into account. So in (Chen and Cook, 2012), where Chen et al. propose an STLF approach on individual level that additionally uses the activity in the building. They

Table 5: Forecast Aggregation Level and Exogenous Data [%].

	%	Load	Weather	Calendar	Day	Socio-economic	Demographic	Others
Indiv.	13	15	23	23	23	0	0	15
Local	5	33	33	17	17	0	0	0
Utility	82	31	25	19	19	4	2	0

spread several motion sensors and monitored a whole residential area. The recorded motion data allowed to derive a pattern that provides extended information to the consumption, which they used to enrich the input variables for the power forecast. Similarly, Wang et al. proposed an approach that considers the occupancy information to forecast the energy usage of educational buildings (Wang et al., 2018). Furthermore, Tascikaraoglu and Sanandaji looked for relational patterns among correlations of time series of surrounding houses (Tascikaraoglu and Sanandaji, 2016). The proposed approach showed a considerable improvement for short term forecasts against various benchmark models using real and high-quality data.

Other than in STLF, forecasts with long time windows use exogenous data related to higher-level consumption influences. This information does not comprise changes having a direct influence on single end user's behaviour but represent large-scale events. In fact, load forecasts ahead for several years are more affected by long term changes such as economic, demographic or climatic movements. Hence, LTLF approaches as in (Chui et al., 2009; Khatoun et al., 2014) include information about the population, Gross Domestic Product (GDP) or geographical related changes. Huang et al. showed in their study (Huang et al., 2016a) that the rapidly increase of the power demand in a city in northeast China matches with the population growth and the development of the local society. The End-use model, a detailed modelling approach for LTLF, breaks down the energy consumption to single consumers. This model is applied to estimate the energy consumption for long time windows. It is based on extensive information

about the end user that ranges till to the device level. The approach is discussed by Ghods and Kalantar in (Ghods and Kalantar, 2008).

#### 4 VALUE OF EXOGENOUS DATA FOR DIFFERENT LOAD FORECAST PARAMETERS

This passage narrates on the *correlation between exogenous data types for different forecast parameters* (e.g. forecast horizon, location, climate). Higher-level conditions have different influence on the energy consumption. Derived from statistics in section 3, it is obvious that the usable information contained in a specific exogenous data type mainly depends on the forecast parameters. Fast or slowly changing values, for instance, have more or less impact on the accuracy depending on the forecast time window. While economic or climate aspects have a significant influence in long-term forecasts, they provide barely any useful information for short-term forecasts. In fact, they appear to be basically stationary for the forecasting time period considered. Therefore, their contribution in increasing the accuracy in STLF is very small. In the study (Gul et al., 2011) Gul et al. investigate the relationship between electrical power demand and the slowly evolving economic and demographic variables. On a country level case study for Pakistan, they discovered a high correlation between the selected variables and the power demand. Generally, it can be shown that long-term forecasts use averaged data or slow changing variables with low sampling rates. In contrast, fast changing data with high sampling rates such as sky coverage are primary meaningful for short-term forecasts.

In (Wang et al., 2018), a study from Wang et al. on the energy prediction of two educational buildings, they show a variation of the most influential factors from one semester to another. The study discuss the variable importance in a case study of an hourly energy prediction. As a consequence of the observed variation in the influence, they conclude that the energy usage of these educational buildings follow rather a semester than an annual basis pattern.

Equally to the effect on the forecast time window, exogenous data also differently influences the forecast accuracy at various aggregation levels. The lower the forecast level (e.g: room or apartment), the higher the observable impact generated by single energy consuming appliances. Information about the energy consumption for charging a battery of an e-vehicle, for instance, highly affect the consumption

behaviour of a single end user. On the other hand, the aggregated load on a city or country level (i.e. on a utility level) is not considerable affected by the consumption of a single battery charge. The large number of end users cancels out significant changes in a single load behaviour, due to the so called averaging effect. This effect of the aggregation level is analysed by Mirowski et al. in (Mirowski et al., 2014) and Gerwig et al. in (Gerwig, 2015) and seems to have a relevant impact in terms of the prediction accuracy.

