Autonomous Correction of Sensor Data Applied to Building Technologies Using Filtering Methods

Charles C. Castello[1] and Joshua R. New[2]
Energy Transportation and Science Division
Oak Ridge National Laboratory
Oak Ridge, TN USA
[1]castellocc@ornl.gov and [2]newjr@ornl.gov

Matt K. Smith
Department of Electrical and Computer Engineering
The University of Alabama
Tuscaloosa, AL USA
mksmith3@crimson.ua.edu

***Abstract*— Sensor data validity is extremely important in a number of applications, particularly building technologies. An example of this is Oak Ridge National Laboratory's ZEBRAlliance research project, which consists of four single-family homes located in Oak Ridge, TN. The homes are outfitted with a total of 1,218 sensors to determine the performance of a variety of different technologies integrated within each home. Issues arise with such a large amount of sensors, such as missing or corrupt data. This paper aims to eliminate these problems using: (1) Kalman filtering and (2) linear predictive coding (LPC) techniques. Simulations show the Kalman filtering method performed best in predicting temperature, humidity, pressure, and airflow data, while the LPC method performed best with energy consumption data.**

Keywords—sensor systems, data processing, data analysis, filtering algorithms, Kalman filters

I. INTRODUCTION

Residential and commercial buildings consume approximately 41% of electrical energy used in the U.S. [1]. Retrofitting inefficient buildings with new and innovative technologies that help to curb energy consumption will reduce environmental impacts and enhance the ability to optimize use of our energy distribution infrastructure. Oak Ridge National Laboratory (ORNL) conducts research on these types of technologies by using a wide range of sensors (e.g., humidity, temperature, and wind speed) to develop and characterize performance.

In 2008, ORNL, Schaad Companies, BarberMcMurry Architects, Tennessee Valley Authority, the Department of Energy, and several dozen ORNL industry partners embarked on the ZEBRAlliance research project. Four homes were built in Oak Ridge, TN that integrates sensors with energy-efficient technologies to gauge the integral success and affordability of components and houses [2]. The ZEBRAlliance dataset is composed of data from four residential homes, each with approximately 300 sensors measuring temperature, humidity, electrical usage, and many other variables. 9,352 data points are collected in an hour, 224,448 in a day, and 81,699,072 in a year. Many concerns arise with this amount of data being collected, specifically data corruption and sensor or data acquisition system failure.

Sensor data validation is a very important concern due to the implications on energy and control of energy efficiency research for buildings. Even with the most sophisticated instruments, sensors, and control systems, ORNL's experience with real-world demonstration facilities has logged a plethora of entropy-related failures including power backup outages, data acquisition kernel panics, multiplexer timing failures, sensor wire disconnects, oversaturated/failing sensors, and many other problems. While many analyses treat sensor data as "ground truth", the reality is that this sensor data is often manually corrected in a non-rigorous manner (usually by averaging the time before and after data loss for small blocks and copying from previous hours/days for large blocks). Analysis and decisions based on faulty data could lead to inaccuracies when analyzing components, systems, and whole-buildings.

Previous research by the authors of this paper [3] focused on using independent prediction to minimize needed resources. We accomplished this by using data for a given sensor to form a model and predict corrupt or missing data using interpolation and extrapolation. We investigated statistical processing methods to autonomously validate and correct data from sensors. We investigated four techniques: (1) least squares regression; (2) maximum likelihood estimation; (3) segmentation averaging; and (4) threshold-based methods. Data from the ZEBRAlliance research project such as temperature, humidity, and energy usage data, were used to determine the performance of the four methods. The results of this study showed that threshold-based statistical processing method produced the highest accuracy in predicting temperature, humidity, and energy data.

This paper means to extend the concept of using independent prediction for sensor data validation by using two filter processing methods: (1) Kalman filtering and (2) linear predictive coding (LPC). Results show the Kalman filtering technique performed best in predicting temperature, humidity, pressure, and airflow data while LPC performed best with energy consumption data.

II. FILTERING METHODS

We apply two filtering techniques to sensor data validation (1) Kalman filtering and (2) LPC. We determined the accuracy of both algorithms by removing random data points to create artificial gaps in the data set. Points that are not removed are used to build the model; they are called the "training set." The removed points are called the "testing set" and are used to test the accuracy of the trained models. The training and testing sets are 70% and 30% of the original data set, respectively. We

treat each sensor as an independent variable where missing (testing) data is determined (interpolation and/or extrapolation) using a prediction model (e.g., Kalman or LPC) generated using present (training) data. Inaccurate prediction occurs at data points where there is significant error between the recorded and predicted values.

