Social data to enhance typical consumer energy profile estimation on a national level

Amr Alyafi, Pierre Cauchois, Benoit Delinchant, Alain Berges

ELECTRIMACS 2022 – Nancy, France, 16th-19th May 2022

Social data to enhance typical consumer energy profile estimation on a national level

Amr Alyafi · Pierre Cauchois · Benoit Delinchant · Alain Berges

Abstract Since the electrical grid creation, assessing the electricity demand is essential as we need to match the energy production/demand at all times. Load analysis is essential in improving the reliability and efficiency of the grid. Beside regular human activities, the main impact factor which explains consumption variations is the outside temperature. But there are still unpredictable variations that are mainly coming from arising social events. To build a better understanding of these variations, this work will focus on how to detect these events from social media and how to quantify their impact on residential and professional typical profiles for energy demand.

1 What is profile estimation

Electricity market settlement is a challenge. On one hand, there is a need to allocate energy consumed by each customer in the portfolio of a specific actor (Balance responsible) every half hour as it is the granularity of this market. On the other hand, it is only possible to get measurements from meters every 6 months or worse for the manual reading, and every day for smart meters. There is a tool that allows projecting this 6-month energy into every half-hour within, load Profiling. More than 37 million customers are concerned in France, representing about 42% of the energy passing through the electricity network [21]. Between 2008 and 2018, measuring half-hourly load curve for the mass market was expensive. Models were calibrated once and for all with a small dataset (half-hourly curves of users) and used to extrapolate future load curves for several years. This estimation method was suffering from any change in behavior, renovations, specific events and was not the best in terms of accuracy. Nevertheless, it was very easy to examine the coefficients of the model that were shared. It was therefore a deterministic equation for the Balance Responsible Entities [14]. Data sources and models have changed in 2018 [15]. The new method is called dynamic load profiling. The model uses a complex weighted average of the half-hourly load curves of a representative sample of 10 000 users [16]. These half-hourly load curves are collected every day from Linky smart meters.

Therefore, no modelization/extrapolation of what is going to happen in the future is needed to create these dynamic profiles.

This new method is more precise/accurate as it allows to capture a whole new set of events in the dynamic profiles since 2018, such as a football world cup match or the president talking on the news. The drops or raises in the load series are measured or noticed and are reflected naturally in the profile. If it makes the system fairer, the right energy allocated to the right actor, it makes its estimation a real challenge for Balance Responsible Entities. From the distributor point of view (as in France, the metering mission belongs to the DSO), it makes it also harder to guarantee the quality of the automated calculation that creates the dynamic profile. Thus, there is a need to control the quality of those calculations and create a reference. A confidence interval reference could help put the system under control (i.e validation of profiles calculated, emergency profiles if load curves are not available). The reference was first created through basic calculations using Day-7 or Day-1 values. But improving this reference means creating the best possible forecast.

Amr Alyafi · Benoit Delinchant
G2ELab - Laboratoire de Génie Electrique de Grenoble
Batiment GreEn-ER - 21 avenue des Martyrs -38031 Grenoble
e-mail: Amr.Alzouhri-Alyafi@grenoble-inp.fr
e-mail: benoit.delinchant@grenoble-inp.fr

Pierre Cauchois · Alain Berges
Enedis
345 avenue G. Clemenceau , Nanterre
e-mail: pierre.cauchois@enedis.fr,e-mail: alain.berges@enedis.fr

This is very challenging to forecast people's behavior, and small, rare events were not taken into account, though they have a clear impact on the load.

As an example to present the impact of social events on the energy consumption, Figure 1 presents the national load estimation (not consumer energy profile estimation). It is done by RTE (Electricity Transmission Network Operator of France), one day before, and on the same day vs actual energy demand for 2018 July the 15th. The figure shows an error gap between the estimation and the actual energy. This error gap of 2264 MW of energy at 5,30 pm and for a few hours later. After checking and verifying this error gap was not related to changes in the outside weather, but it is related to a world cup final football match. This example presents the importance of public events on energy load consumption and how identifying them can help to enhance the system and reduce error.

