Machine Learning Techniques Applied to Sensor Data Correction in Building Technologies

Matt K. Smith
Department of Electrical and Computer Engineering
The University of Alabama
Tuscaloosa, AL 35486
Email: mksmith3@crimson.ua.edu

Charles C. Castello, Joshua R. New
Energy and Transportation Science Division
Oak Ridge National Laboratory
Oak Ridge, TN 37830
Email: castellocc@ornl.gov, newjr@ornl.gov

***Abstract*—Since commercial and residential buildings account for nearly half of the United States' energy consumption, making them more energy-efficient is a vital part of the nation's overall energy strategy. Sensors play an important role in this research by collecting data needed to analyze performance of components, systems, and whole-buildings. Given this reliance on sensors, ensuring that sensor data are valid is a crucial problem. Solutions being researched are machine learning techniques, namely: artificial neural networks and Bayesian Networks. Types of data investigated in this study are: (1) temperature; (2) humidity; (3) refrigerator energy consumption; (4) heat pump liquid pressure; and (5) water flow. These data are taken from Oak Ridge National Laboratory's (ORNL) ZEBRAlliance research project which is composed of four single-family homes in Oak Ridge, TN. Results show that for the temperature, humidity, pressure, and flow sensors, data can mostly be predicted with root-mean-square error (RMSE) of less than 10% of the respective sensor's mean value. Results for the energy sensor are not as good; RMSE are centered about 100% of the mean value and are often well above 200%. Bayesian networks have RSME of less than 5% of the respective sensor's mean value, but took substantially longer to train.**

I. Introduction

Commercial and residential buildings are the largest consumers of energy in the United States, accounting for 41% of the nation's total energy consumption [1]. Improving building energy efficiency is one of the most important energy challenges we face and better sensor technologies can help accomplish this goal. Sensors are used in buildings to analyze variables affecting energy use and efficiency. Given our reliance on sensors, ensuring that sensor data are valid emerges as a crucial problem.

Previous research at Oak Ridge National Laboratory (ORNL) has applied statistical techniques to populate and replace missing and corrupt data, respectively [2]. This was accomplished using a sensor's past data to generate a model through: (1) least-squares; (2) maximum likelihood estimation; (3) segmentation averaging; and (4) threshold-based averaging. In this paper, we incorporate data from other building sensors into the predictions. The two machine learning techniques we investigate here are artificial neural networks (ANNs) [3] and Bayesian Networks [4].

In the remainder of this work, we first provide a brief introduction of relevant building research and methods for sensor validation. We then describe our experimental methods and results. Finally, we examine the conclusions that may be drawn from the work.

II. Background

There are many studies regarding the efficient use of energy in residential and commercial buildings [5], [6], [7]. Christian [5] compares three single-family homes in Knoxville, TN, part of the Campbell Creek subdivision. Work by Norton and Christensen [6] review a case study of a 1,200 square foot, 3-bedroom zero energy home in the cold climate of Denver, CO. Miller and Kosny [7] discuss the experimentation of prototype residential roof and attic assemblies to determine heat transfer during peak day irradiance. A commonality among this research is the use of sensor data to help understand the impact on energy usage on component, system, and whole-building levels. With a reliance on sensor data, techniques are needed to ensure data completeness and accuracy.

There are currently two approaches used to validate data: hardware redundancy and analytical redundancy [8]. Hardware redundancy uses an increased amount of resources such as redundant sensors, data acquisition channels/systems, installation and maintenance labor, etc. to ensure missing and corrupt data are eliminated. However, this increases the cost substantially. Analytical redundancy uses mathematical models between measurements to predict a target sensor's values. The disadvantage of analytical redundancy is decreased efficiency when the number of sensors and complexity of the model increases. Feature selection techniques can be used to decrease the amount of sensors needed to generate the models [9].

III. Methods

A. Data acquisition

The sensors we used for this research came from ZEBRAlliance homes [10]. ZEBRAlliance is a collaboration by ORNL, Department of Energy (DOE), Tennessee Valley Authority (TVA), and industry partners to develop four of the most energy-efficient houses on the market. The houses were unoccupied during the study, but were made to simulate the average American according to Building America benchmarks through the use of robotically emulated occupancy via autonomously controlled temperature settings, lighting,

appliances, and human emulators. Each of the houses were equipped with over 250 sensors [2].

