Occupancy Estimation Using Sensor Data Analytics

Diksha Bathula[1], Divya Bansal[2], E Biphul Krishna[3], Divya Lohani[4]

Department of Computer Science & Engineering

Shiv Nadar University

Gautam Buddha Nagar, U.P., India

Email: [1]db452@snu.edu.in, [2]db435@snu.edu.in, [3]ek364@snu.edu.in, [4]divya.lohani@snu.edu.in

***Abstract*—IoT devices generate data at a remarkable speed which requires near real-time processing. Such need has inspired a new computing paradigm that advocates moving computation to the edge, closer to where data is generated for ensuring low-latency and responsive data analytics. This research aims at estimating the occupancy in a university classroom along with monitoring some environmental parameters inside the classroom in real time. The estimation is achieved by successfully combining edge analysis and cloud communication. This is done by measuring different environmental parameters like temperature, CO_2 etc., and then by performing regression on the collected data. For the purpose of real time monitoring an android application is used which uses MQTT for communication with the edge node. The estimation is then verified against the actual attendance in the classroom at that period of time and assessed for prediction and error.**

Keywords—occupancy estimation; CO_2; temperature; edge node; cloud communication; regression; MQTT; android application

I. INTRODUCTION

People observation and counting is of interest in many commercial and non-commercial scenarios. The number of people entering and leaving shops, occupancy of office buildings or the passenger count of commuter trains provide useful information to shop merchants and marketers, security officials or train operators. Occupancy of a building is an important and useful parameter for efficient building operations, such as HVAC control, lighting control, security monitoring, and emergency evacuation. Building security can be improved as well as building energy can be saved by utilizing the occupancy information. Demand controlled ventilation (DCV) is an energy-saving way to provide fresh air according to ventilation demand. If the ventilation inside a room is very less, the occupants feel suffocated and it affects their health. Xu and Wang (2007) [1] reported that energy consumption can be reduced by as much as 8-33% by utilizing various DCV schemes. Klein (2011) [2] conducted multi-agent simulations and operated an air-conditioning system according to the number of occupants. He reported that energy consumption was saved by 12-17%, and that the indoor comfort level was improved by 5% [3].

The motivation behind this project lies in understanding and predicting the occupancy patterns in a classroom owing to the data from environmental parameters like temperature, humidity, carbon dioxide composition etc. The aim is to understand how the number of people affect these parameters and vice-versa. The knowledge of the occupancy in a room will help the students to make informed choices about where to study and meet, and the university can make informed decisions about a classroom's usage for e.g. room popularity, energy consumption, scheduling of examinations etc. Moreover, monitoring the environment will help in controlling the indoor lighting and the Heating, Ventilation and Air Conditioning systems (HVAC) in the rooms. Hence, this project would also help in increasing efficiency, saving electricity and reducing costs. Studies have shown that around one-third of the energy consumed in buildings can be saved using occupancy-based control. Therefore, indoor occupancy estimation has also become a popular area of research in recent years [4][5].

In this paper, the authors have conducted a regression-based study for estimating occupancy of a university classroom using sensor data analytics. Environmental parameters like temperature, humidity, CO_2 have been monitored in real time with the help of non-intrusive environmental sensors inside the classroom. Once the estimation is accomplished, the data is used to establish correlations between the different parameters and the estimated occupancy.

The paper is organized as follows: Section II reviews various techniques for occupancy estimation. Section III discusses the experimental setup with focus on hardware and software used. Section IV discusses the proposed regression models and their results. Section V concludes the paper and Section VI discusses the scope for future work.

II. LITERATURE REVIEW

Initial research literature indicated that many technologies used are either too costly (e.g. video cameras) or apply wireless sensor networks (WSN) primarily to in- network data analysis and distributed counting algorithms rather than the communication of actual people count. In consequence, such systems hold an inherent level of complexity that increases the cost of deployment.

A lot of other sensing techniques for occupancy estimation were studied before finalizing the use of IR sensors for the training data and the multi-sensor fusion of environmental sensors and analytics for the final prediction. The technologies were wearable sensors [6], vision-based sensors [7], ultrasonic sensors [8], IR sensors [9], sensor fusion methodology [10]-[12], and CO_2 sensors [13].

