International Journal of Forecasting 32 (2016) 948–956

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Probabilistic anomaly detection in natural gas time series data

Hermine N. Akouemo *, Richard J. Povinelli

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

ARTICLE INFO

Keywords:
Data cleaning
Energy
Outlier detection
Linear regression
Bayesian classifier
Gaussian mixture models

ABSTRACT

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

1. Introduction

Anomaly detection, which is the first step of the data cleaning process, improves the accuracy of forecasting models. Data sets are cleaned for the purpose of being used to train forecasting models. Training a forecasting model on time series that contain anomalous data usually results in an erroneous model, because the parameters and variance of the model are affected (Chang, Tiao, & Chen, 1988). There are various anomalies in historical natural gas time series, due to factors such as human reporting error, data processing error, failure of a natural gas delivery subsystem due to extreme weather, or faulty meter measurements. Examining natural gas time series manually for all causes of anomalies 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 proposes a two-stage method for the detection of anomalies. In the first stage, the probability of a data point being anomalous is determined, using a linear regression model derived from natural gas domain knowledge and a geometric probability distribution of the residuals. The second stage consists of training a Bayesian maximum likelihood classifier based on the types of anomalies identified at the first stage. For a test set, the classifier calculates the maximum likelihood of the data points given the prior classes, and uses the likelihood values to distinguish between false positives and true anomalies. If a data point is anomalous, the classifier is able to report the type of the anomaly. The contribution of the proposed method is its ability to incorporate domain knowledge in the techniques developed for the efficient detection of anomalies in natural gas time series.

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 natural gas domain. A detailed description of our method is presented in Section 4. The experiments and results are presented and analyzed in Section 5.

* Corresponding author.
E-mail addresses: hermine.akouemokengmokenfack@marquette.edu (H.N. Akouemo), richard.povinelli@marquette.edu (R.J. Povinelli).

http://dx.doi.org/10.1016/j.ijforecast.2015.06.001
0169-2070/

2. Previous work

Anomalous data are data that we do not have (missing data), that we had and then lost (manual reporting error, bad query), or that deviate from the system expectations (natural gas consumption during outages due to extreme weather) (McCallum, 2012). Markou and Singh (2003) presented a survey of anomaly detection techniques, ranging from graphical methods such as box plots to more complex techniques such as neural networks. Statistical approaches to anomaly detection are based on the idea of modeling data using different distributions and looking at how probable it is that the data under test belong to these distributions. The method presented in this paper combines linear regressions and distribution functions for the detection of anomalies in natural gas time series, then uses Gaussian mixture models (GMM) for modeling training subsets that contain anomalous features (Barber, 2012). The likelihood of a test data point belonging to a prior subset is calculated using the GMM distributions, and the data point is classified.

Regression analysis is a statistical method that is used widely for electricity and natural gas demand forecasting (Aras & Aras, 2004; Hong, 2014; Hong, Wilson, & Xie, 2014; Hyndman & Fan, 2010; Lyness, 1984; Nedellec, Cugliari, & Goude, 2014). It has also been used in combination with a penalty function for outlier detection (Zou, Tseng, & Wang, 2014). The disadvantage of using a penalty function is that the design of the tuning parameters has to be precise, and is often quite subjective. Therefore, penalty function strategies do not always guarantee practical results. The advantage of linear regression is that, with the dependent variables being well defined, the technique is able to extract time series features (Magld, 2012). Lee and Fung (1997) showed that linear and nonlinear regressions can also be used for outlier detection, but they used a 5% upper and lower threshold limit for choosing outliers after fitting, which yielded many false positives for very large data sets. Linear regression has also been combined with clustering techniques for the detection of outliers (Adnan, Setan, & Mohamad, 2003). In this paper, linear regression is used for extracting weather features from the time series data and computing the residuals of the data.