We generate a model for each observation window of size *c,* where each datum is predicted for each time-step within the observation window. The observation window moves forward by *c* time-steps (no overlap), and prediction for each sample within the observation window is calculated. This process occurs for every possible window within a given set of time-series sensor data. We calculate absolute error (AE), relative error (RE), and root-mean-square error (RMSE), shown in Equations (6), (7), and (8) respectively for each prediction to determine algorithm performance.

A. Kalman Filter

The basic concept of the Kalman filter [4] is used to estimate the state $x \in \Re^n$ using a linear stochastic difference equation for a discrete-time controlled process which is shown as,

$$x_n = Ax_{n-1} + Bu_{n-1} + w_{n-1}, \quad (1)$$

with a measurement $x \in \Re^n$ that is represented by

$$z_n = Hx_n + v_n. \quad (2)$$

Here, *x* represents the state vector estimate, *z* is the observation vector, and *u* is the input control vector. The *A* variable signifies the state transition matrix, *B* is the input matrix, and *H* is the observation matrix. The *n* value represents the time-step.

The *w* and *v* variables denote the process noise and measurement noise, respectively. Both are assumed to be independent of each other, white noise, and Gaussian where

$$p(w) \sim N(0, Q), \quad (3)$$

$$p(v) \sim N(0, R). \quad (4)$$

B. Linear Predictive Coding

A LPC filter [5] is a finite impulse response filter that uses past samples to predict a signal's behavior, which is characterized as:

$$\hat{k}(n) = -a(2)k(n-1) - a(3)k(n-2) - \ldots - a(p+1)k(n-p) \quad (5)$$

where the multiplicative coefficients are represented by a = [1, $a(2)$, …, $a(p+1)$], k is the signal being filtered, n is the time-step, and p is the length of k minus 1. The autocorrelation method of autoregressive modeling is used to calculate the coefficients. This encompasses the use of least squares, Yule-Walker equations, and the Levinson-Durbin algorithm.

III. Experimental Dataset

ORNL's ZEBRAlliance project [2] is used as the data source for this research. Specifically, we used temperature, humidity, energy usage, pressure, and airflow sensor data from house #2 (of four) during the 2010 calendar year. House #2 consists of many high efficiency technologies, specifically advanced framing for its envelopes, high-efficiency florescent lighting, and Energy Star appliances. Space conditioning is delivered using a water-to-air heat pump. Hot water is supplied by a specially built water-to-water heat pump. Table I lists characteristics of the five sensors investigated in this paper.

TABLE I. SENSOR CHARACTERISTICS.

Quantity	Units	Sensor Name	Sensor Location
Temperature	°F	Z09_T_ERV_IN_Avg	Energy recovery ventilation
Humidity	%RH	Z09_RH_ERVin_Avg	Energy recovery ventilation
Energy Usage	Wh	A01_WH_fridge_Tot	Refrigerator
Pressure	psi	H21_PR_LiqBrine_Avg	Integrated heat pump
Airflow	ft^3	A19_AF_DryerEx_Avg	Dryer exhaust

Statistic characteristics for all five types of data are shown in Table II. We used Campbell Scientific's CR1000 measurement and control datalogger for data collection. Resolution for all four data types is 15 min., giving a total number of samples for each sensor, N = 35,040.

TABLE II. STATISTICAL CHARACTERISTICS OF DATA.

Data Type	Average	STD	Min	Max
Temperature (°F)	62.30	13.86	17.03	91.00
Humidity (%RH)	64.91	19.84	17.00	106.00
Energy Usage (Wh)	12.77	11.32	0.00	125.50
Pressure (psi)	273.09	77.34	0.41	500.90
Airflow (ft^3)	0.46	2.43	0.00	70.72

IV. Experimental Setup

We performed experimental trials applying Kalman filter and LPC methods to temperature, humidity, energy, pressure, and airflow data. The performance metrics used for filter processing methods are absolute error, relative error, and RMSE. Absolute error, e_{abs}, is calculated by

$$e_{abs,c,n} = \sum_{s=n-c}^{n-1} \left| \frac{r(s)}{y_{max} - y_{min}} \right|, \quad (6)$$

where y_{max} and y_{min} are the maximum and minimum sensor data values respectively within the specific sensor's dataset, *Y*. Relative error, e_{rel}, is calculated by

$$e_{rel,c,n} = \sum_{s=n-c}^{n-1} \left| \frac{r(s)}{y(s)} \right|, \quad (7)$$

where n is the current time-step, c is the observation window size, s represents the time-step in the observation window, $y(s)$ is actual sensor data, and $r(s)$ is the residual corresponding to $y(s)$. RMSE is calculated by

$$\varepsilon = \sqrt{\frac{1}{N}(r_1^2 + r_2^2 + \cdots + r_N^2)} \quad (8)$$

where r_s^2 represents a residual difference between the actual sensor value and the predicted value, and N is the total number of data points in Y.