The geographic location of the considered energy consumption is one of the variables that define the higher-level conditions. The specific climate, economic or lifestyle among others depend on the location and influence the energy consumption significantly. The effect of the geographic location, how it defines the information value of exogenous variables, becomes evident by observing the weather variables. Howe features in his thesis (Howe, 2010) the dependency of the region and the influence of the temperature on the energy consumption. So, in Philadelphia for instance, the temperature swings are limited thanks to the proximity to the ocean although the city experiences large temperature ranges than Chicago, from bitter cold winter to hot summer days. This has a certain impact on the energy consumption. In fact, the temperature influence varies among the regions due to a different use of cooling and heating devices. This effect is also discussed by De Felice et al. (De Felice et al., 2013) and Silva et al. in (Silva et al., 2019b) for the climate in Italy and South America respectively. Furthermore, in their study De Felice et al. discovered that from all included weather variables only the temperature demonstrated an evident influence on the daily load variation. On the other hand, they also conclude that in some cases the use of weather information does not show an evident benefit for daily load forecasting. A similar statement was made by Kandil et al. in (Kandil et al., 2006). For Hydro-Quebec they investigated the effect of various weather variables on the power load for the province of Quebec, in Canada. Thereby, they discovered that only temperature has a serious influence. Other weather variables like sky condition (cloud cover) and wind velocity showed no relation to the load. A brief overview of the influence of the weather factors on the electrical demand is given in (Janicki, 2017).

Using weather variables in load forecasting mainly means the inclusion of some weather variable forecasts in the input variable list. Douglas et al. investigated in (Douglas et al., 1998) the effects of the uncertainty of the variable forecasts on STLF. The performed analysis presents differing impacts of the temperature forecast errors in the vari-

ous annual seasons. Also López et al. discovered a non-linear dependency of the load on temperature in (López García et al., 2013). According to this study, the non-linearity makes raw temperature data insufficient for using in load forecasting. The data need to be contextualised for a meaningful use. However, they also proposed a STLF approach for Balearic Islands that considers the solar radiation, cloudiness and wind velocity without temperature and claimed, that the use of all variables in combination outperforms all other variants (López et al., 2017).

As exemplary shown on weather variables, the various exogenous data provide information with a different value. It highly depends on the target of the forecast - the location, time window, aggregation level and so on. Therefore, it is not possible to uniquely order the variables according to a specific relevance or prevalence rank. A proper selection of the including exogenous data and thereof the applying features is indispensable. The following section give an overview about the various feature selection methods and discusses the need for identifying the most appropriate forecast methodology.

## 5 FEATURE SELECTION METHODS

In this part it is concisely explored the *role of feature selection and its relation to exogenous variables* for different settings of load forecasting. Using exogenous data in load forecasting makes a careful feature selection fundamental. The data should provide a considerable added value to the basic information from historic load data for reaching more accurate forecasting results. The power demand is a high complex system that depends on many external factors; and some provide information with a higher value than others. Typically, it is straight forward to assume, that more information would lead to a higher forecast accuracy. However, the forecast accuracy does not strictly monotonous increase with a growth in feature number. Too many and redundant features may even drag down the forecasting performance, either by introducing inessential information or by accumulating the effects of the noise present in any real measurement. This behaviour is discussed by Cheng et al. in (Cheng et al., 2017). As a consequence, it is crucial to include only features producing a considerable, non-negligible increase in the forecast performance. However, those selected features do not necessarily show always a high correlation with the target data. In the case of sudden changes in the contextual situation, even features without a high correlation lead to

a more robust forecaster, as described by Drezga and Rahman in their study (Drezga and Rahman, 1998).