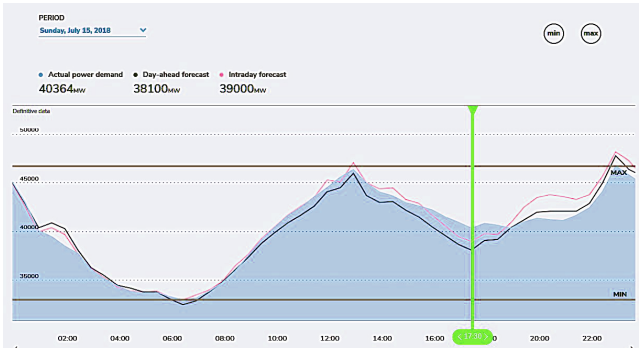

Fig. 1 National Load estimation for 2018 July the 15th at 5,30 pm [20]

Imperfect electricity load estimation leads to significant financial losses and increases the chance of having a blackout. Depending on T. Hong research [1], the predictive analytic could save a significant amount of operational cost even by %1 improvement in the load forecast accuracy.

For this many scientific research papers focused on how to estimate the electricity load on different horizons using different techniques[19]. Many solutions were proposed from knowledge based approaches to machine learning based technics [17],[18]. But the majority of research papers focused on using mostly only the weather data, and they neglect the events data. This is due to the difficulties relays to how to get these data. How to identify important events and how to quantify them to enhance the load estimation. In equivalent reduce the error when an important event appears again.

The purpose of this paper is to help design new techniques that allow to explain, and model, the error between a typical energy profile forecast and the realized series, without / with social networks data.

This work will demonstrate different techniques used for modeling the energy profile estimation using weather data with and without the events.

This paper will begin by presenting the social data. Which data to look for and how to get it?

Then it shows the proposed solution to identify the important events from social media, and at last, compares different models with and without the events data. To validate the proposed approach, this research will take the electricity load profiles in France as a case study.

2 Social media data

2.1 Twitter

Twitter is a social networking service also referred to as a micro-blogging service. The term micro-blogging service is used because Twitter only allows users to share short messages of 140 characters. It does not allow customers to play games or other advanced features other social networks sometimes offer [6].

Unlike other social networks, Twitter accounts are public by default and it is very common for users to keep their accounts public. Users can become followers of others, when one user follows other users, he is able to sees their messages called Tweets. Twitter is used by a lot of public figures like politicians and celebrities. It is quite common to follow these public figures, this is different from other social networks where people mainly connect with real friends and family.

In addition to sharing the latest happenings, tweets often contain a hashtag, i.e., a tagging mechanism allowing users to attach a word or phrase with the hash (#) symbol to a tweet; hashtags can facilitate searching on Twitter [2].

Like other social networks, Twitter is quite popular in France. According to a study, there are about 12 million french users of Twitter and around 50 million Tweets per day [5].

2.2 Social data collection

The main difficulty is how to collect the data. Due to the GDPR (General Data Protection Regulation), and other economical and regulatory reasons the API (Application Programming Interface) that allows access to Twitter data in automatic way was deprecated. Another issue to answer is when and which data to collect? There are around: 500 million tweets per day [3]. The size of each tweet is around: 560 Bytes. This means that around 280 GB of data is transmitted daily. It means in the period between 2014 and 2019 (our data set) we need to collect more than 608 TB of data which is a lot.

To overcome the first issue a tool is developed to scrape the data from Twitter. This tool allows getting the tweets depending on a specific hashtag. Generally, Web data scraping is defined as the process of extracting and combining

contents of interest from the Web in a systematic way [7]. In such a process, a software agent, also known as a Web robot, mimics the browsing interaction between the Web servers and the human in a conventional Web exchange. Step by step, the robot accesses the Twitter Web site as needed, parses its contents to find and extract data of interest, and structures those contents as desired [7].

Still to answer which data and when to look for it? A method proposed in figure 2 helps achieving this objective. To begin, an algorithm for anomaly detection is deployed; Anomaly detection is finding patterns in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different application domains [8].

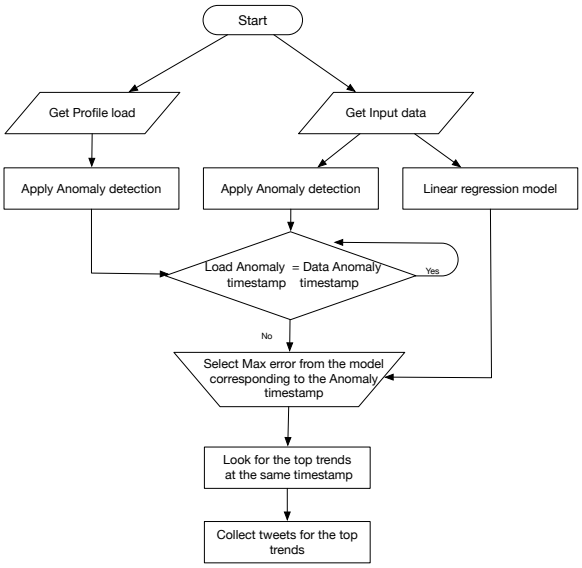

Fig. 2 Which data to look for and how to get it.

