

Probabilistic Anomaly Detection in Energy Time Series Data

Hermine N. Akouemo^{a,*}, Richard J. Povinelli^a

^a*Department of Electrical and Computer Engineering, Marquette University, Milwaukee WI, USA*

Abstract

This paper introduces probabilistic approaches for anomaly detection, specifically in natural gas time series data. In the energy field, there are various types of anomalies, each induced by diverse causes and sources. The causes of a set of anomalies are examined and categorized. A Bayesian maximum likelihood classifier learns the temporal structure of known anomalies. Given a previously unseen time series data, the system detects anomalies using a linear regression model with weather inputs. Then the anomalies are tested for false positives and classified using the Bayesian classifier. The system also can identify anomalies of unknown origin. Thus, the likelihood of a data point being anomalous is given for anomalies of both known and unknown origins. The anomaly detection algorithms are tested on a reported natural gas consumption data set.

Keywords: anomaly detection, data cleaning, Bayesian classifier, linear regression, natural gas data

1. Introduction

Anomaly detection, which is the first step of the data cleaning process, improves the accuracy of forecasting models. The data sets are cleaned for the purpose of being used to train forecasting models. Training a forecasting model on time series containing anomalous data usually results in an erroneous model because the parameters and variance of the model are affected (Chang et al., 1988). There are various anomalies in energy demand time series such as human reporting error, data processing error, failure of an energy demand delivery subsystem due to power outages or extreme weather, or faulty meter measurements. Manually examining energy time series for all anomaly causes is a tedious task, and one that is infeasible for large data sets. Thus, there is a need for automated and accurate algorithms for anomaly detection.

This paper combines two approaches for the detection of anomalies. The first approach determines the probability of a data point being anomalous. This is done using an energy domain knowledge derived linear regression model and a geometric probability distribution of the residuals. The second approach is to train a Bayesian maximum likelihood classifier based on the categories of anomalies identified by the first approach. For new data points, the classifier calculates the maximum likelihood of the data points given the prior classes and use the likelihood values to distinguish between false positives and true anomalies. If a data point is anomalous, the classifier

*Corresponding author

Email addresses: hermine.akouemokengmokenfack@marquette.edu (Hermine N. Akouemo), richard.povinelli@marquette.edu (Richard J. Povinelli)

is also able to report the category of anomalies to which the point belongs. The contribution of the proposed algorithms is their ability to incorporate domain knowledge in techniques developed to detect anomalies efficiently in energy time series data sets.

Previous work in anomaly detection using probabilistic and statistical methods is discussed in Section 2. Section 3 presents the types of anomalous data encountered in the energy domain. A detailed description of our models and algorithms is presented in Section 4. The experiments and results are presented and analyzed in Section 5.

2. Previous Work

Anomalous data is data that we do not have (missing data), that we had and then lost (manual reporting error, bad query), or that deviates from the system expectations (extreme weather and power outages consumption days) (McCallum, 2012). Markou and Singh (2003) present a survey of anomaly detection techniques ranging from graphical (box plot method) to more complex techniques such as neural networks. Statistical approaches for anomaly detection are based on modeling data using distributions and looking at how probable it is that the data under test belongs to the distributions. The approaches presented in this paper combine linear regression and distribution functions for the detection of anomalies in energy time series. Then, Gaussian mixture models (GMM) are used for modeling training subsets containing anomalous features. The likelihood of a testing data point of belonging to a prior subset is calculated using the GMM distributions, and the data point is classified.

Linear regression is a statistical method widely used for energy forecasting (Charlton and Singleton, 2014; Hong, 2014). It also has been used in combination with a penalty function for outlier detection (Zou et al., 2014). The disadvantage of using a penalty function is that the design of the tuning parameters needs to be precise. Therefore, penalty function strategies do not always guarantee practical results. The advantage of linear regression is that with dependent variables well defined, the technique is able to extract time series features (Magld, 2012). Lee and Fung (1997) show that linear and nonlinear regressions also can be used for outlier detection, but they used a 5% upper and lower threshold limit to choose outliers after fitting, which yielded many false positives for very large data sets. Linear regression also has been combined with clustering techniques for the detection of outliers (Adnan et al., 2003). In this paper, linear regression is used to extract weather features from the time series data and compute the residuals of the data.

Bouguessa (2012) proposed a probabilistic approach that uses scores from existing outlier detection algorithms to discriminate automatically between outliers and inliers on data sets to avoid false positives. Statistical approaches such as the Gaussian mixture model (Yamanishi et al., 2000), distance-based approaches such as the k -nearest neighbor (Ramaswamy et al., 2000), and density-based approaches such as the Local Outlier Factor (LOF) (Breunig et al., 2000) are existing techniques that Bouguessa (2012) uses for his ensemble model. Each techniques provide a score to every data point, and the results are combined to decide which outliers are true and which are false. Yuen and Mu (2012) proposed a method that calculates the probability for a data point being an outlier by taking into account not only the optimal values of the parameters obtained by linear regression but also the prediction error variance uncertainties.

Gaussian mixture model approaches also have been used for outlier detection and classification. Tarassenko et al. (1995) studied the detection of masses in mammograms using Parzen windows and Gaussian mixture models. The authors showed that GMMs do not work well when the number of training samples is very small and that using Parzen windows yielded false positives. Gaussian

mixture models also were used by Tax and Duin (1998) to reject outliers based on the data density distribution. They showed that the challenge to using GMMs is selecting the correct number of kernels. However, an approach developed by Povinelli et al. (2006) demonstrated that transforming the signal from a time domain into a phase space improves the GMM classifier. The approach also works well for small training samples and for multivariate data. Gaussian mixture models are a common descriptor of data, but the outliers need to be well defined. This is why standard methods such as linear regression and statistical hypothesis testing are used to first detect the anomalies in a time series.

