Appendix E

Paper E - Evaluating Anomaly Detection Algorithms through different Grid scenarios using k-Nearest Neighbor, iforest and Local Outlier Factor

Nils Jakob Johannesen and Mohan Kolhe and Morten Goodwin

Faculty of Engineering and Science, University of Agder, PO Box 422, NO 4604 Kristiansand, Norway.

***Abstract* - Detection of anomalies based on smart meter data is crucial to identify potential risks and unusual events at an early stage. The available advanced information and communicating platform and computational capability renders smart grid prone to attacks with extreme social, financial and physical effects. The smart network enables energy management of smart appliances contributing support for ancillary services. Cyber threats could affect operation of smart appliances and hence the ancillary services, which might lead to stability and security issues. In this work, an overview is presented of different methods used in anomaly detection, performance evaluation of 3 models, the k-Nearest Neighbor, local outlier factor and isolated forest on recorded smart meter data from urban area and rural region.**

Keywords cybersecurity, anomaly detection, smart grid, local outlier factor, isolated forest

E.1 Introduction

The smart electrical energy network grid requires more accurate demand and prediction for control and managing the demand in coordination with intermittent renewable energy

sources [1]. The smart grid will require advanced control and management, including reliable forecasting to anticipate the events involved in dispatching, control and management of the operating grid. The accurate load prediction can help in managing peak demand and to reduce overall capital cost investment [2].

From the field of Artificial Intelligence (AI) a tool to process meaningful relation of complex big data by uncovering structures and patterns is learned through training with Machine Learning (ML). When presented with new data the machine can learn to perform a task without the need of re-programming [9]. ML can provide electrical load demand forecasting, giving information about future loads, which provides essential input to other applications such as Demand Response, Topology Optimization and Anomaly Detection, facilitating the integration of intermittent clean energy sources. Anomaly detection can been used as a first step in data cleaning process and has been known to enhance any forecasting algorithm [4][5].

The data used is of such an amount, that it is not possible to do so manually or by visual inspection, and there is a need for an efficient, automated and accurate anomaly detection methods [48]

An anomaly is defined as a deviation from an established normal pattern. Spotting an anomaly depends on the ability to defy what is normal. Anomaly detection systems aim at finding these anomalies. Anomaly detection systems are in high demand, despite the fact that there is no clear validation approach. These systems rely on deep domain expertise. Cyber threats could affect the ancillary services that are being delivered from the aggregators, which might lead to stability and security issues resulting in brownout or massive blackouts [7]. Large scale monitoring using the supervisory control and data acquisition (SCADA) makes it vulnerable to cyber attacks. Anomaly detection can be used for preventing possible cyber-attacks.

The buses in a power system is in normal operation in the same state, it is reasonable that an anomaly exists if one bus deviates from the others [8] The implementation of two way communication by the use of sensors and intelligent agents such as advanced metering infrastructure as well as load aggregation, make these attractive objects for cyber attacks. Sensors can be penetrated using a Trojan Horse, to manipulate the adversary inside the control platform, and change reference inputs in controllers of components. The attacker can here change acquisition gains, that create bias in the measurements report.

In the distributed power network the attack can disrupt the frequency regulation, voltage stability and the power flow management [9].

It is necessary to investigate different computing methods, and their applications in anomaly detection. In this work the performance evaluation of 3 models is analysed on recorded smart meter data from urban area and rural region.

This article is organised in sections: Section E.2 the literature review. Theory in Section E.3, user scenarios in Section E.4, results in Section E.5, and conclusion in Section E.6.

E.2 Review on Anomaly Detection

Anomaly detection is done on any time series data. Various anomalies can be detected in historic time series data, due to human error, false meter measurement, inaccuracies in data processing and failure of delivery due to extreme weather or other failures. A two-stage method is proposed in reference [48] combining two probabilistic anomaly detection approaches for identifying anomalies in time series data of natural gas. Exogenous variables are known to influent the electrical load consumption [10], and loads are identified accordingly as baseload, intermediate load and peak load [11]

An autoregressive integrated moving average with exogenous inputs (ARIMAX) model is used to extract weather dependency to find the residuals, then through hypothesis testing the extremities, maximum and minimums are found [49]. This procedure was reproduced, with linear regression finding the residuals and a Bayesian maximum likelihood classifier to identify anomalies [48].

A data-mining based framework using DBSCAN was used to detect anomalies in office buildings. The framework is aimed to identify typical electricity load patterns and gain knowledge hidden in the patterns and to potentially be used in an early fault detection of anomalous electricity load profiles[50]. Also to detect anomalies of electricity consumption in office buildings an improved kNN is proposed, ikNN, to automatically classify consumption footprints as normal or abnormal [51].

Dynamic Bayesian Networks and Restricted Boltzman Machine has been proposed for anomaly detection in large-scale smart grids. Simulated on the IEEE 39, 118, and 2848 bus systems the results were verified [52]. Real-Time Mechanism for detecting FDIA analyzed the change of correlation between two phasor measurement units parameters using Pearson correlation coefficient on IEEE 118 and 300-bus sytems [53]. Machine learning techniques have been highlighted for theire ability to differentiate between cyberattacks and natural disturbances. By a simulating a variety of scenarios the ability for One R, Random Forest, Naive Bayes and J-Ripper to recoginze attacks was investigated: Short Circuit faults; location is represented by the percentage range, Line maintenance; identified through remote relay trip command, Remote tripping command injection; the attacker operates the relay remotely that causes a breaker to open, Relay setting change; the attacker misconfigures the relay settings to cause maloperation of relays, FDIA; attacker manipulates measurements sensors. The simulated scenarios was grouped into classes; natural events, attack events, and no events [54].

