



BUILDING UTILISATION ANALYTICS:
HUMAN OCCUPANCY COUNTING AND
THERMAL COMFORT PREDICTION
WITH AMBIENT SENSING

A thesis submitted in fulfilment of the requirements for
the degree of Doctor of Philosophy

IRVAN BASTIAN ARIEF ANG

Bachelor of Science in Resources Engineering 《 BSc(Eng) 》

[National Cheng Kung University - 國立成功大學]

Master of Business Information Systems (Professional) 《 MBIS Prof 》

[Monash University]

School of Science

College of Science, Engineering and Health

RMIT University

Melbourne, Australia

July, 2018

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed. I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Irvan Bastian Arief Ang (翁道益)

School of Science

RMIT University

2nd July 2018

Acknowledgement

All praise, honour and glory to my Lord Jesus Christ for His richest grace and mercy for the accomplishment of this thesis. I am humbled by and grateful for the company of great colleagues, friends and mentors that I enjoyed for many years.

It is with great pleasure that I express my sincere gratitude to my academic supervisors, Dr Flora Dilys Salim and Associate Professor Margaret Hamilton, for their undying support during the last three and a half years. Their patience, motivation, direction and immense knowledge were tremendous and made my graduate experience manageable, invaluable and fulfilling. I am indebted to them for their encouragement and inspirational advice throughout our association. I appreciate their enthusiasm for my research and their assistance in writing this thesis. Specifically, I would like to thank Flora for being incredibly patient while teaching me valuable research techniques. She pushed me beyond my comfort zone and, for the first time, I learnt the real meaning of the words resilience and perseverance. Her never-ending guidance in designing my research was crucial, and I could not imagine having a better advisor and mentor for my PhD study. I profoundly thank Margaret for her patience, encouragement and guidance. I have learnt a lot from her friendly and calm personality in the last four years of my research. Not only did she advise me on the scientific aspects of my research, she also diligently critiqued my academic writing. Margaret was also the one who kept a sense of humour when I had lost mine. Additionally, she enhanced my confidence to accomplish this research.

I take this opportunity to express my appreciation to the special people who provided me with a fulfilling research candidature at RMIT University. Those people start with the members of the Context Recognition and Urban Intelligence (CRUISE) research group. Dr Mohammad Saiedur Rahaman, my longlasting and loyal officemate, for always accompanying me when we were moving between three different research spaces during the last 3.5 years. By supporting one another in the gravest moment, we pushed our limits and became better

researchers. Jonathan Liono, for staying together night by night, a brother who has a similar mentality who understands both academic and industry mindsets. Your brotherly love and care will always be treasured. Samuel Shao, for always reminding us that PhD life is not just about research and always being able to start conversations on a variety of topics. Other CRUISE group members were Amin Sadri, Hui Song and Rumi Shakila Khan, and with them I shared sweat, tears, joy, love and many wonderful moments during the course of my doctorate. It has been a great pleasure to meet each one of them, and I hope our lives will cross paths again in the future. The time spent with every CRUISE group member will always be remembered and cherished. The group meetings advanced my understanding in the many areas that you guys work in.

I would also like to extend my appreciation to my church community in general and to the CIA cell group members specifically: Stanley S., Priska, Allegra, Nicky, Candy, Janice, Stanley C., Jeannifer, Delicia, Meiga, Yovita, Rebecca, Nathania, Cynthia, Daniel, Ravarrell, Amanda, Alfred, Jesslyn and Kenan. You were my second family to me while I was far from home. I would also like to mention my close friends Roby Lie, Ferryan Khodiat and Juwita Susanto for being there for me whenever I needed moral support. I am also grateful to my friends from PPIA who indirectly supported me throughout my PhD journey. I thank RMIT University, Siemens and Honeywell for providing me with financial support in the form of Sustainable Urban Precincts Program (SUPP) double scholarships, and the Government of Australia for an Australian Government Research Training Program (RTP) scholarship. Thank you to Jörg Wicker for providing the cinema dataset and Mikkel Baun Kjærgaard for providing the study zones and the classrooms dataset.

On another note, I am deeply grateful to Nobuo Uematsu (植松 伸夫), Yasunori Mitsuda (光田 康典) and Michiko Naruke (成毛 美智子) for creating the mesmerising music from the Final Fantasy, Xeno-Chrono and Wild Arms series. That music accompanied me during my research time. Thank you to Keiji Inafune (稲船 敬二) for creating fantastic games in Mega Man series (ロックマン) where I learnt the true meaning of determination and endurance since I was small. I thank Arsène Wenger for managing my favourite team, Arsenal Football Club, to win the British Premier League without a single defeat in a season with *The Invincibles*. Finally, I would like to express my deep appreciation to my father, my mother, my brothers and my sister for supporting me throughout my PhD study and my life in general.

Dedication

I would like to dedicate the hard work of this thesis to my loving and supportive parents, Arief Hartono (翁志鴻) and The Swie Tjo (鄭瑞珠). Without their support during this lengthy journey, I would have never been able to go through the uphill battles of hard work and produce this tome. I would love to say that I have finally fulfilled our family dream of finishing my PhD and making my parents proud of their son in becoming the first in the Arief family to hold the title “Dr”.

I would like to express my sincere gratitude to my father, who always told me “Love what you do, to do what you love”. Indeed, I love technology, and I loved what I did during my PhD. His encouragement has been a constant source of inspiration in my journey and is what made me able to continue with all the hard work and finish my thesis. He gives me the strength and courage to persevere and has made me a better person. My mother’s never-ending love and care makes me strong. She is the one who always reminds me to take a break and take care of my health. My mother always believed in me and never doubt me for a single moment. This belief was my motivation to continue. I learn what real kindness is from my mother. My father improves me, and my mother makes me complete.

To my eldest brother, Arief Handoko (翁道漢), you always give me as much advice as possible. I really appreciate all of that advice and implement it in my life. Thank you ko Han. To my second elder brother, Arief Gunawan Halim (翁道坤), as middle siblings we always have a rebellious attitude and are more straightforward to our family. I like the straightforwardness and learn a lot from you. Thank you ko Gun. To my cutest little sister, Lusiana Arief (翁惠玲), we all love you; but, if we feel bored, we like to tease you. That’s the fate of the little sister. But you are very dependable and mentally strong. Thank you nonik. To all of my siblings, thank you for the endless prayers and words of support.

This thesis is also dedicated to the memory of my beloved grandfather, Arief Hersanto (翁德遠/翁青山). I will never forget all the life advice that you have given since I was a kid. Your

untimely departure from us has left me in sadness. May you find peace while you rest. In memory of Poniem (Mbek-Mbek), you taught me many lessons about life since I was small. Those invaluable skills are still relevant today, and I will always treasure each one of them.

To all my family and friends, without your continuous support and motivation, I would not have been able to get through these hard times. You were always extremely strong supporters of me and my passion for research. Although I have completed my tertiary studies away from home, my friends, as the second family here, always made feel at home.

I dedicate this thesis to all of you.

Music is a mysterious thing. Sometimes it makes people remember things they do not expect. Many thoughts, feelings, memories... things almost forgotten... Regardless of whether the listener desires to remember or not.

Dr Citan Uzuki

Credits

Portions of the material in this thesis have previously appeared in the following publications:

- Irvan B. Arief Ang, Flora D. Salim, and Margaret Hamilton. “Human Occupancy Recognition with Multivariate Ambient Sensors”. *In the Proceedings of the fourteenth IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom 2016)*, pages 1-6. IEEE, Sydney, Australia. 14-18 March 2016. DOI:[10.1109/PERCOMW.2016.7457116](https://doi.org/10.1109/PERCOMW.2016.7457116) [[Ang et al., 2016](#)] (**Workshop at a CORE Rank A* conference**)
- Irvan B. Arief-Ang, Flora D. Salim, and Margaret Hamilton. “SD-HOC: Seasonal Decomposition Algorithm for Mining Lagged Time Series”. *Communications in Computer and Information Science*, volume 845, pages 1-19. ISBN 978-981-13-0292-3. Springer International Publishing. 2018. DOI:[10.1007/978-981-13-0292-3_8](https://doi.org/10.1007/978-981-13-0292-3_8) [[Arief-Ang et al., 2018b](#)]
- Irvan B. Arief-Ang, Flora D. Salim, and Margaret Hamilton. “DA-HOC: Semi-Supervised Domain Adaptation for Room Occupancy Prediction using CO₂ Sensor Data”. *In the Proceedings of the fourth ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys 2017)*, pages 1-10. Delft, The Netherlands. 8-9 November 2017. DOI:[10.1145/3137133.3137146](https://doi.org/10.1145/3137133.3137146) [[Arief-Ang et al., 2017](#)] (**CORE Rank A**)
- Irvan B. Arief-Ang, Margaret Hamilton, and Flora D. Salim. “RUP: Large Room Utilisation Prediction with Carbon Dioxide Sensor”. *Pervasive and Mobile Computing*. Volume 46, June 2018, pages 49-72. DOI:[10.1016/j.pmcj.2018.03.001](https://doi.org/10.1016/j.pmcj.2018.03.001) [[Arief-Ang et al., 2018a](#)] (**Impact Factor:2.349, SJR: Q1**)
- Irvan B. Arief-Ang, Margaret Hamilton, and Flora D. Salim. “A Scalable Room Occupancy Prediction with Transferable Time Series Decomposition of CO₂ Sensor Data”.

