ISEIS

Journal of Environmental Informatics 33(1) 1-16 (2019)

Journal of Environmental Informatics

www.iseis.org/jei

Supervised Machine Learning and Heuristic Algorithms for Outlier Detection in Irregular Spatiotemporal Datasets

K. P. Chowdhury[1] *

[1]*University of California - Irvine, Paul Merage School of Business, CA 92697-3125, USA*

Received 9 September 2014; revised 11 August 2016; accepted 8 October 2016; published online 3 October 2017

ABSTRACT. A central problem in time series analysis is the detection of outliers, with further complications presented by irregular time series data measured having spatiotemporal components. This paper presents one Heuristic and two Supervised Machine Learning algorithms for the detection of outliers in this context in univariate time series data, with comparison of results to Chen and Liu's (1993) automatic outlier detection methodology. Due to the recent trend of set up of large environmental databases across many states in the US and around the world, which allow submission of pollutant measurement data from virtually any source, these procedures are applied to the measurements of various surface water pollutants in the California Environmental Data Exchange Network (CEDEN) for understanding and exploring the viability of such databases and the proposed methods. The proposed methodologies though not as robust, give similar results to existing methodologies given the nature of the data, but can be far less time intensive to implement providing interesting insights into the database. Thus, the algorithms presented can be widely used with minimal computing resource requirements with very tractable results even with very large datasets. The methodologies have wide applicability in a variety of contexts and a wide variety of databases with similar measurement challenges across many disciplines, specifically in the environmental setting. In particular, the results have large potential regulatory impact on accepted levels of different pollutants in California water bodies, as well as the amounts to be charged for industrial discharge into those water bodies, and is intended to provide direction for further research and regulatory investments. Based on the results it seems reasonable to assume that there is further room for the inclusion of nongovernmental agency pollutant measurements in the debate of environmental pollution, specifically in California. However, the results also indicate that the use of such databases in a more inclusive way for regulatory matters must be carefully evaluated on an individualized basis. That is to ensure that poorly collected/handled measurements, do not inundate the database over and above those collected with more rigor, thus potentially making inference on the true population distribution of the pollutants more difficult; being especially relevant for those pollutant measurements, which require more delicate sampling procedures.

Keywords: time series, irregular spatiotemporal time series, outlier detection, water pollution, CEDEN

1. Introduction

We live in a world that is increasingly aware of the environmental impact of human activities from the large negative impacts of climate change to the destruction of our natural resources (Diamond, 2005; Xia et al., 2015). To better track this impact there are extensive regulations in most US states, including California, and many countries around the world, for tracking the amount of pollution in natural resources such as rivers and lakes (water bodies). In a time when historic droughts severely affect countries around the world and states such as California in particular, in countless ways, from water conservation to spending on new infrastructure, it has never been more important to track the quality of precious water resources of the world. In California, for example, the California Water Code Section 13260, states that anyone discharging or proposing to discharge waste that could affect the quality of water of the State, other than into the sewer system, are required to file a Report of Waste Discharge (ROWD) to the Regional Water Quality Control Board (RWQCB) (CA, Waste Discharge Form). In addition, all such dischargers, regulated under Waste Discharge Requirements (WDR) and National Pollutant Discharge Elimination System (NPDES) are subject to an annual fee (except dairies which pay a filing fee only) (NPDES Reporting Requirement Handbook). It is, therefore, also vitally important to have accurate data on hand to set the prices and allowable amounts for such discharge activities. To that end California has set up the California Environmental Data Exchange Network (CEDEN), tracked and maintained by various regional data centers as part of CEDEN (CEDEN website).

The utility of such a database is without question. Yet the setup of the exchange is such that anyone fulfilling certain filing criteria can contribute data they have collected on the

* Corresponding author. Tel.: +(818) 857-9325. *E-mail address:* kpchowdh@uci.edu (K. P. Chowdhury).

ISSN: 1726-2135 print/1684-8799 online
 doi: 10.3808/jei.201700375

California water bodies, including academic, nonacademic and governmental reporting agencies. A natural question then presents itself as to the quality of the database in terms of accuracy of the measurements submitted to it. Furthermore, as the data that are collected by the various sources, are often, uncoordinated it is irregular in nature and can be unstandardized in methodology of collection, as well as, units of measurements reported. Thus, the aim here is to understand the outlier issues in this context of data generation process and sampling inconsistencies through existing statistical techniques and some extensions engendering new techniques. In addition, the success and viability of this data source can in turn be used as a model for further regulatory guidance in other states and countries around the world.

The nature of the data means that it is highly correlated to past observations for each type of pollutants or analytes measured, and is therefore ideal for analysis as time series data. However, while some of the measurements by regulatory agencies can and are carried out at regular intervals, those measurements as carried out by other agencies need not be, making outlier detection nontrivial. Thus, it becomes important to understand how this correlation may change not only due to the presence of outliers, but how these outliers may change inference on our model of choice, acceptable levels of the pollutants in relevant water bodies and how much to charge potential dischargers. It is therefore, important to understand the leverage in regression terms that these measurements can have on both inference and prediction on the observed ensemble.

As such, there are multiple avenues by which the time series outlier detection question has been answered in the literature, and this article presents another approach that takes advantage of the error term in the models having certain distributional properties and considers three new approaches, for detecting outliers, one for general modeling purposes and the other two for irregular time series data with irregular spatiotemporal aspects. To that end, it is well known that simulation can be of particular importance in such a search. For example, there are multiple recent papers that rely on simulations to identify outliers while looking at the ordered properties of the random sample (Basu and Meckesheimer, 2007; Harvey et al., 2013), for a summary see (Gupta et al., 2013), however when the time series data is irregular such tests may give erroneous results. Another common methodology is to refer to the asymptotic distributions of such outliers (see for example, Fox, 1972; Tsay, 1988; Chen and Liu, 1993; Chen and Tiao, 1990), and thus based on these distributions and an index value, a particular observation is identified as one type of outlier or another. In addition, there have been further extensions made through Bayesian Analysis as well (see for example, West and Harrison, 1997; Gelman et al., 2014). In the spatiotemporal setting various outlier detection methodologies in a multitude of contexts have been put forward including (Jun et al., 2005; Lasaponara, 2006). Such methodologies all have their strengths and weaknesses, with one of the major problems being that when there is high dimensionality of particular data, most procedures with simulation or otherwise, can be particularly computer intensive even in the current era of abundant cheap computing power, and therefore, it is itself time relevant.

Accordingly, while machine learning in environmental science has been seeing more and more applications (Marvuglia et al., 2015), this work seeks to present new machine learning approaches for the outlier detection literature, which have been shown to be very effective in analyzing large environmental databases when an appropriate training dataset is available for the right spatiotemporal context. Therefore, what is presented here is an application of some of these select methodologies along with these new algorithms, to understand the underlying reliability of the CEDEN datasets. As such, one heuristic and two machine learning methodologies are introduced, two of which are based on inference and the other on the predictive ability of the model applied. While simple, the algorithms still give very tractable results which for comparison is used against results found under Chen and Liu's 1993, algorithm for automatic detection of outliers. Thus, in what follows, the Materials and Methods section explains the models used in the analyses with some further summary of existing outlier methodologies. The Data section specifies the particular issues in dealing with a dataset of the size and complexity of CEDEN, and talks about the specific transformations and the subsets of data on which each methodology is carried out and why. The Results section presents the outcomes of the various methodologies applied in the context of the CEDEN dataset. The Discussion section, compares and contrasts the results and expands on possible future extensions of the methodologies in general and on the CEDEN data in particular, and finally the conclusion gives a brief summary and some further thoughts.

2. Materials and Methods

Time series data can be explained in various ways, but the model that is considered here is the additive model. In particular, consider:

$$Y_t = T_t + Z_t + S_t + R_t \tag{1}$$

where, T_t is a monotone function of time, Z_t is the long term nonrandom cyclical trend function, S_t is the nonrandom short-term cyclical influence such as a seasonal component and R_t is a random variable accounting for all deviations from the non-stochastic model (Falk, 2012). Of course, if we assume that the model is instead multiplicative or partially multiplicative, we may safely make selective transformations to arrive back at the desired additive model, after sequential demeaning of the data or transformed data, with the accompanying change in the interpretation of the parameter estimates. In addition, the seasonal component itself may have multiple variations in this, with the inclusion of multiple seasonal terms (Taylor, 2003), i.e.

$$Y_t = T_t + Z_t + S_t^1 + S_t^2 + R_t \tag{2}$$

where S_t^1 and S_t^2 are two seasonal series with integer periods. (De Livera et al., 2011) extends this seasonal framework to also include non-integer periods with infinitely many seasonal patterns such that

$$Y_t = T_t + Z_t + \Sigma_{i=1}^{n} S_t^i + R_t \ \forall\ i \in \{1,...\infty\} \tag{3}$$

On the other hand, an outlier in time series data can be of multiple types (Fox, 1972; Tsay, 1988):

When a gross error effects an observation (measurement error), called an Additive Outlier (AO).

When a single innovation is extreme. These outliers effect all subsequent observations as well and are called Innovational Outliers (IO).

Level Shifts when there is a change in the structure of the time series' underlying data such that it can either have a permanent effect, Level Shift (LS) or a transient consequence.