In concept drift, models are inaccurate due to change in the underlying data [56]. Thus the observation can be a result of an improved energy system, and not anomaly[57].

E.3 Anomaly Detection using Machine Learning Algorithms

3 different models is compared for anomaly detection in the different grid scenarios:

E.3.0.1 k-Nearest Neighbor

The k-nearest neighbor (kNN) regressor, which is non-paramteric, relying on its own table look-up and mathematical foundation, and highly non-linear.

$$y_{knn}(x) = \frac{1}{K}\sum_{k=1}^{K} y_k \text{ for K nearest neighbours of x} \tag{E.1}$$

The kNN-classifier is illustrated in Fig. E.1, where the left diagram with a small encirclement options for k = 1, where simply the nearest neighbor decides the class of prediction, whilst in the right diagram in Fig. E.1, the number of k is increased to more then one [70].

Using k = 1 can lead to false prediction, and a set of kNNs is often used. When classifying the dependent variable is categorical, it can easily be made numerical by regression. The kNN regressor makes a regression based on the number of kNNs to minimize false predictions. The model considers a range of different kvalues to find the optimal value. The kNN regressor needs thorough pre-processing and feature engineering to limit the effect of noise caused by irrelevant features, and is, therefore, dependent on finding the appropriate distance model [71].

E.3.0.2 Isolation Forest

The Isolation Forest algorithm is composed of several isolation trees (iTres) Isolation forest takes advantage of the nature of anomalies which are less frequent than regular observations and different from those in terms of values to isolate those. Iforest can deal with large scale data quickly in a simplified way. It builds an ensemble of decision trees (iTrees) for a given data set. Clustering is done using binary tree clustering. Anomalies tend to be isolated closer to the root of the binary tree. Partitions are created using a split value between the minima and maxima of a randomly selected feature. The algorithm then tries to separate each point in the data [82] [83] [84] [85].

E.3.0.3 Local Outlier Factor

Local Outlier Factor (LOF) is a density based anomaly detection algorithm introduced in 2000 [26]. LOF compares the local density of a point to the local density of k of its neighbors. By comparing the local density of a point to the local density of its neighbors one can identify point that have substantially lower density than its neighbors. These

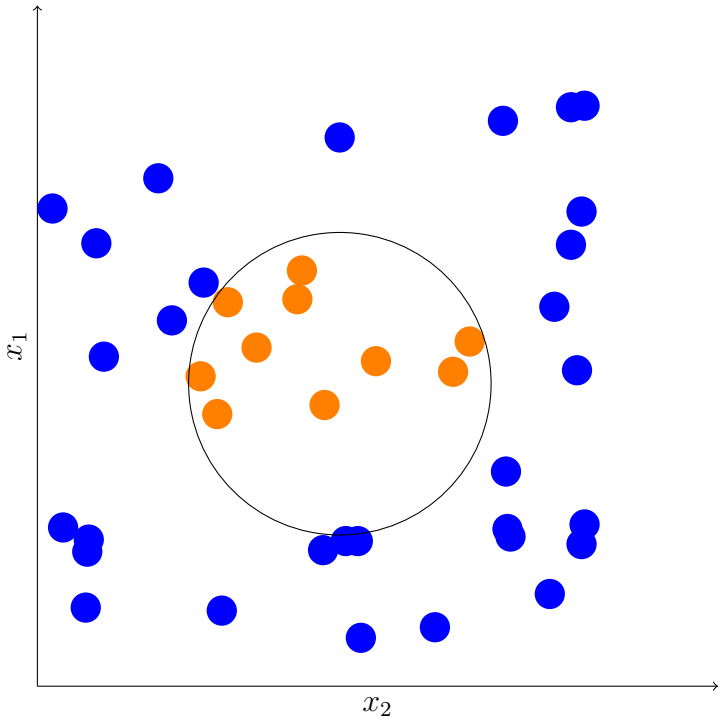

Figure E.1: k-Nearest Neighbour classifying based on the k'th observation.

points are considered outliers. LOF uses the k-distance to a point as in kNN, to find the Local Reachability Density (LRD), where a point is most likely to be found. The sum of LRD is then used to find LOF for the point z, as in Equation (E.2):

$$\mathrm{LOF_k}(\mathbf{z}') = \sum_{\mathrm{z}\in \mathrm{N_k}(\mathbf{z}')} \frac{lrd_{\mathrm{k}}(\mathbf{z})}{lrd_k(\mathbf{z}')} / \left\| N_k(\mathbf{z}') \right\| \tag{E.2}$$

[86]

E.4 User case scenarios

In this work 3 different models is used to detect anomalies in two different grid scenarios:

E.4.1 Scenario 1

New South Wales, Sydney region electrical load profile data set [105] includes meteorological parameters (e.g. DryBulb and WetBulb Temperature, Humidity, Electricity price and time of use) [106]. Data is gathered from 2006-2011. The overall energy mix in New South Wales consists mainly of Coal, Natural Gas, Hydro and other renewable energy sources. Fig. E.2 illustrates the New South Wales distribution network.