ACM Transactions on Sensor Networks (TOSN). **Accepted**. May 2018.

DOI:[10.1145/3217214](https://doi.org/10.1145/3217214) (**Impact Factor:2.322, SJR: Q1**)

- Irvan B. Arief-Ang, Margaret Hamilton, and Flora D. Salim. “THERMO: Thermal Comfort Prediction and Adjustment in Shared Office Environment”. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*. **Under Review (major revision)**. May 2018. IMWUT is the journal of the *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2018)* (**CORE Rank A***)

This research is supported by the Australian Government Research Training Program Scholarship and two RMIT and Siemens Sustainable Urban Precinct Project (SUPP) grants: “iCo2mmunity: Personal and Community Monitoring for University-wide Engagement towards Greener, Healthier, and more Productive Living” and “The Greener Office and Classroom”.

Contents

Declaration	ii
Acknowledgement	iii
Dedication	v
Statistical Summary	vii
Credits	viii
Contents	x
List of Figures	xvii
List of Tables	xxi
Abstract	2
1 Introduction	3
1.1 Background and Motivations	5
1.2 Research Challenges	5
1.3 Research Questions	7
1.4 Research Contributions	8
1.5 Thesis Organisation	10
2 Sensors: Preliminary Analysis	13
2.1 Introduction	13
2.2 Related Works	14

2.3	Methodology	15
2.3.1	Feature Engineering	15
2.3.2	Machine Learning Algorithm	17
2.4	Data Collection and Preprocessing	18
2.4.1	Type of Sensor Devices	18
2.4.2	Occupancy Web-based Application	19
2.4.3	Data Preprocessing	20
2.5	Experiments and Results	21
2.5.1	Experiment 1 (Normalised Raw Data)	22
2.5.2	Experiment 2 (Ten-Minute Time Segments)	22
2.5.3	Experiment 3 (Part of the Day Segments)	23
2.5.4	Most Dominant Sensor in Determining Human Presence	23
2.5.4.1	CO ₂ Rate vs Indoor Human Occupancy	23
2.5.4.2	Illumination Level vs Indoor Human Occupancy	24
2.5.4.3	Sound Rate vs Indoor Human Occupancy	24
2.6	Conclusion	25
3	SD-HOC: Seasonal Decomposition Algorithm for Mining Lagged Time Series	26
3.1	Introduction	26
3.2	Background and Related Work	28
3.3	Problem Definition	28
3.3.1	Scenario Assumption	29
3.3.2	Time Series	29
3.4	The Features	30
3.4.1	Time Delay Components	31
3.4.2	Cross-Correlation and the Line of Best Fit	31
3.4.3	Time Lag	33
3.5	The Framework	34
3.5.1	Seasonal-Trend Decomposition (STD)	34
3.5.2	Correlation Models	37
3.5.2.1	Correlation Model for Trend Feature (T_t)	37
3.5.2.2	Correlation Model for Seasonal Feature (S_t)	38
3.5.2.3	Correlation Model for Irregular Feature (e_t)	39

3.5.3	Zero Pattern Adjustment	39
3.6	Experiments and Results	40
3.6.1	Experiment Setting	40
3.6.1.1	Small Room A	40
3.6.1.2	Big Room B	40
3.6.1.3	Experimental Tool	41
3.6.2	Experimental Parameters	41
3.6.2.1	Experiment for Small Room A Dataset	41
3.6.2.2	Experiment for Big Room B Dataset	42
3.6.3	Experimental Results with Other Data Mining Algorithms	42
3.6.4	Experimental Results with Support Vector Regression (SVR) on Different Number of Prediction Days	43
3.6.4.1	Evaluation and Baseline	43
3.6.4.2	Experimental Results for Small Room A Dataset	44
3.6.4.3	Experimental Results for Big Room B Dataset	44
3.7	Discussion	45
3.8	Conclusion	45
4	RUP: Large Room Utilisation Prediction with Carbon Dioxide Sensor	47
4.1	Introduction	47
4.1.1	Research Motivation	49
4.1.2	Research Contribution	50
4.2	Background and Related Work	51
4.2.1	Background Study of Human Occupancy Calculation	51
4.2.2	Simulation-Based Indoor Human Occupancy Detection	52
4.2.3	Radio-Based Indoor Human Occupancy Detection	53
4.2.4	Indoor Human Occupancy Detection with Sensors	53
4.2.5	CO ₂ -Based Indoor Human Occupancy Detection	58
4.2.6	Summary of Related Work	59
4.3	Problem Definition	60
4.3.1	Scenario Assumption	61
4.3.2	Problem Definition	61
4.4	Method	62
4.4.1	Data Collection Experimental Setup	62

4.4.1.1	Academic Staff Room	62
4.4.1.2	Cinema Theatre	63
4.4.1.3	Experimental Tool	63
4.4.2	Data Preprocessing	64
4.4.2.1	Autocorrelation and the Line of Best Fit	65
4.4.2.2	Time Lag	66
4.4.3	Room Utilisation Prediction Algorithm	66
4.4.3.1	Decomposition Methodologies	67
4.4.3.2	Correlation Models	70
4.4.3.3	Occupancy Model	73
4.5	Experiments, Results and Discussion	73
4.5.1	Experiments	73
4.5.1.1	Experimental Parameters for Academic Staff Room Dataset	74
4.5.1.2	Experimental Parameters for Cinema Dataset	74
4.5.1.3	Evaluation and Baselines	74
4.5.2	Experimental Result	75
4.5.2.1	Experimental Result for Academic Staff Room Dataset	75
4.5.2.2	Experimental Result for Cinema Dataset	76
4.5.3	Discussion	77
4.6	Conclusion	78
5	DA-HOC++: A Scalable Room Occupancy Prediction with Transfer- able Time Series Decomposition of CO₂ Sensor Data	79
5.1	Introduction	79
5.1.1	Research Motivation	81
5.1.2	Research Contribution	82
5.2	Background and Related Works	82
5.2.1	Related Work	82
5.2.2	Semi-Supervised Domain Adaptation	84
5.2.3	Carbon Dioxide - Human Occupancy Counter Model	84
5.3	Problem Definition	86
5.3.1	Scenario	87
5.3.2	Domain Adaptation	88
5.4	Methodology	89

5.4.1	Preprocessing DA-HOC++	89
5.4.2	Main Algorithm DA-HOC++	91
5.4.2.1	Domain Adaptation Method for the Trend Feature (TF_t^{DA})	92
5.4.2.2	Domain Adaptation Method for the Seasonal Feature (SF_t^{DA})	93
5.4.2.3	Domain Adaptation Method for the Irregular Feature (IF_t^{DA})	94
5.4.2.4	Domain Adaptation Method for the Zero Pattern Adjustment Feature (ZF_t^{DA})	95
5.4.3	Post Adjustment DA-HOC++	95
5.5	Location Detail, Settings and Parameters	98
5.5.1	Location Settings and Parameters	99
5.5.1.1	Source Domain Location	99
5.5.1.2	Target Domain Location	100
5.6	Experiments and Results	101
5.6.1	Experiment Tools	102
5.6.2	Baselines	102
5.6.3	Experimental Results	105
5.6.3.1	Domain Adaptation for Cinema Dataset	105
5.6.3.2	Domain Adaptation for Study Zones Dataset	108
5.6.3.3	Domain Adaptation for Classrooms Dataset	108
5.6.4	Evaluation Metrics	109
5.7	Conclusion	111
6	THERMO: Thermal Comfort Prediction and Adjustment in Shared Office Environments	112
6.1	Introduction	112
6.1.1	Research Motivation	115
6.1.2	Research Contribution	116
6.2	Background and Related Works	116
6.2.1	Background Study of Indoor Thermal Comfort	117
6.2.2	Related Works	119
6.3	Problem Definition	122
6.3.1	Scenario Assumption	122
6.3.2	Problem Definition	122
6.4	Methodology	124

6.4.1	The Main Group of Features in THERMO	125
6.4.1.1	Ambient Sensor	125
6.4.1.2	Daily Survey Data	125
6.4.1.3	Background Survey	126
6.4.2	Data Pre-processing	127
6.4.2.1	Data Cleansing	127
6.4.2.2	THERMO Weight Level	128
6.4.3	THERMO Prediction Algorithm	129
6.4.3.1	Pre-model	130
6.4.3.2	THERMO Main Model	130
6.4.4	THERMO Adjustment Algorithm	131
6.5	Evaluation	133
6.5.1	Datasets	133
6.5.2	Experiment Tool	134
6.5.3	Evaluation and Baseline	135
6.6	Experimental Results	136
6.6.1	Thermal Sensation Results and Analysis	137
6.6.2	Analysis related to Season	137
6.6.3	Adjustment Results	139
6.7	Conclusion	139
7	Conclusion	141
7.1	Research Questions and Answers	142
7.2	Future Directions for Research	145
	Bibliography	147
A	Sensor Devices	164
A.1	List of devices (sensors) and Their Capacities.	164
B	App System Architecture used in Chapter 2	166
C	Machine Learning Techniques and Their Abbreviations	167
D	Completed Ethics Proposal, Ethics Approval and Survey Documents	169
D.1	Ethics Approval Document	184