This algorithm is applied to the load curve and the input data separately. The algorithm is based on a decision tree technique. As presented in figure [8]. If an anomaly is detected in the load curve, and there is no anomaly detected in the input data, we would presume that this anomaly is related to an event and not to a change in the weather data (like a condition never seen before). This means that the scrapper needs to look for the possible events at that moment. This is done by looking for the most popular hashtags at the same moment.

After that, it will try to identify the most relevant event or more among them. First, the algorithm will look for the top three hashtags. It will collect all tweets for each hashtag on the whole period of data. In a second step, these tweets will be processed and integrated into models.

2.2.1 Explore data

As mentioned before, tweets about one subject, determined by a hashtag, are collected and stored in a database.

Fig. 3 A tweet data.

Still, a rapid look at the stored data, each record (tweet) does consist of 40 key/value like in figure 3. The key entitled tweet represents the tweet text. The key entitled language, as its name describe, represents the language of the tweet. In this work, search and use of tweets will be limited to the french language tweets.

Yet, data can not be injected directly into models. Lots of tweets keys are unuseful as they don't bring any utility for our case. At the same time for the social data to be useful, they need to represent people's interests in a subject. An indicator demonstrating the curiosity of french people about an event is needed. This can be done by getting the number of tweets for each event every hour. This indicator should capture both the peak period for an event, as well as the number of people interested. In order to do that, tweets data is aggregated to get the number of tweets for each subject (hashtag) every hour.

Figure 4 depicts the different steps needed to process the tweets data before feeding it to the model.

First, there is a need to identify the tweets. Some of the events can be directly identified from the hashtags and others are not that easily identified.

Table 1 presents the top hashtags of the 29 of April 2019 between 9 am and 10 am:

Event	Number of occurrence
Marathonnantes	99
Giletsjaunes	90
srfcpsg	84
journéedelafemme	75

Table 1 Hashtags count in 29 of April 2019 between 9 am and 10 am.

- First hashtag ”marathonnantes” with the highest ranking represents a sporting event in France.
- The second ”giletsjaunes” represents a political social movement on the ground.
- The third ”srfcpsg” is not known (easily identified). But after analyzing the tweets and their hashtags it can be found that it is a football match in league1 in France.
- The fourth ”journéedelafemme” is social events for the women’s wrights.

This step when defining whether a hashtag represents an event directly or not is done for the moment using human expertise. This work aims to validate the importance of social data before automating the process.

The second issue is that some hashtags refer to the same event like for example ”gameofthrone” and ”got”. Some of the tweets use both hashtags, so when collecting the tweets for each hashtag, the same tweet can be collected twice. The two hashtags are merged (Identifying if the two hashtags refer to the same event is manually done) and the duplicated tweets are deleted.

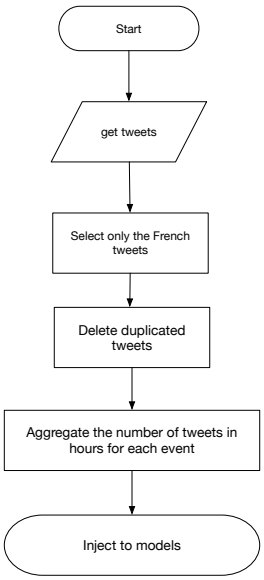

Fig. 4 schema of the different steps to treat the tweets data

Still, some hashtags can hold a different meaning than from the researched event. For example, when collecting the ”Bouchons” (traffic-jam in English) hashtag, it can be stated tweets with different meanings to find that ”Bouchons” is used for corks and as a nickname for Lyon restaurants. This problem is not treated here as NLP Natural Language Processing techniques are needed to solve it.

3 Validation

To validate the proposed approach, this work will take electricity in France as an example.

Data is collected from 32 weather stations distributed all over France from 2014 till 2019. Data contains 313 variables representing different taken measures for temperature, humidity, luminosity, and wind. Around 49178 measurements for each variable are taken in different timestamps different for each variable. In addition, they provide different corresponding profiles estimations (residential/professional). Data from 2014 till 2018 are taken to train a model. Data from 2019 are taken to test and validate the model.