In a common forecast pipeline, valuable features are defined in an extraction and selection process that precedes the definition or training of the forecast model. The forecast output, on the other hand, may have a direct feedback on the preceding feature definition process as illustrated in Figure 1. The state-of-the-art feature selection approaches usually apply a multi-step selection procedure. Additional steps break down the feature set to a few crucial variables. Commonly used distance related metrics for defining feature's information value comprise correlation functions, Mutual Information (MI) or Fisher Information (FI). This distance metrics quantify the feature's similarity to the load data and are used as measure for the influence on it (Rana and Koprinska, 2012a; Cai et al., 2018; Hu et al., 2015; Huang et al., 2016a). However, the relation between two variables can also show a dispersed non-linear characteristics such as it is often shown for the temperature and load. For this case, Silva et al. claim more accurate relation indices when using models that estimate the existence condition and not only the linear dependency as the Pearson Linear Correlation method (Silva et al., 2019a). The approaches for the selection refinement are very broad and comprise methods of Machine Learning (ML) (Niu et al., 2010; Rana and Koprinska, 2012a), wrappers (Hu et al., 2015), minimum Redundancy Maximum Relevance (mRMR) (Huang et al., 2016a), Permutation Importance (PI) (Huang et al., 2016b) or Random Forest (RF) (Cheng et al., 2017).

Finally, a qualitative feature selection requires clean data sets, affected by a minimised noise component. As all raw data, exogenous data usually contain outliers or spikes that do not correlate with target data and distort the valuable information. For a reasonable benefit from the additional information, the anomalies need to be rejected. The advantage of such a preprocessed filtering is shown in several studies as in (Guan et al., 2013; Mustapha et al., 2016; Saleh et al., 2016). The same holds, if only certain parts of exogenous data provide valuable information. In such a case, filters need to cancel out the remaining parts before the model uses the data for training and forecast.

## 6 LOAD FORECASTING METHODOLOGIES

In this section, on top of the basic approaches reported in literature, the *selection importance of some specific features* is detailed. A broad review of various

STLF methodologies is presented in Srivastava (Srivastava et al., 2016). This study, in accordance with other literature analysed, separates the models mainly in two groups: statistical and machine learning approaches. Statistical models usually explicitly describe a mathematical relationship between multiple variables. Therefore their applicability is limited in case of large number of variables as well as for highly non-linear, complex dependencies. Based on this limitation, machine learning approaches gained more attention in recent years. According to comprehensive reviews of multiple applied methods and models by Gerwig et al. (Gerwig, 2015) and Hammad et al. (Hammad et al., 2020), STLF approaches favourably use machine learning models. For STLF on individual level, the results in (Marinescu et al., 2013) indicate equally good performances provided by the analyzed Artificial Intelligence (AI) and Autoregressive (AR) methods. Similarly, (Gerwig, 2015) state comparable results for Artificial Neural Network (ANN), AR and hybrid methods of both for individual households up to 1000 end-users. In contrast, Linear Regression (LR) shows comparable results only for individual users while Support Vector Regression (SVR) works well for more than 32 households. Additionally, this study also remarks a possible accuracy improvement by combining clustering methods with ANN or autoregressive methods.

Independently from the method chosen, each load forecast is based on two main components: a model and a selection of input variables presenting a considerable influence on the target dimension. Thereby, the input elements may include recordings, instant values or other forecast measurements. That holds for exogenous data but also for load data, as illustrated in Figure 1. For weather input variables, models mostly use forecast variables. These seem to provide a higher information value than historical data or instant values. In fact, the use of historic weather data requires that load forecasts also comprise weather models in order to map the information from historical data to forecast load behaviours. The other way round, when using forecasts of the weather variables, customised models do the mapping of the information from historical data to future scenarios separately (Zhu et al., 2018). Yet, in this case, the uncertainty of the weather variable forecast directly enters into the load forecast model. This yields to a forecast, that implicitly comprises the uncertainty of the input variable.

Recursive forecast approaches show a similar effect. Models such as recursive Kalman filters or Bayesian estimations use the previously calculated data point of the same time series as base for the calculation of the subsequent point, as reported in (Dou-

glas et al., 1998). Due to the recursive calculation, the forecast errors of previous data points directly affect the calculation (i.e. the forecast of the next data point). Because of this effect - the accumulation of the prediction errors - all recursive models can have a significant drift in the forecast results.