D.2	Participant Information Document	186
D.3	Consent Form Document	190
D.4	Survey Form Document	192
E	Full Experiment Accuracy Result for both Academic Staff Room and Cinema Theatre	195

List of Figures

1.1	List of Multiple Ambient Sensors.	4
1.2	The Location of Each Dataset Used in This Thesis.	6
1.3	Main Framework Diagram of Building Utilisation Analytics Related to Ambient Sensors.	7
1.4	Thesis Structure and Organisation.	9
2.1	(1) SmartThings SmartSense Open/Closed Sensor; (2) Netatmo Urban Weather Station; (3) Z-Wave Aeon MultiSensor;(4) SmartThings SmartPower Outlet.	18
2.2	Occupancy Web-based Application.	19
2.3	Root-mean-square Error Results of Various Machine Learning Algorithms.	22
2.4	Occupancy Recognition Probability over CO ₂ Level.	23
2.5	Occupancy Recognition Probability over Illumination Level.	24
2.6	Occupancy Recognition Probability over Sound Sensor Rate.	25
3.1	Real-time Prediction Scenario Showing Human Occupancy and CO ₂ Fluctuation. The Fundamental Task is to Predict the Number of Occupants at Time $t+\Delta t$	29
3.2	Data Collection and Analysis Framework.	30
3.3	Ordinary Least Square Regression Normalised Root Mean Square Error Between CO ₂ Data and Actual Occupancy for 60 Minutes Time Lag.	33
3.4	Seasonal Decomposition for Human Occupancy Counting (SD-HOC) Analysis Framework.	35
3.5	Accuracy Results of Various Machine Learning Algorithms.	42
3.6	Small Room A Dataset - Comparison for Indoor Human Occupancy.	43
3.7	Big Room B Dataset - Comparison for Indoor Human Occupancy.	44

4.1	CO ₂ concentration in the movie theatre over time. (a) shows an overview of all measurements; days, times and movies screening can be easily identified. (b) shows the concentration on December 28; movie screenings can be seen and the shape of each screening can be identified differently. (c) shows the CO ₂ concentration during the movie “The Hunger Games 2: Catching Fire” on December 28, 2013 at 13:15.	49
4.2	Simplified presentation of gas exchange in the respiratory chamber where $\varphi_i(t)$ is the flow rate of input air at time t , $\varphi_o(t)$ is the flow rate of output air at time t , V is the volume at time t , and $u^g(t)$ is the gas production rate at time t	52
4.3	Real-time prediction scenario showing human occupancy and CO ₂ fluctuations. The fundamental task is to predict the number of occupants at time $t+\Delta t$	60
4.4	Data Collection and Analysis Framework.	61
4.5	A Netatmo urban weather station (left), a sensor device to gather ambient CO ₂ data that was set up near the window in the academic staff room (right).	63
4.6	Measurement in the cinema theatre. Air is drawn out from the screen room via the ventilation system and is transported to the mass spectrometer. [Wicker et al., 2015]	63
4.7	Correlation between the number of occupants and CO ₂ readings (ppm) with time lag 0.	64
4.8	Ordinary Least Square Regression Normalised Root Mean Square Error (NRMSE) between CO ₂ data and actual occupancy, for 60 minutes time lag.	65
4.9	Large Room Utilisation Prediction (RUP) with Carbon Dioxide Sensor.	67
4.10	Examples of Time Series Decomposition.	68
4.11	Confusion Matrix between Real Occupancy and Prediction Occupancy for Cinema Theatre.	73
4.12	Academic staff room dataset: comparison for indoor human occupancy with zero unit of error tolerance.	75
4.13	Academic staff room dataset: comparison for indoor human occupancy with one unit of error tolerance.	75
4.14	Cinema dataset: comparison for indoor human occupancy with zero units of error tolerance.	76
4.15	Cinema dataset: comparison for indoor human occupancy with ten units of error tolerance.	76

5.1	Illustration for the Main Algorithm DA-HOC++.	80
5.2	Domain adaptation prediction scenario in target domain showing human occupancy and CO ₂ fluctuations. The fundamental task is to predict the number of occupants at time $t+\Delta t$.	86
5.3	Data Collection and Analysis Framework.	87
5.4	Overview of the Semi-supervised Domain Adaptation Learning Method.	88
5.5	Correlation between human occupancy number and carbon dioxide concentration.	89
5.6	Semi-supervised domain adaptation method for Seasonal Decomposition for Human Occupancy Counter Double Plus (DA-HOC++).	91
5.7	The correlation plot between trend, seasonal, irregular and CO ₂ value between vacant (red) and non-vacant (blue) room.	96
5.8	Source Domain and Target Domain.	96
5.9	The DA-HOC++ experiment covers multiple locations in three different countries, in six different rooms with a variety of sizes, surrounding environments and characteristics.	97
5.10	A Netatmo urban weather station (left), a sensor device to gather ambient CO ₂ , data which was set up near the window in the academic staff room (right).	99
5.11	Measurement in the cinema theatre. Air is drawn out from the screen room via the ventilation system and is transported to the mass spectrometer [Wicker et al., 2015].	100
5.12	One of the classrooms located in OU44 building (shown top-right) at the University of Southern Denmark [Sangogboye et al., 2017].	101
5.13	Binary Occupancy Prediction Result for the Cinema Dataset.	103
5.14	Occupancy Counting Accuracy Result for the Cinema Dataset.	103
5.15	Binary Occupancy Prediction Results for the Study Zone 1 Dataset.	104
5.16	Occupancy Counting Accuracy Results for the Study Zone 1 Dataset.	104
5.17	Binary Occupancy Prediction Results for the Study Zone 2 Dataset.	105
5.18	Occupancy Counting Accuracy Results for the Study Zone 2 Dataset.	105
5.19	Binary Occupancy Prediction Results for the Classroom 1 Dataset.	106
5.20	Occupancy Counting Accuracy Results for the Classroom 1 Dataset.	106
5.21	Binary Occupancy Prediction Results for the Classroom 2 Dataset.	107
5.22	Occupancy Counting Accuracy Results for the Classroom 2 Dataset.	107

6.1	ASHRAE 7 Point Scale plots of Friend Center for three participants survey results regarding their general thermal comforts and thermal sensations	113
6.2	Some of the Factors influencing Thermal Comfort.	117
6.3	Real-time prediction scenario for continuous t showing multiple ambient sensor fluctuations. The fundamental task is to predict both general thermal comfort and thermal sensation at time $t+\Delta t$	121
6.4	The data flow in THERMO and the detailed source of groups of features covering ambient sensors, daily survey data and background survey with the distribution of different weighted levels.	123
6.5	The measurement of General Thermal Comfort (From very uncomfortable [1] to very comfortable [6]).	124
6.6	The measurement of Thermal Sensation (From cold [-3] to hot [+3]).	124
6.7	The Complete Framework Structure of THERMO.	126
6.8	Complete THERMO Prediction Algorithm Model.	130
6.9	The Friend Center in Philadelphia, USA.	133
6.10	The Friend Center Floor Map [Langevin et al., 2015].	133
6.11	ROC Curve Plot of General Thermal Comfort for Various Machine Learning Algorithms.	136
6.12	ROC Curve Plot of Thermal Sensation for Various Machine Learning Algorithms.	136
6.13	The distribution of general thermal comfort by each season (From the left to the right: Spring, Summer, Autumn and Winter).	137
6.14	General Thermal Comfort Distribution.	137
6.15	The distribution of thermal sensation by each season (From the left to the right: Spring, Summer, Autumn and Winter).	138
6.16	Thermal Sensation Distribution.	138
A.1	Z-Wave Aeon Multi Sensors	164
A.2	SmartThings SmartSense Open/Closed Sensor	164
A.3	SmartThings SmartPower Outlet	165
A.4	Netatmo Urban Weather Station	165
B.1	App System architecture	166