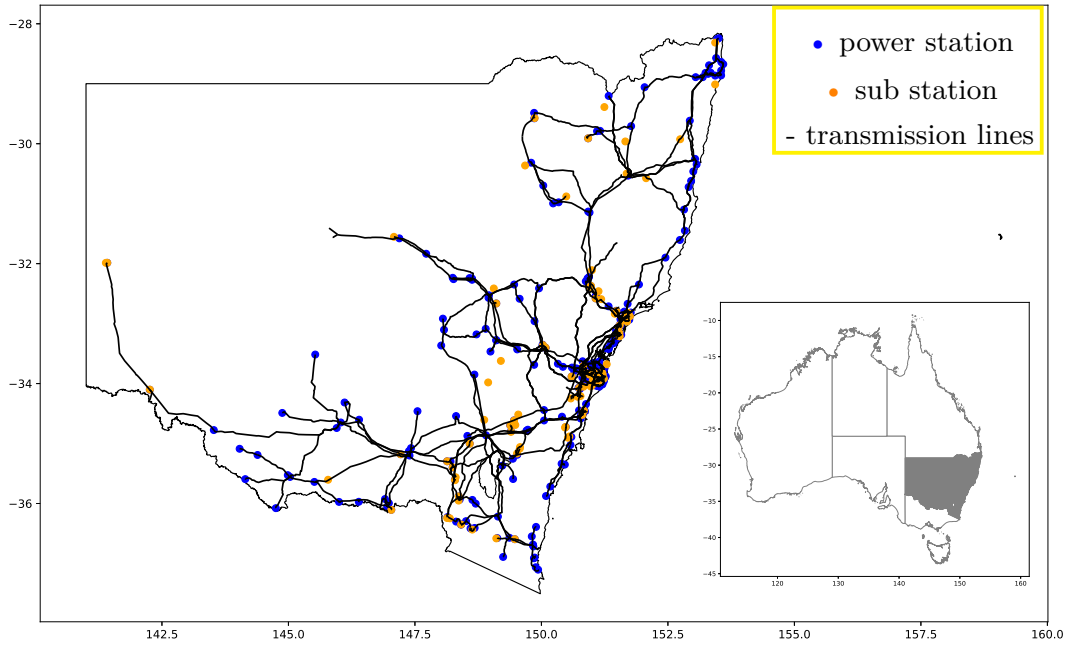

Figure E.2: New South Wales Power system, indicating transmission lines, power stations, and substations

E.4.2 Scenario 2

From rural cabin area in Bjønntjønn, Telemark, Norway, the electrical load demand consumption profile is collected from smart meters. Weather data is collected from surrounding weather information statins in the surrounding area. The land owner of the area wants to realize the project 'Bjønntjønn Grønn' (Bjønntjønn Green). The project seeks through different initiatives to make the cabin area 'green', with power from local hydro power stations, possibility of electric vehicle charging and operation of the load consumption related to the power intensive usages. The land owner has currently an application to get license from The Norwegian Energy Regulatory Authority (NVE) to run hydro power stations in the area, with a total production of 10,08 GWh [108]. In the fall of 2021 NVE approved an appplication for a Tesla Supercharger from Tesla Norway, situated in the center of Treungen, an 8 km drive from the planned Bjønntjønn hydro power station [109] [110].

The rural area network of a typical Norwegian holiday resort cabin area, Bjønntjønn Cabin Area. It comprises 125 cottages with a peak demand of 478 kW. As for today, this cabin area is grid connected, but a microgrid solution involving photovoltaics and energy storage is also considered. In the summer of 2020 the land owner presented plan of building 445 new cabins in the area [111].

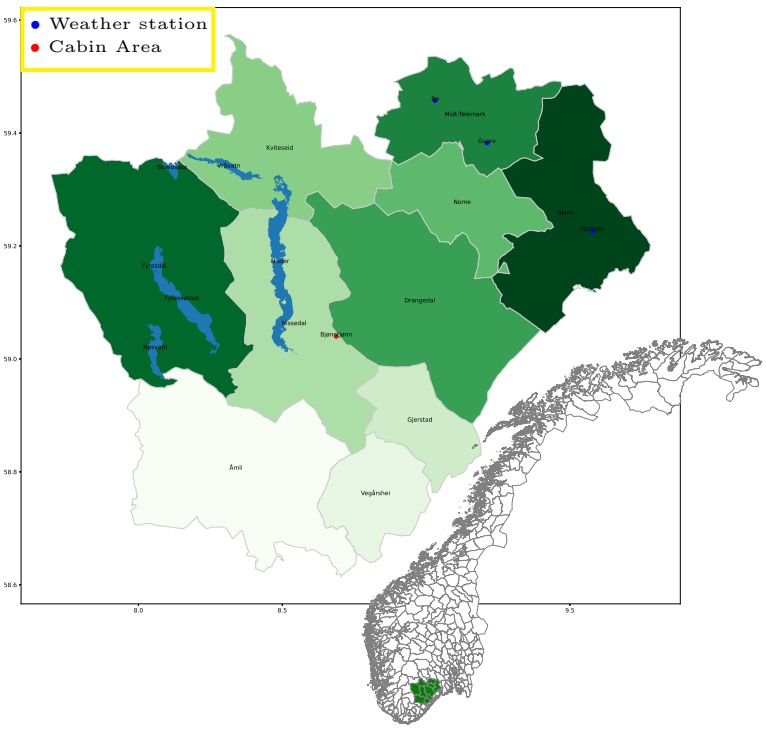

Figure E.3: Rural Area in the South-East of Norway, with the situation of cabin area and nearby weather stations

Rural electrification is very different from the urban area electrical consumption, due to diversified energy mix and overall conditions. A variety of case studies is necessary for a generic approach, although each system requires an independent approach. The Nordic market is much reliant on hydropower, as Norway's share of hydropower is 95.8 % [112]. Norway also has the highest integration of Electric Vehicles, and this faces challenges to the grid. This is especially a case in the rural area, where capacity is low, and the electrical vehicle charging poses a liability to the grid. In these cases, a micro-grid solution can aid the low-capacity network, with implementation of distributed generators, in combination with energy storage.