The effect of forecast errors in weather variables on load forecasting models is discussed in (Douglas et al., 1998; Fay and Ringwood, 2010). Douglas et al. note that a sizeable portion of the load forecast error is due to a lack of accuracy in the weather forecast. In order to minimise this effect, Taylor and Buizza present an approach in (Taylor and Buizza, 2003) that uses an ensemble prediction system. It estimates the midday power demand from the density function of an ensemble of 51 weather-related demand scenarios. Another approach propose Fay and Ringwood with a model fusion technique (Fay and Ringwood, 2010). They claim, that fused forecasts are often more accurate than any individual model forecasts. The model in the applied case study comprises four independent sub-models that feed a fusion algorithm. In a former study they found already, that decomposing load data into parallel series is advantageous due to the degree of independence of parallel series. Equally, Tascikaraoglu points out the positive effect of decomposed forecasting in (Tascikaraoglu and Sanandaji, 2016) and Borojeni et al. proposed a similar approach in (Borojeni et al., 2017). In the latter, they separate the forecast in multiple models by means of an extended Seasonal Auto-Regressive Integrated Moving Average Model (SARIMA). This allows to map multiple seasonality cycles to the power demand forecasting.

The various load forecasting methodologies process the input data in different ways. This has, as shown in literature, a notable impact on the forecast quality. Therefore, it is meaningful to define the forecast methodology depending on the input variable list.

## 7 FURTHER RESEARCH

This part presents *some gaps we noted and our future research directions* in this domain. Knowledge about the future power demand is valuable for planning and controlling the power distribution system. The increasing electrification of consumers devices and the deregulation of the energy market enhance the demand for accurate forecasts. Especially for end-user and prosumer, it is beneficial for participating on deregulated energy markets in the near future. In parallel, the growing data recordings from innumerable consumers devices opens access to new exoge-



nous data. The diffusion of social media platforms, digital social networking services and the electrical enhancement of everyday objects can constitute new sources of exogenous data for energy behaviour prediction. To list only a few, posts on social media platforms, status information from in-house and mobility smart appliances or wearable device data embed information relevant for factor estimation of electricity consumption. Anyway, in literature these new sources represent a minimal fraction of the exogenous variables considered for power demand forecasting, as can be observed from the category *Other data* in Table 1 (see Section 3). For instance, details about a planned home party from social media channels contain direct information of the end-user behaviour and may contribute in describing the upcoming power demand, but are currently rarely considered for this purpose. Another peculiarity is their end-user level granularity. This aggregation scale is better suited for more precisely describing the energy behaviour of individuals.

The literature shows a clear potential for increasing the load forecast accuracy by using exogenous data. Several studies highlight the correlation between these data series and the power demand and show the consequent advantage by their usage. Yet for reasonable accuracy improvements, the interaction of the consumption behaviour and the exogenous variables as well as their influence need to be fully understood.

Despite an intense research on forecast methodologies that consider exogenous data, the value provided by exogenous data is not sufficiently explored. Indeed, studies compare several exogenous variables with the consumption behaviour and analyse the forecast improvements when considering this additional information; but they mostly fail to provide an analysis of the real influence. To date, it seems to be widely unexplored, which information (i.e. which part of the exogenous variables) correlate with anomalies in the consumption or defines the higher-level contextual conditions such as the humidity in combination with the ambient temperature that influence the perception of the temperature. This would especially be of interest for demand forecasts on an individual level, as this level generally presents a high volatile consumption behaviour.

In that issue, we discovered a certain research demand in the analysis of the information value of exogenous and historical data on the individual forecast level. We see a demand in research on the impact of the describing variables and expect to have a weighting for the variables; possibly based on a fuzzification of the information value. This will play

a twofold role: on one side, structurally considering and account for the uncertainty, while, on the other side, allowing the use of a more fine grained reasoning, such as for categorical dimensions and their mapping from/to continuous values, using membership degrees. A detailed knowledge about the influence of exogenous data on the power load as well as the information value from historical data on a future load can be a key value for further research on forecast methods.