3. Energy Time Series Anomalies

Understanding the sources of anomalies in energy time series data plays an important role in their detection and classification because the definition of false positives depends on the context. The energy data in this paper are the reported natural gas consumption for residential and commercial (offices, schools, administrative buildings, and hospitals) customers. For those categories of customers, sources of anomalous data include:

- **Missing data or missing components of aggregated data** occur when there are no data values for a specific observation in a univariate data set or when there are no data values for a particular variable of a multivariate data set.
- **Electric power generation** occurs when the natural gas load used for the generation of electric power is included in the residential or commercial customers' consumption load.
- **Main breaks** are unplanned events that occur to the normal consumption of energy, such as a backhoe hitting a pipeline, power outages, and heavy snow days.
- **Naïve disaggregation** or a **stuck meter** occurs when a normally variable energy load does not vary across several meter reporting periods.
- **Negative energy consumption** typically is the result of a system misconfiguration. An energy consumption can be zero but not negative. Negative energy consumption can be reported because different pieces of the system (pipelines, types of customers, or corrections) have been mistakenly merged together.
- **Human error** yields unexpected data values resulting from a bad query or incorrect manual entry reporting.
- **Mismatched meter factor** or **mismatched units of aggregated data** occur when the meter factor is switched during data collection (usually, the energy load for an operating area is composed of energy loads from various territories) without applying the adjustment factor to previous data (for example decatherms to therms). It also occurs when the units between subsets of the data are different, and the proper conversion is not applied when merging the data.
- **Outliers** are data points that are dissimilar to the remaining points in the data set (Hawkins, 1980). If there is no correlation between an energy consumption and the factors driving the consumption, the data point is simply considered an outlier if no other cause is identifiable.

The causes of anomalies mentioned above are used to divide a training set into subsets. Each subset contains a specific type of anomalous features and is used to train a Bayesian maximum likelihood classifier.

4. Anomaly Detection Methods

This section presents the energy time series anomaly detection algorithm and the Bayesian maximum likelihood classifier developed for anomaly detection. Because the energy consumption for residential and commercial customers is influenced by weather, a linear regression model is used to extract weather features from the time series data. The residuals of the time series data form a data set that can be studied with distribution functions.

4.1. Linear Regression

The energy (electric and natural gas) demand for residential and commercial customers is driven mainly by weather. Any energy time series can be divided into three parts, a base load that does not depend on the temperature, but is related to everyday usages of energy such as cooking, water heating, and drying clothes; a heating load and a cooling load that vary with temperature (Hong, 2014; Vitullo et al., 2009).

Figure 1 shows an example of relationship between electric load consumption and temperature for operating area 1. The explanatory variables for the linear regression model, derived from operating area 1 data, are weather-related inputs.

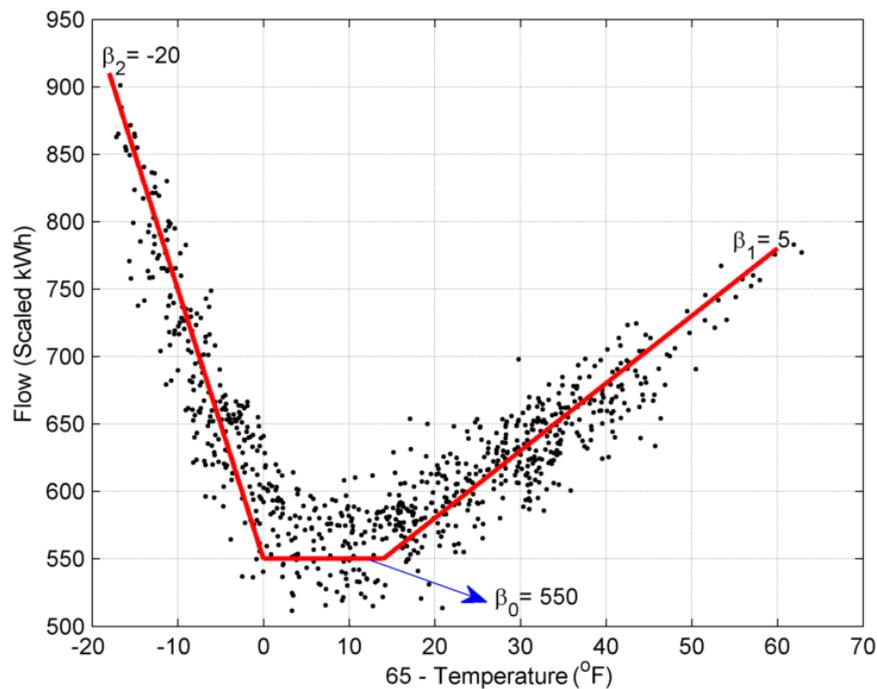


Figure 1: Relation between electric load consumption and temperature for operating area 1. The red function captures the trend lines of the linear regression model for operating area 1, given by $y_t = \beta_0 + \beta_1 \text{HDD}_{65} + \beta_2 \text{CDD}_{75}$.

The general linear regression model used to extract features and find residuals on any energy time series data set in this paper is

$$y_t = \beta_0 + \beta_1 \text{HDDW}_{T_{ref_H}} + \beta_2 \Delta \text{HDDW} + \beta_3 \text{CDD}_{T_{ref_C}} + \beta_4 y_{t-1}, \quad (1)$$

where T_{ref_H} and T_{ref_C} are the reference temperatures below or above which heating or cooling is needed, respectively (Beccali et al., 2008). The reference temperatures usually vary per climate regions. $\text{HDDW}_{T_{ref_H}}$ and $\text{CDD}_{T_{ref_C}}$ are the daily wind-adjusted heating degree days and cooling degree days calculated at reference temperatures T_{ref_H} and T_{ref_C} , respectively. ΔHDDW is the difference of heating degree days between two consecutive days. It captures the temperature variation from one day to the next. If T is an average daily temperature,

$$\text{HDDW}_{T_{ref_H}} = \max(0, T_{ref_H} - T) \times (\text{wind factor}), \text{ and } \text{CDD}_{T_{ref_C}} = \max(0, T - T_{ref_C}). \quad (2)$$

After the coefficients of the linear regression are calculated, they are used to compute the residuals of the data by taking the difference between actual and estimated energy values. The energy time series anomaly detection algorithm is applied to residuals to find anomalies.