Bouguessa (2012) proposed a probabilistic approach that uses the scores from existing outlier detection algorithms to discriminate automatically between outliers and the remaining points in the data set. Statistical approaches such as the GMM (Yamanishi, Takeuchi, & Williams, 2000), distance-based approaches such as k-nearest neighbors (Ramaswamy, Rastogi, & Shim, 2000), and density-based approaches such as the Local Outlier Factor (LOF; see Breunig, Kriegel, Ng, & Sander, 2000) are existing techniques that Bouguessa (2012) used for his ensemble model. Each technique provides a score for each observation, and the results are combined to decide whether the observation is an outlier or not. Yuen and Mu (2012) proposed a method that calculates the probability of 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 have also been used for outlier detection and classification. Tarassenko, Hayton, Cerneaz, and Brady (1995) studied the detection of masses in mammograms using Parzen windows and GMMs. 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 were also used by Tax and Duin (1998) to reject outliers based on the data density distribution. They showed that the challenge when using GMMs is selecting the correct number of kernels. However, the approach developed by Povinelli, Johnson, Lindgren, Roberts, and Ye (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 first for detecting the anomalies in a time series.

3. Natural gas time series anomalies

Understanding the sources of anomalies in natural gas time series data is important for their detection and classification, because the definition of false positives depends on the context. The time series data in this paper are the reported natural gas consumption levels for residential and commercial (offices, schools, administrative buildings, and hospitals) customers. For these categories of customers, the possible 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 interfere with the normal consumption of natural gas, such as a backhoe hitting a pipeline or heavy snow days.
- **Naïve disaggregation** or a **stuck meter** occurs when a normally variable natural gas load does not vary across several meter reporting periods.
- **Negative natural gas consumption** is typically the result of a system misconfiguration. A natural gas consumption can be zero but not negative. A negative consumption can be reported because different pieces of the system (pipelines, types of customers, or corrections) have been merged together mistakenly.
- **Human error** yields unexpected data values as a result of a bad query or incorrect manual entry reporting.
- **Mismatched meter factors** or **mismatched units of aggregated data** occur when the meter factor is switched during data collection (usually, the natural gas load for an operating area is composed of loads from various territories) without applying the adjustment factor to previous data (for example decatherms to therms). It also occurs when the units of subsets of the data are different, and the proper conversion is not applied when merging the data.

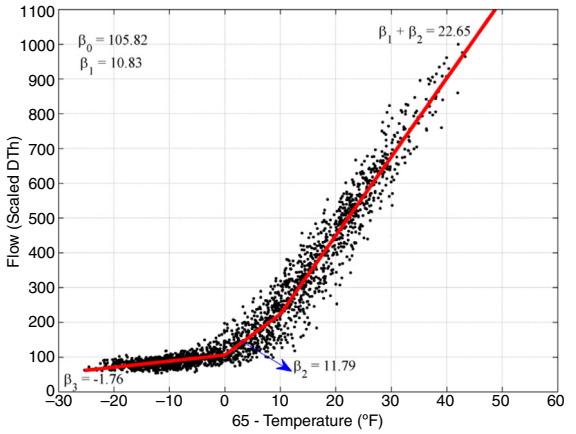

Fig. 1. The relationship between natural gas consumption and temperature for operating area 1. The red function captures the trend lines of the linear regression model for operating area 1, given by $y_t = \beta_0 + \beta_1 \mathrm{HDD}_{55} + \beta_2 \mathrm{HDD}_{65} + \beta_3 \mathrm{CDD}_{65}$. (For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- **Outliers** are data points that are dissimilar to the remaining points in the data set (Hawkins, 1980). If there is no correlation between natural gas consumption and the factors driving the consumption, and the cause is not identifiable, the data point is simply considered an outlier. In this paper, outliers refer to anomalies that do not fit into any of the cases defined above.

These causes of anomalies are used to divide a training set into subsets. Each subset contains a specific type of anomalous feature, and is used to train a Bayesian maximum likelihood classifier.

4. Anomaly detection method

This section presents the natural gas time series anomaly detection algorithm and the Bayesian maximum likelihood classifier developed for anomaly detection. Because the consumption of natural gas by residential and commercial customers is influenced by the 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 using distribution functions.

4.1. Linear regression

Any natural gas 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 natural gas, such as cooking, water heating, and drying clothes; and heating and cooling loads that vary with the temperature (Vitullo, Brown, Corliss, & Marx, 2009).