Many studies [14] [13] have used regression as a technique of estimation. In [14], authors have used Least Square Regression estimation to find the relation between CO_2 and number of people in a room. CO_2 acted as the predictor variable and the occupancy estimated was 95% accurate in comparison to other others in literature. We have drawn inspiration from this model and expect to achieve almost or more accuracy. Since our classroom has HVAC running at all times, CO_2 levels may or may not be as accurate as defined by [14].

III. Experimental Setup

A. Assumptions

- There is only one entry or exit in a room.
- People walk at a normal speed. No person is walking in a hurry or in a leisure manner.
- All the people who enter or leave the room, walk parallel to the IR sensors.
- People should enter or leave the room one at a time.
- Metabolic rate for usual classroom activities is in the range of 1-2 MET.

B. Environmental Parameters

The following parameters have been considered for this study:

- **Temperature:** Room occupants add heat to the room since the normal body temperature is much higher than the room temperature. ASHRAE guidelines [15] recommend 20°C to 24 °C in the winter and 23 °C to 27 °C in the summer as a comfort level.
- **Relative Humidity:** Room occupants add considerable moisture to the room through exhaled air which is at 100% relative humidity. ASHRAE guidelines [15] recommend RH of 30 - 60 %.
- **Luminosity:** The sole purpose of this parameter is to reduce the energy consumption. When the room is occupied, the light detected by the sensors slightly decreases due to the shadows of moving people.
- **Carbon dioxide (CO_2):** CO_2 concentration in an occupied indoor space indicates if the building's air exchange balance is appropriate – that is, if the optimal amount of outside air is being mixed with air that has been circulating in the building [8]. The CO_2 levels in the air outside a building are usually 380 ppm or higher. An elevated indoor CO_2 concentration is directly related to the number of occupants in the building, the kind of physical activity they are doing, the building's ventilation rate, and the CO_2 level in the outside air. ASHRAE recommends that indoor CO_2 levels should not exceed 1000 ppm [15]

C. Data Collection Setup

To measure the environmental parameters, two Sensordrones [Fig.1] were place two meters apart on the window sill on the left side of the classroom before the start of each class. This place was chosen so that the sensors could be placed as close as possible to the breathing level of the students. The collected data was transferred to a Raspberry Pi via Bluetooth at discrete intervals of 20-30 seconds for each drone. The implementation was carried out by a Java program deployed on the Raspberry Pi.

Fig. 1. Sensordrone with CO_2 Module

For counting the number of people entering the classroom, a setup [Fig. 2] consisting of two IR sensors, Arduino Uno R3 and Raspberry Pi was placed on a small table just near the entrance of the classroom. The orientation was such that all the people entering the room had to pass in parallel to the front of the IR sensors.

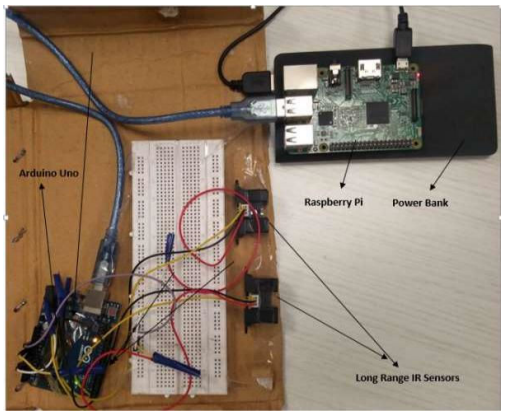

Fig. 2. Hardware Setup

To understand the working, let us assume that the sensor nearer to the door was IR1 and the sensor farther to the door was IR2. So, if a person crossed IR1 first and then IR2, it means that they were entering the room and the people count was increased by 1. On the contrary, if IR2 detected a person before IR1, it meant that the person was leaving, and people count was decreased by 1. These IR sensors had an analog output, so an Arduino board was used to collect the data which was transferred to the Pi using the serial port.