4.2. Energy Time Series Anomaly Detection

The linear regression model only extracts the weather dependency of the time series. Therefore, the residuals form a data set that can be modeled with probability distribution functions. The extrema (maximum and minimum) of the set of residuals are used to find anomalies. An extremum is an anomaly if its probability of belonging to the same distribution as the remaining points in the residual data set is less than the probability of committing a type I error at a specified level of significance, typically 1% (Akouemo and Povinelli, 2014).

The data need to be imputed at each iteration of the anomaly detection process to reduce masking (Grané and Veiga, 2010). The estimated coefficients are possibly erroneous at the beginning of the process because it is uncertain whether the data set contains anomalies. After an anomaly is found, the linear regression model coefficients are re-calculated on cleaner data at each iteration of the algorithm. The algorithm stops when no more anomalies are identified. The ‘‘MATLAB-like’’ pseudo-code of the energy time series anomaly detection algorithm is presented in Algorithm 1.

The replacement values in this paper are calculated using the same linear regression model used for anomaly detection. However, that model only provides a naïve imputation of the anomalous data because it does not include trends and seasonality components of energy time series. The replacement values are sufficient for the anomaly detection purposes, but complex forecasting models are suitable for data imputation because they include domain knowledge necessary to model the particularities of each data category or utility system.

After the anomalies are detected, each category of anomalous features identified using natural gas domain knowledge is used to train a Bayesian maximum likelihood classifier.

4.3. Bayesian Maximum Likelihood Classifier

A Bayesian maximum likelihood classifier is used to learn anomalous features found from a training set using Algorithm 1. The features are used to test and classify unseen data points. A classifier is an algorithm with features as input and that produces a label but also confidence values as output (Palaanen, 2004). The probability that a feature vector x belongs to a class c_i is $p(c_i|x)$,

Algorithm 1 ENERGY-TS-ANOMALY-DETECTION

Require : energy time series Y , temperature, wind, α , T_{ref_H} , T_{ref_C} , assumed distribution $\text{Dist}(X, \beta)$

potentialAnomalies \leftarrow true
anomalies $\leftarrow \emptyset$

% Calculate the non-varying inputs to the anomaly detection linear regression model
weatherLRInputs $\leftarrow [1 \text{ HDDW}_{T_{ref_H}} \Delta\text{MHDDW CDD}_{T_{ref_C}}]$

while (potentialAnomalies) **do**

 % Include the first lag of Y as input and calculate the model coefficients
 LRInputs $\leftarrow [\text{weatherLRInputs } Y_{-1}]$
 $\beta \leftarrow Y/\text{LRInputs}$

 % Use the coefficients to calculate estimated values and residuals
 $\hat{Y} \leftarrow \beta \times \text{LRInputs}$
 residuals $\leftarrow Y - \hat{Y}$

 % Select the minimum and maximum values of the residuals as potential anomalies
 maxResiduals $\leftarrow \max(\text{Residuals})$
 minResiduals $\leftarrow \min(\text{Residuals})$

 % Calculate the probability that each potential anomaly belongs to the underlying distribution
 % of the remaining data points
 $p_{max} \leftarrow \text{Probability}(\max\text{Residuals} \sim \text{Dist}(\{\text{residuals}\} \setminus \{\max\text{Residuals}\}))$
 $p_{min} \leftarrow \text{Probability}(\min\text{Residuals} \sim \text{Dist}(\{\text{residuals}\} \setminus \{\min\text{Residuals}\}))$

 % Determine if the extrema are anomalous based on the level of significance α
 $g_{min} \leftarrow 1 - (1 - p_{min})^n$
 $g_{max} \leftarrow 1 - (1 - p_{max})^n$

if $(g_{max} > \alpha) \vee (g_{min} > \alpha)$ **then**

 % Exit condition for the algorithm, because there are no more anomalies
 potentialAnomalies \leftarrow false

else

 % Test whether the minimum or the maximum is the anomaly

if $p_{max} < p_{min}$ **then**
 anomalies $\leftarrow \{\text{anomalies}, \max\text{Residuals}\}$
 else
 anomalies $\leftarrow \{\text{anomalies}, \min\text{Residuals}\}$
 end if

 % Re-impute all anomalies found and keep iterating
 Re-forecast(anomalies)
 Re-impute anomalies in signal Y

end if

end while

return anomalies, Y

and it is often referred to as the *a posteriori* probability, which is derived using Bayes theorem. If x is a feature vector and c_i is the i^{th} class, the probability $p(c_i|x)$ is given by equation (3).

$$p(c_i|x) = \frac{p(x|c_i)p(c_i)}{p(x)}, \quad (3)$$

where $p(x)$ is the unknown probability of the feature variables ($x = \{x_1, \dots, x_j, \dots, x_n\}$) and does not depend on the class c_i . The prior of the i^{th} class is $p(c_i)$. The prior is assumed to be equiprobable across all classes ($p(c_i) = p(c)$).