Fig. 1 shows an example of the relationship between natural gas consumption and temperature for operating area 1. The explanatory variables for the linear regression model are weather-related inputs.

The general linear regression model that is used to extract features and calculate residuals on the natural gas time series data sets in this paper is

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

where T_{ref_H} and T_{ref_C} are the reference temperatures below or above which heating or cooling is needed, respectively (Beccali, Cellura, Brano, & Marvuglia, 2008). The reference temperatures usually vary by climatic regions. $\mathrm{HDDW}_{T_{ref_H}}$ and $\mathrm{CDD}_{T_{ref_C}}$ 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 in heating degree days between two consecutive days, and captures the temperature variation from one day to the next. If T is an average daily temperature,

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

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

4.2. Natural gas 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 using 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 & Povinelli, 2014).

The data need to be imputed at each iteration of the anomaly detection process to reduce masking (Grané & Veiga, 2010). The estimated coefficients may be erroneous at the beginning of the process because it is uncertain whether the data set contains anomalies. After an anomaly has been identified, 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 natural gas time series anomaly detection algorithm is presented in Algorithm 1.

The replacement values in this paper are calculated using the same linear regression model as is used for anomaly detection. However, the model only provides a naïve imputation of the anomalous data because it does not include the trends or seasonality components of the natural gas time series. The replacement values are sufficient for anomaly detection purposes, but complex forecasting models are more suitable for data imputation because they include the domain knowledge that is necessary for modeling the particularities of natural gas data sets or utility systems.

Algorithm 1 NATURAL-GAS-TS-ANOMALY-DETECTION

Require : natural gas time series Y, temperature, wind, α, T_{ref_H}, T_{ref_C}, assumed distribution Dist(X, β)

potentialAnomalies $\leftarrow$ true
anomalies $\leftarrow \emptyset$

% Calculate the non-varying inputs to the anomaly detection linear regression model
weatherLRInputs $\leftarrow \left[1 \;\; \text{HDDW}_{T_{ref_H}} \;\; \Delta\text{HDDW} \;\; \text{CDD}_{T_{ref_C}}\right]$
while (potentialAnomalies) **do**

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

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

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

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

% 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 $\leftarrow$ {anomalies, maxResiduals}
else
anomalies $\leftarrow$ {anomalies, minResiduals}
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

After the anomalies have been detected, they are divided into subsets according to the types of anomalies, as defined in Section 3. Each type of anomaly constitutes an anomalous feature, and each subset is used to train the Bayesian maximum likelihood classifier.

4.3. Bayesian maximum likelihood classifier

A Bayesian maximum likelihood classifier is used to learn the anomalous features found in a training set using Algorithm 1. The features are used to test and classify unseen data points. A classifier is an algorithm which includes features as inputs and produces both a label and confidence values as outputs (Palaanen, 2004). The probability that a feature vector x belongs to a class c_i is $p(c_i|x)$; this is often referred to as the *a posteriori* probability, which is derived using the Bayes theorem. If x is a feature vector and c_i is the ith class, the probability $p(c_i|x)$ is

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

where $p(x)$ is the unknown probability of the feature variables ($x = \{x_1, \ldots, x_j, \ldots, x_n\}$), and does not depend on the class c_i. The prior of the ith 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 scaling factors, and $p(c_i|x)$ becomes a non-normalized

probability,

$$p(c_i|x) \propto p(x|c_i). \tag{4}$$

GMMs are used to model the density of the data belonging to each class. A GMM is a parametric probability distribution function that consists 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

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

where $p(x_j|c_i)$ is the probability of the feature vectors in the jth mixture assuming the ith 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 likelihood that a feature vector is from a label or class c_i is

$$\widehat{c_i} = \operatorname{argmax} p(x|c_i) = \sum_j \operatorname{argmax} p(x_j|c_i). \tag{6}$$

The likelihood 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 are transformed from the time domain to a phase space in order 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 are embedded in a time series signal by studying the signal against delayed versions of itself (Sauer, Yorke, & Casdagli, 1991). The RPS is formed as

$$\begin{aligned} Y &= [y_k\, y_{k-\tau} \cdots y_{k-(d-1)\tau}] \\ &\text{with } k = (1 + (d-1)\tau) \cdots N, \end{aligned} \tag{7}$$

where Y is the dimensional phase space vector of features, y_k is the kth 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 is therefore an effective mechanism for representing 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 are assumed to come from a specific class. The k-means technique can be used for the efficient detection of the numbers of lags and mixtures necessary for representing a data set. In practice, it is also found that the Bayesian maximum likelihood classifier trained on phase space features works well for as few as two mixtures (Povinelli et al., 2006).