List of Tables

2.1	List of coefficients for each parameter for non-linear regression formula	21
2.2	List of coefficients for each parameter for regression line polynomial order of 4. . .	23
2.3	R ² value and Pearson correlation r-value for the three most dominant sensors for recognising indoor human occupancy.	24
3.1	Accuracy results of various machine learning algorithms.	43
4.1	Models, parameters and reported accuracies for radio-based occupancy detection research.	53
4.2	Models, parameters and reported accuracies for sensor-based occupancy detection research.	54
4.3	Algorithms, devices and reported accuracies for CO ₂ sensor-based occupancy detection research.	58
4.4	Example of Indoor Human Occupancy History Data.	60
4.5	Academic staff room indoor human accuracy result.	75
4.6	Cinema theatre indoor human occupancy accuracy result.	77
5.1	Algorithms, devices and reported accuracies for CO ₂ -sensor-based occupancy detection research.	85
5.2	Detailed statistical information on the datasets for small and large rooms.	98
5.3	SVR (+TB), SD-HOC (+TB), SVR (-TB), SD-HOC (-TB) and DA-HOC++ Human Binary Occupancy and Occupancy Counting Prediction Accuracy Results for the Cinema Dataset.	103
5.4	SVR (+TB), SD-HOC (+TB), SVR (-TB), SD-HOC (-TB) and DA-HOC++ Human Binary Occupancy and Occupancy Counting Prediction Accuracy Results for the Study Zone 1 Dataset.	104

5.5	SVR (+TB), SD-HOC (+TB), SVR (-TB), SD-HOC (-TB) and DA-HOC++ Human Binary Occupancy and Occupancy Counting Prediction Accuracy Results for the Study Zone 2 Dataset.	105
5.6	SVR (+TB), SD-HOC (+TB), SVR (-TB), SD-HOC (-TB) and DA-HOC++ Human Binary Occupancy and Occupancy Counting Prediction Accuracy Results for the Classroom 1 Dataset.	106
5.7	SVR (+TB), SD-HOC (+TB), SVR (-TB), SD-HOC (-TB) and DA-HOC++ Human Binary Occupancy and Occupancy Counting Prediction Accuracy Results for the Classroom 2 Dataset.	107
5.8	The Accuracy, Precision, Recall and F-Score for Cinema Dataset.	108
5.9	The Accuracy, Precision, Recall and the F-Score for Study Zones 1 and 2.	109
5.10	The Accuracy, Precision, Recall and F-Score for Classrooms 1 and 2.	110
6.1	Five different level of weighting in THERMO.	127
6.2	General Thermal Comfort and Thermal Sensation Prediction Accuracy Results from THERMO and other Machine Learning Algorithms.	135
6.3	A General Thermal Comfort Comparison Snapshot between the Original and after Adjustment.	139
C.1	Machine Learning Techniques and their abbreviations that are used in related works.	168
E.1	Academic staff room indoor human occupancy accuracy result for 1 day and 2 days prediction with 7-13 days training data	195
E.2	Academic staff room indoor human occupancy accuracy result for 3 days and 4 days prediction with 7-11 days training data	196
E.3	Academic staff room indoor human occupancy accuracy result for 5 days and 6 days prediction with 7-9 days training data	196
E.4	Academic staff room indoor human occupancy accuracy result for 7 days prediction with 7 days training data	196
E.5	Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 1 day and 2 days prediction with 12-22 days training data	197
E.6	Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 3 days and 4 days prediction with 12-20 days training data	197
E.7	Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 5 days and 6 days prediction with 12-18 days training data	198

E.8 Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 7 days and 8 days prediction with 12-16 days training data 198

E.9 Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 9 days and 10 days prediction with 12-14 days training data 198

E.10 Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 11 days prediction with 12 days training data 198

Abstract

With advancement in sensors and the Internet of Things, gathering spatiotemporal information from one's surroundings has become more convenient. There are multiple phenomenological behaviours, such as indoor comfort and occupancy trends, that can be inferred from this information. There are multiple advantages to having an accurate indoor occupancy prediction, including better understanding of space-room utilisation, which can be used to further inform energy consumption reduction, human indoor comfort optimisation and security enhancement.

We use non-intrusive ambient sensors to infer indoor occupancy patterns. Non-intrusive ambient sensors are utilised because they are commonly available in building management systems (BMSs). Machine learning techniques are applied and data-driven approaches are implemented to identify indoor human occupancy and predict comfort. These facilitate the decision-making tasks for building management professionals and are used in real-time monitoring. Our preliminary study with multiple ambient sensors reveals that carbon dioxide is one of the best predictors of indoor human occupancy.

We design a seasonal trend decomposition algorithm by implementing pervasive sensing and leveraging carbon dioxide data from BMS sensors. The first model is seasonal decomposition for human occupancy counting (SD-HOC), a customised feature transformation decomposition prediction model. This provides a novel way to estimate the number of people within a closed space, using one carbon dioxide sensor. SD-HOC integrates a time lag and line of best fit model in the preprocessing algorithms and customises different regression algorithms for each subcomponent, to predict each respective human occupancy component value. Utilising several machine learning techniques, a set of prediction values for each component is obtained. Finally, additive decomposition is used to reconstruct the prediction value for human indoor occupancy.

We improve the algorithm to cover multiple buildings with different contexts and locations and develop a large Room Utilisation Prediction with carbon dioxide sensor (RUP). RUP improves SD-HOC and is able to predict a larger number of occupants, up to three hundred, using

data from a single carbon dioxide sensor. RUP de-noises and pre-processes the carbon dioxide data. We use multiple variants of seasonal decomposition techniques and feature factorisation for both occupant and carbon dioxide datasets, and develop a zero pattern adjustment model to increase the accuracy. We run our model in two different locations that have different contexts. The prediction accuracy results outweigh the state-of-the-art techniques for time series decomposition and regression.

RUP is a reliable model for any building with adequate historical data. In the real world, this condition is not always feasible, due to several limitations such as a new building only having limited historical data, or government/military buildings that have strictly controlled access to historical ambient sensor data. One way to solve this problem is by implementing a transfer learning technique with SD-HOC. We design a semi-supervised domain adaptation method for carbon dioxide - human occupancy counter (DA-HOC) to estimate the number of people within one room, by using a carbon dioxide sensor with a limited number of training labels (as little as one day of historical data). The DA-HOC model is trained using data from a source domain that has a more complete set of training labels, and transferred to predict the occupancy of a much larger room of the target domain, with very little training data. We enhance DA-HOC into DA-HOC++ and successfully experiment with the model to transfer the knowledge from one room to five different rooms in different countries.

Moving beyond indoor human occupancy, each occupant's comfort is also a crucial problem that needs to be considered. Indoor comfort prediction is crucial for energy efficiency cost adjustment, human productivity and non-wastage of resources. Maintaining human indoor comfort levels at acceptable values is one of the primary goals in any building and room utilisation. The main problem is that everybody has a different level of acceptance of what is comfortable. We implement a machine learning algorithm to predict the thermal comfort for each occupant. Our model successfully achieves a respectable accuracy of comfort prediction to help the BMS adjust the temperature.

This thesis presents several contributions in machine learning for indoor human occupancy and comfort prediction. This research implements and extends existing data mining techniques to solve problems on time series prediction. The solutions are scalable and can also work with minimal sets of historical training data with a transfer learning method. The research contributions in this thesis present multiple occupancy algorithms for both indoor human occupancy and thermal comfort. We believe that this research provides a big step towards building a robust solution for smart homes and smart buildings, in which the buildings are more aware of their occupants and can adapt to their needs.

Chapter 1

Introduction

Building and housing are an integral part of human life. From education and work to entertainment and leisure, every aspect of life activity needs some shelter. The study of building-related domains has been underway for many years [Dodier et al., 2006, Lu et al., 2010, Feige et al., 2013, Sangogboye et al., 2017]. Due to its importance, there are many studies around building management, which includes smart buildings [Lam et al., 2014, Shin et al., 2017], indoor analytics [Huang et al., 2012, Jiang et al., 2012] and energy efficient buildings [Delaney et al., 2009, Jradi et al., 2017]. As buildings are becoming smarter every day [Agarwal et al., 2010], concern about how smart buildings deliver more benefits and better experiences to their owners and inhabitants is also growing [Zhang et al., 2011, Beltran et al., 2013, Clear et al., 2013, Ascione et al., 2014].

With advances in ambient sensors around us, gathering ambient information and translating it to data has never been easier [Ekwevugbe et al., 2013b, Khan et al., 2014]. Smart buildings utilise multiple sensors with different capabilities to increase the way they communicate with their inhabitants. With numerous different types of sensors that can be installed in a building, the possibilities are limitless. There are multitudinous benefits from sensor installation in buildings, such as intelligent building security [Luo et al., 2003], indoor air quality [Jiang et al., 2011], the Internet of Things [Swan, 2012] and pervasive and ubiquitous computing [Erickson et al., 2009]. Variety types of ambient sensors are shown in Figure 1.1.