Because $p(x)$ and $p(c_i)$ are constants, they can be treated as a scaling factor, and $p(c_i|x)$ becomes a non-normalized probability,

$$p(c_i|x) \propto p(x|c_i). \quad (4)$$

Gaussian mixture models (GMM) are used to model the density of the data belonging to each class. A GMM is a parametric probability distribution function composed of a weighted sum of Gaussian densities. If the number of Gaussian mixtures chosen to represent a data set is M , the probability $p(x|c_i)$ is

$$p(x|c_i) = \prod_{j=1}^M p(x_j|c_i), \quad (5)$$

where $p(x_j|c_i)$ is the probability of the feature vectors in the j^{th} mixture assuming the i^{th} class. The GMM parameters are estimated using Expectation Maximization (EM). The estimation fits the distribution to the training features (Reynolds, 2008). If the GMM is used for modeling the data, the likelihoods that a feature vector is from a label or class c_i is

$$\hat{c}_i = \operatorname{argmax} p(x|c_i) = \sum_j \operatorname{argmax} p(x_j|c_i). \quad (6)$$

The likelihoods of a data feature is calculated for every class. The data feature belongs to the class that yields the maximum likelihood. Because time series data are not the outcomes of a random process, Bayesian techniques are difficult to apply to time series data. Therefore, the data is transformed from the time domain to a phase space to extract the multidimensional features of the data using a Reconstructed Phase Space (RPS) (Povinelli et al., 2006). A RPS is a way of extracting the multidimensional features of the data that is embedded in a time series signal by studying the signal against delayed versions of itself (Sauer et al., 1991). The RPS is formed as

$$Y = [y_k \ y_{k-\tau} \ \dots \ y_{k-(d-1)\tau}] \text{ with } k = (1 + (d-1)\tau) \dots N, \quad (7)$$

where Y is the dimensional phase space vector of features, y_k is the k^{th} d -dimensional time series vector feature, τ is the time lag, d is the phase space dimension, and N is the number of features or observations in the time series. For the experiment presented in this paper, $y_k = (\text{flow}_k, \text{temperature}_k)$. A RPS is equivalent in a topological sense to the original system (Sauer et al., 1991) and therefore is an effective mechanism to represent the data.

The classifier is trained on RPS training features instead of time series features. Training a classifier is a supervised learning process because the data is assumed to come from a specific class. The k -means technique can be used to detect efficiently the number of lags and mixtures necessary to represent a data set. It is also found in practice that the Bayesian maximum likelihood classifier trained on phase space features works well for as few as two mixtures (Povinelli et al., 2006).

A data point is anomalous with certainty if both the energy time series anomaly detection algorithm and the Bayesian maximum likelihood classifier detect and classify it as anomalous. The next section presents the experiments made, the results obtained, and an analysis of the results.

5. Experiments and Results

The energy time series anomaly detection algorithm and Bayesian maximum likelihood classifier are tested on a natural gas data set. The data set represents the daily reported natural gas consumption of operating area 2. The data set ranges from 01 January 1996 through 31 August 2009 for a total of 4992 data points. The data is scaled to maintain confidentiality, but the scaling is done in such a manner that it preserves the time series characteristics.

5.1. Anomaly Detection Results

For this data set, the HDDW are calculated at both reference temperatures 55°F and 65°F, and CDD are calculated at both reference temperatures 65°F and 75°F. Therefore, the linear regression model used for anomaly detection is a 7-parameter model. ΔHDDW is the difference between the mean HDDW of two consecutive days,

$$\Delta\text{HDDW} = 0.5[\text{HDDW}_{55} + \text{HDDW}_{65}] - 0.5[(\text{HDDW}_{55})_{-1} + (\text{HDDW}_{65})_{-1}]. \quad (8)$$

Figure 2 shows the result of Algorithm 1 for the natural gas data set of operating area 2. It

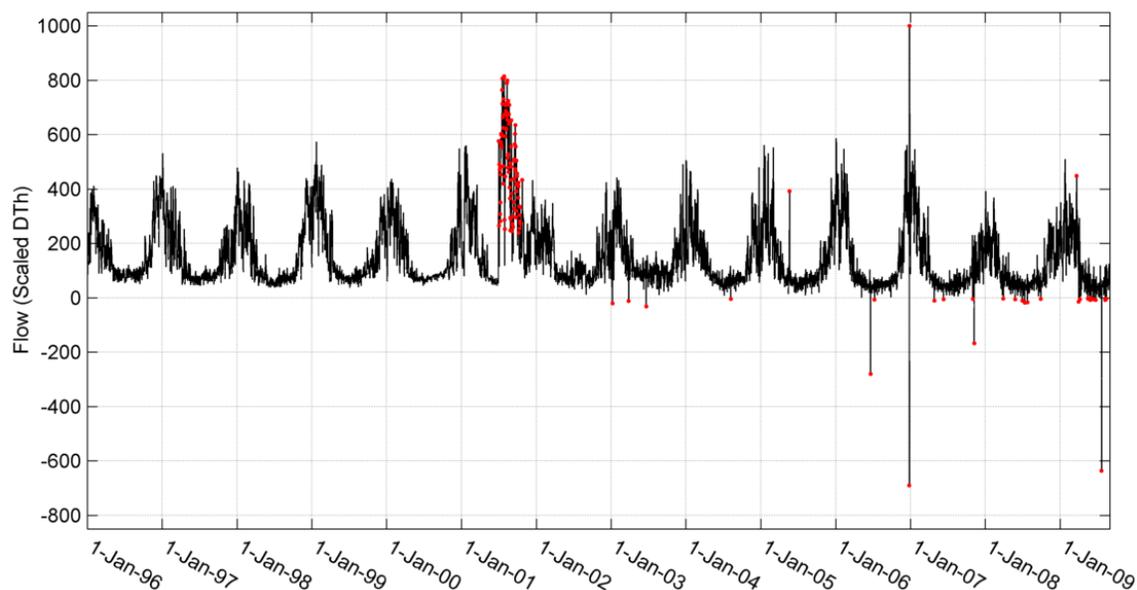


Figure 2: Anomaly detection result for the natural gas time series of operating area 2. The red dots represent the anomalies identified by the energy time series anomaly detection algorithm.