We can be certain that a data point is anomalous if both the natural gas time series anomaly detection algorithm and the Bayesian maximum likelihood classifier detect and classify it as anomalous. The next section presents the experiments, the results, and an analysis of the results.

5. Experiments and results

The natural gas time series anomaly detection algorithm and the 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 covers the period from 01 January 1996 to 31 August 2009, with a total of 4992 data points. The data are scaled so as 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 the CDD are calculated at both reference temperatures 65° F and 75° F. Therefore, the linear regression model used for anomaly detection is a seven-parameter model. ΔHDDW is the difference between the mean HDDWs of two consecutive days:

$$\begin{aligned} \Delta\text{HDDW} = {} & 0.5[\text{HDDW}_{55} + \text{HDDW}_{65}] \\ & - 0.5[(\text{HDDW}_{55})_{-1} + (\text{HDDW}_{65})_{-1}]. \end{aligned} \tag{8}$$

Fig. 2 shows the results of Algorithm 1 for the natural gas data set of operating area 2. It 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 2006), and outliers (all other types of anomalies that are not recognized by domain knowledge). The data set is divided into a training set from 01 January 1996 to 31 December 2008, and a test set from 01 January 2009 to 31 August 2009, as depicted in Fig. 3. The training set is divided further into three subsets. The first subset, from 01 January 1996 to 30 June 2001, corresponds to the portion of the data set where no anomalies were found. In the second subset, from 01 July 2001 to 15 October 2001, all anomalies are due to power generation. The third subset, from 16 October 2001 to 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 represent the class of "clean" data. The classifier is also trained on 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 on a class of only negative flow values is impossible because it yields non-positive semi-definite covariance matrices. Therefore, the third subset, representing the "outlier" class, contains all of the other types of anomalies that have not been trained yet. The classifier is trained with one time lag and two Gaussian mixtures. Each data feature consists of the pair (flow, temperature). These "clean", "power generation", and "outlier" classes are used to test the last year of the data set.

The anomaly detection results on the test set are presented in Fig. 4. The maximum likelihoods of the monthly subsets of the data are calculated, and the results are presented in Table 1. Table 2 presents the maximum likelihoods of the anomalies found using the natural gas

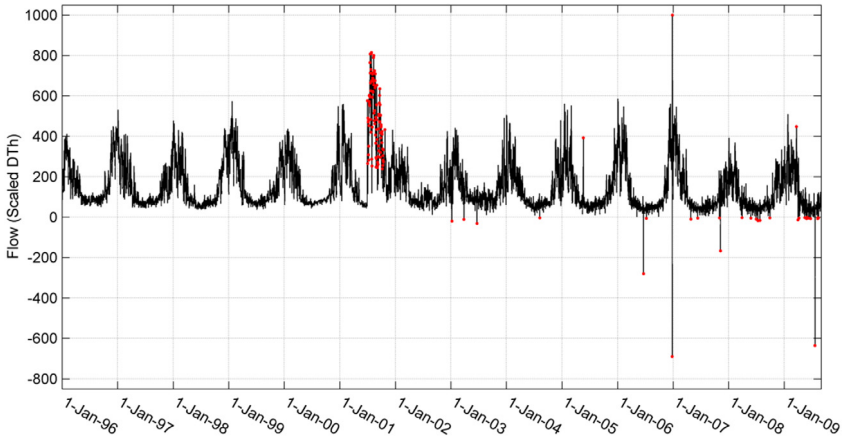

Fig. 2. Anomaly detection result for the natural gas time series of operating area 2. The red dots represent the anomalies identified by the natural gas time series anomaly detection algorithm. (For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