For a better use of the inclusion of exogenous data in load forecasting, the influence of these information needs to be analysed in more detail. Thereby, the choice of a suitable metric for the information value builds the base for a proper analysis of their influence. Secondly, an information value quantification for each candidate exogenous variable is required, by examining its point-wise correlation with respect to the power demand time series. Finally, weighting factors are necessary to control the exogenous variables effect on the power consumption forecast. To achieve this, it is crucial to identify the factors averaged quantified influence on the power behaviour. In this way, it should be possible to extract and apply only the key information from additional variables to power forecasts. It is expected that selective inclusion of exogenous variables in forecasting achieves higher accuracy, by removing redundant and irrelevant noise. This will be of paramount importance for improvements on energy management at the local community level, where individual power influx estimation plays a major role.

## 8 CONCLUSION

Exogenous data provides additional information usable for increasing the load forecast accuracy. This review shows the significantly additional value that exogenous data can contribute to better predictive performances. Additionally, it demonstrates the need for a detailed analysis of the extracted features. Depending on several factors such as the aggregation level, area or the forecast time horizon, the variables show different influence on the load. Additional variables from exogenous data show a noteworthy information value for increasing forecast accuracy. Especially weather variables, but also others, show a high correlation with power demand. This emerges from several studies ranging from short-term to LTLF and considering high aggregation levels. In contrast, there is still a significant lack of attention about the use of exogenous data in load forecast for residential and individual level. In fact, forecast on individual level is

generally affected by the high volatility issue. Nevertheless, this last category is increasingly becoming relevant due to the energy market deregulation and the possibility for end-users to actively participate in smart grids. As a consequence a detailed research on the feature information value from historic or exogenous data show an increased interest.