2.1. Generalized Methodologies and Algorithms

Outlier detection is a central problem in any data analysis. As such, it has been the topic of voluminous publications in academia. At a high level they may be segregated into multiple methodologies (Agarwal, 2013), with a certain amount of overlap, such as Extreme Value Analysis (Pickands, 1975), Probabilistic and Statistical Models (Agarwal, 1996; Dempster et al., 1977; Gao et al., 2006), Linear and Nonlinear Models (Agarwal, 2001; Jolliffe, 2002; Rousseeuw et al., 2003), Proximity-based Models (Knorr et al., 1998; Ramaswamy et al., 2000), Information Theoretic Models (Keogh et al., 2004; Lee et al., 2001) etc., with each having their own strengths and weaknesses based on the assumptions that underlie them. For example, Extreme Value Analysis, as the name implies is dependent on any realization of an ensemble being too large or small, based on the "statistical tail of the underlying distribution" (Agarwal, 2013). A Probabilistic and Statistical model, relies on an established distribution for the data, with the key assumption being that of the distribution, for example Expectation-Maximization algorithm. Linear Models, segregate the data into optimal subspaces of smaller dimensions as determined by minimizing deviations of observed data from the subspace, for example Linear Regression and PCA analysis. Proximity-based Models rely on some form of distance measure, based on either an assumption as to actual number of groups observed in the data or an algorithm that determines the optimal number of groups, by minimizing some distance measure, for example, Clustering or Nearest Neighbor methods. Somewhat separately, Information Theoretic Models look at methods of summarizing the data with any significant deviations from the summary considered outliers and measures such as Kolmogorov Complexity being used as guides to understanding such deviations within the data.

The key in all such methods is the assumptions underlying the Data Generating Process (DGP), for the observed ensemble. As such, the guiding principle of the algorithms presented is based on the insight that outliers are a function of the model fitted to the observed ensemble. That is, different models should identify different outliers if fitted on the same dataset. Therefore, outlier detection in a limited sense can be seen as a proxy for model selection and vice versa (note that it has been demonstrated for many other automatic outlier detection methodologies that transcription errors, improper units in measurement, irregular sample handling as well as problems with equipment recalibration etc. may give erroneous results, as they may not always be caught if too numerously prevalent in the underlying data). Similarly, further complications arise when the time series data is irregular and spatiotemporal in nature, with large datasets, the norm for todays' data oriented world, presenting an added layer of complexity. Therefore, the algorithms presented do not rely on fitting any particular model, but rather on the appropriateness of the model fit to the data under certain assumptions of the error terms. The best fitted model, as such, will give the least amount of outliers, which by transitivity implies that model fit estimates such as minimized sum of squared error, Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Deviance Information Criteria (DIC) or Crossvalidation etc. can be used to that end.

The reasoning for three particular outlier detection methodologies presented is multi-fold. The Heuristic Algorithm is intended to be a first step in data analysis, to understand what may be the extent of outliers in the time series or non-time series data and if indeed more detailed analyses are warranted of any potential outlier problems. On the other hand, the Supervised Machine Learning Algorithms are presented as an extension of the Heuristic Algorithm, in the presence of the proper spatiotemporal training set, to overcome irregular temporal sampling patterns, when a training set is present which may be used as a benchmark being relatively free from outliers (such as government sampling agencies especially within the environmental context). Thus, both of the machine learning algorithms can be used in the presence of irregular sampling patterns, if properly demeaned for many different sources, such as say, for environmental pollutant measurements, done by mandated government agencies, as well as, non-mandated government agencies. The second Supervised Algorithm, however, does require more consistency in the data and therefore, is more susceptible to measurements being irregular, especially if the number of observations in the training data is small. In addition, as with any modeling scheme, without understanding the nature of the data, it is hard to irrevocably assert the validity of one or more of the assumptions for the algorithms presented.

As such, any combination of the algorithms can be used to understand potential issues in the data under question, with the application of the Heuristic Algorithm recommended to be the initial step in the process. Thus, as the algorithms do not assume the fit of any particular model, and rely only circumstantially on certain assumptions on the error terms of the model fits, which is entirely consistent with many modeling schemes, it is imperative that the modeler be cognizant of the many ways to determine the best model fit.

The various model fitting criteria in the literature lay forth the different advantages and disadvantages of using these criteria in different contexts. Therefore, it is assumed that in determining which model to use, the researcher will take such information into account before applying the heuristic and machine learning algorithms presented. As such, no focused attempt is made to identify which of these statistics are the most applicable in any given modeling context, to determine the applicability of the model, but it is rather left up to the researcher to decide from the many excellent sources which attempt to answer some of these questions (see for example, James et al., 2013; Agresti, 2013; Gelman et al., 2014 among many others). That is, given the nature of the data which the researcher has chosen to investigate, even before applying the outlier detection methodologies discussed, s/he is assumed to be cognizant of the appropriateness of the model fit criteria being used to determine the best model. Furthermore, the modeler should also be aware of the assumptions underlying the algorithms, before deciding on which model may be appropriate, as if the assumptions of the algorithms cannot be supported, neither can the application of any one particular model, building on those assumptions. Inherent in that line of reasoning is that the researcher must be careful not to over fit the data, and may use a host of existing machine learning algorithms, applying them separately, or in conjunction with the algorithms presented, to understand the extent of outliers in their respective datasets. This is because the goal of the algorithms, is not to identify the best model fit for a particular dataset, but rather, understanding what the extent of outliers may be within that dataset given the application of the model chosen.

Of course, in general for large datasets finding outliers can be especially time consuming no matter what type of outlier detection methodology is employed. The focus for this paper, thus, is to employ a methodology that can be used efficiently regardless of the amount of data with little iteration needed for quick convergence, under certain regularity assumptions. As such, the algorithms presented serve different purposes, to ascertain the applicability of the model applied to the data, so that the appropriate outlier detection methodology can be used. The initial step in the analysis is the application of a general heuristic methodology that determines the appropriateness of the model being considered. Upon which if the model is found to be suitable and an appropriate spatio-temporal training dataset is present, the supervised machine learning algorithms can be used to understand with greater accuracy the outlier problems in the data. Therefore, in the absence of an appropriate training dataset, while the heuristic algorithm may be used as a guide to understanding the outlier problem within the data, more time consuming algorithms such as Chen and Liu's (1993) algorithm must be used.

The real benefit of the supervised algorithms, come into play when the nature of the dataset is irregular. In what follows, the supervised algorithms consider the long-term and short-term trend of the training and test datasets to be independent with only the stochastic component assumed to come from the same population distribution or DGP. Thus, when the algorithms are implemented on the appropriately demeaned time series data, the error terms from the fitted model and/or between predictions and observed, can be used as a guide to understanding the outlier problem in the ensemble (the observed realization of the true population time series). In essence, the methodologies as mentioned here are mainly used to find the AO type outliers. The inherent assumption is that if there is a permanent or temporary level shift on the observed ensemble, the measurement errors should reflect this if the underlying model that is fit, correctly takes into account the level shift. Furthermore, given the assumptions of the error structure which will be further elaborated below, the Supervised Machine Learning Algorithms may be looked at as specific applications of the Heuristic Algorithms under certain assumptions on the sampling schemes and of course, the distribution of the error term as elaborated above.

2.1.1. Heuristic Algorithm (Heuristic)

1. Fit any model of choice to the ensemble.

2. On the best-fitted model's error, under assumptions of it being distributed Gaussian with finite mean and variance, perform a LjungBox test of autocorrelation. If the errors are uncorrelated, then they are also independent, under assumptions of normality, and thus proceed to step 3 otherwise stop.

3. Therefore, upon standardizing the errors, under assumption of it being distributed Gaussian, should now be identically distributed with $\epsilon_i \sim N(0,1)$. Thus, now we have an i.i.d ensemble.

4. Identify those errors, which are more than $\pm X_\alpha$ standard deviations away from the expected mean of 0 in absolute value, as $((1-\alpha)*100)\%$ of the density of the Standard Normal should lie in this range.

5. Apply ordinal ranking to these potential outliers according to how many standard deviations away from 0 they are, in absolute value from furthest away to least-furthest, and let $|\tau|$ represent the cardinality of this set. If there are ties, go back and determine the best model fit by sequentially excluding the ties, to determine outlier index based on best/worst model fit as appropriate.

6. Since at any alpha level, there is $(\alpha/100)\%$ chance that some outliers would be detected just by randomness, identify how many such potential random outliers may be in the ensemble through, $(\alpha/100)*|D| = |x| = \upsilon$, where $|D|$ is the cardinality of the set of errors and υ is the integer ceiling.

7. Identify the lowest υ ranked outliers as randomness and not outliers, if and only if they are within $X_\alpha \pm \sigma_\upsilon$, where $\sigma_\upsilon \in R^+$. Let the cardinality of this set be υ' such that $\upsilon' \leq \upsilon$.

8. Thus, identify the indices with the highest ranked $|\tau| - \upsilon'$ potential outliers as the required outliers if and only if $|\tau| - \upsilon' > 0$. Note that any other appropriate interval range and cutoff point can also be used as seen fit by the modeler with the corresponding changes made to the algorithm.