We chose to study data from five sensors from House #2 (out of four): (1) energy recovery ventilator (ERV) temperature, (2) ERV humidity, (3) refrigerator energy, (4) integrated heat pump (IHP) water flow, (5) and IHP liquid line pressure. These are called our target sensors. We used data from the month of January, 2011. Basic information about the sensors is given in Tables I and II.

TABLE I
SENSOR DESCRIPTIONS

Quantity	Units	Name	Location
Temperature	°F	Z09_T_ERV_IN_AVG	ERV
Humidity	%	Z09_RH_ERVin_AVG	ERV
Energy Usage	Wh	A01_WH_fridge_tot	Refrigerator
Pressure	psi	H01_PR_LiqAir_Avg	IHP
Liquid flow	gal	Hxx_WF_IHPtoTank_Tot	IHP

B. Setup

For both ANNs and Bayesian Networks, our basic method was to predict the data of our target sensors using other sensors' data as input. For each of our five target sensors, we chose five to eight other sensors to use as inputs for our networks using domain knowledge. Information regarding these sensors is shown in Tables III, IV, V, VI, and VII. These input sensors were chosen by using Joint Mutual Information (JMI) feature selection [9].

An observation window of size *n* is used to predict each time-step within that window as if the data were missing. The observation window moves forward by *n* time-steps with no overlap. This occurs for every possible window within a given set of time-series sensor data. In this study, we use 1-minute resolution data with an observation window of size 90, yielding 496 windows per sensor for the month of January. This window size was selected in order to decrease the complexity of the networks to be trained. For each window, we use 70% of the sensor data as the training set and the remaining 30% as the testing set. For both ANNs and Bayesian networks, we used the root-mean-square error (RMSE) between network predictions and actual data as our performance metric.

We used ANNs to predict data with polynomial regression models. For each network, we had one hidden layer consisting of ten nodes. The Levenberg-Marquardt backpropagation algorithm [11] was used to train the ANN.

For Bayesian Networks, we used each sensor as a node, with each input node x_i connected to the target node t. This is shown in Figure 1. The networks produce a multivariate

TABLE II
STATISTICAL SENSOR PROPERTIES

Data Type	Mean	Std. Dev.	Min.	Max.
Temperature	43.9	9.28	16.1	69.3
Humidity	59.7	18.6	16.0	106.0
Energy Usage	0.880	0.992	0.000	10.380
Pressure	268.0	71.9	91.7	443.9
Liquid Flow	0.429	1.293	0.000	4.679

TABLE III
LIST OF SENSORS USED TO CONSTRUCT Z09_T_ERV_IN_AVG

Priority	Sensor Name	Type	Description
1	H20_PR_LiqAir_Avg	Pressure	IHP Liquid Line
2	Z12_Tm_ERV_OUT_Avg	Temperature	ERV Exhaust
3	Z09_RH_ERVin_Avg	Humidity	ERV Intake
4	Z12_RH_ERVout_Avg	Humidity	ERV Exhaust
5	OD_Air_Avg	Temperature	Outside
6	WindDir_D1_WVT	Direction	Wind Direction
7	E22_SF_Avg	Temperature	Wall
8	E23_SF_Avg	Temperature	Wall

TABLE IV
LIST OF SENSORS USED TO MODEL Z09_RH_ERVIN_AVG

Priority	Sensor Name	Type	Description
1	Z10_RH_ERVsup_Avg	Humidity	ERV Supply
2	OD_AIR_Avg	Temperature	Outside
3	H21_PR_LiqBrine_Avg	Pressure	IHP Liquid Line
4	E10_SF_Avg	Energy/Area	Whole House
5	Z09_Tm_ERV_IN_Avg	Temperature	ERV Intake

TABLE V
LIST OF SENSORS USED TO MODEL A01_WH_FRIDGE_TOT

Priority	Sensor Name	Type	Description
1	WindDir_SD1_WVT	Direction	Stan. Dev.
2	OD_AIR_Avg	Temperature	Outside
3	Z11_Tm_ERV_Return_Avg	Temperature	ERV Return
4	E10_SF_Avg	Energy/Area	Whole House
5	A04_WH_ERV_Tot	Energy	ERV
6	WindDir_D1_WVT	Direction	Wind Direction
7	Z10_RH_ERVsup_Avg	Humidity	ERV Supply