The data collected from the Sensordrones is stored in a csv file and transferred to Raspberry Pi via Bluetooth. Multiple programs were deployed on the Pi for the purpose of counting the people and taking the average over time. These calculated

average values were uploaded to the cloud (ThingSpeak) with the help of HTTP communication protocols. While calculating these average values, if any anomalies were found, he user was alerted via an android app. The communication between the Pi and the app is done with the help of MQTT protocol. User could also visualize the data on the cloud through this app.

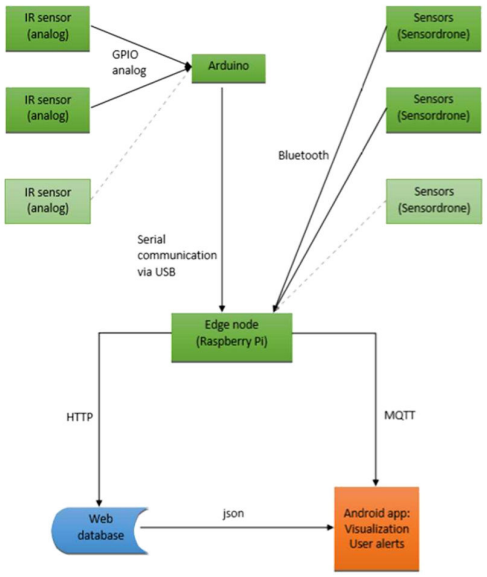

Fig. 3. Flow Chart for Network Implementation

D. MQTT Publish

The data being collected at the Raspberry Pi was simultaneously being published via MQTT after every 30 seconds. An MQTT broker called Mosquito was set up on the Raspberry Pi to facilitate this communication. A python script was used to publish this data using the Paho MQTT library.

E. Smart Class Monitor

An Android application [Fig. 4] for monitoring the indoor parameters of the classroom was developed using the Android Studio. The app attaches a service to subscribe to the MQTT channel to stay connected to the Raspberry Pi. When the app is closed, the device can still get push notifications just in case something goes wrong. The user has an option to unsubscribe to the MQTT service if he does not wish to receive any notifications.

IV. Data Analysis And Regression On Raspberry Pi

The primary reason for choosing regression analysis as the technique for occupancy estimation is that it is widely used for predicting, forecasting, finding correlation between the dependent and independent variables and analyzing the causal relationships for each of them. Once a regression model is developed and estimated, it is important to check the goodness of fit of the model and statistical significance of the model. The commonly used check is R-Squared (percent of variance explained by the model). In general, the higher the R-Squared, the better the data. But this is not true at all times. In various scenarios, R Square values are expected to be lower, especially in the field of psychology, human prediction model has an R Squared lower that 50%. This attributes to the fact that humans are harder to predict [16]. Once regression is achieved, prediction of the model is checked by plugging the average values in the regression equation. Once predicted values are obtained, they are compared against actual values to estimate error and accuracy and also compared against RMSE values of independent models.

The collected data is cleaned where all the non-integer values and most of the recognized outliers are eliminated from the database. The database is then made uniform by eliminating the data collected in only one class in a three-week period and the data which was taken after more than 30 minutes after the class had started. Data cleaning has improved the efficiency of the model and can account for all the correlations in the regression estimation.