2.1.2. Supervised Machine Learning Algorithm I (Supervised I)

1. Fit any model of choice to the long term and seasonally demeaned training ensemble.

2. On the best-fitted model's error, under assumptions of it being distributed Gaussian with finite mean and variance, perform a LjungBox test of autocorrelation. If the errors are uncorrelated, then they are also independent, under assumptions of normality, and thus proceed to fit this model on the long-term and seasonally demeaned Test data on the same spatiotemporal context (the long-term and seasonal trends allowed to be different between the Training and Test datasets), and go to step 3 otherwise stop.

3. If the model fits were accurate then for the Test data fit, the error term should have a mean of 0 with some variance, which can be approximated by the error variances in the sample realization. Therefore, upon standardizing the errors, under assumption of it being distributed Gaussian, they should now be identically distributed with $\epsilon_i \sim N(0,1)$. Thus, now we have an i.i.d ensemble.

4. Identify those errors, which are more than $\pm X_\alpha$ standard deviations away from the expected mean of 0 in absolute value, as $((1-\alpha)*100)\%$ of the density of the Standard Normal should lie in this range.

5. Apply ordinal ranking to these potential outliers according to how many standard deviations away from 0 they are, in absolute value from furthest away to least-furthest, and let $|\tau|$ represent the cardinality of this set. If there are ties, go back and determine the best model fit by sequentially excluding the ties, to determine outlier index based on best/worst model fit as appropriate.

6. Since at any alpha level, there is $(\alpha/100)\%$ chance that some outliers would be detected just by randomness, identify how many such potential random outliers may be in the ensemble through, $(\alpha/100)*|D| = |x| = \upsilon$, where $|D|$ is the cardinality of the set of errors and υ is the integer ceiling.

7. Identify the lowest υ ranked outliers as randomness and not outliers, if and only if they are within $X_\alpha \pm \sigma_\upsilon$, where $\sigma_\upsilon \in R^+$. Let the cardinality of this set be υ' such that $\upsilon' \leq \upsilon$

8. Thus, identify the indices with the highest ranked $|\tau| - \upsilon'$ potential outliers as the required outliers if and only if $|\tau| - \upsilon' > 0$. Note that any other appropriate interval range and cutoff point can also be used as seen fit by the modeler with the corresponding changes made to the algorithm.

2.1.3. Supervised Machine Learning Algorithm II (Supervised II)

1. Fit any model of choice to the long term and seasonally demeaned training ensemble.

2. On the best fitted model's error, under assumptions of it being distributed Gaussian with finite mean and variance, perform a LjungBox test of autocorrelation. If the errors are uncorrelated, then they are also independent, under assumptions of normality, and thus proceed to predict the long term and seasonally demeaned Test data on the same spatiotemporal context (the long term and seasonal trends allowed to be different between the Training and Test datasets), and go to step 3 otherwise stop.

3. Standardize the difference in prediction vs. actual such that they should now be distributed with $\epsilon_i \sim N(0,1)$. Thus, now we have an i.i.d. ensemble.

4. Identify those errors, which are more than $\pm X_\alpha$ standard deviations away from the expected mean of 0 in absolute value, as $((1-\alpha)*100)\%$ of the density of the Standard Normal should lie in this range.

5. Apply ordinal ranking to these potential outliers according to how many standard deviations away from 0 they are, in absolute value from furthest away to least-furthest, and let $|\tau|$ represent the cardinality of this set. If there are ties, go back and determine the best model fit by sequentially excluding the ties, to determine outlier index based on best/worst model fit as appropriate.

6. Since at any alpha level, there is $(\alpha/100)\%$ chance that some outliers would be detected just by randomness, identify how many such potential random outliers may be in the ensemble through, $(\alpha/100)*|D| = |x| = \upsilon$, where $|D|$ is the cardinality of the set of errors and υ is the integer ceiling.

7. Identify the lowest υ ranked outliers as randomness and not outliers, if and only if they are within $X_\alpha \pm \sigma_\upsilon$, where $\sigma_\upsilon \in R^+$. Let the cardinality of this set be υ' such that $\upsilon' \leq \upsilon$.

8. Thus, identify the indices with the highest ranked $|\tau| - \upsilon'$ potential outliers as the required outliers if and only if $|\tau| - \upsilon' > 0$. Note that any other appropriate interval range and cutoff point can also be used as seen fit by the modeler with the corresponding changes made to the algorithm.

2.2. Specific Applications

Given the goals of the generalized algorithms outlined above specific applications are attempted on the CEDEN data. To that end, models of a state space, where an error is assumed to come from a single source (De Livera et. al., 2011) are considered. For the single source of error model, the Box-Cox transform, Autoregressive Moving Average Error (ARMA), Trend and Seasonal components (BATS) model and Trigonometric Box-Cox transform, Autoregressive Moving Average Error (ARMA), Trend and Seasonal components (TBATS) model (De Livera et al., 2011) were considered. For a more general approach the Seasonal, Autoregressive Integrated Moving Average model with drift (SARIMAd) was also considered.

Thus, for the heuristic application, the TBATS, BATS and SARIMAd models are applied to the variance adjusted data by individual pollutants measured across the state for all available time periods, with the best fitted model being considered based on the minimized sum of squared error statistic. Once fit, the outliers are detected according to Application 1 (Heuristic) given below (Application 1). Secondly, the automatic approach to detecting outliers as given in Chen and Liu's 1993 paper (Chen and Liu, 1993; Lacalle, 2014b) is applied, a summary of which is given below in

(Application 2). Next for the Supervised Algorithm I, (Application 3) it is assumed that the government mandated programs such as the Surface Water Monitoring Program (SWAMP) tasked with official measurements of the water bodies of California to be a representative sample from the underlying DGP for each type of analyte (training data) and apply the model parameters retrieved to Non-SWAMP data and subsequently assessing the fitted model for outliers according to Application 3 (Supervised I) given below. The key assumption made in order to apply this methodology dealt with the underlying irregularity of the measurements as mentioned in outlining the first general supervised algorithm above. While the Heuristic methodology considered all the data for each analyte without regard to the source of the measurements, in considering both the source and the geospatial location it is possible to subset the data so as to restrict the sample to a non-dispersed population set and achieve better model fits and inference.

As such, this approach considered that because of irregularity of the measurements between the different sources, the underlying long term and seasonal trends will vary between the two samples. Therefore, these parameters for the SARIMAd models are allowed to vary between the two samples, however, the underlying stochastic model from which the data is assumed to be generated (DGP) is kept fixed. That is the ARMA errors are assumed to have the same order (p, q) for measurements recorded by both SWAMP and Non-SWAMP data at the analyte by county level. In this way, not only is it possible to correct for the irregularity of the measurements, but also to test whether the two samples can be considered to be coming from the same population density (any trend or seasonality in the datasets are ascertained using spectral decomposition).

Finally, one more supervised analysis on the SWAMP vs. Non-SWAMP data is carried out based on the forecast of the model fitted to demeaned SWAMP data by analyte at the county level, and comparing that forecast to the demeaned Non-SWAMP data. The assumptions for this were the same as in Application 3, that is given the irregular nature of the data, while the trend and seasonal components can vary between SWAMP and Non-SWAMP data, the stochastic component comes from the same population density. This is given in Application 4 (Supervised II). Thus, a summary of the specific applications, are given below.

2.2.1. Specific Application Methodology 1 (Application 1)

For a mathematical exposition of the BATS, TBATS and SARIMAd models described above please see Appendix 1. The complete algorithm for the specific heuristic algorithm application is given below.

Application 1 (Heuristic):

1. Fit BATS, TBATS and SARIMAd models to each analyte and find the best model fit, by minimizing sum of squared error.

2. Apply the Heuristic Algorithm to the error terms of the best-fitted model to identify outliers with $\sigma_v = 1$ and $\chi_\alpha = 2$.

2.2.2. Existing Outlier Detection Methodology - Application 2

The second methodology applied was the one described in (Chen and Liu, 1993) with implementation through (Lacalle, 2014a, b) in *R*. A comprehensive methodology for detection of outliers using the ARIMA models, the major disadvantage of which being that when the dimensionality of the data gets high, convergence can indeed be difficult in a reasonable amount of time. Nevertheless, the ease with which it can be applied given existing resources is still very attractive. As such the model can be described as follows:

$$Y_t = Y_t^* + \omega \frac{A(B)}{G(B)H(B)} I_t(t_1)$$
$$where\ I_t(t_1) = Indicator\ \ for\ time\ t = t_1 \quad (4)$$
$$\frac{A(B)}{G(B)H(B)}\ and\ \omega = Outlier\ Effect\ at\ Each\ Time\ Period\ t.$$

They further differentiate an outlier effect into 4 types:

1. Innovative outlier (IO):

$$\frac{A(B)}{G(B)H(B)} = \frac{\theta(B)}{\alpha(B)\phi(B)} \quad (5)$$

2. Additive outlier (AO):

$$\frac{A(B)}{G(B)H(B)} = \frac{\theta(B)}{\alpha(B)\phi(B)} = 1 \quad (6)$$

3. Temporary change outlier (TC):

$$\frac{A(B)}{G(B)H(B)} = \frac{\theta(B)}{\alpha(B)\phi(B)} = \frac{1}{\delta B} \quad (7)$$

4. Level shift outlier (LS):

$$\frac{A(B)}{G(B)H(B)} = \frac{\theta(B)}{\alpha(B)\phi(B)} = \frac{1}{1-B} \quad (8)$$

where B is the backward shift operator and θ, φ and α are the Autoregressive, Moving average and Difference polynomial operators in B, with the parameter δ modeled to control the pace of the dampening effect of an extraordinary event in the ensemble. Thus, the authors can model the behavior of error terms and determine based on their assumptions for the various types of errors the specific indices which may be considered as outliers (please see Appendix 2 for further details). More specifically, a reduced form summary of the algorithm can be given as follows.