TABLE VI
LIST OF SENSORS USED TO MODEL H20_PR_LIQAIR_AVG

Priority	Sensor Name	Type	Description
1	H32_Wh_IHP	Energy	IHP Total
2	H20_PR_LiqAir_Avg	Pressure	IHP Liquid Line
3	OD_AIR_Avg	Temperature	Outside
4	Z11_Tm_ERV_Return_Avg	Temperature	ERV Return
5	E10_SF_Avg	Energy/Area	Whole House
6	E70_Wh_IHP_tot_Tot	Energy	IHP Total Energy
7	Z12_RH_ERVout_Avg	Humidity	ERV Exhaust

TABLE VII
LIST OF SENSORS USED TO MODEL HXX_WF_IHPTOTANK_TOT

Priority	Sensor Name	Type	Description
1	A04_WH_ERV_Tot	Energy	ERV
2	NI_DAQ_Tot	Energy	Pulse Counter
3	Z13_WF_Shower_Tot	Water Flow	Shower
4	H32_Wh_IHP_comp_Tot	Energy	IHP Compressor
5	E71_Wh_Housetot_Tot	Energy	Whole House
6	E70_Wh_IHP_tot_Tot	Energy	IHP Total Energy
7	H35_WH_WHelem_Tot	Energy	Heater Elements

Gaussian distribution for a range of likely target values. We used the means of these distributions as the predicted value.

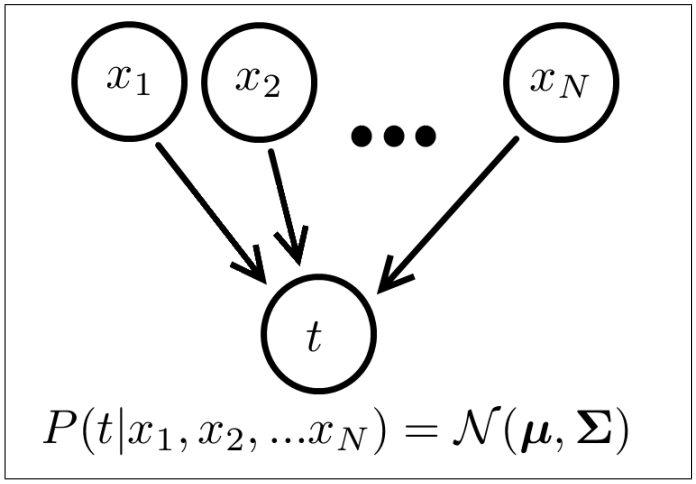

Fig. 1: The Bayesian Network we used. Each input sensor x_i is connected to the target sensor t. The resultant probability distribution for t is a multivariate Gaussian distribution, $\mathcal{N}$. We took the mean of $\mathcal{N}$ as the predicted value of the target sensor.

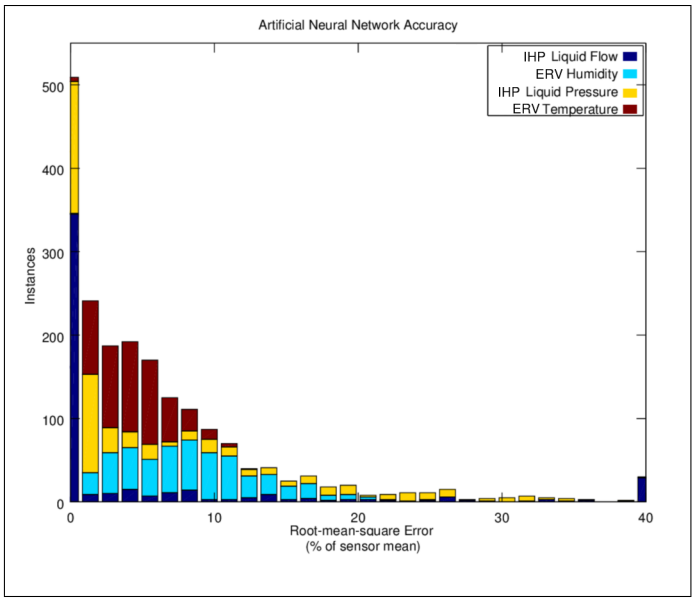

Fig. 2: Histogram of root-mean-square error (RMSE) of 496 ANNs each for heat pump liquid flow, temperature, humidity, and pressure sensors. The RMSE values are expressed as a percentage of their respective sensor's mean value.