When examining the general load profile of all Norwegian Holiday Cabins, a clear trend is observed in the user behaviour. The load demand for Norwegian Cabins has increased their total consumption from 0.7 TWh in 1993 to 2.3 TWh in 2016. Although the consumption tripled and has been only 1.8 % of the total Norwegian load demand in 2016 [35]. Statistics Norway concludes in the 2018 report, that the increasing trend is due to the general development, and that more Norwegians have bought cottages in rural areas, such as mountains and seaside. Also, more cottages have been electrified in this period [112].

In the Bjønntjønn Cabin Area, to deal with the ever-increasing penetration of electric vehicles, photovoltaic system together with energy storage could be a scenario for the future rural electrification. For the Nordic rural area network, a microgrid solution can improve the electrical network capacity of the rural area, despite challenges from power demanding operations as electric vehicle charging. Since the electric vehicle will not be used mostly of the holiday resort area, the battery pack of the vehicle is be considered as the battery bank for the microgrid. When the state of charge (SOC) of the battery reaches a certain threshold level, it will be considered as a prosumer for the micro grid and be able to contribute to electrical supply and stability.

E.5 Results and Discussion

The results of kNN, iforest and LOF on urban area data, are shown in Fig. E.4, E.5 and E.6, and from rural region data in Fig. E.7, E.8 and E.9. The results are depicted with a 0.0005 amount of contamination of the data set, this is the proportion of outliers in the data set. Used when fitting to define the threshold on the decision function [36].

It is observed that the anomaly detection for the two grid scenarios are different, for the rural region most of the anomalies where observed in the latter timeline of the data concentrated in the last year of the collected data. For the urban area data the anomalies are spread out over the entire timeline. In Table E.1, it is shown that the frequency of detected anomalies where considerably higher for the rural area load demand than for the urban area load demand. When observing the anomalies detected based on the algorithm the results in Table E.1 are consistent.

algorithm	urban	rural
kNN	44	10
iforest	35	25*
lof	44	21

Table E.1: Results using fraction 0.0005, except * = 0.0006

Observing from these case scenarios the incidents of detected anomalies are more data driven, then exceptions in the algorithms. It is observed that there are 3 anomalies, where the recorded electrial load demand is zero, in the rural region dataset that the iforest and LOF did not detect. This was only detected by kNN, see Fig. E.7.

When comparing the 3 algorithms tested on the urban area data it is observed that kNN and isolated forest finds a threshold value, based in the mentioned fraction of contamination, and separates a lower and upper bound, whilst the density based LOF finds anomalies at several ranges of the dataset, see Fig. E.4, E.5 and E.6.

When visually inspecting results in Fig. E.4, E.5, E.6, E.7, E.8 and E.9, it is observed that from the domain knowledge of smart energy systems the LOF is able to detect observations that could not have detected by visual inspection alone, in contrast to kNN and iforest. Whereas kNN and iforest excludes an upper and lower bound, the LOF is density based and separates out anomalies amidst in the data. The capability that LOF has to identify anomalies amidst the data will together with the deep domain knowledge is an advantage when detecting anomalies in smart meter data.

E.6 Conclusion

Detection of anomalies based on smart meter data is crucial to identify potential risks and unusual events at an early stage. An anomaly is defined as a deviation from an established normal pattern. Spotting an anomaly depends on the ability to defy what is normal.Cyber threats could affect operation of smart appliances and hence the ancillary services, which might lead to stability and security issues. In this work is evaluated the performance of 3 models, the k-Nearest Neighbor, local outlier factor and isolated forest on recorded smart meter data from urban area and rural region. Observed that from the domain knowledge of smart energy systems the LOF is able to detect observations that could not have detected by visual inspection alone, in contrast to kNN and iforest. Whereas kNN and iforest excludes an upper and lower bound, the LOF is density based and separates out anomalies amidst in the data. The capability that LOF has to identify anomalies amidst the data will together with the deep domain knowledge is an advantage when detecting anomalies in smart meter data. The anomaly detection based on machine learning algorithms gives a fast response to potential anomalies.