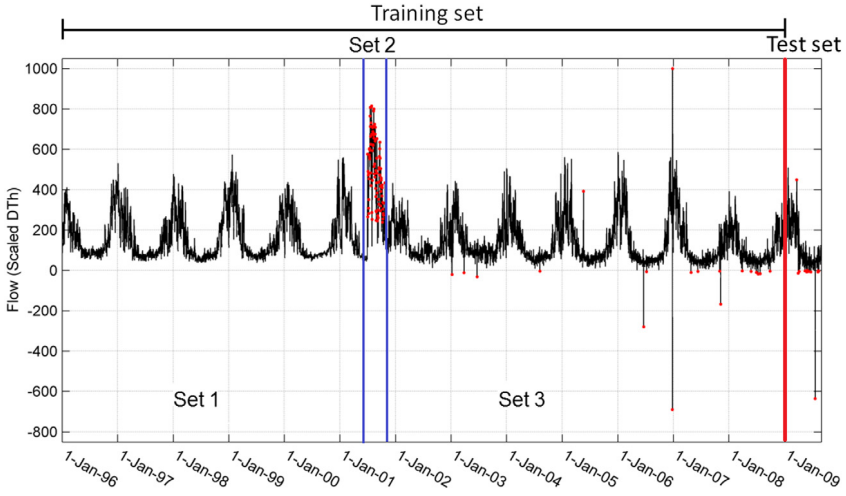

Fig. 3. Anomaly detection results for the natural gas time series of operating area 2, depicting the set used to train the Bayesian classifier and the test set. (For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Bayesian maximum likelihood classifier results on monthly subsets.

Months	Estimated classes			Actual
	Clean	Outlier	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

anomaly detection algorithm, labeled B to M. In addition, the maximum value of the time series data set, labeled A, is also classified. The point A is tested to show that the extremum of the time series data set is not necessarily an anomaly. Confusion matrices of the Bayesian maximum likelihood classifier results are also built and presented in Tables 3 and 4. The maximum likelihoods measure how 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 agrees with the data set of Fig. 4, with the exception of March 2009. In Table 1, January and February 2009 are clean data sets, while the data set from April to August 2009 contains some anomalous negative flow values. March 2009 is labeled "clean", but its actual label according to Algorithm 1 was "outlier". The classifier accuracy calculated on monthly subsets is 87.5%, as is shown in the confusion matrix of Table 3.

Table 2 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 probabilities of being anomalous, and the Bayesian maximum likelihood classifier results. According to the output of Algorithm 1, 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

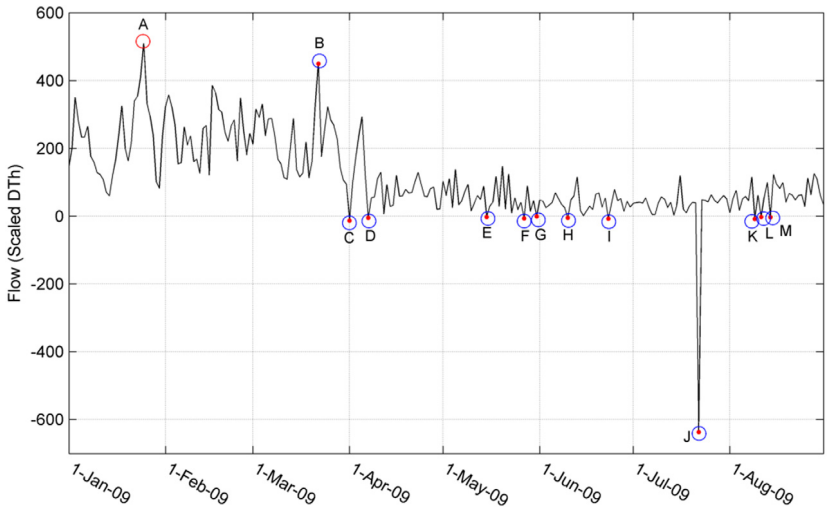

Fig. 4. Test set of operating area 2, from 01 January 2009 to 31 August 2009. The blue circles represent the anomalies identified by the natural gas time series anomaly detection algorithm. The red circle is the maximum value of the time series that is tested for being a false positive. The points are annotated with letters for ease of representation in Table 2. (For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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

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

Table 3
Confusion matrix of the Bayesian maximum likelihood results presented in Table 1.