There are some important steps to clean the data before being used to learn a model:

- Re-sample the data to have the same granularity (30 min, 1 hour, ...) for all variables to obtain the equivalent number of records for each sensor.In this study the data were re-sampled each hour.
- Extract information from the timestamp like hours, the day of the week, the month of the year, year, working days, holidays, and weekends.
- Apply the ”one-hot” coding on the adapted data to prepare them for learning. This is a method to quantify categorical data. In short, this method produces a vector with a length equal to the number of categories in the data set. This way each category is completely independent of the other categories. For example, working days, weekends, and holidays are represented in three columns each day will have one column that corresponds to it and zero for the others.

The method that can be used for cross-validation with time-series data is continuous cross-validation. Start with a small subset of data for learning purposes, predict the subsequent data points, and then check the accuracy of the predicted data points. The predicted data points are then included in the next learning data set and subsequent data points are predicted. To make things intuitive, as illustrated in the figure 5.

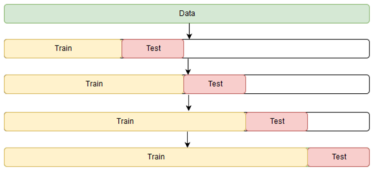

Fig. 5 Continuous cross validation

We take data between 2014 and 2018 for learning and data of 2019 for testing. The prediction horizon is one week (168 hours), then it is integrated into learning data and the process repeats itself.

Evaluating a machine learning model is an essential part of any project. A model can perform well when evaluated against one metric, such as accuracy score, but can perform poorly when evaluated against other metrics, such as logarithmic loss or any other metric. Two evaluation metrics are used:

– Max Error

$$\delta X_i = |X_{observed,i} - X_{actual,i}| \quad (1)$$

$$\forall i \in [1,n] \quad (2)$$

$$\max \delta X_i \quad (3)$$

Where n is the total number of measured vlaues.

– Root of the root mean square error RMSE:

$$RMSE = \sqrt{\Sigma_{i=1}^{n} \left(X_{observed,i} - X_{actual,i}\right)^2} \quad (4)$$

The evaluation metrics are used to compare the precision of different models before integrating the social data and after. This allows confirming the importance of the social data to improve the precision of estimating the load profile and the proposed approach.

3.1 Modeling

Three types of models are retained:

– Linear regression:
Linear regression is a regression model that seeks to establish a linear relationship between a variable, called the explained variable, and one or more variables called the explanatory variable [11].

– XGboost:
xgboost is short for eXtreme Gradient Boosting package. It is an efficient and scalable implementation of gradient boosting framework by Friedman [13]. The package includes efficient linear model solver and tree learning algorithm. It supports various objective functions, including regression, classification and ranking [12].

– Deep Learning:
Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture [9]. Unlike standard feed-forward neural networks, LSTM has feedback connections. It can process not only single data points (such as images), but also entire sequences of data (such as speech or video). It is well used for time series data [10].

Table 2 Accuracy scores without social data

	Max Error	RMSE
Linear Regression	0.403	0.0947
XGBoost	0.374	0.0918
LSTM	0.310	0.0441

Applying these three models only on the weather data, give the evaluation metrics presented in the table 2.

Collecting more than 182 events and more than 100 million tweets took many computing months. Those events are various (sport, political, culture, historical, tv shows, festivals, ...) and for different periods.

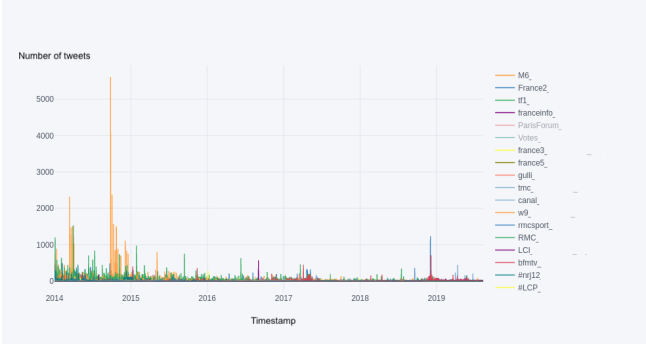

Fig. 6 Aggregated tweets for different Tv channels in France during 6 years

As an example, figure 6 presents different collected tweets for tv channels in France. It presents the number of tweets for each channel each hour in the period between 2014-2019.

After providing all these events plus the weather data and applying the models with the same conditions as before, we obtain the accuracy presented in the table 3.