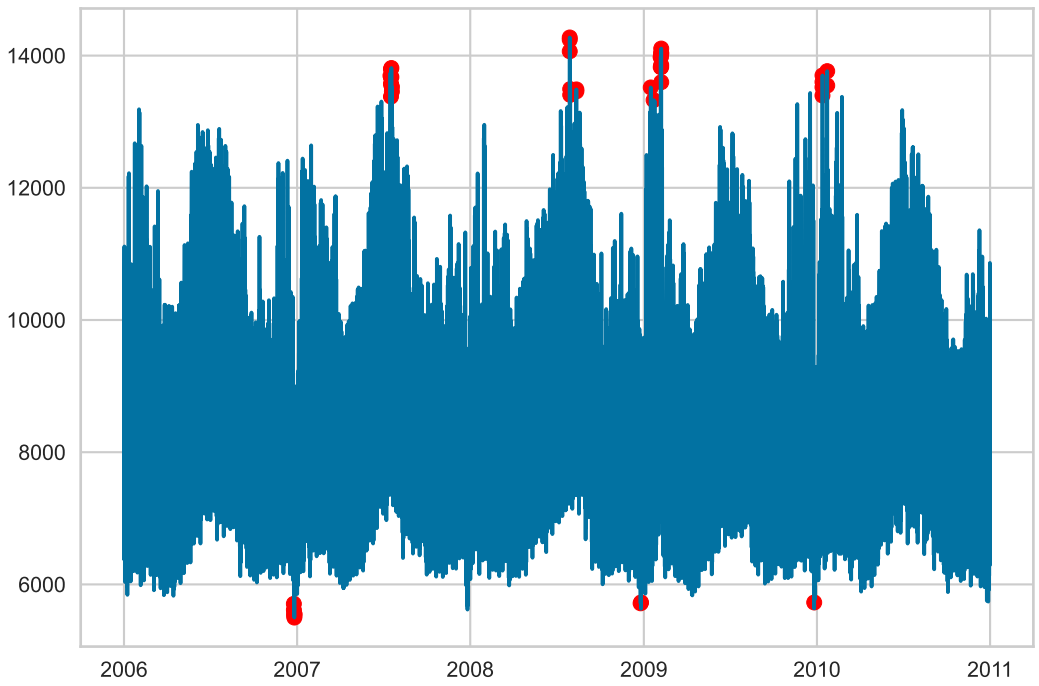

Figure E.4: Anomaly detected outliers marked in red using kNN, fraction = 0.0005

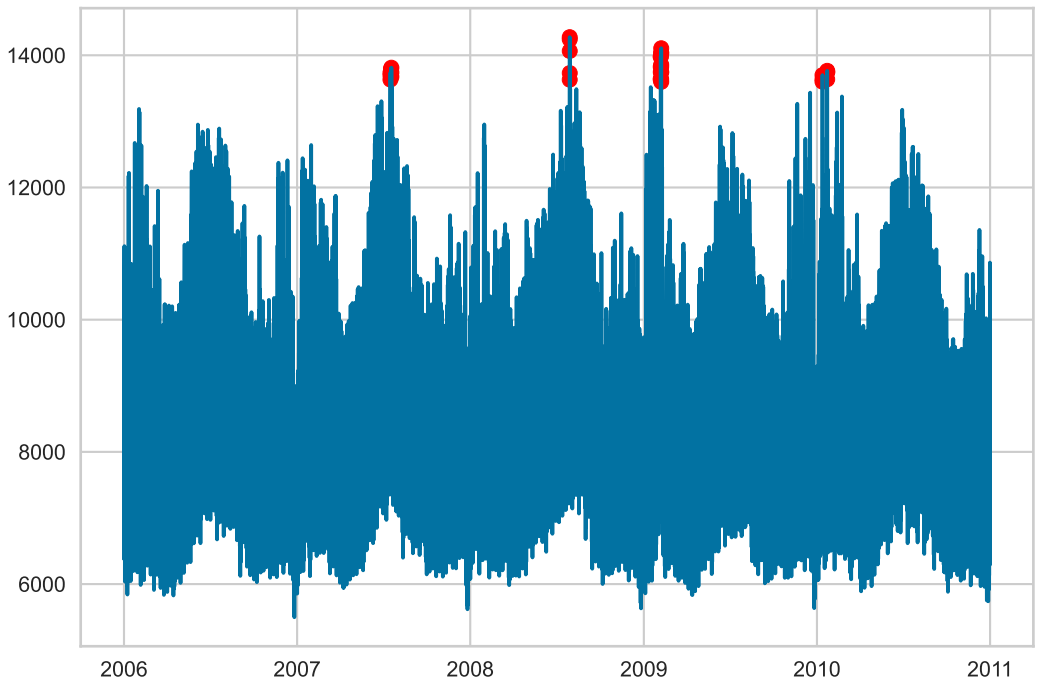

Figure E.5: Anomaly detected outliers marked in red using iforest, f = 0.0005

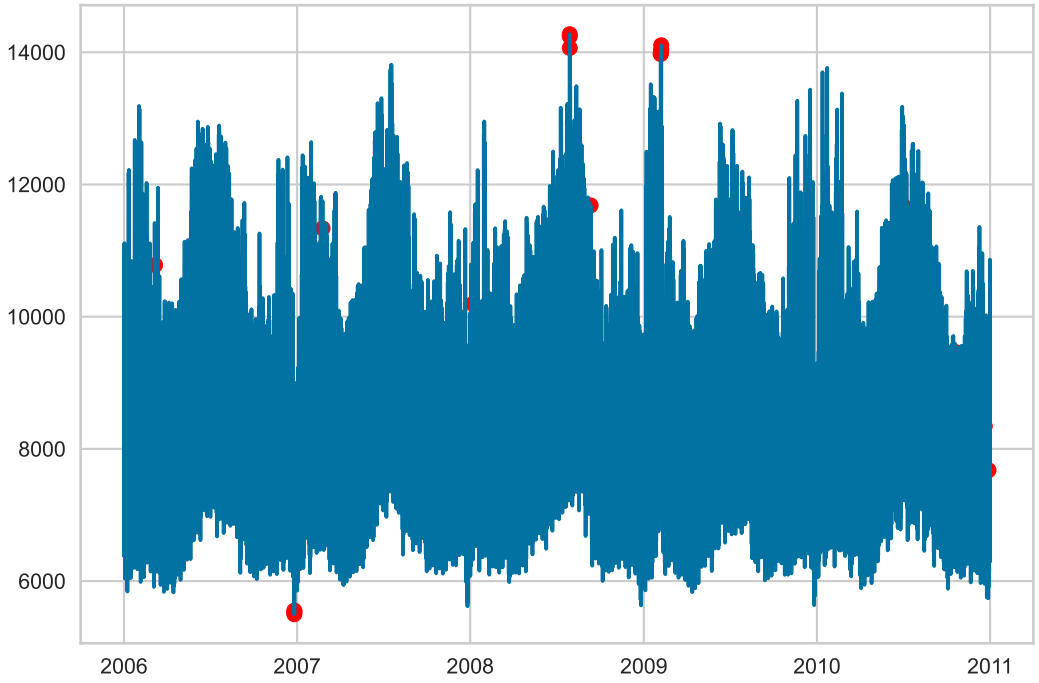

Figure E.6: Anomaly detected outliers marked in red using LOF, f= 0.0005

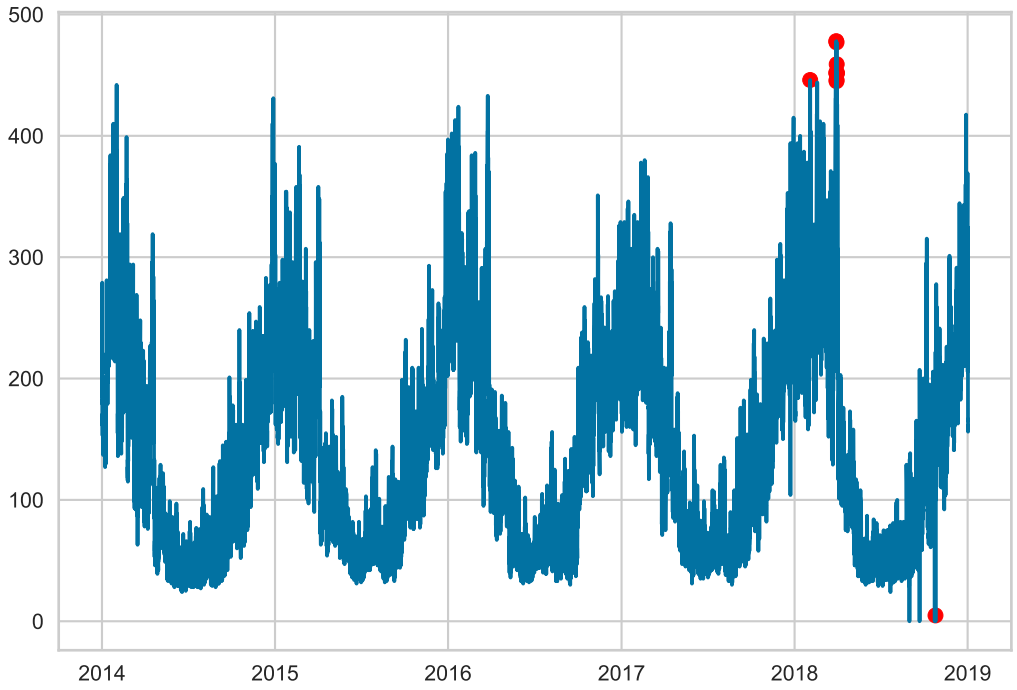

Figure E.7: Anomaly detected outliers marked in red using kNN, fraction = 0.0005

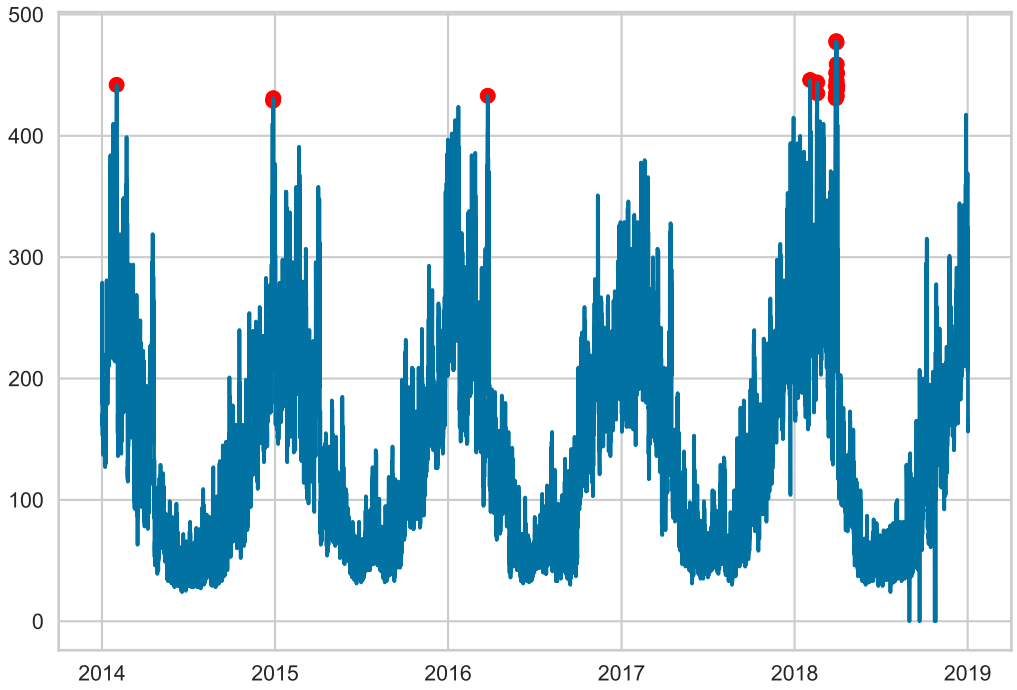

Figure E.8: Anomaly detected outliers marked in red using iforest, f = 0.0006

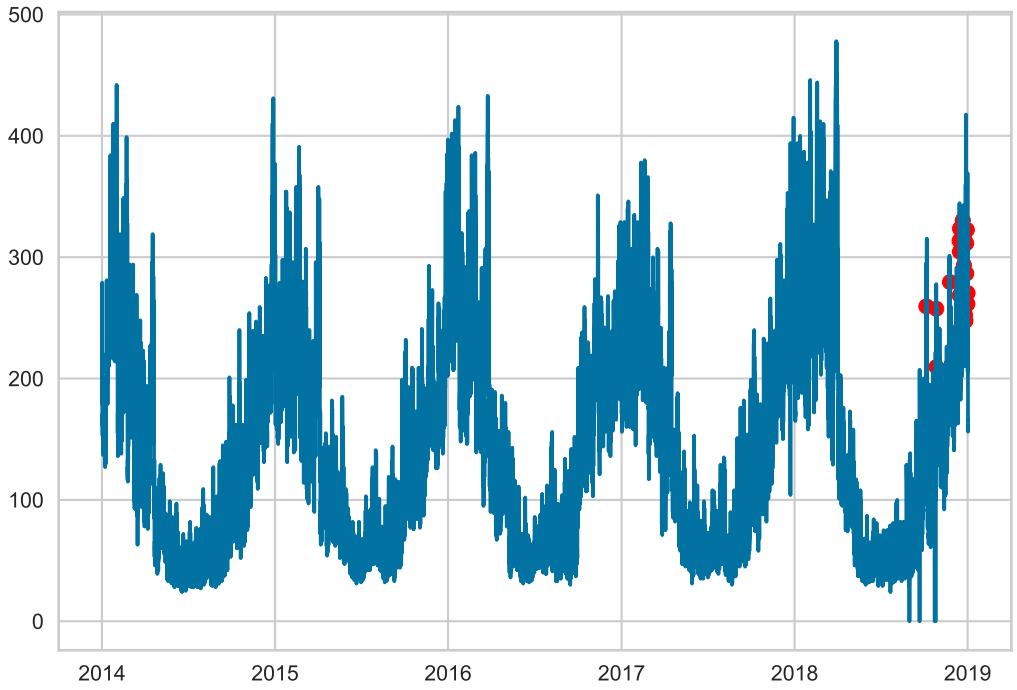

Figure E.9: Anomaly detected outliers marked in red using LOF, f= 0.0005

Bibliography

[1] Pierluigi Siano. Demand response and smart grids—a survey. *Renewable and Sustainable Energy Reviews*, 30:461–478, 2014.