depicts four types of natural gas anomalies: power generation (in the summer of 2001), negative flow values, main break (extreme high and low flow values in December of 2006), and outliers

(all other types of anomalies not recognized by domain knowledge). The data set is divided in a training set from 01 January 1996 through 31 December 2008 and a test set from 01 January 2009 through 31 August 2009, depicted in Figure 3. The training set is divided further into three

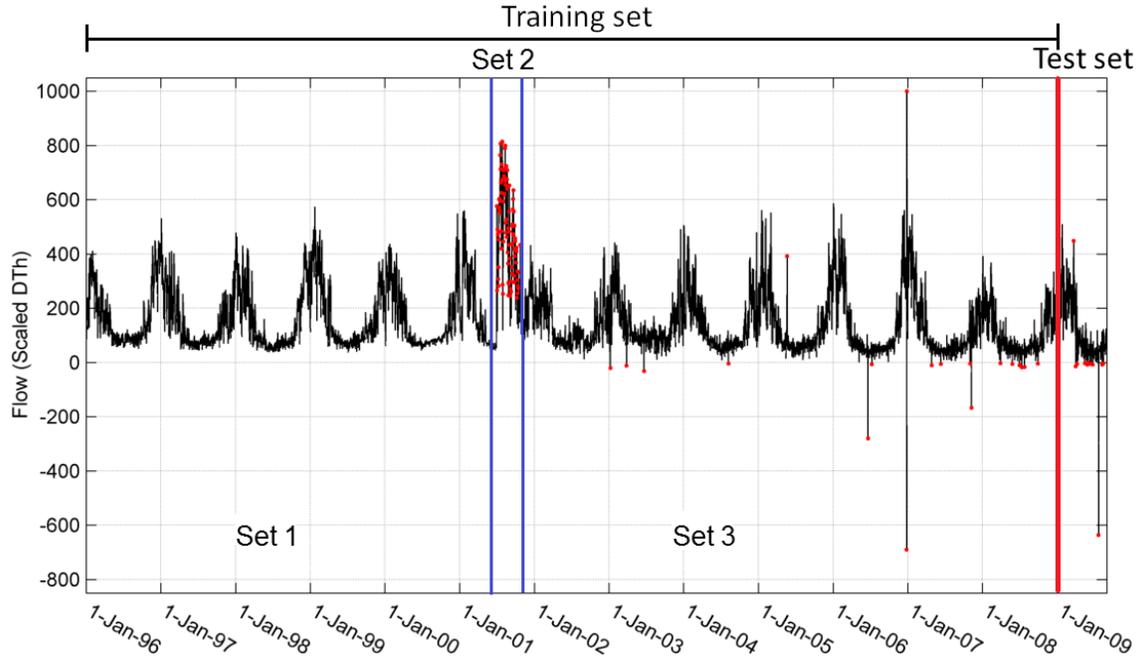


Figure 3: Anomaly detection result for the natural gas time series of operating area 2, depicting the sets used to train the Bayesian classifier and the test set.

subsets. The first subset from 01 January 1996 through 30 June 2001 corresponds to the portion of the data set where no anomalies were found. In the second subset from 01 July 2001 through 15 October 2001, all anomalies are due to power generation. The third subset from 16 October 2001 through 31 December 2008 contains all other types of anomalies. The classifier is trained on each subset. Because no anomalies were found in the first subset, it is considered to be representing the class of “clean” data. The classifier is also trained for the power generation anomalies set because there are enough samples. The main break phenomena in December 2006 cannot be trained as a class because of the lack of training samples. Also, training a class of only negative flow values is impossible because it yields non-positive semi-definite covariance matrices. Therefore, the third subset representing the “outliers” class contains all other types of anomalies not yet trained. The classifier is trained with one time lag and two Gaussian mixtures. Each data feature is composed of the pair (flow, temperature). The “clean”, “power generation”, and “outliers” classes described above are used to test the last year of the data set.

The anomaly detection results on the test set are presented in Figure 4 and Table 3 with the anomalies found labeled from B to M. The maximum value of the test set labeled A is tested for being a false positive. The maximum likelihoods of the monthly subsets of the data are calculated, and the results are presented in Table 1. Confusion matrices of the Bayesian maximum likelihood classifier results are built and presented in Tables 2 and 4. The maximum likelihoods measure how