Algorithm 2 (Chen and Liu, 1993):

1. Inner Loop 1: Calculate maximum likelihood estimates

and 44 - 45. If the maximum of 44 - 45 is greater than a predetermined cutoff value (3.5), there may be an outlier effect for that time period for that particular type of outlier with the maximum estimate, if no such outlier effect found stop. Otherwise, recalculate maximum likelihood estimates after removing the outlier effect and repeat procedure. If the total number of outliers in all of the inner loops is greater than 0, and no extra outliers are detected go to next step.

2. Inner Loop 2: Say m outliers in time periods $t_1,\cdots,t_m$ are detected, then ω_j's can be estimated using 48. Then compute $\hat{\tau} = \omega / std(\omega_j), \forall j \in \{1,\cdots,m\}$:

$$\hat{\tau} = \frac{\omega}{std(\omega_j)}, \forall j \in \{1,...,m\} \tag{9}$$

If:

$$\min_j \hat{\tau} = \hat{\tau}_v \le C \tag{10}$$

with C being the same critical value as before, delete this outlier at time point t_v, and recalculate this step. Otherwise obtain the adjusted series by removing the outlier effects using the latest estimate of ω_j. Then calculate the maximum likelihood estimates again, and if the standard deviation from previous estimates is greater than ε recalculate this particular step till it is less than ε.

3. Outer Loop: Compute the residuals by filtering the original series based on the parameter estimates obtained in the last step. Use the residuals obtained in this step and recalculate steps 1 and 2 with parameters estimates in step 1 being fixed and the last two sub-steps of step 2 being omitted. Thus, the estimated ω_j's of the last iteration of step 2 are the final estimates of the effect of the detected outliers.

2.2.3. Specific Application Methodologies 3 and 4 (Application 3/Application 4)

In addition, as mentioned previously the nature of the data lead to two further avenues for comparison of outliers among the considerable amounts of data gathered from different sources. Through conversation with various governmental sources it was apparent that the SWAMP monitoring program tended to have the most rigorous data collection and validation methodology. As such, the above-mentioned models were fit to those analytes which had sufficient sample sizes (greater than or equal to 15 observations) and were collected by both the SWAMP and Non-SWAMP programs. The model parameters were then applied to Non-SWAMP based collected data. If the model fit was appropriate, the two supervised machine learning analyses of outliers based on the errors were done to ascertain the number of outliers and are given in Application 3 and Application 4 below (Casella and Berger, 2002; James et al., 2013):

Application 3 (Supervised I):

1. Fit the TBATS, BATS and SARIMAd models to the properly variance adjusted data by each analyte on a maximum of 5,000 latest observations for all years other than the latest two years of data collected.

2. On the best fitted model, carry out Supervised Algorithm I in the same spatial context on all Non-SWAMP data in the latest two years of data collected, to identify all potential outliers with:

$$\sigma_v = 1 \; and \; \chi_\alpha = 2 \tag{11}$$

Finally, a forecast based machine learning algorithm was used to identify the error and it is given below.

Application 4 (Supervised II):

1. Fit the TBEST, BEST and SARIMAd models to the LOESS de-seasonalized and de-trended SWAMP data by analyte at the county level for data for all years other than the two latest years of data collected.

2. Based on the best-fitted model carry out Supervised Algorithm II to identify all potential outliers for all Non-SWAMP observations for the latest two years of samples collected in the same spatial context.

2.3. Ground Truth Comparison of Selected Algorithms

Furthermore, a first step in comparing any two algorithms is to run them independently on a ground truth dataset with known position of outliers to understand the strengths and weaknesses of each. To compare the large number of iterations for the algorithms, in order to simulate the size of datasets on which the algorithms would be run, a novel approach was used to run the comparison based on True Positives (TP) and False Positives (FP) on the same datasets and looking at the FP over TP statistics for not only cardinal comparison, but also ordinal comparison. As such, let:

$$True\ Positive = \frac{|S(t) \cap G|}{|G|} \tag{12}$$

$$False\ Positive = \frac{|S(t) - G|}{|D - G|} \tag{13}$$

where t is a given threshold, which in the current cases are the standard deviations away from the mean value; $S(t)$ is the declared set of outliers by an algorithm given t; G is the true set of outliers, the Ground Truth Set; and D is the entire data set (Agarwal, 2013). Given this construct, often called the Receiver Operating Characteristics Curve (ROC), it is common practice to graph the data such that any curve that strictly dominates the other, is identified as the superior algorithm. However, such a procedure is only effective as a graphical representation if the values are consistently away from 0, which may or may not be the case in general, and in the presence of inconsistent results that vary drastically, were there to be many iterations run, the graph would be very much un-interpretable. Consequently, a new metric for comparing the algorithms was considered:

$$Statistic\ 1 = \frac{\frac{|S(t) - G|}{|D - G|}}{\frac{|S(t) \cap G|}{|G|}} \quad (14)$$

Thus, for each iteration of the algorithms run, this statistic can be calculated individually for each and compared to understand if the performance for any one or the other was "Superior", "Inferior", "Same" or "Cannot Be Compared". With the algorithm having the smaller numeric value considered to be "Superior" to the other. On the other hand, the reason for there being instances of non-comparison is that the statistic may be subject to instances of division by 0. In such cases, while we may still be able to make inference as to the applicability of the algorithms in comparison to each other, this may not be universally the case. As such, consider the following algorithm.

2.3.1. General Algorithm for Comparing Outlier Detection Algorithms

Without loss of generality let the first algorithm for comparison be A1 and the second, A2. Thus,

1. For each iteration for comparison, ensure that any ensemble with more than half of the data artificially constructed to be outliers is discarded. That is if more than half of the data are randomly chosen to be outliers that sample cannot be used as a comparison and should be thrown out.

2. For all remaining iterations proceed as follows:

(a) For any iteration, if A2 could not be fit on the data yet A1 could be, then consider A1 as the "Superior" algorithm as long as:

$$|S(t) \cap G| \neq 0 \quad (15)$$

If:

$$|S(t) \cap G| = 0 \quad (16)$$

then no comparison should be made and thus we "Cannot Compare" A1 to A2 (as it becomes subjective as to what is considered an acceptable measure of comparison, to the researcher).

(b) There are quite a few combinations of possibilities for Statistic 1, for the algorithms under consideration, thus they must be compared separately. For each algorithm, Statistic 1 may be either 0, ∞ or a number x in the positive reals, for both A1 and A2. Thus, there are 9 permutations to consider:

i. Case 1: Statistic 1 is ∞ for both A1 and A2.

A. We "Cannot Compare" A1 to A2.

ii. Case 2 and 3: Statistic 1 is ∞ for one algorithm and 0 for the other.

A. If Statistic 1 for A2 is ∞, with Statistic 1 = $x_2/0$, where, x_2 belongs to the positive reals. Where as for A1 it is 0, then A1 is "Superior" to A2, and vice versa.

iii. Case 4 and 5: Statistic 1 is ∞ for one algorithm and some, x in the positive reals, for the other.

A. If Statistic 1 for A2 is ∞, with Statistic 1 = $x_2/0$, where, x_2 belongs to the positive reals. Where as for A1 it is some x_1, in the positive reals, then A1 is "Superior" to A2, and vice versa.

iv. Case 6: If Statistic 1 for both A1 and A2 are 0's.

A. If Statistic 1 for both algorithms is 0, then:

$$S(t) \cap G \text{ for } A1 = S(t) \cap G \text{ for } A2 \quad (17)$$

Thus, they are the "Same".

v. Case 7 and 8: If Statistic 1 for one of the algorithms is 0 while the other is some x in the positive reals.

A. If Statistic 1 for A1 is 0 then A1 is "Superior" to A2 and vice versa.

vi. Case 9: If Statistic 1 for both A1 and A2 is some $x \in R^+$.

A. If x_1 represents Statistic 1 for A1 and x_2 for A2, then if $x_1 < x_2$ and:

$$\{x_1, x_2\} \in R^+ \quad (18)$$

then A1 is "Superior" to A2 and vice versa. If $x_1 = x_2$ then they are the "Same".

2.3.2. Specific Application of Outlier Detection Algorithm Comparison

This particular analysis was broken up into two parts. In the first, a ground truth time series dataset was created where the underlying distribution for the ensemble was Gaussian, aligning well with the assumptions of Heuristic, Supervised I and Supervised II algorithms (GT I). In the second case, the ensemble was created from a Poisson distribution (GT II). For both GT I and GT II, the time series model was an ARMA(2, 2), where the Autoregressive coefficients were (0.9, 0.1) and the moving average coefficients being (-0.8879, 0.1). For positions of the outliers a separate Poisson distribution was used with varying values for the parameter λ of the distribution, with a range of λ = 0.02 to 12.023 in an increment of .04. That is, for varying values of the parameter if a value of 1 was drawn from this Poisson distribution for any corresponding position for the univariate ensemble, that position was considered to be the position of an outlier. In addition, at these positions varying magnitude of outliers were created, from 0.1 to 90.1 units, in an increment of 10, away from the average of the ensemble, to understand the effectiveness of the algorithms in terms of the magnitude of the outliers.