## REFERENCES

- Boroojeni, K. G., Amini, M. H., Bahrami, S., Iyengar, S. S., Sarwat, A. I., and Karabasoglu, O. (2017). A novel multi-time-scale modeling for electric power demand forecasting: From short-term to medium-term horizon. *Electric Power Systems Research*, 142:58–73. ZSCC: 0000109.
- Cai, S., Liu, L., Sun, H., and Yan, J. (2018). Fisher Information Based Meteorological Factors Introduction and Features Selection for Short-Term Load Forecasting. *Entropy*, 20(3):184. ZSCC: 0000001.
- Chen, C. and Cook, D. J. (2012). Behavior-Based Home Energy Prediction. In *2012 Eighth International Conference on Intelligent Environments*, pages 57–63. ZSCC: 0000038.
- Chen, Y., Luh, P. B., and Rourke, S. J. (2008). Short-term load forecasting: Similar day-based wavelet neural networks. In *2008 7th World Congress on Intelligent Control and Automation*, pages 3353–3358. ZSCC: 0000030.
- Cheng, Y., Xu, C., Mashima, D., Thing, V. L. L., and Wu, Y. (2017). PowerLSTM: Power Demand Forecasting Using Long Short-Term Memory Neural Network. In Cong, G., Peng, W.-C., Zhang, W. E., Li, C., and Sun, A., editors, *Advanced Data Mining and Applications*, Lecture Notes in Computer Science, pages 727–740, Cham. Springer International Publishing. ZSCC: NoCitationData[s0].
- Chu, W.-C., Chen, Y.-P., Xu, Z.-W., and Lee, W.-J. (2011). Multiregion Short-Term Load Forecasting in Consideration of HI and Load/Weather Diversity. *IEEE Transactions on Industry Applications*, 47(1):232–237. ZSCC: 0000000.
- Chui, F., Elkamel, A., Surit, R., Croiset, E., and Douglas, P. (2009). Long-term electricity demand forecasting for power system planning using economic, demographic and climatic variables. *European J. of Industrial Engineering*, 3(3):277. ZSCC: 0000026.
- De Felice, M., Alessandri, A., and Ruti, P. M. (2013). Electricity demand forecasting over Italy: Potential benefits using numerical weather prediction models. *Electric Power Systems Research*, 104:71–79. ZSCC: 0000053.
- Douglas, A., Breipohl, A., Lee, F., and Adapa, R. (1998). The impacts of temperature forecast uncertainty on Bayesian load forecasting. *IEEE Transactions on Power Systems*, 13(4):1507–1513. ZSCC: 0000152.
- Drezga, I. and Rahman, S. (1998). Input variable selection for ANN-based short-term load forecasting. *IEEE Transactions on Power Systems*, 13(4):1238–1244. ZSCC: 0000223.
- Fay, D. and Ringwood, J. V. (2010). On the Influence of Weather Forecast Errors in Short-Term Load Forecasting Models. *IEEE Transactions on Power Systems*, 25(3):1751–1758. ZSCC: 0000063.
- Gerwig, C. (2015). Short Term Load Forecasting for Residential Buildings—An Extensive Literature Review. In Neves-Silva, R., Jain, L. C., and Howlett, R. J., editors, *Intelligent Decision Technologies*, Smart Innovation, Systems and Technologies, pages 181–193, Cham. Springer International Publishing. ZSCC: NoCitationData[s0].
- Ghods, L. and Kalantar, M. (2008). Methods for long-term electric load demand forecasting: a comprehensive investigation. In *2008 IEEE International Conference on Industrial Technology*, pages 1–4. ZSCC: 0000110 ISSN: null.
- Gross, G. and Galiana, F. (1987). Short-term load forecasting. *Proceedings of the IEEE*, 75(12):1558–1573. ZSCC: 0000921.
- Guan, C., Luh, P. B., Michel, L. D., Wang, Y., and Friedland, P. B. (2013). Very Short-Term Load Forecasting: Wavelet Neural Networks With Data Pre-Filtering. *IEEE Transactions on Power Systems*, 28(1):30–41. ZSCC: NoCitationData[s0].
- Gul, M., Qazi, S. A., and Qureshi, W. A. (2011). Incorporating economic and demographic variables for forecasting electricity consumption in Pakistan. In *2011 2nd International Conference on Electric Power and Energy Conversion Systems (EPECS)*, pages 1–5. ZSCC: 0000015 ISSN: null.
- Hammad, M. A., Jereb, B., Rosi, B., and Dragan, D. (2020). Methods and Models for Electric Load Forecasting: A Comprehensive Review. *Logistics & Sustainable Transport*, 11(1):51–76. ZSCC: 0000000.
- Heinemann, G. T., Nordmian, D. A., and Plant, E. C. (1966). The Relationship Between Summer Weather and Summer Loads - A Regression Analysis. *IEEE Transactions on Power Apparatus and Systems*, PAS-85(11):1144–1154. ZSCC: 0000127.
- Hernández, L., Baladrón, C., Aguiar, J. M., Calavia, L., Carro, B., Sánchez-Esguevillas, A., Cook, D. J., Chinarro, D., and Gómez, J. (2012). A Study of the Relationship between Weather Variables and Electric Power Demand inside a Smart Grid/Smart World Framework. *Sensors*, 12(9):11571–11591. ZSCC: NoCitationData[s0].
- Hong, T. (2010). *Short Term Electric Load Forecasting*. PhD thesis, North Carolina State University.
- Howe, K. J. (2010). *AN ANALYSIS OF WEATHER FORECASTS IN THE CONTEXT OF ELECTRICITY USE*. PhD thesis, Pennsylvania State University. ZSCC: 0000000.
- Hu, Z., Bao, Y., Xiong, T., and Chiong, R. (2015). Hybrid filter-wrapper feature selection for short-term load forecasting. *Engineering Applications of Artificial Intelligence*, 40:17–27. ZSCC: 0000106.
- Huang, N., Hu, Z., Cai, G., and Yang, D. (2016a). Short Term Electrical Load Forecasting Using Mu-