[2] Stephen Haben, Siddharth Arora, Georgios Giasemidis, Marcus Voss, and Danica Vukadinovic Greetham. Review of low-voltage load forecasting: Methods, applications, and recommendations. *arXiv preprint arXiv:2106.00006*, 2021.

[3] Peyman Razmi, Mahdi Ghaemi Asl, Giorgio Canarella, and Afsaneh Sadat Emami. Topology identification in distribution system via machine learning algorithms. *PLOS ONE*, 16(6):1–20, 06 2021.

[4] Ih Chang, George C. Tiao, and Chung Chen. Estimation of time series parameters in the presence of outliers. *Technometrics*, 30(2):193–204, 1988.

[5] Nils Jakob Johannesen, Mohan Kolhe, and Morten Goodwin. Deregulated electric energy price forecasting in nordpool market using regression techniques. In *2019 IEEE Sustainable Power and Energy Conference (iSPEC)*, pages 1932–1938, 2019.

[6] Hermine N. Akouemo and Richard J. Povinelli. Probabilistic anomaly detection in natural gas time series data. *International Journal of Forecasting*, 32(3):948–956, 2016.

[7] Kaixing Huang, Chunjie Zhou, Yu-Chu Tian, Shuanghua Yang, and Yuanqing Qin. Assessing the physical impact of cyberattacks on industrial cyber-physical systems. *IEEE Transactions on Industrial Electronics*, 65(10):8153–8162, 2018.

[8] Shengyuan Liu, Yuxuan Zhao, Zhenzhi Lin, Yilu Liu, Yi Ding, Li Yang, and Shimin Yi. Data-driven event detection of power systems based on unequal-interval reduction of pmu data and local outlier factor. *IEEE Transactions on Smart Grid*, 11(2):1630–1643, 2020.

[9] S. Sahoo, T. Dragičević, and F. Blaabjerg. Cyber security in control of grid-tied power electronic converters–challenges and vulnerabilities. *IEEE Journal of Emerging and Selected Topics in Power Electronics*, pages 1–1, 2019.

[10] Nils Jakob Johannesen, Mohan Kolhe, and Morten Goodwin. Relative evaluation of regression tools for urban area electrical energy demand forecasting. *Journal of Cleaner Production*, 218:555–564, 2019.