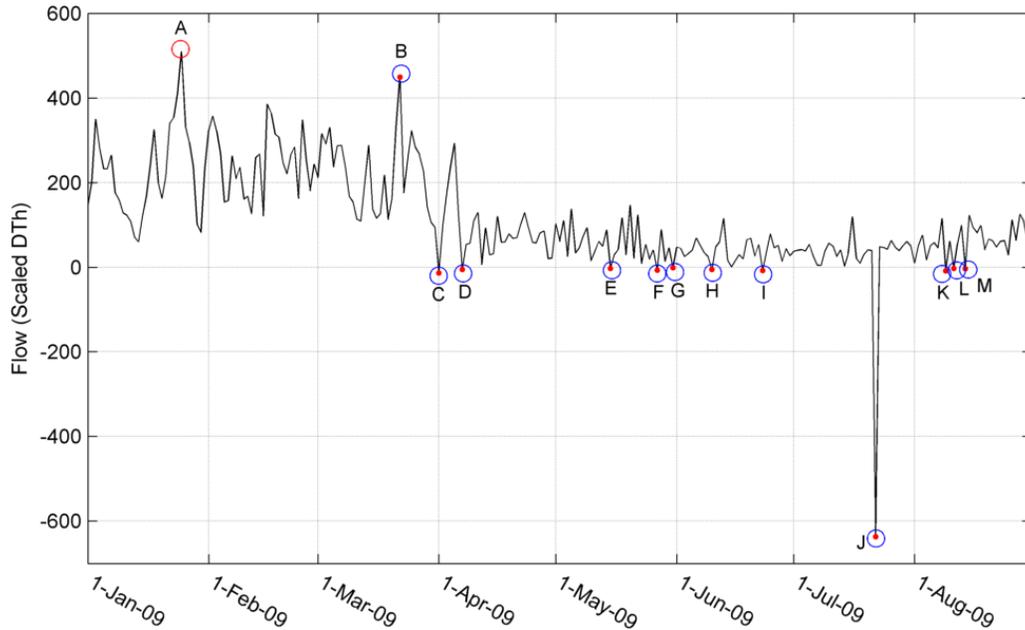


Figure 4: Test set of operating area 2, ranging from 01 January 2009 through 31 August 2009. The blue dots represent the anomalies identified by the energy time series anomaly detection algorithm. The red dot is the maximum value of the time series that is tested for being a false positive. The points are annotated with letters for ease of representation in Table 3.

confident we are that a particular point is anomalous. Because the maximum likelihood is not a normalized probability, the output of the algorithm is a Boolean variable (0 or 1).

Table 1: Bayesian maximum likelihood classifier results on monthly subsets.

Months	Estimated classes			Actual
	Clean	Outliers	Power generation	
January 2009	1	0	0	clean
February 2009	1	0	0	clean
March 2009	1	0	0	outlier
April 2009	0	1	0	outlier
May 2009	0	1	0	outlier
June 2009	0	1	0	outlier
July 2009	0	1	0	outlier
August 2009	0	1	0	outlier

Table 1 agrees with the data set of Figure 4 with the exception of March 2009. In Table 1, January and February 2009 are clean data sets, while the April to August 2009 data set contains some anomalous negative flow values. March 2009 label is “clean”, but its actual label was “outliers” according to Algorithm 1. The classifier accuracy calculated on monthly subsets is 87.5%, as shown by the confusion matrix of Table 2.

Table 2: Confusion matrix of the Bayesian maximum likelihood results presented in Table 1.

Actual \ Predicted	Clean	Outliers	Power generation
Clean	2	0	0
Outliers	1	5	0
Power generation	0	0	0

Table 3: Anomaly detection results for the test set of operating area 2.

Points	Flow values	Probability	Actual label	Estimated classes		
				Clean	Outliers	Power generation
A (Jan-25)	509.74	1.0	clean	1	0	0
B (Mar-22)	449.26	$1.1 \cdot 10^{-3}$	outlier	1	0	0
C (Apr-01)	-13.50	$4.7 \cdot 10^{-13}$	outlier	0	1	0
D (Apr-07)	-5.43	$1.4 \cdot 10^{-4}$	outlier	0	1	0
E (May-15)	-2.93	$6.3 \cdot 10^{-3}$	outlier	0	1	0
F (May-27)	-7.39	$3.2 \cdot 10^{-3}$	outlier	0	1	0
G (May-31)	-1.75	$9.4 \cdot 10^{-3}$	outlier	0	1	0
H (Jun-10)	-5.48	$4.1 \cdot 10^{-3}$	outlier	0	1	0
I (Jun-23)	-8.13	$6.3 \cdot 10^{-4}$	outlier	0	1	0
J (Jul-22)	-636.56	$3.4 \cdot 10^{-102}$	outlier	0	1	0
K (Aug-09)	-8.29	$1.2 \cdot 10^{-5}$	outlier	0	1	0
L (Aug-11)	-3.24	$8.2 \cdot 10^{-3}$	outlier	0	1	0
M (Aug-14)	-3.52	$8.3 \cdot 10^{-4}$	outlier	0	1	0

Table 4: Confusion matrix of the Bayesian maximum likelihood results presented in Table 3.

Actual \ Predicted	Clean	Outliers	Power generation
Clean	1	0	0
Outliers	1	11	0
Power generation	0	0	0

Table 3 presents the anomalies identified and the maximum value of the test set that is tested for being a false positive, along with the values of the data points, their probability of being anomalous, and the Bayesian maximum likelihood classifier results. According to the output of Algorithm 1, the points B to M are anomalous data points, and A is a clean data point. The classifier labels A and B as clean data points, and C to M as anomalous data points. The label output of B is in agreement with March 2009 being labeled a clean data set. The point A, while being the maximum value of the data set, is not an anomaly. The probabilities are calculated at different iterations of the anomaly detection process. The actual labels are derived from the comparison between the probabilities of the data points and the level of significance chosen to be 0.01.

The confusion matrix for individual test data points is presented in Table 4, and the results yield an accuracy of 92.3%. Testing the Bayesian classifier on monthly subsets yields a low accuracy compare to testing individual data points because of the number of samples (8 monthly samples as opposed to 13 data points). A data point is anomalous with certainty if it is labeled by both the energy time series anomaly detection algorithm and the Bayesian maximum likelihood classifier as

anomalous. We conclude that the points C to M are anomalous, while the points A and B are not anomalous.

5.2. Evaluation of Forecasting Improvement

To evaluate the percentage of improvement brought to forecasting accuracy by data cleaning, both the original and clean data sets are used to train the same forecasting model and calculate out-of-samples root mean squared errors (RMSE) on the test set from 01 January 2009 through 31 August 2009. The forecasting model used is derived from Vitullo et al. (2009). The replacement values for all anomalies found are calculated using the same linear regression model used for anomaly detection. The clean data set obtained is presented in Figure 5, and the comparison between RMSE for the models trained on the original and clean data sets is presented in Figure 6.

