Vetting the Energy and Security of Smart Buildings with Data Science

Total Funds Request: $6,000

September 6th, 2018

Student Co_PIs:

Zhen Li
li2215@purdue.edu
Ecological Sciences and Engineering Program & Department of Civil Engineering
Purdue University
[no advisor]

Dinuka Harshana Gallaba
dinuka4@siu.edu
Department of Physics
Southern Illinois University
Advisor: Prof. Aldo D.Migone

MyVan Vo
vom@purdue.edu
Department of Mathematics
Purdue University
Advisor: Prof. Zhilan Feng

Jan Moffett
janlmoffett@gmail.com
Department of Mathematics and Statistics
Eastern Kentucky University
[no advisor]

1. Problem Statement

Cyber Physical Systems have numerous applications in civil, energy, transportation, and manufacturing fields. Meanwhile, critical infrastructures are increasingly interwoven with cyber components such as sensing, computing, and control devices [1][2]. An ad hoc issue in Cyber-Physical Infrastructures (CPIs) is the security of the system; namely, how to protect it from cyber and physical attacks. Research into strategies to vet a CPI and to minimize the influence of attacks is urgent.

With the rapid development of sensing technologies, i.e. Internet of Things (IoT), modern buildings are becoming more intelligent and automated [3]. Considering the gap between the areas of cybersecurity and physical protection and of novel smart and connected sensors, it is necessary to learn information from the data in intelligent buildings to maintain them in a secure environment. This project aims to use a smart building as a platform to develop data mining technologies in intelligent buildings and to apply them to common Cyber Physical Infrastructures, which will have significant influence on system security.

2. Intellectual Merit & Broader Impact

This research is based on a research level residence. The chosen residence for the case study is the ReNEWW House at Purdue, which has been designed specifically for research purposes. In this project, patterns from the different data sources will be modeled and validated. Beyond this, the potential for cyber or physical attacks on smart buildings can be vetted with the data flow in the systems through data mining technologies. The proposed research aims to evaluate and predict the performance of different systems of buildings and infrastructures through data mining, to explore the relationship between different systems, and to develop a platform to vet potential attacks or abnormal events in the infrastructures. It is an example of a learning and statistics project as it will use data science and statistic models. It is related to human working and living conditions, as well as physical and cyber security. Moreover, it is a multidisciplinary study as it requires techniques from different fields of science, such as mechanical engineering, electrical engineering, architecture, statistics, and mathematic, as well as data science and programming.

3. Proposed Activity

Different kinds of data can be collected through the pre-installed control system or the customer installed sensors. Since the large amounts of data have different formats and meanings, it is necessary to use data analysis tools. The team will use R as the main data analysis package along with the other software like Python and SAS.

The installed sensor network provides continuous data regarding the utility (water and energy) consumption in real or near-real time. We have already accumulated data over a period of three years, and there will be more data available in future. Since the volume of data generated in the smart house sensor network is comparably high and is being created in near-real time (i.e. has high “velocity”) from various types of sensors, constitutes it a big data problem. Although the volume of the data is not yet qualified to be in the regime of “big data”, as we continue the project,

we anticipate a growing volume. As a result of this, we will treat our problem as a big data problem, so we will utilize data mining techniques for the cleaning and classification of data.

The data collected from the smart house has a time dependency (Temporal Data), so we will use time dependent data mining techniques (Temporal Data Mining-TDM) [4]. All the standard classification and clustering can be used in analyzing the time domain (time series analysis), and for some cases, a domain change (i.e. Discrete Fourier Transformation-DFT) [5] may be helpful. The behavior of any categorical data available in the dataset will be analyzed using Markov Models. Autoregressive integrated moving average (ARIMA) will be used for the continuous sequence data.

To test our algorithms for potential "attacks", a simulated data set will be generated, and further investigations will be carried out to acquire real-world anomalies. The algorithms that we are going to develop will be tested on real-world smart building data sets available for machine learning research from the University of California Irvine: Machine Learning Repository [6].

For this project, there are three main tasks to perform in order to achieve our goals: 1) data acquisition and interpretation; 2) statistical and mathematical model selection and application; 3) System development. The corresponding and detailed activities are listed as follows:

Preliminary Task: Literature Review

All members have to learn the basic building systems and accessed data, review the studies about the building system relationships, and research the data collection system used in this case. A comprehensive literature review will be needed.

Task 1: Data Collection and Processing

One team member has collaborated with the ReNEWW House, so the team can access the data collected in the building. For water consumption, electricity usage, and solar energy harvested, the data are sampled every five minutes since 2016. That means it is hard to process the data directly. The team plans to review the data and pre-process it before applying data mining. For instance, data can be grouped or a moving average can be utilized before analysis to reduce the size [7]. With data mining, the team will conduct a preliminary statistical analysis, such as the normal range for data in one system and identify any outlier data points.