- tual Information Based Feature Selection with Generalized Minimum-Redundancy and Maximum-Relevance Criteria. *Entropy*, 18(9):330. ZSCC: 0000009.
- Huang, N., Lu, G., and Xu, D. (2016b). A Permutation Importance-Based Feature Selection Method for Short-Term Electricity Load Forecasting Using Random Forest. *Energies*, 9(10):767. ZSCC: 0000027.
- Janicki, M. (2017). Methods of weather variables introduction into short-term electric load forecasting models - a review. *Przegląd Elektrotechniczny*, 1(4):72–75. ZSCC: 0000001.
- Jiang, P., Liu, F., and Song, Y. (2017). A hybrid forecasting model based on date-framework strategy and improved feature selection technology for short-term load forecasting. *Energy*, 119:694–709. ZSCC: 0000065.
- Kandil, N., Wamkeue, R., Saad, M., and Georges, S. (2006). An Efficient Approach for Short-term Load Forecasting using Artificial Neural Networks. In *2006 IEEE International Symposium on Industrial Electronics*, volume 3, pages 1928–1932. ZSCC: 0000203 ISSN: 2163-5145.
- Khatoon, S., Ibraheem, Singh, A. K., and Priti (2014). Effects of various factors on electric load forecasting: An overview. In *2014 6th IEEE Power India International Conference (PIICON)*, pages 1–5. ZSCC: NoCitationData[s0] ISSN: null.
- López, M., Valero, S., Senabre, C., and Gabaldón, A. (2017). Analysis of the influence of meteorological variables on real-time Short-Term Load Forecasting in Balearic Islands. In *2017 11th IEEE International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG)*, pages 10–15. ZSCC: 0000004 ISSN: 2166-9546.
- López García, M., Valero, S., Senabre, C., and Gabaldón Marín, A. (2013). Short-Term Predictability of Load Series: Characterization of Load Data Bases. *IEEE Transactions on Power Systems*, 28(3):2466–2474. ZSCC: 0000011.
- Ma, J. and Ma, X. (2017). State-of-the-art forecasting algorithms for microgrids. In *2017 23rd International Conference on Automation and Computing (ICAC)*, pages 1–6. ZSCC: 0000006 ISSN: null.
- Marinescu, A., Harris, C., Dusparic, I., Clarke, S., and Cahill, V. (2013). Residential electrical demand forecasting in very small scale: An evaluation of forecasting methods. In *2013 2nd International Workshop on Software Engineering Challenges for the Smart Grid (SE4SG)*, pages 25–32. ZSCC: 0000053 ISSN: null.
- Matijaš, M., Cerjan, M., and Krajcar, S. (2011). Features affecting the load forecasting error on country level. In *Proceedings of the 2011 3rd International Youth Conference on Energetics (IYCE)*, pages 1–7. ZSCC: 0000008 ISSN: null.
- Mazzola, L., Denzler, A., and Christen, R. (2020). Towards a Peer-to-Peer Energy Market: an Overview. *arXiv:2003.07940 [physics]*. ZSCC: NoCitation-Data[s0] arXiv: 2003.07940.
- Mirasgedis, S., Sarafidis, Y., Georgopoulou, E., Lalas, D. P., Moschovits, M., Karagiannis, F., and Papakonstantinou, D. (2006). Models for mid-term electricity demand forecasting incorporating weather influences. *Energy*, 31(2):208–227. ZSCC: 0000258.
- Mirowski, P., Chen, S., Ho, T. K., and Yu, C.-N. (2014). Demand forecasting in smart grids. *Bell Labs Technical Journal*, 18(4):135–158. ZSCC: 0000085.
- Mustapha, M., Mustafa, M. W., Khalid, S. N., Abubakar, I., and Abdilahi, A. M. (2016). Correlation and Wavelet-based Short-Term Load Forecasting using Anfis. *Indian Journal of Science and Technology*, 9(46). ZSCC: 0000006.
- Niu, D., Wang, Y., and Wu, D. D. (2010). Power load forecasting using support vector machine and ant colony optimization. *Expert Systems with Applications*, 37(3):2531–2539. ZSCC: 0000305.
- Rahman, S. and Hazim, O. (1993). A generalized knowledge-based short-term load-forecasting technique. *IEEE Transactions on Power Systems*, 8(2):508–514. ZSCC: 0000325.
- Rana, M. and Koprinska, I. (2012a). Electricity load forecasting using non-decimated wavelet prediction methods with two-stage feature selection. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. ZSCC: 0000005 ISSN: 2161-4407.
- Rana, M. and Koprinska, I. (2012b). Shift Invariance and Border Distortion in Wavelet-Based Electricity Load Forecasting. In *21st International Conference on Pattern Recognition (ICPR 2012)*, page 4. ZSCC: 0000000.
- Rana, M. and Koprinska, I. (2013). Wavelet Neural Networks for Electricity Load Forecasting – Dealing with Border Distortion and Shift Invariance. In Mladenov, V., Koprinkova-Hristova, P., Palm, G., Villa, A. E. P., Appollini, B., and Kasabov, N., editors, *Artificial Neural Networks and Machine Learning – ICANN 2013*, Lecture Notes in Computer Science, pages 571–578, Berlin, Heidelberg. Springer. ZSCC: NoCitation-Data[s0].
- Raza, M. Q. and Khosravi, A. (2015). A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. *Renewable and Sustainable Energy Reviews*, 50:1352–1372. ZSCC: 0000334.
- Sahay, K. B. and Tripathi, M. (2014). Day ahead hourly load forecast of PJM electricity market and iso new england market by using artificial neural network. In *ISGT 2014*, pages 1–5. ZSCC: 0000029 ISSN: null.
- Saleh, A. I., Rabie, A. H., and Abo-Al-Ez, K. M. (2016). A data mining based load forecasting strategy for smart electrical grids. *Advanced Engineering Informatics*, 30(3):422–448. ZSCC: 0000034.
- Silva, L. N., Abaide, A. R., Figueiró, I. C., Martinuzzi, D., and Rigodanzo, J. (2017). Development of an ANN model to multi-region short-term load forecasting based on power demand patterns recognition. In *2017 IEEE PES Innovative Smart Grid Technologies Conference - Latin America (ISGT Latin America)*, pages 1–6. ZSCC: 0000001.