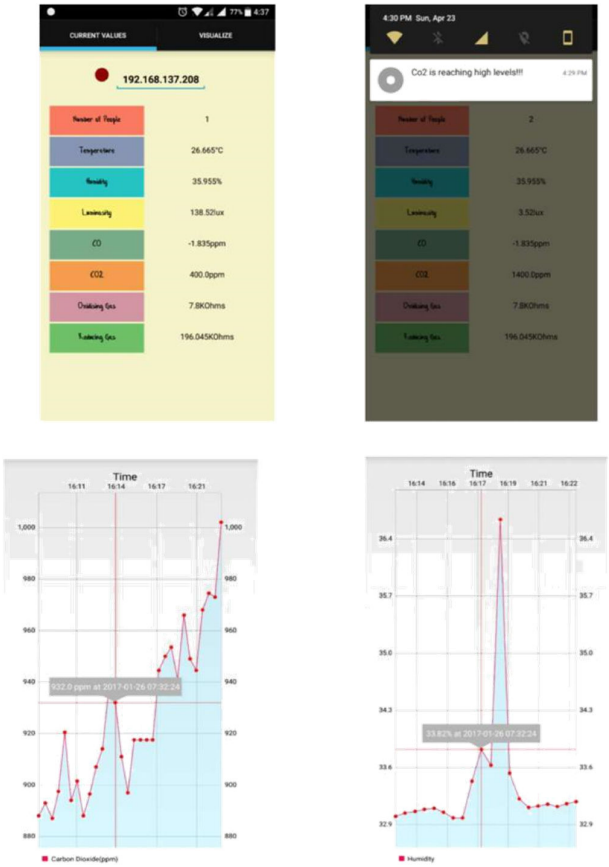

Fig. 4. Screenshots of Smart Class Monitor App: Introduction Page, Notification Pop Up, Visualization (Clockwise)

Dummy Variables: A dummy variable in regression is used when subgroups need to be represented in the analysis. It is a numerical variable and is used to distinguish treatment groups [17]. They act like any other independent variables but are not given importance in terms of significance and multicollinearity [18]. In our regression model, we have used dummy variables for each course. If the data is collected in that course, the dummy is activated to 1 or it is 0 at all other

times. This is done because the numbers of students enrolled in a particular course were inherently different and also the environmental conditions of a particular course were similar depending on time of the day of the lecture or if the classroom was empty before the lecture.

The dummy variables are as follows:

- D0 = Monday, Wednesday and Friday: 9-10 AM
- D1 = Monday, Wednesday and Friday: 10-11 AM
- D2 = Monday, Wednesday: 4-6 PM
- D3 = Monday, Tuesday and Friday: 2-4 PM
- D4 =Wednesday, Thursday: 2-4 PM & 3:30-5:30 PM
- D5 = Saturday: 8-9 AM

Fig. 5. Cloud Visualization for Occupancy, Temperature, Luminosity, Oxidizing Gas, CO_2, Reducing Gas, Precision Gas and Humidity (clockwise)

MODEL 1:

$$\boldsymbol{Occupancy = 55.47790 + 0.6097(Temperature) - 0.8926(Humidity) - 0.0193(Luminosity) + 6.1643(Precisiongas) + 0.7054(Oxidizinggas) - 0.0347(Reducinggas) + 0.0082(CO2) + 26.5589(D0) + 13.4227(D1) + 10.9648(D2) + 17.2214(D3) + 11.5300(D4) - 24.2220(D5)}$$

The independent variables are moderately correlated with each other [Table 1]. Hence, to check for a better fit and gradually even better model with no issues of multicollinearity, we regressed occupancy on independent parameters.

Variable	VIF	1/VIF
D4	10.03	0.099678
D2	7.57	0.132051
D3	6.09	0.164237
D0	4.62	0.216527
humidity	3.88	0.257678
D1	3.73	0.267899
oxidizinggas	3.65	0.274196
reducinggas	3.47	0.288209
co2	3.28	0.305061
precisiongas	2.41	0.414178
luminosity	2.10	0.476580
temperature	1.70	0.586591
Mean VIF	4.38	

Table 1. Multicollinearity Table

MODEL 2:

$$\boldsymbol{Temperature = 27.4948 + 0.0058(Occupancy)}$$

This model has very low R-Squared (0.004), meaning that Temperature can't be explained by Occupancy alone. Hence, we understand the importance of dummies and incorporate them in all our subsequent models.

MODEL 3:

$$\boldsymbol{Temperature = 27.6254 + 0.0121(Occupancy) - 0.4983(D0) + 1.5571(D1) - 2.4454(D2) - 2.2121(D3) + 1.3697(D4) - 0.0355(D5)}$$

This model has good R-Squared (0.528). We can see a minor positive correlation between temperature and occupancy meaning, as the number of people in the room increases, so does the temperature.