Therefore, 300 iterations of this ARIMA process, with a maximum of 5,000 observations for the Heuristic Algorithm and 800 observations for the Chen and Liu algorithm was

considered (for time consideration). Within each, a cutoff value, in standard deviations from 1 to 5 were considered for determining outliers for the Heuristic Algorithm and using the default values as set in the Chen and Liu Algorithm. Furthermore, each iteration was done over varying magnitude of values from either a $N(42, 12)$ or a Poisson(42) distribution. In addition, the amount of outliers, within each iteration was also varied as set forth above using the Poisson (λ) distribution. Accordingly, the identified outlier positions were filled with:

$$\overline{\chi_{ARIMA}} + \left\{ \begin{matrix} .1,\ 10.1,\ 20.1,\ 30.1,\ 40.1, \\ 50.1,\ 60.1,\ 70.1,\ 80.1, \\ 90.1 \end{matrix} \right\} \tag{19}$$

where χ = Iteration Number. In this way, both how the algorithms compare for different data generating processes, and its sensitivity to magnitude of and varying percentage of outliers, within each dataset, could be ascertained. Thus, in total for the Heuristic Algorithm an ARIMA model was fit on $5{,}000 \times 4{,}050 \times 10 = 202{,}500{,}000$ samples and for the Chen and Liu algorithm on $800 \times 4{,}050 \times 10 = 32{,}400{,}000$ samples (for time consideration as running the Chen and Liu algorithm on 5,000 data rows would have taken far too long) for GT1 and $5{,}000 \times 7{,}490 \times 10 = 374{,}500{,}000$ for Heuristic Algorithm and $800 \times 7{,}490 \times 10 = 59{,}920{,}000$ samples for Chen and Liu for GT II.

Finally, two other analyses were done on the results obtained above. First, a K-Means clustering algorithm was done on both the Heuristic and Chen and Liu datasets, based on the percentage of outliers for the ground truth set, as a function of the entire dataset. This way the performance of the algorithms on datasets with similar amounts of ground truth outlier percentages could be compared. Secondly, a simple summary based on the general algorithm was done. A summary of the results can be found in the Results section.

3. Data

The dataset itself was downloaded on March 17th, 2014 and consisted of a total of 2,875,445 million data points, with 1,837 unique analytes, excluding missing values. The number of missing values in the dataset was significant for many of the analytes, however, as any outliers would influence the effect of filling in these missing values, for the purpose of the analysis it was deemed unnecessary to include them. In addition, of these 1,837 analytes, only 1,254 had more than 5 observations, of which, certain analytes had data that were virtually identical, across time periods. Thus, a realistic analysis could only be done on about 924 analytes, or 50.2% of all the analytes present. However, these analytes covered 2,023,324 data rows of the entire 2,875,445 million and comprised of about 70.4% of the entire dataset. Thus, even before any particularly complicated analysis, it can be seen that the database and its data quality are immediately under scrutiny with as much as 30% of the database being too unreliable or sparse to rely on with testable accuracy.

In addition to the general missing data issues, the variance of the dataset by analyte differed considerably. Consequently, before the application of any of the above-mentioned models, the underlying data was adjusted for variance stabilization through a square root transformation. However, the principle concern with the dataset was that the many agencies and individuals submitting data for the same analytes can and do report findings in various units. To correct for this inconsistency, most of the wet sample measurements were converted to picograms per liter or grams per liter and most of the dry weight measurements were converted to nanograms per gram, both to avoid these issues and any possible numerical errors (some of the measurements can be extremely small thus any analysis, to avoid numerical singularity issues where the optimization process does not converge, must be done on appropriately reweighted data).

Furthermore, because dry-weight data could not be readily converted to wet-weight data, due to the absence of relevant information in the database acquired, all models as discussed above were applied to two separate subsets of CEDEN, namely dry-weight data and wet-weight data. As such, the dry-weight data after conversion, consisted of 286,086 observations of 402 analytes and the wet-weight data consisted of 1,637,476 observations of 924 analytes. Consequently, there are 8 specific outlier detection results to compare and contrast.

Also, since in both datasets for the same analytes, there could be multiple measurement units, which could not readily be converted to the standardized units, only those units, after standardization that had the highest frequency of measurement for any particular measurement unit, were used for the final analysis.

4. Results

4.1. Ground Truth Algorithm Comparison

Before moving on to the results seen in the CEDEN dataset, it is worthwhile to discuss the results of the algorithm run on the ground truth dataset. The results on the more summarized algorithm for all cutoff values for the Heuristic Algorithm showed very promising results for the Normal DGP. A more detailed breakdown by outlier identification cutoff value and outlier value can be found in Appendix 3.

Somewhat paradoxically, the results for GT II were in fact better than those for GT I. A possible explanation for this could be the relatively weak assumption made in regards to the type of outliers and the error terms as opposed to distributional assumption for multiple statistics, as done in the Chen and Liu's algorithm.

Furthermore, a detailed breakdown of the percentage of outlier, based cluster analysis can be found in Appendix 5 and 6 for GT I and GT II respectively. The results were comparable, at least for the datasets under consideration as explained in the methodology of the ground truth dataset algorithms. Similarly, for the Poisson DGP the results were

Table 1. Comparison Summary of Ground Truth Algorithm GT I over all Standard Deviations (1-5) Considered as Cutoff

Outcome	Number of Outcomes	Outcome Percentage
Cannot Compare	1,766	43.60%
Superior	1,141	28.17%
Same	776	19.16%
Inferior	367	9.06%
Grand Total	4,050	100.00%

Table 2. Comparison Summary of Ground Truth Algorithm GT II over all Standard Deviations (1-5) Considered as Cutoff

Outcome	Number of Outcomes	Outcome Percentage
Superior	2,449	32.70%
Same	2,443	32.62%
Cannot Compare	1,983	26.48%
Inferior	615	8.21%
Grand Total	7,490	100.00%

again very promising. A more detailed breakdown by outlier identification cutoff value and outlier value can be found in Appendix 4.

4.2. CEDEN Results

For Application 1 (Heuristic), the methodology could be fitted to 799 of the total 924 possible analytes for wet-weight data and on 363 of the possible 402 analytes for the dry-weight data. For Application 2 (Chen and Liu, 1993), the methodology could be fit to 776 of the possible 924 analytes for the wet-weight data and on 359 of the total possible 402 analytes for the dry weight data. However, for the supervised algorithms, the number of model fits were in general smaller, because of geospatial consideration, which subsets the data such that only analytes with enough observation and in the right counties, to make tractable inference are used. Specifically, Application 3 (Supervised I) could be fit to 89 of the total 342 analytes considered on which SWAMP collected data with a slightly higher fit ratio for Application 4 (Supervised II) at 131 out of 342 for the wet-weight data. Of these 342 analytes, only 206 had greater than 15 observations on which a specific, tractable and reliable model could be run, meaning the algorithms could be run on 43.20% and 63.59% of the wet-weight data respectively. For the dry-weight data, Algorithm 3 (Supervised I) could be fit to 40 of the total 216 analytes on which SWAMP specifically collected data, with a fit amount for Application 4 (Supervised II) at 87, out of 216. Of these 216, Supervised I could only be considered on 163 and supervised II on 142 total analytes with a fit percentage of 24.54% and 61.27% respectively. A summary of the result is given in Figure 3.

To that end, for Application 2 (Chen and Liu, 1993) it required 17.11 hours of continuous running on only a maximum of 500 observations per analyte on the wet weight

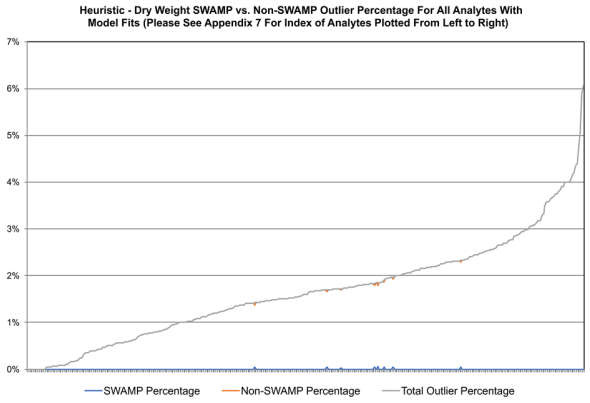

Figure 1. Heuristic dry weight results.

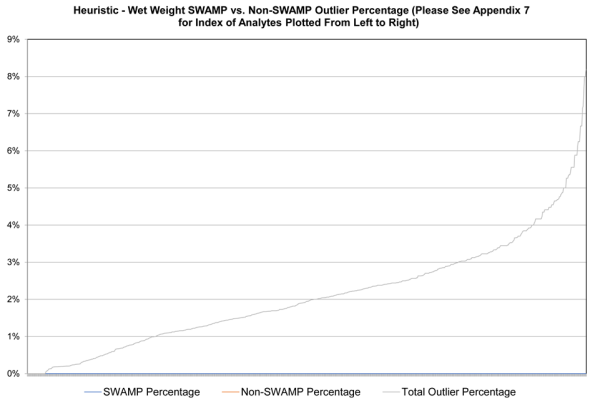

Figure 2. Heuristic wet weight results.