- Silva, L. N., Abaide, A. R., Negri, V. G., Capeletti, M., and Lopes, L. F. (2019a). Impact Evaluation of Feature Selection to Short-Term Load Forecasting Models considering Weather Inputs and Load History. In *2019 54th International Universities Power Engineering Conference (UPEC)*, pages 1–6. ZSCC: 0000000 ISSN: null.
- Silva, L. N., Abaide, A. R., Negri, V. G., Capeletti, M., Lopes, L. F., and Cardoso, G. (2019b). Diagnostic and Input Selection Tool applied on Weather Variables for Studies of Short-Term Load Forecasting. In *2019 8th International Conference on Modern Power Systems (MPS)*, pages 1–6. ZSCC: 0000000 ISSN: null.
- Srivastava, A. K., Pandey, A. S., and Singh, D. (2016). Short-term load forecasting methods: A review. In *2016 International Conference on Emerging Trends in Electrical Electronics Sustainable Energy Systems (ICETEESES)*, pages 130–138. ZSCC: NoCitation-Data[s0].
- Tascikaraoglu, A. and Sanandaji, B. M. (2016). Short-term residential electric load forecasting: A compressive spatio-temporal approach. *Energy and Buildings*, 111:380–392. ZSCC: 0000052.
- Taylor, J. W. and Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, 19(1):57–70. ZSCC: 0000290.
- Wang, Z., Wang, Y., Zeng, R., Srinivasan, R. S., and Ahrentzen, S. (2018). Random Forest based hourly building energy prediction. *Energy and Buildings*, 171:11–25. ZSCC: 0000059.
- Zheng, H., Yuan, J., and Chen, L. (2017). Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. *Energies*, 10(8):1168. ZSCC: 0000123.
- Zhu, G., Chow, T.-T., and Tse, N. (2018). Short-term load forecasting coupled with weather profile generation methodology. *Building Services Engineering Research and Technology*, 39(3):310–327. ZSCC: 0000018.
- Zor, K., Timur, O., and Teke, A. (2017). A state-of-the-art review of artificial intelligence techniques for short-term electric load forecasting. In *2017 6th International Youth Conference on Energy (IYCE)*, pages 1–7. ZSCC: 0000026 ISSN: null.