A key part of smart building management is an efficient *building utilisation* [Akkaya et al., 2015]. Understanding how a building and space are utilised improve the quality of human life. The study of building utilisation focuses on knowing where the inhabitants are at one specific point in time. Understanding how many people are within a building or in one particular room

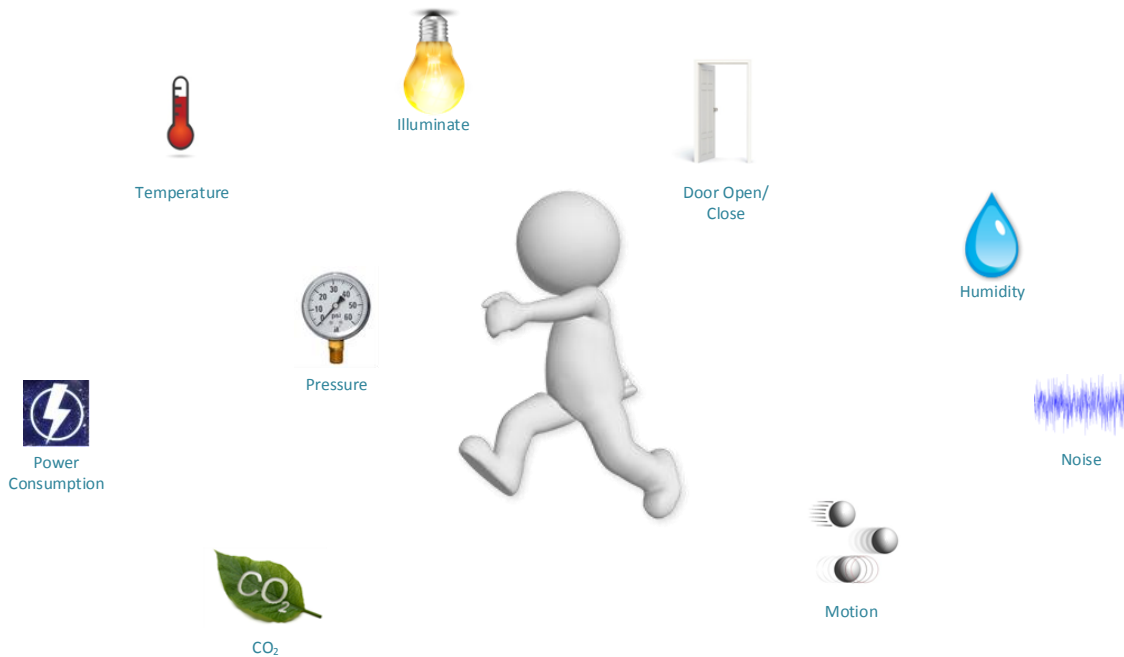


Figure 1.1: List of Multiple Ambient Sensors.

level is an important component for smart heating, ventilation and air conditioning (HVAC) systems. On the other hand, if the number of occupants can be precisely identified, HVAC systems can be adjusted to be more efficient. The room temperature can be adjusted when there is at least one person inside the room, and electricity can be saved when there is nobody inside.

The majority of buildings do not have apposite infrastructure to accurately sense people and where they are within a building. With this limitation, it is difficult to obtain accurate occupancy data inside a building and to get the precise ground truth value for analysis purposes. Existing works mainly focus on the use of simulation models [Goldstein et al., 2010, Page et al., 2008] to reduce energy consumption. Others simulate occupant behaviour and aim for reducing energy consumption based on their behaviour [Richardson et al., 2008, Saelens et al., 2011]. A stable model for transfer learning is needed so that a human occupancy model can be deployed in another context with acceptable accuracy of prediction.

The process of reducing the cost and knowing the number of people should not sacrifice personal indoor comfort, and these domains are strictly related to one another. Indoor comfort includes visual comfort, acoustic comfort, thermal comfort and good air quality. Thermal

comfort analysis is the most complicated of these, as every human has a different sense of what is comfortable. This analysis requires both quantitative sensor data and qualitative survey data. By marrying these data, an indoor thermal comfort and thermal sensation prediction model can be developed.

1.1 Background and Motivations

The research area of human occupancy prediction has become established in recent years due to advances in powerful technologies for data analysis. In the past, only a supercomputer could handle the analysis of millions of data. However, in recent years, this limitation is closing and a standard personal computer can do basic data analysis for a few million data points with an acceptable processing time.

Many practical real-life advantages can be obtained by knowing the number of occupants residing in one building or one room at a particular time. The benefits include energy consumption reduction, human indoor comfort and security. Knowing the number of people in advance can be used to adjust the HVAC to reduce the power if there will be nobody inside a room for a given period. Human comfort quality can be improved by increasing or decreasing the heating or air conditioning based on the crowdedness of a room. For security, if the owner of a building knows that there should be no-one inside the building at a particular time, detecting a person at that point could be linked to a security breach.

Understanding the occupancy pattern is also necessary for space utilisation. If the usage pattern of one room is low, the room can be utilised for other more beneficial purposes. Maximising space and room utilisation creates greater space efficiency and could increase both individual and group productivity. Our research gathered and utilised data from around the world as shown in Figure 1.2 to ensure that our solutions can be implemented everywhere.

1.2 Research Challenges

In this thesis, we aim to utilise ambient sensors to address multiple key challenges related to building analytics. The rapid growth of availability of a variety of sensors, and the possibility of mass production of those sensors in an affordable manner, enables the collection of diverse time series of ambient sensor data.

Advances in artificial intelligence for data analysis in machine learning facilitates a new way of handling massive amounts of data - sometimes referred to as big data, which is data

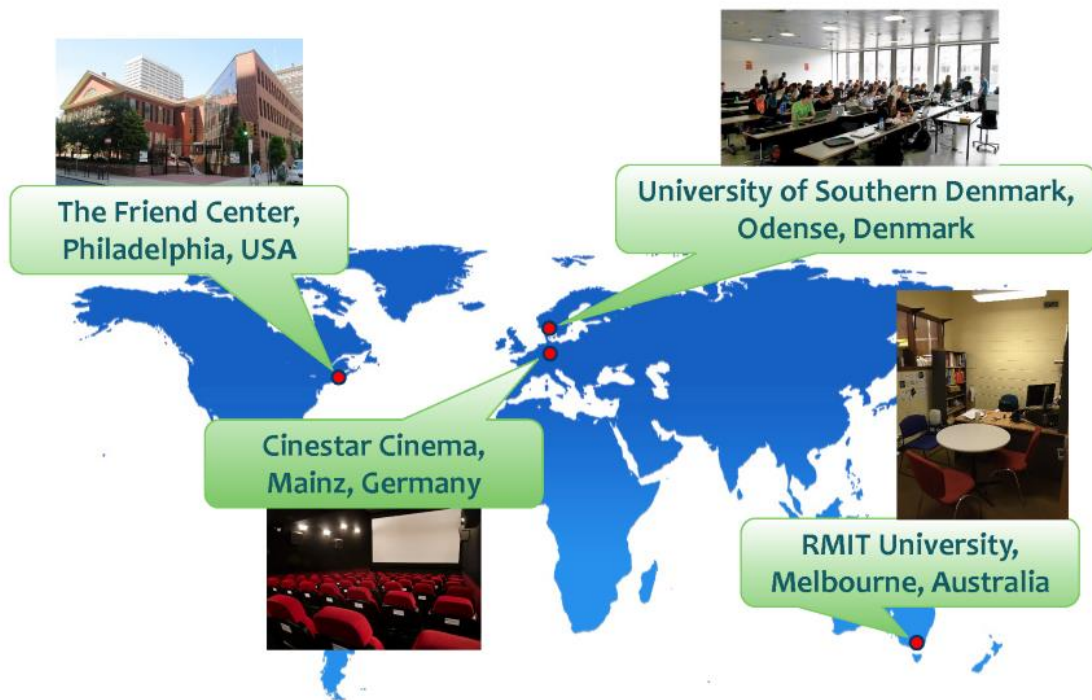


Figure 1.2: The Location of Each Dataset Used in This Thesis.

that contain the 3Vs volume, variety and velocity [Douglas, 2001]) - and the ability to infer the meaning of the data in a useful way.

In summary, the core challenges in building analytics surrounding human occupancy and thermal comfort are:

- Fusion, integration and representation of collected large ambient and survey-related data sources.
- Providing intelligent analysis and prediction on human indoor occupancy and thermal comfort.
- Performing building utilisation analytics prediction without breaching occupants' privacy.
- Inferring sparse sensor and occupancy contexts from limited annotation data (e.g. from post-occupancy surveys).

- Understanding and correlating both quantitative and qualitative data to mine deeper meaning to support building and machine automation.

Indoor human occupancy has been explored for more than a decade [Barandiaran et al., 2008, Erickson et al., 2009, Lee et al., 2011]. One of the most prominent problems is how to study human occupancy without a camera. Image processing techniques from vision-based devices could be very accurate but raise privacy issues. This thesis tries to fill this research gap by introducing a new approach of counting people using a device unrelated to human occupancy and several other new techniques for this domain. The main framework of this thesis in one single image is summarised in Figure 1.3.