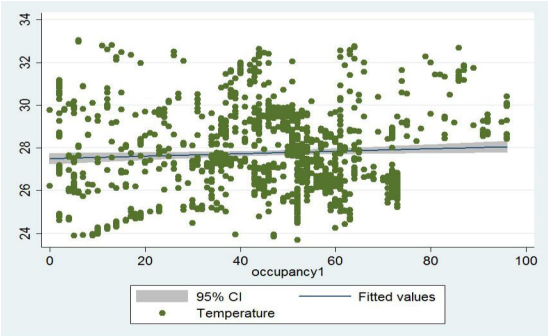

Fig. 6. Regression for Occupancy and Temperature

We can observe that the temperature initially was very random as people moved in and subsequently the HVAC was

on with the mean value around 26-27 degrees until it stabilized to values between 28 and 30 when the classroom reached full occupancy.

MODEL 4:

$$\boldsymbol{Humidity = 30.8523 - 0.0403(Occupancy) + 3.0247(D0) + 3.0247(D1) + 0.2715(D2) + 9.0998(D3) + 5.8656(D4) + 2.9870(D5)}$$

This model has good R -Squared (0.405). The line of best fit informs us that though the values of humidity are scattered over the range of 20 to 50. There is negative correlation between humidity and occupancy, meaning as the number of people in the room increase, the humidity decreases. The 95% confidence level has humidity between 40 and 30 [Fig. 7] and this is observed to be the humidity in the room at most times irrespective of the number of people in the classroom. Also, after a certain point when then number of people and the level of humidity in the classroom is constant, the HVAC kicks in decreasing the value of humidity. This accounts for the negative correlation in the regression model.

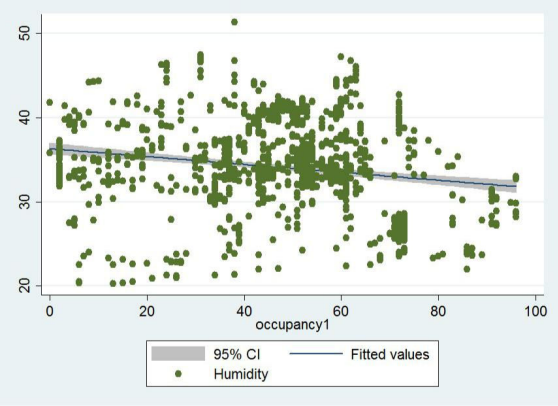

Fig. 7. Regression for Occupancy and Humidity

MODEL 5:

$$\boldsymbol{Luminosity = 207.8167 - 0.7594(Occupancy) + 47.9490(D0) + 31.1601(D1) - 54.9060(D2) + 123.6111(D3) - 2.5266(D4) + 62.5291(D5)}$$

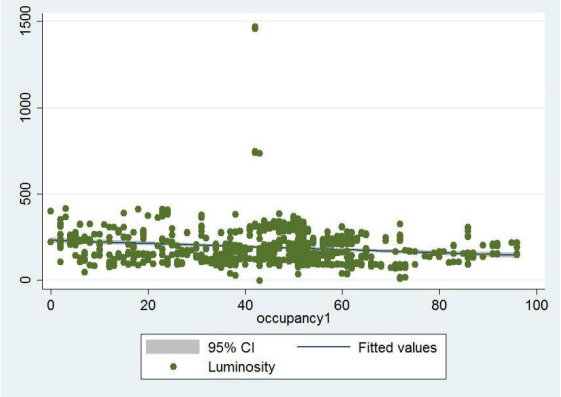

Fig. 8. Regression for Occupancy and Luminosity

This model has good R -Squared (0.382). The line of best fit informs us that though the values of the luminosity are scattered over the range of 40 to 1400. There is a negative correlation between luminosity and occupancy, meaning as the number of people in the room increases, the luminosity decreases. The 95% confidence level has humidity between 0 and 500 [Fig. 8] and this is observed to be the luminosity irrespective of the number of people in the classroom. It is expected that the values above 500 to be outliers due to the placement of the Sensordrone which might be too close to the window. The negative correlation is due to the fact that as the room fills with students, the luminosity is expected to decrease due to the formation of shadows in front of the Sensordrone.