Actual	Predicted		
	Clean	Outlier	Power generation
Clean	2	0	0
Outlier	1	5	0
Power generation	0	0	0

Table 4
Confusion matrix of the Bayesian maximum likelihood results presented in Table 2.

Actual	Predicted		
	Clean	Outlier	Power generation
Clean	1	0	0
Outlier	1	11	0
Power generation	0	0	0

2009 being labeled a clean data set. Point A, while being the maximum value of the data set, is not classified as an anomaly. The probabilities are calculated at different iterations of the anomaly detection process. The actual labels are derived from a comparison of the probabilities of the data points, and the level of significance is 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 compared to testing individual data points because of the number of samples (eight monthly samples as opposed to 13 data points). We can be certain that a data point is anomalous if it is labeled anomalous by both the natural gas time series anomaly detection algorithm and the Bayesian maximum likelihood classifier. We conclude that points C to M are anomalous, while points A and B are not anomalous.

5.2. Evaluation of forecasting improvement

To evaluate the percentage improvement in the forecasting accuracy due to data cleaning, the original and

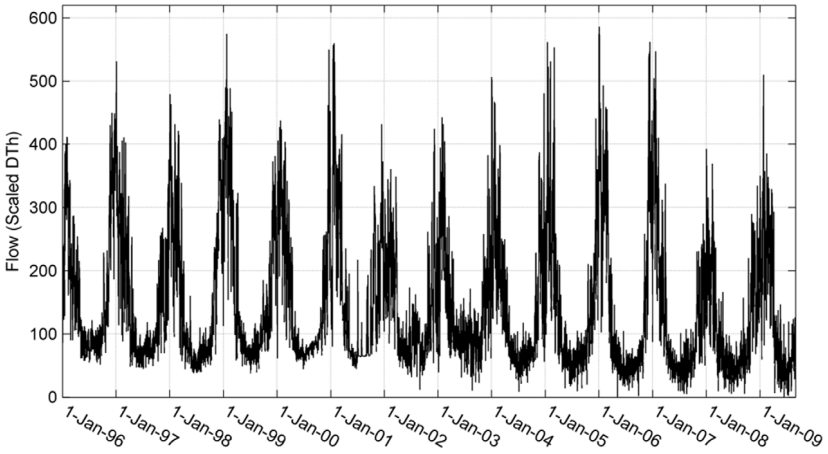

Fig. 5. Clean natural gas time series for operating area 2.

Table 5
RMSEs and MAPEs for all days and by months, calculated on the test set of operating area 2.

Months	RMSE (Scaled DTh)		MAPE (%)	
	Original	Clean	Original	Clean
All days	52.62	32.88	20.27	12.43
January 2009	25.39	24.52	2.36	2.22
February 2009	38.87	38.12	2.79	2.75
March 2009	44.12	27.43	2.70	2.27
April 2009	48.62	39.63	34.11	18.40
May 2009	42.17	38.54	52.88	19.06
June 2009	25.33	23.11	45.41	24.73
July 2009	131.29	21.55	28.01	24.70
August 2009	35.94	29.36	33.40	9.89

cleaned data sets are each used to train the same forecasting model and calculate out-of-sample root mean squared errors (RMSE) and mean average percentage errors (MAPE). The errors are calculated on the test set from 01 January 2009 to 31 August 2009 using Vitullo's natural gas demand forecasting model (Vitullo et al., 2009)

$$\widehat{y_t} = \beta_0 + \beta_1 \mathrm{HDDW}_{T_{ref_H}} + \beta_2 \Delta\mathrm{HDDW} + \beta_3 \mathrm{CDD}_{T_{ref_C}} + \beta_4 \sin\left(\frac{2\pi DOW}{7}\right) + \beta_5 \cos\left(\frac{2\pi DOW}{7}\right) + f(t). \quad (9)$$

The coefficients (β_i, $i = \{0, \ldots, 3\}$) are explained in Section 4.1. β_4 and β_5 are used to model the variation in the natural gas demand by the day of the week (DOW). $f(t)$ is used to model the effects of holidays and days around holidays on the natural gas demand.