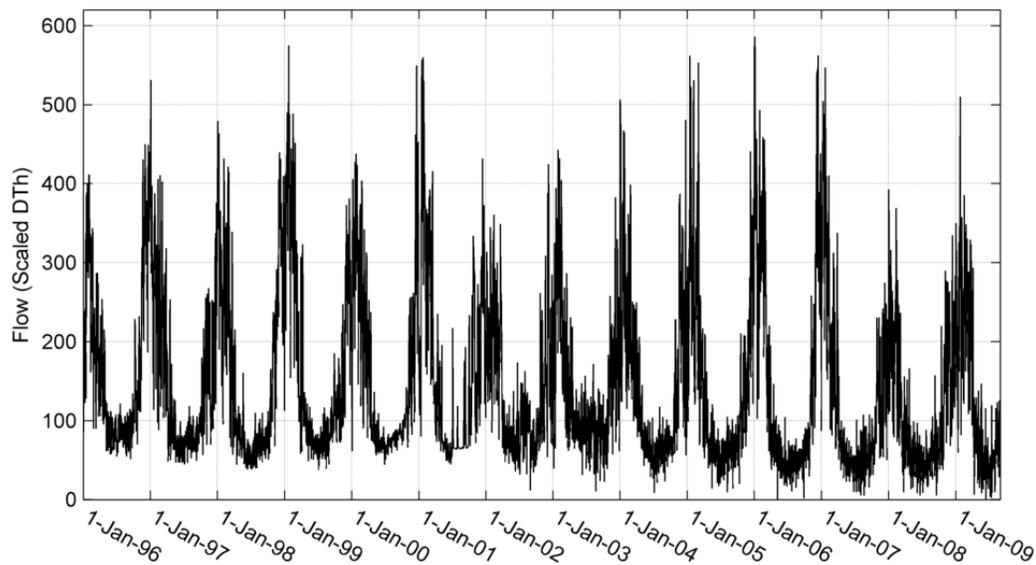


Figure 5: Clean natural gas time series for operating area 2.

Figure 6 depicts the RMSE on average for all days in the test set and the RMSE by months. The RMSE calculated on the clean data set is smaller than the RMSE calculated on the original data set for all months. On average, the RMSE computed on the test set using models trained on the clean data set is 50% smaller than the RMSE computed on the test set using models trained on the original data set. The maximum observed improvement of approximately 84.6% is obtained for the month of July (cleaning of the data point J and the power generation subset shown in Figure 3).

The imputation model used in this case is a naïve model that does not include the particularities of energy time series such as trends and seasonality components. Therefore, using robust forecasting models to improve data imputation can improve further the forecasting accuracy and reduce the errors. The data imputation models can be substituted easily in the energy time series anomaly detection algorithm.

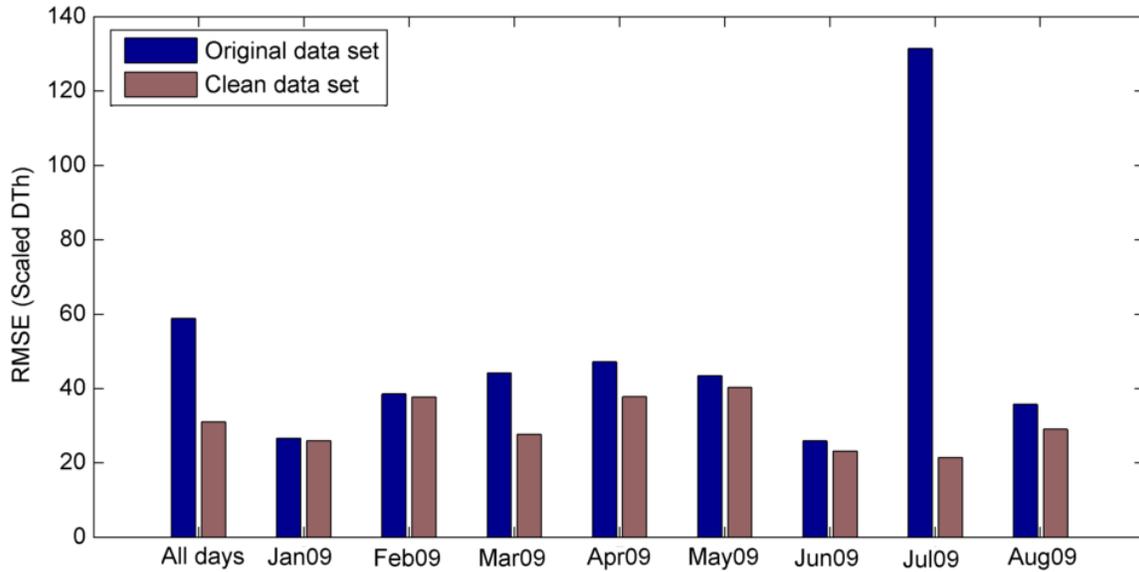


Figure 6: RMSE for all days and by months calculated on the test set of operating area 2 with forecasting models trained on the original and clean data sets.

6. Conclusion

This paper presents the combination of two probabilistic anomaly detection approaches to identify and classify anomalies in historical natural gas time series data. A Bayesian maximum likelihood classifier is trained for every category of anomalous features that has enough training samples. For each data point under test, it is determined if the point is anomalous. Its label is found using the classifier. A data point is anomalous with certainty if it is labeled by both the energy time series anomaly detection algorithm and the Bayesian maximum likelihood classifier as anomalous. The techniques are applied to the daily reported natural gas consumption of an utility and provide good results. The improvement of forecasting accuracy obtained by cleaning the data, with replacement values calculated using a naïve imputation model, is 50% on average. The percentage of forecasting accuracy can be improved further by using robust forecasting models for data imputation. The Bayesian maximum likelihood classifier can be improved by adding exogenous inputs to the reconstructed phase space. Also, the data sets can be normalized using surrogate data techniques to overcome the lack of training samples for some categories of anomalies. The techniques can be expanded to other fields such as electric energy, econometrics or finance, if the exogenous factors of the time series data are known.