1 **Input:** # of observations taken into account (c) and input dataset (Y)
2 **Output:** Relative error, $E_{rel\text{-}mean}$, absolute error, $E_{abs\text{-}mean}$, and RMSE, ε_{mean}
3 **begin**
4 Randomly divide dataset Y into training set Y_{train} (70%) and test set Y_{test} (30%)
5 *// training*
6 $m = 1$ *// variable to keep track of starting point of observations used in prediction*
7 *// loop through all input values where* $Y_{train} = \{y_1, y_2, ..., y_{(0.70)*N}\}$
8 **for** $j = c$ to (0.30)*N **do** *// N=size(Y) and iteration of c*
9 $deg = c$
10 Kalman (Equations [1] and [2]) or LPC (Equation [5]) calculations
11 Calculate residuals, $R = \{r_1, r_2, ..., r_c\}$
12 Calculate absolute error, e_{abs} using Equation (6)
13 Record e_{abs} value in $E_{abs\text{-}coll} = \{e_1, e_2, ..., e_{(0.70)*N-c}\}$
14 Calculate relative error, e_{rel} using Equation (7)
15 Record e_{rel} value in $E_{rel\text{-}coll} = \{e_1, e_2, ..., e_{(0.70)*N-c}\}$
16 Record ε value in $\varepsilon_{coll} = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_{(0.70)*N-v}\}$
17 Calculate relative error, e_{rel} using Equation (8)
18 $m = m + c$ *// iterate*
19 **end**
20 Calculate the mean value of $E_{abs\text{-}coll}$, $E_{abs\text{-}mean}$
21 Calculate the mean value of $E_{rel\text{-}coll}$, $E_{rel\text{-}mean}$
22 Calculate the mean value of ε_{coll}, ε_{mean}
23 $E_{abs\text{-}mean}$, $E_{rel\text{-}mean}$, and ε_{coll} = null
24 *// testing*
25 $m = 1$ *// variable to keep track of starting point of observations used in prediction*
26 *// loop through all input values where* $Y_{test} = \{y_1, y_2, ..., y_{(0.30)*N}\}$
27 **for** $j = c$ to (0.30)*N **do** *// N=size(Y) and iteration of c*
28 Calculate predicted values for the test set using results from Line 10
29 Calculate residuals, $R = \{r_1, r_2, ..., r_c\}$
30 Calculate absolute error, e_{abs} using Equation (6)
31 Record e_{abs} value in $E_{abs\text{-}coll} = \{e_1, e_2, ..., e_{(0.70)*N-c}\}$
32 Calculate relative error, e_{rel} using Equation (7)
33 Record e_{rel} value in $E_{rel\text{-}coll} = \{e_1, e_2, ..., e_{(0.70)*N-c}\}$
34 Record ε value in $\varepsilon_{coll} = \{\varepsilon_1, \varepsilon_2, ..., \varepsilon_{(0.70)*N-v}\}$
35 Calculate relative error, e_{rel} using Equation (8)
36 $m = m + c$ *// iterate*
37 **end**
38 Calculate the mean value of $E_{abs\text{-}coll}$, $E_{abs\text{-}mean}$
39 Calculate the mean value of $E_{rel\text{-}coll}$, $E_{rel\text{-}mean}$
40 Calculate the mean value of ε_{coll}, ε_{mean}
41 $E_{abs\text{-}mean}$, $E_{rel\text{-}mean}$, and ε_{coll} = null

Fig. 1. Pseudo-code of Kalman and LPC filter experimental setup.

A. Kalman Filter

Fig. 1 shows pseudo-code for the experimental setup for the Kalman filter. The inputs are the observation window size (c) and the input dataset (Y). The outputs are the mean absolute error ($E_{abs\text{-}mean}$), mean relative error ($E_{rel\text{-}mean}$), and mean RMSE (ε_{mean}) for the training and testing sets. The input dataset is divided into the training set (70%) and testing set (30%) to simulate artificial gaps in the data. Each observation window in the training set is analyzed by calculating the state vector estimates (x) using Equation (1) and then the observation vectors (z) using Equation (2). The state transition matrix (A) and observation matrix (H) are each set to an identity matrix. The initial state of the Kalman filter is not specified (x and P). The system input (control) functions are set to zero (B and u). The observation vector (z) is sensor data. We are assuming that noise is the only source of error in the observation vector [6]. Therefore, the process noise (Q) and measurement noise (R) are set to the sensor data variance. Temperature, humidity, energy usage, pressure, and airflow are all assumed to be Gaussian in nature to simplify the solution.