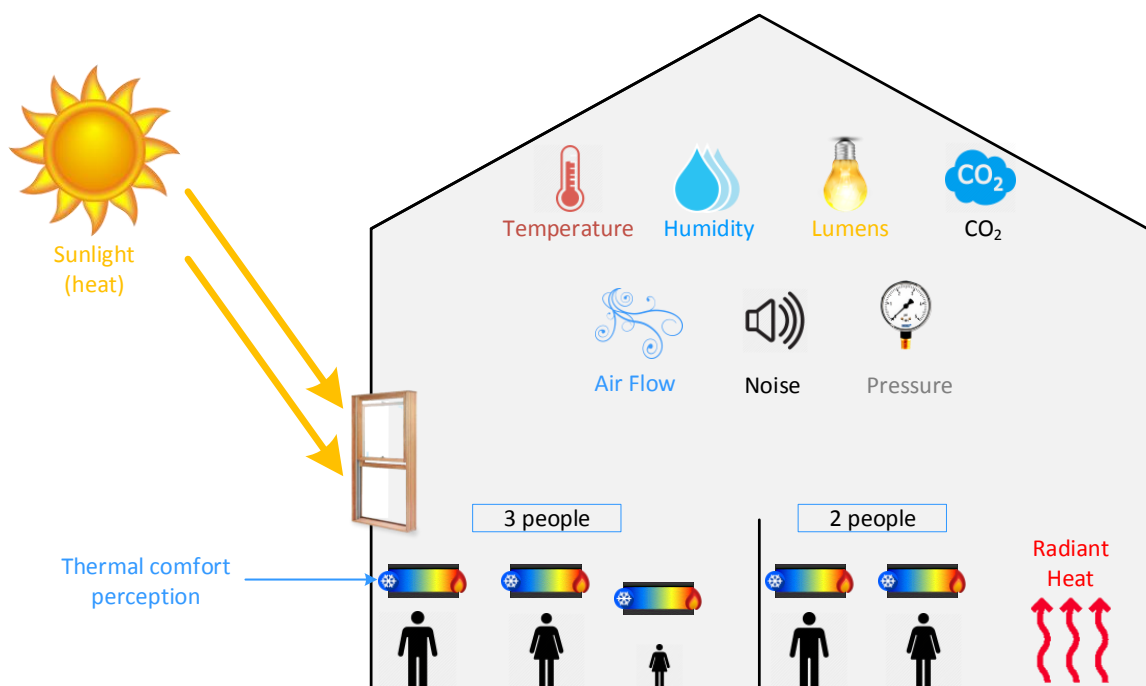


Figure 1.3: Main Framework Diagram of Building Utilisation Analytics Related to Ambient Sensors.

1.3 Research Questions

In order to overcome the aforementioned research challenges, the following research questions (RQs) are defined, with the goals of achieving the most robust building analytics around hu-

man occupancy and thermal comfort.

RQ-1. *How to recognise indoor human occupancy using multivariate ambient sensor data?*

Understanding whether a room is occupied or vacant is the foundation of this research. We utilise various machine learning algorithms to find the best method for binary indoor human occupancy: whether the room is vacant or occupied. Furthermore, we build a model to compare which ambient sensor is the most dominant.

RQ-2. *How to perform room utilisation prediction using carbon dioxide data?*

From the first research question, we found that carbon dioxide is the best ambient sensor predictor for recognising indoor human occupancy. We develop a novel feature engineered method to utilise only carbon dioxide data to predict the number of occupants in a space.

RQ-3. *How to perform transfer learning to use an existing occupancy prediction model to predict the utilisation of another room with limited training data?*

Continuing from the previous research question, we develop a domain adaptation technique using information from one location to other location. A general transfer learning model is examined and customised for the domain of indoor human occupancy.

RQ-4. *How to predict indoor comfort based on ambient sensor and users' survey data?*

Thermal comfort and thermal sensation are two central matrices in understanding the level of satisfaction in one building. Understanding human indoor comfort in general is crucial as part of a complete model of building utilisation analytics.

1.4 Research Contributions

To address the aforementioned research questions, the contributions of this thesis are as follows:

- To investigate and analyse space and room utilisation with the use of only ambient sensor data.

- To develop and model a novel method for indoor human occupancy using only one type of ambient sensor data.
- To design a transfer learning model for human occupancy that can be used across different locations.
- To redefine and redesign the thermal comfort and thermal prediction models with machine learning, and automate the HVAC adjustment model.

This research focuses on counting human occupancy and indoor comfort, and builds a transferable generic model to other locations to avoid over-fitting a design model. The analysis is conducted from a data-driven perspective.

The practicality of this problem can be applied to multiple domains. It is known that suitable indoor comfort improves the overall productivity of inhabitants [McCartney and Humphreys, 2002, Akimoto et al., 2010]. Furthermore, the health and wellbeing of each inhabitant can be maintained in a comfortable environment [Matzarakis and Amelung, 2008, Ortiz et al., 2017].

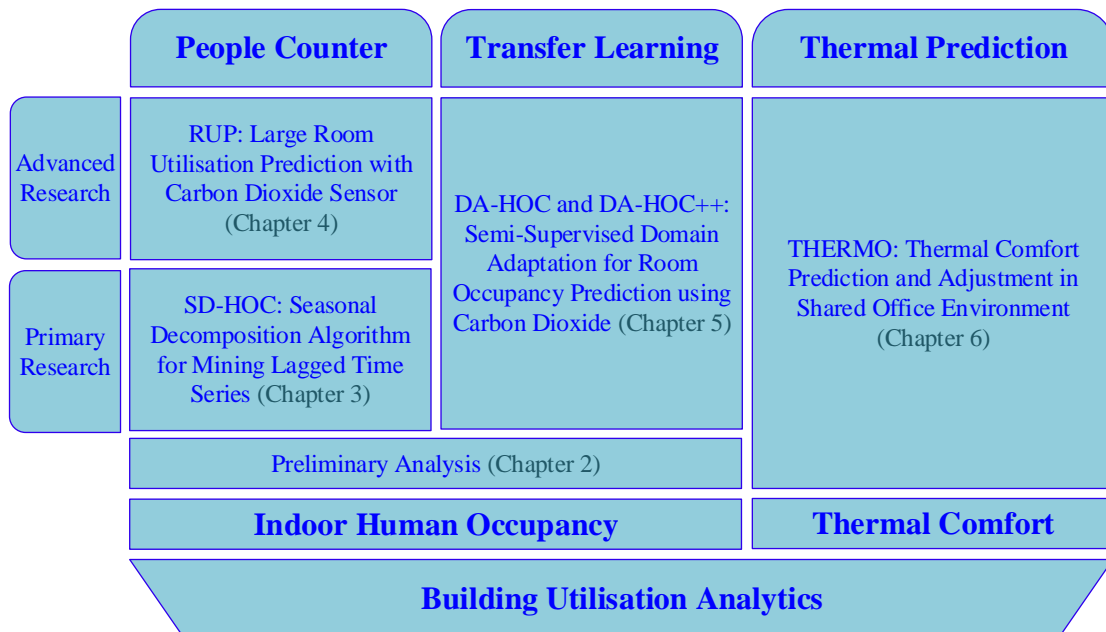


Figure 1.4: Thesis Structure and Organisation.

1.5 Thesis Organisation

Each chapter starts with a chapter introduction, motivation and contribution. Background and related works are discussed to give the reader context for that chapter. The problem is defined and followed by methodology, experiments and results. Each chapter closes with a conclusion. An overarching view of thesis structure and organisation is shown in Figure 1.4. The thesis is organised as follows:

- **Chapter 2 - Sensors: Preliminary Analysis**

A preliminary experiment with a variety of ambient sensors is conducted and recorded. The solution presented in this chapter relates to *RQ-1*. The detailed correlation between sensor data and human presence is modelled, and the three most dominant sensors in determining human presence are identified.

Copyright/credit/reuse notice: The contents of this chapter have been taken and revised as needed from a paper published as: [[Ang et al., 2016](#)]

Irvan B. Arief Ang, Flora D. Salim, and Margaret Hamilton. “Human Occupancy Recognition with Multivariate Ambient Sensors”. *In the Proceedings of the fourteenth IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom 2016)*, pages 1-6. IEEE, Sydney, Australia. 14-18 March 2016.

DOI: [10.1109/PERCOMW.2016.7457116](https://doi.org/10.1109/PERCOMW.2016.7457116)

- **Chapter 3 - SD-HOC: Seasonal Decomposition Algorithm for Mining Lagged Time Series**

SD-HOC is a customised feature transformation decomposition, a novel way to estimate the number of people within a closed space using only a single carbon dioxide sensor. The solution presented in this chapter answered half of *RQ-2*.

Copyright/credit/reuse notice: The contents of this chapter have been taken and revised as needed from a paper published as: [[Arief-Ang et al., 2018b](#)]

Irvan B. Arief-Ang, Flora D. Salim, and Margaret Hamilton. “SD-HOC: Seasonal Decomposition Algorithm for Mining Lagged Time Series”. *Communications in Computer and Information Science*, volume 845, pages 1-19. ISBN 978-981-13-0292-3. Springer International Publishing. 2018.

DOI: [10.1007/978-981-13-0292-3_8](https://doi.org/10.1007/978-981-13-0292-3_8)

- **Chapter 4 - RUP: Large Room Utilisation Prediction with Carbon Dioxide Sensor**

RUP is a novel machine learning technique to estimate the number of people within a closed space from a single carbon dioxide sensor. RUP is the extension work from SD-HOC, and at the time of this thesis being written, is the state-of-the-art machine learning technique for people counting. The solution presented in this chapter completed the answer for *RQ-2*.