References

- Adnan, R., Setan, H., Mohamad, M. N., 2003. Multiple outliers detection procedures in linear regression. *Matematika* 1, 29–45.
- Akouemo, H. N., Povinelli, R. J., July 2014. Time series outlier detection and imputation. In: *PES General Meeting | Conference Exposition, 2014 IEEE*. pp. 1–5.

- Beccali, M., Cellura, M., Brano, V. L., Marvuglia, A., 2008. Short-term prediction of household electricity consumption: Assessing weather sensitivity in a mediterranean area. *Renewable and Sustainable Energy Reviews* 12, 2040–2065.
- Bouguessa, M., 2012. A probabilistic combination approach to improve outlier detection. In: 2012 IEEE 24th International Conference on Tools with Artificial Intelligence. Vol. 1. pp. 666–673.
- Breunig, M. M., Kriegel, H.-P., Ng, R. T., Sander, J., 2000. LOF: Identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Vol. 29. ACM Press, pp. 93–104.
- Chang, I., Tiao, G. C., Chen, C., 1988. Estimation of time series parameters in the presence of outliers. *Journal of Technometrics* 30 (2), 193–204.
- Charlton, N., Singleton, C., 2014. A refined parametric model for short term load forecasting. *International Journal of Forecasting* 30, 364–368.
- Grané, A., Veiga, H., 2010. Wavelet-based detection of outliers in financial time series. *Journal of Computational Statistics and Data Analysis* 54, 2580–2593.
- Hawkins, D. M., 1980. Identification of Outliers. Chapman and Hall, United Kingdom.
- Hong, T., 2014. Energy forecasting: Past, present and future. Foresight issue: *The International Journal of Applied Forecasting* (32), 43–48.
- Lee, A. H., Fung, W. K., 1997. Confirmation of multiple outliers in generalized linear and nonlinear regressions. *Journal of Computational Statistics and Data Analysis* 25 (1), 55–65.
- Magld, K. W., 2012. Features extraction based on linear regression technique. *Journal of Computer Science* 8 (5), 701–704.
- Markou, M., Singh, S., 2003. Novelty detection: A review - part 1: Statistical procedures. *Journal of Signal Processing* 83, 2481–2497.
- McCallum, E. Q., 2012. *Bad Data Handbook: Mapping the World of Data Problems*. O’Reilly Media, Sebastopol, CA.
- Palaanen, P., 2004. Bayesian classification using Gaussian mixture model and EM estimation: implementations and comparisons. Tech. rep., Lappeenranta University of Technology, Lappeenranta, Finland.
- Povinelli, R. J., Johnson, M. T., Lindgren, A. C., Roberts, F. M., Ye, J., 2006. Statistical models for reconstructed phase spaces for signal classification. *IEEE Transactions on Signal Processing* 54 (6), 2178–2186.
- Ramaswamy, S., Rastogi, R., Shim, K., 2000. Efficient algorithms for mining outliers from large data sets. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. Vol. 29. ACM Press, pp. 427–438.
- Reynolds, D., 2008. Gaussian mixture models. Tech. rep., MIT Lincoln Laboratory, Lexington, MA.

- Sauer, T., Yorke, A., Casdagli, M., 1991. Embedology. *Journal of Statistical Physics* 65 (3/4), 579–616.
- Tarassenko, L., Hayton, P., Cerneaz, N., Brady, M., 1995. Novelty detection for the identification of masses in mammograms. In: *Proceedings of the 4th IEEE International Conference on Artificial Neural Networks*. Vol. 4. pp. 442–447.
- Tax, D. M. J., Duin, R. P. W., 1998. Outlier detection using classifier instability. In: *SSPR '98/SPR '98 Proceedings of the Joint International Workshops on Advances in Pattern Recognition*. pp. 593–601.
- Vitullo, S. R., Brown, R. H., Corliss, G. F., Marx, B. M., 2009. Mathematical models for natural gas forecasting. *Canadian Applied Mathematics Quarterly* 17 (4), 807–827.
- Yamanishi, K., Takeuchi, J.-i., Williams, G., 2000. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, pp. 320–324.
- Yuen, K.-V., Mu, H.-Q., 2012. A novel probabilistic method for robust parametric identification and outlier detection. *Journal of Probabilistic Engineering Mechanics* 30, 48–59.
- Zou, C., Tseng, S.-T., Wang, Z., 2014. Outlier detection in general profiles using penalized regression method. *IIE Transactions* 46 (2), 106–117.

Authors' bios

Hermine N. Akouemo is pursuing a Ph.D. degree in Electrical and Computer Engineering at Marquette University. She received her M.S. in Electrical Engineering from Marquette University, Milwaukee WI. She is currently a Graduate Research Assistant at the GasDay Project at Marquette University, focusing on cleaning energy time series data for the improvement of forecasting model accuracy. She is a member of IEEE, Eta Kappa Nu, and Sigma Xi.

Richard J. Povinelli, Associate Professor of Electrical and Computer Engineering at Marquette University, was a software engineer with General Electric (GE) Corporate Research and Development from 1987 to 1990. From 1990 to 1994, he served as a Program Manager and then as a Global Project Leader with GE Medical Systems. Dr. Povinelli's research interests include data mining of time series, chaos and dynamical systems, computational intelligence, and financial engineering. He is a member of the Association for Computing Machinery, American Society of Engineering Education, Tau Beta Pi, Phi Beta Kappa, Sigma Xi, Eta Kappa Nu, Upsilon Pi Epsilon, and the Golden Key. He was voted Young Engineer of the Year for 2003 by the Engineers and Scientists of Milwaukee.