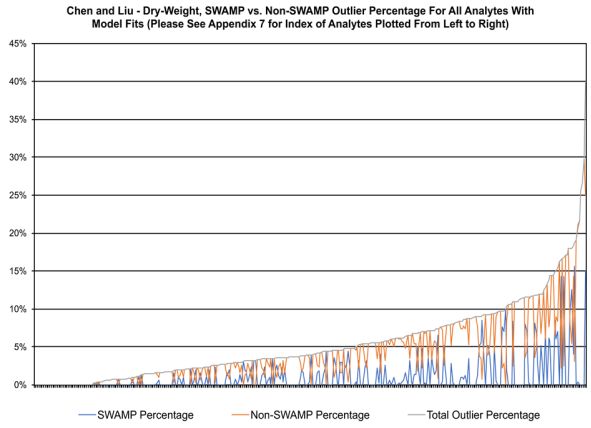

Figure 3. Chen and Liu's dry weight results.

data alone (15.70 hours, for the dry weight data), even after parallelizing, on a MacBook Pro with a 2 GHz quad core processor and 8 GB of RAM for the final results to be ascertained, that too at the lowest number of iterations for the inner loops. If we add to this that the cutoff point for the algorithm, in terms of a critical value itself, can be iterated

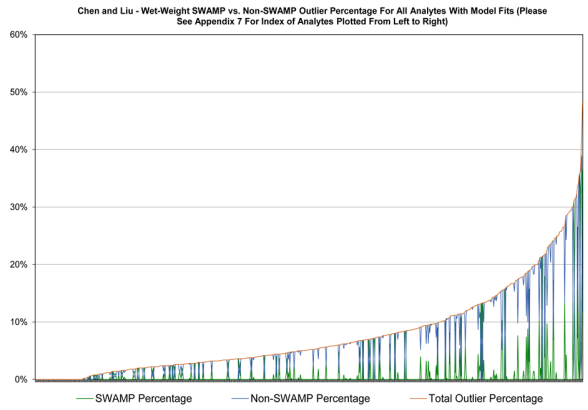

Figure 4. Chen and Liu's wet weight results.

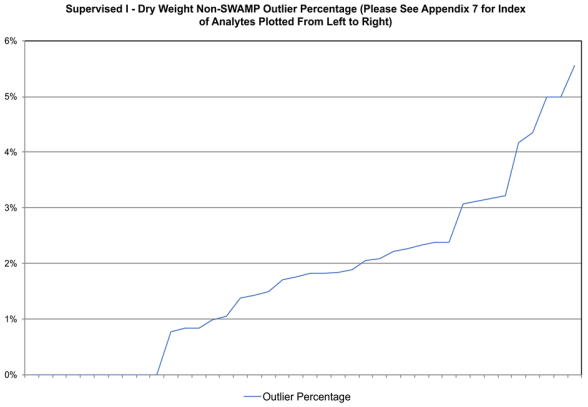

Figure 5. Supervised I dry weight results.

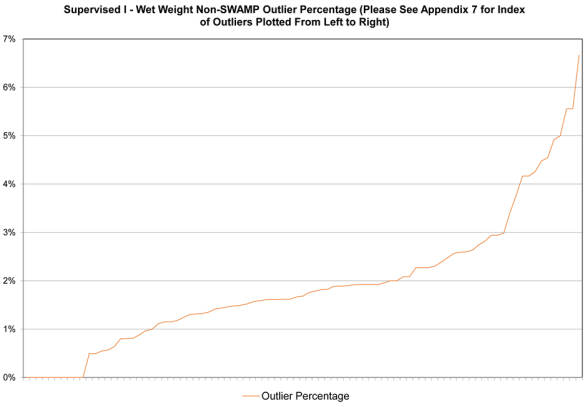

Figure 6. Supervised I wet weight results.

over to find the model that minimizes sum of squared error, it becomes highly unlikely that the optimized model with the list of outliers can be ascertained without very large computational resources or a proportional expenditure of time. In contrast, both Application 3 (Supervised I) and 4 (Supervised II), after parallelizing, could be finished in 6.99 hours (standardized) and 2.73 hours (standardized) respectively, with Application 1 (Heuristic) taking about 4.03 hours (standardized), of continuous running time for the wet-weight data. For the dry-weight data it took about 4.65 hours (standardized) for Application 3, 1.17 hours (standardized) for Application 4 (Supervised II) and 1.36 hours (standardized) for Application 1 (Heuristic).

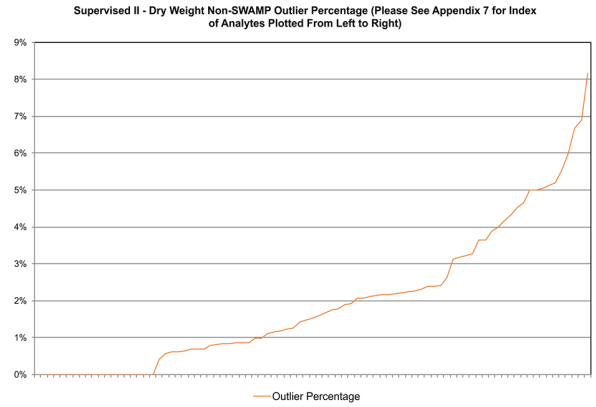

Figure 7. Supervised II dry weight results.

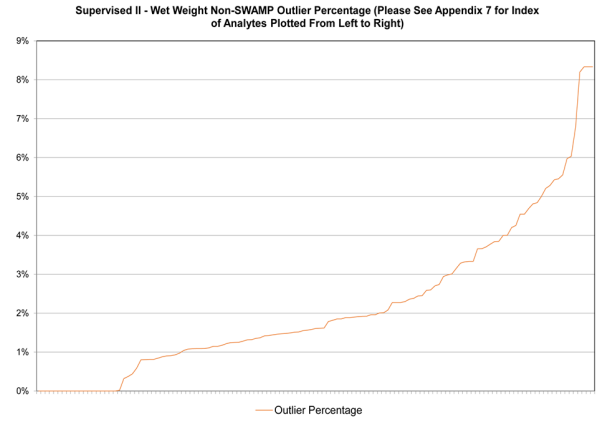

Figure 8. Supervised II wet weight results.

The results overall for all the algorithms on the surface seem comparable, especially given that only a subset of the data was used for the Chen and Liu algorithm, and the algorithm was run for the default number of iterations. Interestingly, as cursory validation of the assumption made regarding the consistency of the data collected by SWAMP, in Application 2 (Chen and Liu, 1993) most of the outliers were from the Non-SWAMP data.

While remarkable, it is worthwhile mentioning that as the algorithm was run on the smaller subset of the data for time consideration, and because the cutoff points as well as the number of iterations in the algorithm are flexible, this particular result can potentially change drastically as it is a function of those particular variables. On the other hand, the cutoff points for the other algorithms are comparatively more fixed. Overall the dry-weight data showed less outliers than the wet-

Table 3. Results Summary for all Algorithms Fit and Outlier Percentages

Summary of Algorithms and Outliers Percentages								
	Heuristic		Algorithm 2 (Chen and Liu)		Algorithm 3 (Supervised I)		Algorithm 4 (Supervised II)	
	Dry	Wet	Dry	Wet	Dry	Wet	Dry	Wet
Fraction of Outliers from SWAMP Data	0.00%	0.00%	1.24%	1.08%	N/A	N/A	N/A	N/A
Fraction of Outliers from Non-SWAMP Data	1.66%	2.09%	4.483%	7.073%	1.80%	1.91%	1.98%	2.18%
Comprehensive Mean of Outlier Percentage	1.67%	2.09%	5.72%	8.15%	1.80%	1.91%	1.98%	2.18%
Mean of Outlier Percentage on Analytes with All Algorithms Fit (Dry-Weight: 105, Wet-Weight: 83)	1.29	1.34	5.80	6.75	1.80	1.92	1.60	2.21
Number of Analytes Fitted	363	799	359	776	40	89	87	131
Number of Analytes Considered	402	924	402	924	163	206	142	206
Total Number of Analytes Available for Analysis	402	984	402	984	216	342	216	342
Percentage of Analytes Fitted	91.90%	86.47%	90.89%	83.98%	24.54%	43.20%	61.27%	63.59%
Actual Computation Time in Hours	1.38	4.15	15.70	17.11	0.52	0.80	0.28	0.46
Standardized Computation Time in Hours	1.36	4.03	15.70	17.11	4.65	6.99	1.17	2.73

weight data, at 1.67 to 5.72%, vs. for the wet-weight data at 1.91 to 8.15%. For Figures 1 through 8, which follow below, please see Appendix 7, for the relevant analysis, for the index of analytes which are plotted on the x-axis from left to right.