Copyright/credit/reuse notice: The contents of this chapter have been taken and revised as needed from a paper published as: [[Arief-Ang et al., 2018a](#)]

Irvan B. Arief-Ang, Margaret Hamilton and Flora D. Salim. “RUP: Large Room Utilisation Prediction with Carbon Dioxide Sensor”. *Pervasive and Mobile Computing*. Volume 46, June 2018, pages 49-72.

DOI: [10.1016/j.pmcj.2018.03.001](https://doi.org/10.1016/j.pmcj.2018.03.001)

- **Chapter 5 - DA-HOC++: A Scalable Room Occupancy Prediction with Transferable Time Series Decomposition of CO₂ Sensor Data**

DA-HOC and DA-HOC++ are the two main techniques that we developed for domain adaptation related indoor human occupancy prediction. The solution presented in this chapter completed the answer for *RQ-3*.

Copyright/credit/reuse notice: The contents of this chapter have been taken and revised as needed from papers published as: [[Arief-Ang et al., 2017](#)]

Irvan B. Arief-Ang, Flora D. Salim, and Margaret Hamilton. “DA-HOC: Semi-Supervised Domain Adaptation for Room Occupancy Prediction using CO₂ Sensor Data”. *In the Proceedings of the fourth ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys 2017)*, pages 1-10. Delft, The Netherlands. 8-9 November 2017.

DOI: [10.1145/3137133.3137146](https://doi.org/10.1145/3137133.3137146)

Irvan B. Arief-Ang, Margaret Hamilton and Flora D. Salim. “A Scalable Room Occupancy Prediction with Transferable Time Series Decomposition of CO₂ Sensor Data”. *ACM Transactions on Sensor Networks (TOSN)*. 2018. (Accepted)

DOI: [10.1145/3217214](https://doi.org/10.1145/3217214)

- **Chapter 6 - THERMO: Thermal Comfort Prediction and Adjustment in Shared Office Environments**

THERMO is a prediction and adjustment model for both general thermal comfort and thermal sensation in a shared office environment. The solution presented in this chapter completed the answer for *RQ-4*.

Copyright/credit/reuse notice: The contents of this chapter have been taken and revised as needed from a paper published as:

Irvan B. Arief-Ang, Margaret Hamilton and Flora D. Salim. “THERMO: Thermal Comfort Prediction and Adjustment in Shared Office Environments”. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*. 2018. (Under review)

DOI: -

- **Chapter 7 - Conclusion**

This chapter concludes the thesis by summarising the main contributions, key findings and limitations of the proposed methods. In addition, the significance of this research and potential future directions in this area are also discussed.

Chapter 2

Sensors: Preliminary Analysis

2.1 Introduction

As discussed in Chapter 1, intelligent analysis and prediction of indoor human occupancy is very important for a variety of purposes, including energy saving. Data obtained from the Department of Energy indicates that the average cost of HVAC [DOE., 2010] accounts for 35-45% of the total maintenance costs within a building. Due to this factor, there are substantial investments made to automate the temperature control systems of buildings, and any type of methodology to improve this area is heavily sought after. Reducing HVAC usage will massively reduce overall energy consumption. One way to improve this is to recognise human occupancy inside each particular space so that a building management system (BMS) can intelligently adjust the HVAC.

The majority of buildings (especially older ones) do not have adequate infrastructure to accurately sense people and to locate where they are within a building, which makes it difficult to obtain accurate occupancy data. Due to this, obtaining a precise ground truth value for analysis will be challenging. Ambient sensors are readily available in most buildings with a BMS, but they have been under-utilised for occupancy recognition.

Ambient intelligence is a vision where technology blends naturally with our everyday life [Basten et al., 2003]. This area has become extremely significant in recent years as more people become aware of their surroundings and want to know more about them. One approach to obtain information about the ambient environment is to use sensor data. With the latest technology, sensor devices are becoming more accurate and economical.

In this chapter, we propose a method of extracting sensor data and correlating this infor-

mation with the number of occupants. Our research contributions are:

1. A feature engineering and selection method to determine the three most dominant feature sets extracted from raw data and using statistical analysis.
2. Using regression and correlation analysis, to determine the most dominant ambient sensor channel for detecting occupancy.

2.2 Related Works

Occupancy detection analysis has existed for some time, and the biggest challenge is to do this without using image recognition. Human occupancy detection using cameras [Erickson et al., 2009] can be accurate, however this method raises privacy issues. Researchers have been undertaking studies to propose various other methods to detect human occupancy.

As occupancy usually relies on motion detection, some researchers [Leephakpreeda, 2005, Dutta et al., 2006] focus on measuring movement. One study used microwave motion sensors to detect motion and then controlled lighting with a time delay to reduce electricity consumption [Leephakpreeda, 2005]. Other research took advantage of ultra-wideband radar to recognise people, because radar can work beyond the line of sight [Dutta et al., 2006]. However, the cost of using radars for occupancy detection is high.

The most popular method for indoor occupancy detection is by using passive infrared (PIR) sensors [Agarwal et al., 2010, Dodier et al., 2006, Garg and Bansal, 2000, Howard and Hoff, 2013, Lu et al., 2010, Wang et al., 2005], as they are cheap and can detect motion easily. Some of these sensors combine PIR sensors with magnetic reed switches and are installed on the doorjamb to detect when the door is opened or closed [Agarwal et al., 2010, Lu et al., 2010, Srinivasan et al., 2010]. A combination of PIR sensor and magnetic reed switches has been used in many types of research, and this has become an unofficial standard to detect indoor human occupancy. SPOT+ uses a Microsoft Kinect sensor, an Arduino microcontroller and an infrared sensor to identify and track the location of user [Gao and Keshav, 2013].

In the last few years, Wi-Fi technology has been used to locate and count the number of occupants in the room. Jiang et al. [Jiang et al., 2012] developed ARIEL, a room localisation system that automatically learns a room's fingerprints based on occupants' indoor movement. Khan et al. [Khan et al., 2015] used smartphones' acoustic, locomotive and location sensors with zero configuration to infer the number of people present at one location. In their research, they required people to keep their smartphones in a pocket or hand, which might not be an

ideal in some cases. Recent work of Depatla et al. [Depatla et al., 2015] in a device-free area used only Wi-Fi signals to estimate occupancy of up to nine people. However, the accuracy dropped significantly when the experiment was conducted indoors. Wi-Fi-based localisation generally requires people to log in to a network or carry their device. There is another Wi-Fi-based localisation technique that does not require users to carry any device, however this only works for a limited number of occupants [Youssef et al., 2007].

Given that occupancy recognition does not require users to be tracked, which would rely on the use of vision-based or wireless-based technology, we ask the following research question ‘Is it possible to detect indoor occupancy using the proliferation of sensor devices that are already widely available in buildings and installed for other purposes, so that we can leverage the functionality of these devices?’ The Building-Level Energy Management Systems (BLEMS) project at the University of Southern California [Mamidi et al., 2012] has considered this possibility and is addressing it using several ambient sensors, such as indoor temperature, relative humidity, illumination, carbon dioxide and sound sensors. The limitations of their research are that they did not use cross-validation to optimise the results and they used limited types of sensors.

With regards to the locations of occupancy detection experiments, a few studies have focused on residential houses [Lu et al., 2010, Srinivasan et al., 2010, Barbato et al., 2009]. Most research is related to user profiling for the purpose of reducing energy consumption. Other research has focused more on single person office rooms [Leephakpreeda, 2005, Wang et al., 2005], as this is a more controlled environment, and supervised learning could be implemented with fewer people engaged.

This research focuses on using contact-free, device-free approach to occupancy recognition using ambient sensors. With ambient sensors, users’ privacy is protected, and the devices are widely available in existing buildings with HVAC systems installed [Agarwal et al., 2011, Tsao and Hsu, 2013].

2.3 Methodology

2.3.1 Feature Engineering

For feature extraction, several types of sensors were used as our main features. The overall set of features were:

- Time: the time information used minutes as the lowest granularity.

- Segment of the day: the days were divided into four equal subsections: morning, afternoon, evening and night (only for experiment 3).
- Indoor temperature (T): the room temperature, in Celsius, which is recorded by the sensor.
- Relative humidity (H): the ratio of the partial pressure of water vapour to the equilibrium vapour pressure of water recorded by the sensor.
- CO₂ rate (CO₂): the carbon dioxide content in ppm recorded by the sensor.
- Sound rate (S): the level of noise within a room in decibels recorded by the sensor.
- Atmospheric pressure (P): the value of pressure exerted by the weight of air recorded by the barometer.
- Illumination (L): the level of brightness within a room in lux recorded by the sensor.

To collect ground-truth data (the number of people who were staying in the room at that time), several other sensors and a web-based app were used:

- Door state: all opening and closing events were recorded.
- Power consumption: computer monitor power usage was monitored.
- Motion sensor: detected and stored data if there was motion within the range of the sensor.
- Occupancy app: our own custom-built web-based app to gather ground-truth data. The room owner stores data about the times when people enter and leave the room, and records the number of people in the room as it changes.