MODEL 6:

$$\boldsymbol{CO2 = 816.8377 + 0.9109(Occupancy) + 134.9739(D0) + 405.6851(D1) + 84.6851(D2) + 323.7445(D3) + 249.7198(D4) - 381.5043(D5)}$$

This model has a low to moderate R-Squared (0.270) with all significant variables. The low R Squared is due to the multicollinearity of CO_2 with oxidizing gas and temperature; as temperature increases, so does CO_2 in a room. The line of best fit informs us that though the values of the oxidizing gas are scattered over the range of 400 to 1800 [Fig. 9], there is positive correlation between CO_2 and occupancy and as the number of people in the room increase, so does the CO_2.

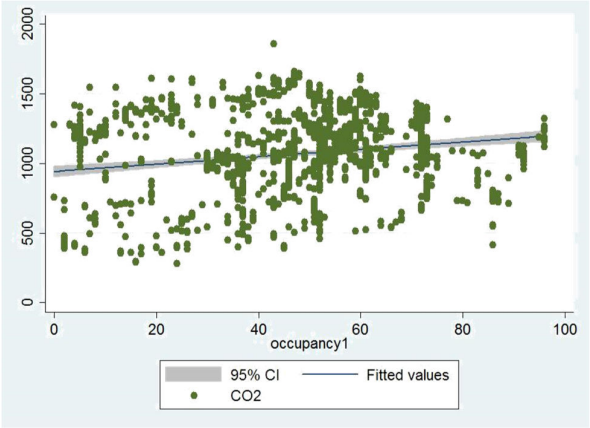

Fig. 9. Regression for Occupancy and CO_2

Prediction and Estimation: For any regression model, as R-Squared is a relative measure of fit, RSME is an absolute measure of fit. RMSE can be interpreted as the square root of variability of unexplained residuals, meaning lower the value of RMSE, the better the model. RMSE is a good measure of how accurately the model predicts the response and is the most important criterion for fit if the main purpose of the model is prediction [20]. Also, these models were trained on the data collected in the first three weeks and tested for errors on the data of the last week. The actual and predicted values were compared for checking the accuracy.

The main sources of error can be attributed to the erroneous nature of low cost off the shelf sensors, students entering in a non-uniform manner in the classroom, students entering in groups, data loss due to communication issues.

Table 2. RMSE Values for Each Model

Model	Root Mean Square Error (%)
1	17.960355
2	1.8885
3	1.7392
4	3.9307
5	75.682
6	25.218

Taking Models 3,4 and 6 into consideration, which are the models for temperature, humidity and carbon dioxide, it can be concluded that the average percentage of accuracy is 90%. The three environmental parameters are taken into consideration given their importance in variability in a due classroom due to HVAC. Since luminosity doesn't vary with HVAC, the model is ignored.

V. CONCLUSION

This paper focused on occupancy estimation of an indoor space using sensor data analytics. Predictive modelling has been performed which exhibited an accuracy of up to 90 – 95% for individual models. Though the R-squared values in the established model were low, previous studies show us that, in a human pattern prediction model, if we have significant variables then it is a good model. There were many reasons behind the insignificance of particular variables, but the main reason was the HVAC. This can be justified by looking at the low root mean square error values for the proposed models. Edge analytics using various regression models have been carried out on Raspberry Pi and cloud communication have been established for data storage and visualization. MQTT has been used as an application layer protocol to facilitate scalability and reduce latency.

VI. FUTURE SCOPE

In future, the authors aim to use the results obtained from this work in energy management in the university premises and studying the effect of occupancy on students' productivity. Additionally, studying other indoor scenarios like accessing indoor air quality, monitoring gaseous levels in a vehicle etc. can give a good insight into the feasibility of using the proposed models. Since data collected in each of these applications will increase manifold, our proposition will be to perform analytics on the cloud and keep a copy of the estimated equation, which will change weekly, on the Pi to predict the values in real time. This way we will be able to make the most efficient use of the cloud and the edge. Finally, to understand how HVAC affects the variability in the values of temperature, humidity and carbon dioxide, the authors would look into predictive and forecasting models to verify the suitability of regression models.