The residuals vector (R) is then calculated for each observation window by finding the difference between the vector estimates and training data points. The absolute error (e_{abs}), relative error (e_{rel}), and RMSE (ε) are calculated using Equations (6), (7), and (8) respectively and stored into collection vectors $E_{rel\text{-}coll}$, $E_{abs\text{-}coll}$, and ε_{coll} respectively. Once all of the observation windows have been analyzed, the mean values for absolute error ($E_{abs\text{-}mean}$), relative error ($E_{rel\text{-}mean}$), and (ε_{mean}) are calculated. The models generated during the training phase are used in the testing phase to predict values that are artificially missing (i.e., testing set). Absolute error, relative error, and RMSE are also calculated to determine the performance.

B. Linear Predictive Coding (LPC) Filter

Fig. 1 also shows pseudo-code for the LPC filter experimental setup. The LPC filter algorithm is similar to Kalman filter except for the modeling methodology which uses Equation (3) to calculate predicted values. The number of coefficients (p), is equal to the number of data points used to generate the coefficients (i.e., observation window size, c) minus 1.

V. Results

The results based on Kalman and LPC filters are shown below for temperature, humidity, energy usage, pressure, and airflow. We analyzed a variety of different observation window sizes: c=4 (1 hour), c=6 (1 ½ hours), c=12 (3 hours), c=24 (6 hours), c=48 (½ day), and c=96 (1 day). Our calculated performance metrics are absolute error, relative error, and RMSE.

A. Kalman Filter

Table III shows testing results for the Kalman filter. Based on these results, observation window sizes that produced the lowest error are c=12 for temperature, c=96 for humidity, c=48 for energy usage, c=12 for pressure, and c=6 for airflow (colored red). Even though a minimum error is easily identifiable for each data type, error is relatively low for all observation window sizes that were tested for temperature, humidity, pressure and airflow. The absolute error range is 0.6% - 5.6%, relative error range is 3.3% - 8.9%, and RMSE range is 0.435 – 21.66. This is not the case however for energy data where absolute error, relative error, and RMSE are significantly higher. The absolute error range is 9.7% - 10.5%, the relative error range is 468.5% - 495.0%, and the RMSE range is 13.16 – 13.92. One possible reason for the significant error is spikes (deviations from predominant values ranging from 0Wh – 35Wh) in the energy data which accounts for approximately 33% of the total data points. The large error indicates the Kalman filter is having difficulty predicting these spikes.

B. Linear Predictive Coding (LPC) Filter

Table IV shows testing results for the LPC filter. Based on these results, observation window sizes that produced the

lowest error are c=96 for temperature, c=96 for humidity, c=4 for energy usage, c=96 for pressure, and c=48 for airflow (colored red). The minimum absolute error range for all data types is 0.6% - 12.0% where pressure data has the only error above 10.0%. The minimum relative error is significantly higher than that of absolute error, with a range of 10.0% - 109.1%. The RMSE range for all data types is 0.98 – 91.76.

Comparing LPC filter with Kalman filter results, the LPC filter performs better with energy data while the Kalman filter performs better with temperature, humidity, pressure, and airflow data.

TABLE III. TESTING RESULTS USING KALMAN FILTERING TECHNIQUE.

	Temperature			**Humidity**			**Energy**			**Pressure**			**Air Flow**		
c	*AE*	*RE*	*RMSE*	*AE*	*RE*	*RMSE*	*AE*	*RE*	*RMSE*	*AE*	*RE*	*RMSE*	*AE*	*RE*	*RMSE*
4	3.9%	5.9%	2.92	5.6%	10.0%	5.07	10.5%	481.3%	13.39	3.7%	5.6%	18.81	**0.6%**	**3.3%**	**0.43**
6	3.8%	5.8%	2.89	5.6%	9.9%	5.16	10.0%	495.0%	13.16	3.6%	6.5%	18.73	0.3%	29.4%	0.24
12	**3.6%**	**5.3%**	**2.75**	5.3%	9.3%	4.99	9.8%	493.2%	13.40	**3.5%**	**7.4%**	**21.66**	0.9%	75.0%	0.79
24	3.6%	5.6%	2.85	5.3%	9.3%	5.22	9.9%	479.5%	13.90	4.2%	7.7%	29.93	0.3%	35.6%	0.34
48	3.6%	5.6%	2.91	5.3%	9.3%	5.46	**9.7%**	**468.5%**	**13.75**	4.3%	7.3%	36.25	0.4%	54.0%	0.74
96	3.6%	5.4%	2.95	**5.2%**	**8.9%**	**5.35**	9.7%	476.7%	13.92	4.4%	6.9%	38.01	0.7%	78.8%	1.38