Task 2: Model Development

The team will implement critical studies with time series and correlation analyses. This approach will relate the statistical and mathematical model. With the model, the working condition of each system can be accessed, and "attacks" will be detected. In this study, 70% of the data will used for model training and 30% of the data will be used for model validation.

Task 3: Algorithm/ System Design

An algorithm or system is expected to be designed then applied to general buildings with similar properties like the ReNEWW House. It should be able to provide the data pattern for systems and detect issues or attacks in the systems by comparing the real time data with predicted data. The validation of the algorithm/ system will be conducted with simulated data from a

constructed simulation program or from real-world datasets from research institutes as mentioned above.

Team Schedule

- Bi-weekly online meeting: The team will have a meeting every other week via Skype or WebEx to discuss the progress and problems.
- Two in person meeting: The team will have two in-person meetings. One around spring break to finalize the model and one at Purdue during the conference.
- Conference meeting: The team plans to attend one conference to present our research, where the team can meet in-person to work on the project face to face.

4. Goals and Outcomes

The primary goals are to 1) find the data patterns of different building systems in the research house, which includes energy systems, water systems security systems, etc; 2) explore the correlations of different building systems with data mining methods; and 3) develop an algorithm to predict system performance with history data. The eventual goal of this project is to design algorithms that will predict the performance of most of the systems infrastructures, diagnose abnormal events, and protect different systems.

The team will present the result in Center's NSF site visit in December. In addition, the team expects to 1) attend one of the well-known high-performance building related conferences such as the annual meeting of American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) in 2019 or 2020; 2) participate in related poster sessions around campus; 3) convert the result to a paper for publication.

5. Proposed Work Statement

To meet the goals and accomplish the outcomes, each team member will have specific roles and tasks based on their interests and backgrounds.

Zhen Li: Her research expertise is environment sensors and data acquisition. Apart from data interpretation and data mining, as an engineer, she will collaborate with the ReNEWW House and monitor the sensor network.

MyVan Vo: As a mathematician, her expertise is mathematical modelling, programming, and data analysis. She will contribute in adapting machine learning algorithms for data mining and system predictions.

Dinuka H. Gallaba: His expertise is data mining and machine learning techniques in scientific computation. As a physicist, he is also interested in data acquisition and time series analysis.

Jan Moffett: Her research area is statistical data analysis and data visualization. She will contribute in time series analysis and incorporate statistical programing in the big data analysis.

Our successful collaboration during the workshop on "*Introduction to Data Science & Interdisciplinary Research Teams"* will lead us to successful completion of this proposed research project.

6. Diversity Statement

The team stands to support diversity through a four-member team with mixed genders, ethnicities, regions of residence, disciplines and education levels. Through this diversity, the team will strengthen the Center's mission to support women, United States citizens, and other members of underrepresented groups in the science of information field:

- Three members of the research team are female.
- Two members are US citizens and two members are international students.
- Four members are in four different Engineering and Science majors.

7. Budget and Justification

As mentioned in Section 3, we plan to meet in person to discuss progress, debug codes, and present our project at the related leading conference in data science or high-performance buildings. The following is an estimate of expenses associated with the proposed activities:

Table 1. Budget and justification

Items	**Cost**	**Justification**
Conference		
Air Travel	4 x $520	Average flight cost for four team members
Registration Fee	4 x $200	Registration fee for four team members
Lodging	4 x $400	Hotel cost for four team members
In-Person Meeting*		
Dinuka-Lodging	3 x $150	On-campus hotel cost for three nights at Purdue
Dinuka-Air Travel	$315	Driving cost for 290 miles to and from Purdue
Jan-Lodging	3 x $150	On-campus hotel cost for three nights at Purdue
Jan-Air Travel	$305	Driving cost for 280 miles to and from Purdue
One-Year Total	**$6000**	

*Zhen and MyVan are at Purdue and will not need support for travel and lodging for the in-person meeting.

Reference

[1] He, H., & Yan, J. (2016). Cyber-physical attacks and defences in the smart grid: a survey. *IET Cyber-Physical Systems: Theory & Applications, 1*(1), 13-27.

[2] Berkeley, A. R., Wallace, M., & COO, C. (2010). A framework for establishing critical infrastructure resilience goals. *Final Report and Recommendations by the Council; National Infrastructure Advisory Council: Washington, DC, USA.*

[3] Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future generation computer systems, 29*(7), 1645-1660.

[4] Mamoulis N. (2009) Temporal Data Mining. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

[5] Lin, W., Orgun, M. A., & Williams, G. J. (2002, December). An overview of temporal data mining. In *Proceedings of the 1st Australian data mining workshop* (pp. 83-90).

[6] UCI Machine Learning Repository. (n.d.). Retrieved September 7, 2018, from https://archive.ics.uci.edu/ml/index.php

[7] Wei, W. W. (2006). Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*.