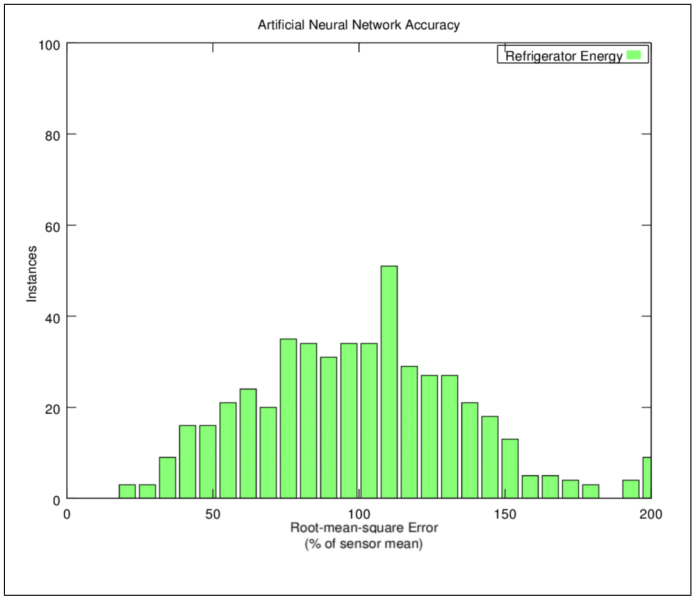

Fig. 3: Histogram of root-mean-square error (RMSE) of 496 ANNs for the refrigerator energy sensor. The RMSE values are expressed as a percentage of the sensor's mean value.

IV. Results

A. Artificial Neural Networks

ANNs proved to be quite effective at predicting data for four of the five sensors: (1) temperature; (2) humidity; (3) pressure; and (4) liquid flow. As Figure 2 shows, for these four sensors, the RMSE between the network output and the actual target data for 496 networks was less than 8% of the respective sensor's mean value.

This was not the case for the other sensor, refrigerator energy, as shown in Figure 3. For this sensor, the RMSEs are centered around 100%. We suspect this poor result is due to the refrigerator's large spikes in energy consumption when the compressor is operating and when the refrigerator doors are opened.

B. Bayesian Networks

Our Bayesian Network experiments had similar results. The humidity, pressure, temperature, and liquid flow sensors performed well. The majority of their RMSEs were under 10% of each sensor's mean value, as shown in Figure 4. The energy sensor also performed similarly for Bayesian Networks as it did for ANNs, as seen in Figure 5.

C. Comparison

A comparison of the average RMSE between ANNs and Bayesian Networks is shown in Table VIII. For all sensors, Bayesian Networks had superior accuracy. The other significant difference between the two methods was that the time required to generate results was substantially different. Each Bayesian Network took approximately eighteen seconds to train, while the average time to train each ANN was less than two seconds.

V. Conclusion

In this paper, we applied two machine learning techniques, artificial neural networks (ANN) and Bayesian Networks, to a sensor validation problem in building technologies research. Five types of sensors were investigated: (1) temperature; (2) humidity; (3) energy use; (4) liquid flow; and (5) liquid pressure. Both techniques obtained root-mean-square errors (RMSE) below 10% for temperature, humidity, liquid flow, and liquid pressure sensors. The fifth sensor, refrigerator energy,

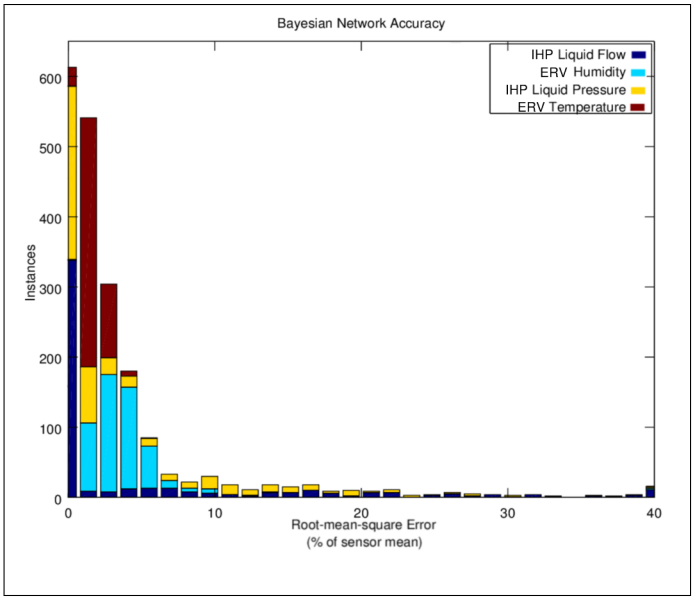

Fig. 4: Histogram of root-mean-square error (RMSE) of 496 Bayesian Networks each for heat pump liquid flow, temperature, humidity, and pressure sensors. The RMSE values are expressed as a percentage of their respective sensor's mean value.