REFERENCES

[1] Xu, X. and Wang, S. (2007). An Adaptive Demand-Controlled Ventilation Strategy with Zone Temperature Reset for Multi-Zone Air-Conditioning Systems. Indoor and Built Environment, 16, 426-437.

[2] Klein, L. (2011). Coordinating Occupant Behavior for building Energy and Comfort Management using Multi-Agent Systems. Automation in Construction, 22, 525-536.

[3] Han, Hwataik, et al. "Occupancy Estimation Based On Co_2 Concentration Using Dynamic Neural Network Model."

[4] F. Oldewurtel, D. Sturzenegger, and M. Morari, "Importance of occupancy information for building climate control," Appl. Energy, vol. 101, pp. 521–532, 2013.

[5] J. Brooks, S. Goyal, R. Subramany, Y. Lin, T. Middelkoop, L. Arpan, L. Carloni, and P. Barooah, "An experimental investigation of occupancy-based energy-efficient control of commercial building indoor climate," 53rd IEEE Conf. Decision Control (CDC), 2014, pp. 5680–5685.K.

[6] LI, N., CALIS, G. & BECERIK-GERBER, B. 2012. Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations. Automation in Construction, 24, pp. 89-99.

[7] CHEN, T., CHEN, T. & CHEN, Z. 2006. An Intelligent People-Flow Counting Method for Passing through a Gate. In Proceedings of IEEE Conference on 244 Robotics, Automation and Mechatronics, 1-3 June, Bangkok, Thailand, pp. 1- 6

[8] Quanbin Chen, Min Gao, Jian Ma, Dian Zhang, Lionel M. Ni, and Yunhao, Liu.Moving Object Counting using Ultrasonic Sensor Networks. International Journal of Sensor Networks, 2008, pages 55–65, volume 3.

[9] K. Hashimoto, K. Morinaka, N. Yoshiike, C. Kawaguchi, and S. Matsueda. People Count System using Multi-Sensing Application. In Proceedings of International Conference on Solid State Sensors and Actuators, 1997. TRANSDUCERS, pages 1291 – 1294, volume 2.

[10] L. M. Candanedo and V. Feldheim, "Accurate occupancy detection of an office room from light, temperature, humidity and CO_2measurements using statistical learning models," Energy Build., vol. 112, pp. 28–39, 2016.

[11] B. Dong, B. Andrews, K. P. Lam, M. Hoynck, R. Zhang, Y.-S. Chiou, ¨ and D. Benitez, "An information technology enabled sustainability test-11 bed (ITEST) for occupancy detection through an environmental sensing network," Energy Build., vol. 42, no. 7, pp.1038–1046,2010.

[12] M. K. Masood, Y. C. Soh, and V. W.-C. Chang, "Real-time occupancy estimation using environmental parameters," Int. Joint Conf. Neural Netw. (IJCNN), 2015, pp. 1–8.

[13] Jiang, Chaoyang, et al. "Indoor occupancy estimation from carbon dioxide concentration." Energy and Buildings 131 (2016): 132-141.

[14] Basu, Koehlr, et all. "Per CCS: Person Count From Carbon dioxide Using Sparse Non-Negative Matrix Factorization (UBICOMP), 2005, September 17-11, Osaka, Japan

[15] ASHRAE Standard, Ventilation For Acceptable Indoor Air Quality. Approved by the ASHRAE Standards Committee on June 28, 2003; by the ASHRAE Board of Directors on July 3, 2003; and by the American National Standards Institute on January 8, 2004.

[16] http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

[17] http://statisticalhorizons.com/multicollinearity

[18] Mendell, Mark J., and Garvin A. Heath. "Do indoor pollutants and thermal conditions in schools influence student performance? A critical review of the literature." Indoor air 15.1 (2005): 27-52.

[19] http://onlinestatbook.com/2/estimation/confidence.html

[20] https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/