[11] Nils Jakob Johannesen and Mohan Lal Kolhe. Application of regression tools for load prediction in distributed network for flexible analysis. In *Flexibility in Electric Power Distribution Networks*. CRC Press, 2021.

[12] Hermine N. Akouemo and Richard J. Povinelli. Time series outlier detection and imputation. *2014 IEEE PES General Meeting — Conference & Exposition*, pages 1–5, 2014.

[13] Xue Liu, Yong Ding, Hao Tang, and Feng Xiao. A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. *Energy and Buildings*, 231:110601, 2021.

[14] Yassine Himeur, Abdullah Alsalemi, Faycal Bensaali, and Abbes Amira. Smart power consumption abnormality detection in buildings using micromoments and improved k-nearest neighbors. *International Journal of Intelligent Systems*, 36(6):2865–2894, 2021.

[15] Hadis Karimipour, Sandra Geris, Ali Dehghantanha, and Henry Leung. Intelligent anomaly detection for large-scale smart grids. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pages 1–4, 2019.

[16] Youbiao He, Gihan J. Mendis, and Jin Wei. Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism. *IEEE Transactions on Smart Grid*, 8(5):2505–2516, 2017.

[17] Manikant Panthi. Anomaly detection in smart grids using machine learning techniques. In *2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*, pages 220–222, 2020.

[18] Giuseppe Fenza, Mariacristina Gallo, and Vincenzo Loia. Drift-aware methodology for anomaly detection in smart grid. *IEEE Access*, 7:9645–9657, 2019.

[19] Alexandra L'heureux, Katarina Grolinger, Hany F Elyamany, and Miriam AM Capretz. Machine learning with big data: Challenges and approaches. *Ieee Access*, 5:7776–7797, 2017.

[20] D. Patidar, Bhavin C. Shah, and M. Mishra. Performance analysis of k nearest neighbors image classifier with different wavelet features. *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*, pages 1–6, 2014.

[21] Sholom M. Weiss and Nitin Indurkhya. Rule-based machine learning methods for functional prediction. *J. Artif. Int. Res.*, 3(1):383–403, December 1995.

[22] Ehdieh Khaledian, Shikhar Pandey, Pratim Kundu, and Anurag K. Srivastava. Real-time synchrophasor data anomaly detection and classification using ¡italic¿isolation forest, kmeans, and loop¡/italic¿. *IEEE Transactions on Smart Grid*, 12(3):2378–2388, 2021.

[23] Wei Mao, Xiu Cao, Qinhua zhou, Tong Yan, and Yongkang Zhang. Anomaly detection for power consumption data based on isolated forest. In *2018 International Conference on Power System Technology (POWERCON)*, pages 4169–4174, 2018.

[24] Armin Aligholian, Mohammad Farajollahi, and Hamed Mohsenian-Rad. Unsupervised learning for online abnormality detection in smart meter data. In *2019 IEEE Power Energy Society General Meeting (PESGM)*, pages 1–5, 2019.

[25] Saeed Ahmed, Youngdoo Lee, Seung-Ho Hyun, and Insoo Koo. Unsupervised machine learning-based detection of covert data integrity assault in smart grid networks utilizing isolation forest. *IEEE Transactions on Information Forensics and Security*, 14(10):2765–2777, 2019.

[26] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, page 93–104, New York, NY, USA, 2000. Association for Computing Machinery.

[27] Yifan Ou, Bin Deng, Xuan Liu, and Ke Zhou. Local outlier factor based false data detection in power systems. In *2019 IEEE Sustainable Power and Energy Conference (iSPEC)*, pages 2003–2007, 2019.

[28] AEMO. National electricity market data - nem, 2021.

[29] Vasudev Dehalwar, Akhtar Kalam, Mohan Lal Kolhe, and Aladin Zayegh. Electricity load forecasting for urban area using weather forecast information. In *2016 IEEE International Conference on Power and Renewable Energy (ICPRE)*, pages 355–359, 2016.

[30] NVE. Bjønntjønn kraftverk og pumpestasjon. https://www.nve.no/konsesjon/konsesjonssaker /konsesjonssak?id=8386type=V-1, 2021.

[31] NVE. Tesla ladestasjon i treungen. https://www.nve.no/konsesjon/konsesjonssaker/konsesjonssak/ 1, 2021.

[32] Jarle Pedersen. Tesla bygger ny supercharger i telemark. https://www.tvedestrandsposten.no/tesla-bygger-ny-supercharger-i-telemark/s/5-50-1252227, 2022.

[33] Camilla Moen. Planlegger gigantutbygging med 445 nye hytter på gautefallheia: - noen mener det er kontroversielt. https://www.kv.no/planlegger-gigantutbygging-med-445-nye-hytter-pa-gautefallheia-noen-mener-det-er-kontroversielt/s/5-63-271204, 2021.

[34] M. Holstad T. Aanensen. Supply and consumption of electricity in the period 1993-2016. https://www.ssb.no/en/energi-og-industri/artikler-og-publikasjoner/supply-and-consumption-of-electricity-in-the-period-1993-2016, 2017.

[35] Nils Jakob Johannesen, Mohan Lal Kolhe, and Morten Goodwin. Load demand analysis of nordic rural area with holiday resorts for network capacity planning. In *2019 4th International Conference on Smart and Sustainable Technologies (SpliTech)*, pages 1–7, 2019.

[36] Moez Ali. *PyCaret: An open source, low-code machine learning library in Python*, April 2020. PyCaret version 1.0.0.