The replacement values for all anomalies found are calculated using the same linear regression model as is used for anomaly detection. The cleaned data set obtained is presented in Fig. 5.

The RMSEs and MAPEs calculated using both the original and clean data sets are presented in Table 5. Table 5 depicts the RMSEs and MAPEs both on average for all days in the test set and by month. The RMSEs and MAPEs calculated on the clean test set are smaller than those calculated on the original test set for all months. On average, the RMSEs computed on the test set using models trained on the clean data set are 37.5% smaller than those computed on the test set using models trained on the original data set. The MAPEs are also improved by 7.84%. The maximum observed improvement in RMSE, 83.6%, is obtained for the month of July (due to cleaning of the data point J and the power generation subset shown in Fig. 3). The maximum observed improvements in MAPEs, 33.8%, 20.6%, and 23.5%, are obtained for the months of May, June, and August, respectively. The high MAPE values are due primarily to the negative flow values that occur in the summer.

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

6. Conclusion

This paper presents a two-stage method that combines two probabilistic anomaly detection approaches in order to identify and classify anomalies in historical natural gas time series data. First, a natural gas time series anomaly detection algorithm is used to identify anomalies; then a Bayesian maximum likelihood classifier is trained for each type of anomalous feature that has enough training samples. For each test data point, it is determined whether the point is anomalous, and its label is obtained using the classifier. We can be certain that a data point is anomalous if it is labeled anomalous by both the natural gas time series anomaly detection algorithm and the Bayesian maximum likelihood classifier. The techniques are applied to the daily reported natural gas consumption of a utility, and provide good results. The improvement in forecasting accuracy obtained by cleaning the data, with replacement values calculated using a naïve imputation model, is 37.5% on average for RMSEs, and 7.84% for MAPEs. The percentage forecast accuracy could be improved further by using robust forecasting models for data imputation. The Bayesian maximum likelihood classifier could be improved by adding exogenous inputs to the reconstructed phase space, and also, the data sets could be normalized using surrogate data techniques, to overcome the lack of training

samples for some types of anomalies. This method could also be extended to other fields such as electric energy, econometrics, or finance, if the exogenous factors of the time series data are known.

Acknowledgments

The authors would like to thank the GasDay Laboratory at Marquette University for providing both financial support and data for this research. They would also like to thank Dr. T. Hong and the reviewers for their valuable comments, feedback, and suggestions on this paper.

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. (2014). Time series outlier detection and imputation. In *PES general meeting—conference exposition, 2014 IEEE* (pp. 1–5).

Aras, H., & Aras, N. (2004). Forecasting residential natural gas demand. *Energy Sources*, *26*(5), 463–472.

Barber, D. (2012). *Bayesian reasoning and machine learning*. United Kingdom: Cambridge University Press.

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* (pp. 93–104). ACM Press.

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.

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*. United Kingdom: Chapman and Hall.

Hong, T. (2014). Energy forecasting: past, present and future. *Foresight: The International Journal of Applied Forecasting*, *32*, 43–48.

Hong, T., Wilson, J., & Xie, J. (2014). Long term probabilistic load forecasting and normalization with hourly information. *IEEE Transactions on Smart Grid*, *5*(1), 456–462.

Hyndman, R. J., & Fan, S. (2010). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems*, *25*(2), 1142–1153.

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.

Lyness, F. K. (1984). Gas demand forecasting. *The Statistician*, *33*(1), 9–21.

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*. Sebastopol, CA: O'Reilly Media.

Nedellec, R., Cugliari, J., & Goude, Y. (2014). GEFCom2012: electric load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting*, *30*(2), 375–381.

Palaanen, P. (2004). *Bayesian classification using Gaussian mixture model and EM estimation: implementations and comparisons*. Tech. rep., Lappeenranta, Finland: Lappeenranta University of Technology.

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 Process*, *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* (pp. 427–438). ACM Press.

Reynolds, D. (2008). *Gaussian mixture models*. Tech. rep., Lexington, MA: MIT Lincoln Laboratory.

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* (pp. 320–324). ACM Press.

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.

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 the 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.