TABLE IV. TESTING RESULTS USING LINEAR PREDICTIVE CODING (LPC) FILTERING TECHNIQUE.

	Temperature			**Humidity**			**Energy**			**Pressure**			**Air Flow**		
c	*AE*	*RE*	*RMSE*	*AE*	*RE*	*RMSE*	*AE*	*RE*	*RMSE*	*AE*	*RE*	*RMSE*	*AE*	*RE*	*RMSE*
4	69.8%	90.1%	52.52	58.6%	90.1%	53.27	**9.6%**	**109.1%**	**12.92**	48.5%	90.6%	247.85	0.9%	91.6%	0.70
6	56.0%	72.6%	45.55	46.9%	72.1%	46.17	9.6%	140.9%	13.52	39.3%	74.2%	214.89	0.7%	72.7%	0.86
12	33.3%	43.3%	33.37	28.9%	44.8%	34.44	9.3%	187.7%	13.83	26.2%	49.2%	167.17	0.7%	53.7%	0.65
24	17.7%	23.4%	23.14	16.8%	26.6%	24.28	9.1%	257.9%	13.93	17.7%	31.8%	131.58	0.8%	86.0%	0.98
48	10.5%	14.1%	16.33	11.2%	18.2%	17.25	8.8%	311.6%	13.65	14.2%	26.4%	108.22	**0.6%**	**66.0%**	**0.98**
96	**7.0%**	**10.0%**	**11.17**	**9.2%**	**15.1%**	**14.36**	8.7%	331.6%	13.32	**12.0%**	**22.5%**	**91.76**	0.8%	113.3%	1.32

VI. CONCLUSIONS

Sensor data validation is of great importance, particularly in regards to building technologies research where data are used to determine performance, analyze efficiency technologies, validate energy simulation engines, and calculate optimal retrofit packages. Kalman and linear predictive coding (LPC) filtering methods were used in this paper to predict missing or corrupt data for temperature, humidity, energy usage, pressure, and airflow sensors. A concept of observation windows is used which divides a set of data into subsets. Each subset is modeled using a filtering technique to predict missing or corrupt data points through interpolation and/or extrapolation. Data used in this study was taken from ORNL's ZEBRAlliance project which consists of four homes equipped with a variety of different energy-efficient technologies outfitted with hundreds of sensors to help in understanding technology impacts. Results from this study shows the Kalman filter performed best with temperature, humidity, pressure, and airflow data using observation window sizes of c=12 (3 hours), c=96 (24 hours), c=12 (3 hours), and c=4 (1 hour) respectively. The LPC filter performed best with energy usage data using an observation window size of c=4 (1 hour). Future work includes investigating machine learning techniques such as: (1) artificial neural networks; (2) fuzzy clustering; (3) Bayesian networks; and (4) hierarchal mixture of experts for sensor data validation and correction applied to building technologies.

ACKNOWLEDGMENT

Research sponsored by the Laboratory Directed Research and Development (LDRD) Program (WN12-036) of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy (DE-AC05-00OR22725).

REFERENCES

[1] U.S. Department of Energy. (2008). "Building energy data book," [Online]. Available: http://buildingsdatabook.eren.doe.gov/

[2] ZEBRAlliance. (2008). "ZEBRAlliance: building smart," [Online]. Available: http://www.zebralliance.com/index.shtml

[3] C.C. Castello and J.R. New, "Autonomous correction of sensor data applied to building technologies utilizing statistical processing methods," *Energy Informatics*, Atlanta, Georgia, October 6, 2012.

[4] G. Welch and G. Bishop. "An introduction to the Kalman filter," University of North Carolina-Chapel Hill, TR 95-041, July 24, 2006.

[5] L.B. Jackson, "Digital filters and signal processing," 2nd Edition, Kluwer Academic Publishers, 1989.

[6] VectornavFilter. (Date Unknown). *Tuning the performance of the VN-100* [Online]. Available: http://www.vectornav.com/