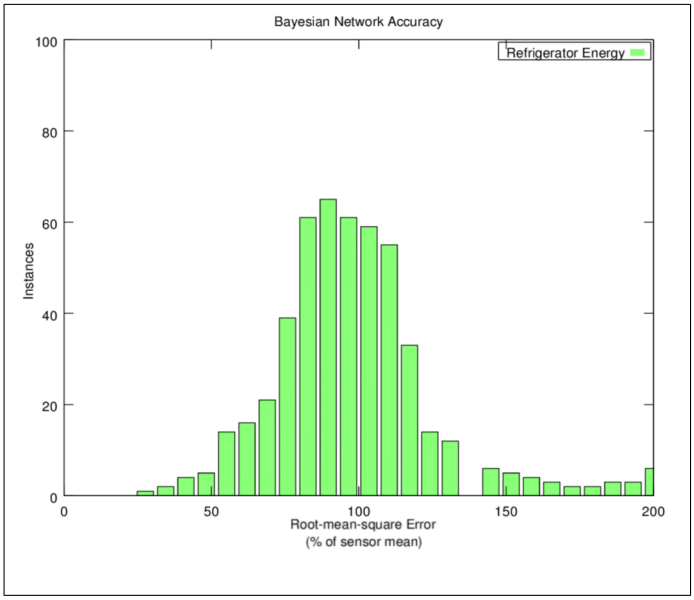

Fig. 5: Histogram of root-mean-square error (RMSE) of 496 Bayesian Networks for the refrigerator energy sensor. The RMSE values are expressed as a percentage of the sensor's mean value.

did not perform well with resulting RMSEs averaging around 100%. The large inaccuracies of using ANNs and Bayesian Networks on the refrigerator energy sensor may be due to the refrigerator's large spikes in energy consumption when the compressor is operating and when the refrigerator doors are opened. Future work will involve additional machine learning techniques such as hierarchical mixture of experts to decrease error, particularly involving energy data. Other sensors will also be investigated along with data from other buildings (e.g., commercial).

TABLE VIII
AVERAGE RMSE BETWEEN ANNS AND BAYESIAN NETWORKS.

	psi	%RH	°F	gal	Wh
ANN	17.7	5.0	1.9	0.03	0.89
Bayes	13.5	2.1	1.1	0.02	0.85
Difference	4.2	2.9	0.8	0.01	0.04

VI. ACKNOWLEDGEMENTS

Research sponsored by the Laboratory Directed Research and Development Program (WN12-036) of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy (DE-AC05-00OR22725). This work was also supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists under the Science Undergraduate Laboratory Internships (SULI) program.

REFERENCES

[1] Department of Energy, "Buildings energy data book," [Online], Available: http://buildingsdatabook.eren.doe.gov/ChapterIntro1.aspx.
[2] C. C. Castello and J. New, "Autonomous correction of sensor data applied to building technologies utilizing statistical processing methods," in *Energy Informatics*, Atlanta, Georgia, October 2012.
[3] M. Hassoun, *Fundamentals of Artificial Neural Networks*. A Bradford Book, 2003.
[4] A. Darwiche, *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.
[5] J. Christian, "Comparison of retrofit, advanced and standard builders homes in campbell creek," in *Thermal Performance of the Exterior Envelopes of Buildings*, Clearwater, Florida, 2010.
[6] P. Norton and C. Christensen, "A cold-climate case study for affordable zero energy homes," in *American Solar Energy Society*, Denver, Colorado, 2006.
[7] W. A. Miller and J. Kosny, "Next generation roofs and attics for homes," in *American Council for an Energy Efficient Economy*, Pacific Grove, California, 2008.
[8] P. H. Ibarguengoytia, L. E. Sucar, and S. Vadera, "Real time intelligent sensor validation," *Transactions on Power Systems*, vol. 16, no. 4, pp. 770–775, 2001.
[9] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.
[10] ZEBRAlliance, "Zebralliance: building smart," [Online], Available: http://www.zebralliance.com/index.shtml.
[11] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the marquardt algorithm," *Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, November 1994.