The real difference, in the results, however, become apparent when we compare the results of those analytes on which all 4 algorithms could be fit. For these analytes, the means of the outlier percentages varied substantially in comparison to Algorithm 2 (Chen and Liu, 1996), with Algorithm 1 (Heuristic) giving the lowest percentage of outliers for the wet-weight data (1.34%) and the dry-weight data (1.29%). The mean for both datasets were higher in all instances for Algorithm 2 (Chen and Liu, 1993). However, what is more relevant is that multiple analytes according to both Algorithm 3 (Supervised I) and 4 (Supervised II) showed no outliers at all for the dry-weight data.

In essence, when we compare the result of analytes on which all algorithms could be fit, the range becomes much more pronounced. For the dry-weight data, for Algorithm 2 (Chen and Liu, 1993), the outlier percentage was 5.80%, yet the range according to the other algorithms was between 1.29 to 1.80%; for the wet-weight data, Algorithm 2 (Chen and Liu, 1993) indicated about 6.75% outliers as opposed to a range of 1.34 to 2.21% for the other algorithms. These differences for both the overall analysis and the subset of the data on which all four algorithms could be fit, are a direct result of the assumptions that go into considering the entire ensemble for outlier detection as opposed to a supervised subsetted approach. As such, I attempt to provide some detailed explanations for this in the Discussion (5) section.

5. Discussion

Given the results seen above, especially on those analytes on which all four algorithms could be fit, it begs to question why there is a discrepancy between the outlier results based on the different methodologies. One explanation is that as from the previous discussions regarding the nature of outlier detection, an outlier is a function of the model that is fitted to the underlying data. Therefore, as the models are different in each of the applications, we should expect them to identify different outliers. In addition, a methodology that considers the entire ensemble without differentiating between the spatiotemporal nature of the various subsamples within the sample, such as Algorithm 1, should on average show less outliers than methodologies that do take that into consideration. In addition, even if such a methodology could correct for this, one would expect convergence to such an outcome to take longer than an approach that does consider this from the onset such as Algorithm 3 and 4.

Figure 9. Summary of dry-weight and wet-weight analytes on which all algorithms could be fit (see Appendix 8 and 9 respectively for the underlying data with the analytes in alphabetical order plotted from left to right).

In fact, this is exactly what we see in the result with the required time needed for convergence in Algorithm 2 (Chen and Liu, 1993) in general being far higher than the other algorithms. Yet the greatest difference between the algorithms in terms of the time difference needed for convergence is the sequential nature of Algorithm 2 as opposed to non-sequential for the others. Furthermore, considering that three different models were considered before the application of the appropriate model, for Algorithm 1, 3 and 4, which minimized sum of squared error, the potential time saving in a single model context can be even greater. In addition, there are several other factors to be considered which are given below.

5.1. Complexity of Large Environmental Databases

Consider, first and foremost that a database of the size and complexity of CEDEN (and many other similar environmental measurement databases) presents particular challenges, because the type of measurement by analyte can be very disparate with the source and spatiotemporal nature of each analyte varying significantly even within the same water body and thus, even more so across larger geographical regions. Therefore, it becomes increasingly difficult to apply one all-encompassing outlier detection methodology which does take into consideration all these specifics. Secondly, even if we make corrections to the dataset to only pass the relevant subsetted data to a procedure such as Algorithm 2 (Chen and Liu, 1993), as the dimensionality and size of the data increases, it becomes extremely difficult to decide as to which subset of the data has a representative sample on which the procedure can be reasonably successful. For example, it seems hardly tenable that such an algorithm could be used one at a time on all the potential 924 analytes and over 1.9 million or so observations above, not knowing the bias-variance trade off that will be necessitated by this implementation.

On the other hand, through a simple heuristic algorithm we can make very reasonable initial inference on the data, treating it as an infimum/supremum of the number of possible

outliers in our dataset, which then can be further augmented by supervised learning to arrive at a much clearer picture regarding the outlier problem in our dataset, given the underlying data used to train our model is relatively accurate and consistent with the test dataset. Since SWAMP and other such pollutant measuring government entities worldwide are mandated to measure pollutant levels in natural resources, it seems highly likely that these measurements should be on average less variable and more consistent data. As such, they present a highly versatile data against which other measurements can be compared in the correct spatiotemporal context.

Thirdly, this becomes even more relevant when we consider that some analyte measurement procedures can be extremely sensitive to the methodology of measurement that is used, and as such, are more likely to not fit into the same ensemble measured using a different methodology. Consequently, even with large amounts of data, outliers if already embedded in our dataset, can skew our results not only in terms of final tractability, but also in terms of the time required for convergence. While the exact time required will vary based on many criteria, including computing resources available and convergence criteria selected, the time required for the proposed algorithms was substantially less than that for Algorithm 2 (Chen and Liu, 1993). In addition, the actual proposed algorithm run times are more highly correlated with the model used to fit the data rather than the outlier detection steps of the algorithms. As such, this is one potentially large advantage of the heuristic and machine learning algorithms in this context over traditional methodologies. Furthermore, the versatility of the algorithms presented mean that they may be used on any fitted model to understand the model's error structure, under presumptive Gaussian errors and can be extended to any number of exploratory data analysis contexts accordingly.

5.2. Assumption Regarding Sampling Consistency

In addition, both Application 1 (Heuristic) and 2 (Chen and Liu, 1993) consider multiple samples coming from different sources as essentially independent and representative of the population density (in this case for each analyte). Yet when the data is irregular in nature, this may or may not hold even if the samples came from the same population. Thus, in such circumstances a supervised machine learning algorithm can really be useful to purse out any differences in the samples. This is because through these algorithms we are making no assumptions about the dependence or independence of the different datasets and can simply look at the model fits to draw unbiased inference. In addition, upon fitting such and other variants, thereof, of the supervised algorithms, if there are many outliers, or high variability between the methodologies, then the application of the model and its fit can and should be questioned. Thus, the results then can be used as a guide to implement other methodologies for outlier detection as deemed necessary even if at the cost of more time required to run the analysis.

5.3. Limitations, Applicability and Extensions

However, the algorithms are not without their drawbacks, as without a proper training set, of course, the supervised algorithms may not be used. That is, if in the spatial context the sampling distribution of the training and test datasets vary considerably this approach cannot be applied with accuracy. Furthermore, the temporal considerations between the training and test datasets must also be consistent (though more so for Supervised II than Supervised I). Thus, by no means is it an omnibus test, nor does it specifically identify outliers into any of specific type as has been mentioned above through Application 2 (Chen and Liu, 1993).

This is the reason why in both Application 3 (Supervised I) and 4 (Supervised II), the models were fit to Non-SWAMP data in the same counties in which the SWAMP measurements were taken. In addition, of course, the algorithms cannot be used unless the errors themselves are independent. Though if this is not the case, then the algorithm can be used as a heuristic indicator of model fit itself, since if the errors from our fitted time series model are not independent, the model is unlikely to be the correct one for the dataset being considered and alternatives must be evaluated and considered. Furthermore, the test dataset chosen on which Application 4 (Supervised II) is fitted can give results that show more variability if the training dataset has any particular peculiarities that set it apart, over some unknown prediction window than otherwise. However, Application 3 (Supervised I) above should be less sensitive to these changes because it can be fit on all available test data, given that the test dataset is in the similar time frame and spatiotemporal existence as that of the training dataset.

In addition, there is potentially a very pernicious problem with the sampling consistency assumption in any automatic outlier detection methodology. Consider, for example, when the underlying distribution from which a sample is taken varies even if for the same analyte. This is hardly, an unusual possibility, given the vast geographical regions over which different sampling agencies may take their measurements even if for the same analyte. In such a case, the presence of measurements that vary substantially essentially makes it far harder to identify actual outliers if they outnumber correctly collected and measured samples. In the present case, this is especially relevant, because for CEDEN, measurements are taken statewide for the same analytes and such data can be contributed by anyone, as many times as they may deem fit, by fulfilling reasonably easy filling requirements.

Therefore, utility of such databases must be guarded and applied with extreme care. This is because, if we have overall measurements which are "incorrect" (for whatever reason from non-sanctioned entities, without the proper training, equipment or expertise or indeed from sanctioned entities) which inundate and outnumber properly collected measurements, then analysis on the entire ensemble will undoubtedly give the wrong outlier results, no matter how good the algorithm is. In fact, this was very much the case for many of the analytes on which the algorithms were fit showing far variable results for SWAMP vs. Non-SWAMP measurements.

Thus, the results obtained from all the algorithms must be approached with a reasonable amount of caution, and should be looked at as a first step only and not a final say in the actual outlier problem within the dataset, without more in-depth analysis.

In fact, this is perhaps one of the largest contributions of this article, though it may not be entirely novel. That is, it is simply not enough to have data. It must be reliable data. While a small amount of variability is acceptable and warranted, when measurements from different agencies vary by orders of magnitude for the same analyte, using such measurements along the assumptions of the algorithms presented and in fact many other automatic outlier detection methodologies may be a slippery slope.

Despite this however, in certain contexts, especially for many analytes in the CEDEN context considered here, a supervised machine learning approach can give significant advantages over some other traditional methodologies that consider the model fit sequentially or otherwise, for outlier detection, in terms of model accuracy and time needed for convergence. As an example, in the CEDEN dataset this is because when the spatiotemporal distribution can vary considerably within an analyte across water bodies and counties over time, it seems inappropriate to consider all measurements of that analyte as coming from a single population density. When the data is irregular this becomes even less defensible.