Table 3 precision with social data

	Without Events		With Events		Improvement	Improvement
	Max Error	RMSE	Max Error	RMSE	Max Error in %	RMSE in %
Linear model	0.403	0.0947	0.279	0.080	12,4	1.4
XGBoost	0.374	0.0918	0.342	0.079	3.2	1.2
Deep LSTM	0.310	0.0441	0.271	0.041	3.9	0.3

By comparison between table 2 and table 3 An improvement of 12% on the max error and 1.4% on the RMSE can be remarked with the linear regression model. Moreover, this work insists more on the reduction of the max error. Indeed this method allows a priori to correct punctual calendar effects driven by events that the model does not know. There are not yet structural improvements (like trend capture) that improve the model globally. The global improvement of the model occurs because the algorithm drastically improved some very bad instants, and that removed the ”pernicious

inclusion" of a part of the data which has a very different variance from the rest of the data. This is extremely penalizing on linear regression type of algorithms, which explains the improvement of this model in particular.

The max error was reduced and RMSE was improved for the three models, which proves the importance of the social data and validate the proposed approach to profit from them.

4 Conclusions

Profile load energy estimation is essential for the electrical grid. That is why this work tries to tackle this problem by providing a new source of information from social media, especially Twitter to optimise the typical profile load estimation. The proposed method identifies the main events that impact energy consumption for each profile. This work offers a complete method to acquire the tweets, treat them, and explore them to enhance the estimations. The proposed approach is validated using the French electrical consumption data. After validating, the linear regression model is improved by 12% the max error, and all the three models show an improvement with the accuracy metrics.

References

1. T. Hong, Crystal ball lessons in predictive analytics, EnergyBiz Mag. 12 (2) (2015) 35–37.
2. Trends in twitter hashtag applications: Design features for value-added dimensions to future library catalogues, Chang, Hsia-Ching and Iyer, Hemalata, Library trends, 2012, Johns Hopkins University Press
3. https://www.internetlivestats.com/twitter-statistics/
4. Social media onderzoek 2013. http://www.newcom.nl/socialmedia. Accessed: 2013-05-20.
5. https://www.statista.com/forecasts/1144232/twitter-users-in-france
6. A tale of tweets: Analyzing microblogging among language learners,Lomicka, Lara and Lord, Gillian, System,40-1, pages:48–63, year:2012 Elsevier
7. Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. Briefings in bioinformatics, 15(5), 788-797.
8. Chandola, V., Banerjee, A. and Kumar, V., 2009. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), pp.1-58.
9. Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.
10. Hua, Y., Zhao, Z., Li, R., Chen, X., Liu, Z. and Zhang, H., 2019. Deep learning with long short-term memory for time series prediction. IEEE Communications Magazine, 57(6), pp.114-119.
11. Weisberg, Sanford. Applied linear regression. Vol. 528. John Wiley & Sons, 2005.
12. Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, and Hyunsu Cho. "Xgboost: extreme gradient boosting." R package version 0.4-2 1, no. 4 (2015): 1-4.
13. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)." The annals of statistics 28, no. 2 (2000): 337-407.
14. Eid, C., Codani, P., Perez, Y., Reneses, J., & Hakvoort, R. (2016). Managing electric flexibility from Distributed Energy Resources: A review of incentives for market design. Renewable and Sustainable Energy Reviews, 64, 237-247.
15. Duquesne, X., & Thaon, S. (2021). THERMOSENSIBILITY ESTIMATION OF MASS MARKET WITH SMART METERS IN FRANCE.
16. Karsenti, Laurent, and Philippe Daguzan. "Enedis approach for the roll-out of technical smart grid industrial solutions." CIRED-Open Access Proceedings Journal 2017, no. 1 (2017): 1077-1080.
17. Wang, Zeyu, and Ravi S. Srinivasan. "A review of artificial intelligence based building energy prediction with a focus on ensemble prediction models." 2015 Winter Simulation Conference (WSC). IEEE, 2015.
18. Zhang, Liang, et al. "A review of machine learning in building load prediction." Applied Energy 285 (2021): 116452.
19. Wang, Zeyu, and Ravi S. Srinivasan. "A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models." Renewable and Sustainable Energy Reviews 75 (2017): 796-808.
20. RTE-open-data: https://www.rte-france.com/en/eco2mix/electricity-consumption-france.
21. Alyafi, Amr Alzouhri, et al. "Differential explanations for energy management in buildings." 2017 computing conference. IEEE, 2017.