Due to the nature of the algorithms as mentioned for CEDEN, it would have been preferable to fit SWAMP analyte measurements by each water body in each county to Non-SWAMP institutions on those same water bodies in each county, in the same temporal context. However, on such subsetted data there were just not enough observations to fit a proper model. On the other hand, as more data is collected across the state there will come a time when such an application may be feasible. At which point the increasing volume of data will make the algorithms even more appealing to be used. Furthermore, if and when that time comes, the comparison of sampling distributions can then potentially be done among multiple agencies and will also provide a basis for understanding, if there has been any particular unobserved phenomenon (variable) in regression terms, that would not be easily picked up if we used an algorithm such as in Application 2 (Chen and Liu, 1993) in a timely manner.

Therefore, overall these algorithms can provide results equivalent to some outlier detection procedures at a fraction of the time and provide critical insights in to the data that may not be readily apparent in certain contexts. This is especially relevant when our datasets are large, with significant advantages in time to get tractable results. As such, it presents another tool in the researcher's arsenal for model fitting and outlier detection, though they may not be as robust as other algorithms. In addition, there are multiple potential extensions of these algorithms through an iterative outlier detection methodology along with specific identification of the type of outlier as in Algorithm 2. As such, other outlier handling techniques such as smoothing may also be used through extensions of the methodologies, as one of the first steps in these methodologies is the identification of the potential outliers in the ensemble.

Thus, as environmental regulations become more stringent and data becomes more ubiquitous, the importance of such algorithms and their variants can only increase. This is because environmental databases are unique in that they sample similar geospatial locations overtime making comparisons across sampling sources more viable as the amount of data from each unique source increases over time, as they are likely to do.

In light of the results one final point must be made. That is, as more and more states and countries start using databases such as CEDEN and use the measurements to guide policy, the potential outlier issues must be considered more vigorously. While a 1% - 8% outlier range on average may not seem significant, the amount of variability with each analyte can be significant given any particular method used, and was certainly witnessed within particular analytes varying by sampling agencies. Thus, a simple tertiary use of the data as presented in the database will undoubtedly lead to very wrong conclusions regarding the pollution level of our natural resources.

In fact, given the large geographic regions over which various authorities may sample for the same analytes, it is essential that those with the proper collecting and measurement expertise record and submit to databases such as CEDEN. As otherwise, poorly collected samples may inundate the measurements, far and above the proper samples collected by environmental agencies such as SWAMP. At such an event, the true underlying distribution for each analyte would be extremely difficult to evaluate with any outlier detection methodology.

This point is especially relevant for those analytes that require more stringent collection methodologies. This in turn, can have large impacts on how much to charge potential dischargers of environmental waste into our habitats. Therefore, another possible extension of this methodology may be an attempt at quantifying any such monetary impact of considering outliers in large environmental databases.

6. Conclusion

In conclusion, it is evident that regardless of the methodology applied, the CEDEN dataset contains some outlier problems. Such issues are hardly unexpected given the size, complexity and type of measurements being considered. As such, the concerned parties must be particularly careful in applying any ad hoc inference on the data without taking this into consideration. In addition, the methodologies that are proposed here with their various extensions, can be used in a variety of circumstances in modeling, as a guideline for outlier detection, both for this purpose and others regardless of the models chosen to fit the data. Consequently, these methodologies allow the analyst to arrive at reasonable conclusions about his/her data without relying solely on com-

puting power, as valid inferences can be made even with very few Monte Carlo iterations, in a reasonable amount of time. Thus, regulators and lawmakers relying on such data must be particularly vigilant to make the right assertions before committing billions of dollars to a supposed over pollution or under pollution of our natural resources.

Acknowledgments. A special thank you to the Southern California Coastal Waters Research Project group for allowing me generous access and guidance in analyzing the data.

Supporting Material. This paper contains supporting materials which are available in its online version.

References

Agarwal, C.C. (2013). Outlier Analysis. Yorktown Heights, New York.

Astill, S., Harvey, D.I., and Taylor, A.M.R. (2013). A Bootstrap Test for Additive Outliers in Nonstationary Time Series. J. Time Ser. Anal, 34(4), 454-465. https://doi.org/10.1111/jtsa.12033

Basu, S., and Meckesheimer, M. (2007). Automatic Outlier Detection for Time Series: An Application to Sensor Data. Knowl. Inf. Syst., 11(2), 137-154. https://doi.org/10.1007/s10115-006-0026-6

Box, G.E.P., Jenkins, G.M., and Reinsel, G.C. (1994). Time Series Analysis: Forecasting and Control, 3rd Edition, Prentice-Hall, Englewood Cliffs, NJ.

CA, Waste Discharge Form. California Waste Discharge Form. http://www.waterboards

California Environmental Data Exchange Network (CEDEN). http://www.sccwrp.org/data/DataSubmission/

Cassela, G., and Berger, R.L. (2002). Statistical Inference. 2nd Edition, Thompson Press, India.

Chatfield, C. (2004). The Analysis of Time Series: An Introduction, 6th Edition, Chap man and Hall.

Chen, C., and Tiao, G.C. (1990). Random level-shift time series models, ARIMA approximations, and level-shift detection. J. Bus. Econ. Stat., 8(1), 83-97.

Chen, C., and Liu, L.M. (1993). Joint estimation of model parameters and outlier effects in time series. J. Am. Stat. Assoc., 88(421), 284-297.

De Livera, A.M., Hyndman, R.J., and Snyder, R.D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. J. Am. Stat. Assoc., 106(496), 1513-1527. https://doi.org/10.1198/jasa.2011.tm09771

Diamond, J.M. (2005), Collapse: How Societies Choose to Fail or Succeed. New York.

Falk, M. (2012). A First Course on Time Series Analysis with Examples with SAS, University of Wurzberg, GNU Free Documentation License.

Fox, A.J. (1972). Outliers in time series. J. Roy. Stat. Soc. Ser. B. (Method.), 34(3), 350-363.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2014). Bayesian Data Analysis, Third Edition, Chapman and Hall, Boca Raton, Florida.

Givens, G.H., and Hoeting, J.A. (2013). Computational Statistics, 2nd Edition, John Wiley and Sons, Hoboken, New Jersey.

Gupta, M., Gao, J., Aggarwal, C., and Han, J. (2013). Outlier detection for temporal data presentation. SDM, Austin, Texas. http://www.siam.org/meetings/sdm13/gupta.pdf

Harvey, A. (1989). Forecasting Structural Time Series Models and the Kalman Filter, Cambridge University Press, Cambridge, 1515-1516.

Hill, T., O'Connor, M., and Remus, W. (1996). Neural network models for time series forecasts. Manage. Sci., 42(7), 1082-1092. https://doi.org/10.1287/mnsc.42.7.1082

Hyndman, R. (2015). Forecast Package in R, https://cran.r-project.org/web/packages/forecast/forecast.pdf

James, G., Witten, D., Hastie, T., and Tibshirani, R., (2013). An Introduction to Statistical Learning, Springer, New York. https://doi.org/10.1007/978-1-4614-7138-7

Jun, M.C., Jeong, H., and Kuo, C.J. (2005). Distributed spatio-temporal outlier detection in sensor networks, Proc. SPIE 5819, Digital Wireless Communications VII and Space Communication Technologies, pp. 273-284. https://doi.org/10.11 17/12.604764

Lamport, L. (1994). LATEX: A Document Preparation System, 2nd edition, Addison Wesley, Massachusetts.

Lasaponara, R. (2006). On the use of principal component analysis (PCA) for evaluating interannual vegetation anomalies from SPOT/VEGETATION NDVI temporal series. Ecol. Model., 194(4). 429-434. https://doi.org/10.1016/j.ecolmodel.2005.10.035

López-De-Lacalle, J. (2014a). tsoutliers R Package for Automatic Detection of Outliers in Time Series. (Draft Version).

López-De-Lacalle, J. (2014b). tsoutliers package in R. https://cran.r-project.org/web/packages/tsoutliers/tsoutliers.pdf

Marvuglia, A., Kanevski, M., and Benetto, E. (2015). Machine learning for toxicity characterization of organic chemical emissions using USEtox database: Learning the structure of the input space, Environ. Int., 83, 72-85. https://doi.org/10.1016/j.envint.2015.05.011

Taylor, J.W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. J. Oper. Res. Soc., 54(8), 799-805. https://doi.org/10.1057/palgrave.jors.2601589

Tsay, R.S. (1988). Outliers, level shifts, and variance changes in time series. J. Forecast., 7(1), 1-20. https://doi.org/10.1002/for.3980070102

USEPA (2004). NPDES Reporting Requirement Handbook, United States. http://www.deq.state.ok.us/wqdnew/forms/DMR-Manual.pdf

West, M., Harrison, J. (1997). Bayesian Forecasting and Dynamic Models, 2nd Edition, New York: Springer-Verlag.

Xia, X. H., Wu, Q., Mou X.L., and Lai., Y.J. (2015). Potential impacts of climate change on the water quality of different water bodies. J. Environ. Inf., 25(2), 85-98. https://doi.org/10.3808/jei.201400263