Door state, power consumption, and motion sensor data were for ground truth and were not to be used for the data mining analysis.

Other than normalised raw data analysis, there were several derived features, as follows:

1. Ten-minute time segment
 - a) Ten-minute time window with maximum value
 - b) Ten-minute time window with minimum value

- c) Ten-minute time window with mean (μ) value
 - d) Ten-minute time window with variance (σ) value
2. Segment of the day: we divided the data into time segments (morning, afternoon, evening and night), as follows:
- Morning (6 a.m. to 12 p.m.)
 - Afternoon (12 p.m. to 6 p.m.)
 - Evening (6 p.m. to 12 a.m.)
 - Night (12 a.m. to 6 a.m.)

To summarise, eight main features were used in our experiment.

2.3.2 Machine Learning Algorithm

In this chapter, we use multi-layer perceptron, Gaussian processes with radial basis function (RBF) as core, support vector machine (SVM), random forest and naïve Bayes to recognise the level of accuracy.

To identify the most dominant sensor, we fit our data to linear regression in Equation 2.1 and perform regression analysis.

$$\hat{O} = \alpha T + \beta H + \gamma CO_2 + \delta S + \epsilon P + \zeta L + \eta \quad (2.1)$$

where \hat{O} is the number of occupants, T is temperature, H is humidity, CO_2 is the level of CO_2 , S is sound, P is pressure and L is light (illumination). Our formula provides a straightforward but effective way to identify the dominant sensors for human occupancy, as they will have higher coefficients than the other sensors.

For each dominant sensor, we fit a fourth order polynomial and perform a logistic regression analysis to check the goodness of fit and R^2 , a statistical measure of how close the data are to the fitted regression line. For our final evaluation, we compute the Pearson product-moment correlation coefficient to check the correlation between each sensor and the occupancy data.



Figure 2.1: (1) SmartThings SmartSense Open/Closed Sensor; (2) Netatmo Urban Weather Station; (3) Z-Wave Aeon MultiSensor;(4) SmartThings SmartPower Outlet.

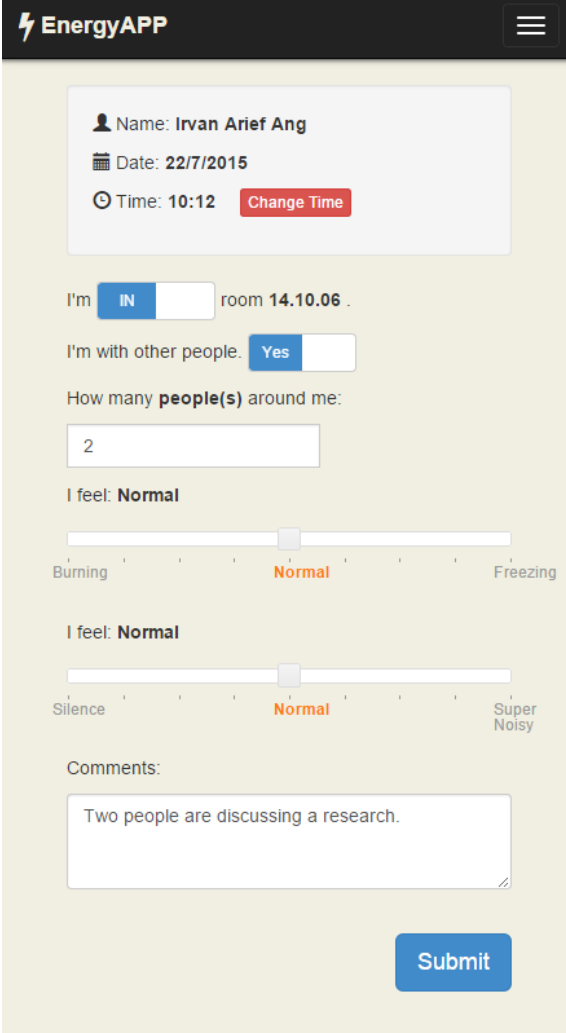
2.4 Data Collection and Preprocessing

2.4.1 Type of Sensor Devices

For this experiment, we deployed a Z-Wave Aeon Multi Sensor, a SmartThings SmartSense Open/Closed Sensor, a SmartThings SmartPower Outlet and a Netatmo Urban Weather Station, as shown in Figure 2.1. We ran the experiment continuously for two weeks and gathered data for indoor temperature, relative humidity, the rate of carbon dioxide (CO_2), sound level, atmospheric pressure, level of illumination, whether the door was open or closed, the power consumption of the monitor, and the motion within the room.

We deployed the SmartSense Open/Closed Sensor on the door, to detect the opening or closing of the door. The Urban Weather Station consists of multivariable sensors that can detect indoor temperature, relative humidity, atmospheric pressure, the rate of CO_2 in the air, and the sound level. The SmartPower Outlet was combined with a step-down converter due to the difference in voltage between the device's country of origin (USA) and our country's

voltage (Australia), and installed with the monitor. The MultiSensor, which is a combination of four sensors (indoor temperature, relative humidity, illumination level and motion sensor), was deployed in front of the user and below the monitor so that it would detect motion every time a user sits in front of the PC.



The screenshot shows the EnergyAPP interface. At the top, there is a header with a lightning bolt icon and the text "EnergyAPP" and a menu icon. Below the header, there is a user profile section with the following information:

- Name: **Irvan Arief Ang**
- Date: **22/7/2015**
- Time: **10:12** with a **Change Time** button.

Below the profile section, there are several input fields and sliders:

- "I'm **IN** room **14.10.06** ."
- "I'm with other people. **Yes**"
- "How many **people(s)** around me:" with a text input field containing **2**.
- "I feel: **Normal**" with a slider ranging from **Burning** to **Freezing**, with **Normal** in the center.
- "I feel: **Normal**" with a slider ranging from **Silence** to **Super Noisy**, with **Normal** in the center.
- "Comments:" with a text area containing **Two people are discussing a research.**

At the bottom right, there is a blue **Submit** button.

Figure 2.2: Occupancy Web-based Application.

2.4.2 Occupancy Web-based Application

We developed an occupancy application to collect user annotations of their room's actual occupancy. This is a web-based application for the room owner to complete every time an

event occurs with the door, where “event” means that a person comes in or out. The interface of this application is shown in Figure 2.2.

2.4.3 Data Preprocessing

To be able to recognise indoor human occupancy and correlate it with ambient sensor data, ground truthing of about the number of people within a room at a specific time is required. On the other hand, privacy must be respected and therefore the use of cameras was not an option. Thus, multiple methods are applied to secure a solid ground truth.

Human occupancy ground truth data was collected from a web-based application, which provides a form to be filled in by the resident every time a person enters or leaves the room. As ground truth data from human input may be unreliable, we added three more validation layers to ensure high-quality ground truth data. The first validation layer is a power consumption sensor that connects to the monitor. If the power consumption goes up after a while, it is assumed that there is a person in the room. The second validation layer is a motion sensor, and the last is a door open and close sensor. Every time the door is opened and closed, there is a possibility that the number of people in the room changes. With those three additional validation layers of ground truth, we have developed a rigorous evaluation method to obtain an accurate measure of indoor human occupancy without using a camera. To handle the missing occupancy value problem, we designed a simple logic shown in Algorithm 1.

Algorithm 1 Fixing missing occupancy value algorithm

```

1: procedure MISSING_OCCUPANCY(SensorData[t])
2:   Occ[t] ← 0
3:   PC[t] ← low
4:   M[t] ← low
5:   DS[t] ← 0
6:   for each node i ∈ SensorData[t] do
7:     if (Occ[i] = 1) AND (PC[i] = high) then
8:       if (M[i] = high) AND (DS[i] = 1) then
9:         SensorData[i].Occupancy ← 1
10:      end if
11:    end if
12:  end for
13: end procedure

```

▷ *Occ*[*t*]: Occupancy
▷ *PC*[*t*]: Power Consumption
▷ *M*[*t*]: Motion
▷ *DS*[*t*]: Door State

For data collection, we set up background devices to gather the data from various different sensors. As we gathered the data using a variety of devices, data integration is challenging.

Table 2.1: List of coefficients for each parameter for non-linear regression formula

Experiment	Temp (α)	Humidity (β)	CO ₂ (γ)	Sound (δ)	Pressure (ϵ)	Illumination (ζ)	Coefficient (η)
Exp. 1	-29.55	-6.20	144.80	54.98	-12.15	83.59	-0.06
Exp. 2 - Max	-26.96	-6.90	137.66	51.82	-12.32	83.00	-0.48
Exp. 2 - Min	-40.50	-5.11	132.09	49.12	-15.81	74.82	6.43
Exp. 2 - Avg	-31.59	-6.23	128.85	55.03	-13.28	82.40	1.57
Exp. 2 - Var	26.39	0.00	53.79	159.23	68.86	213.29	5.98
Exp. 3	0.00	0.00	0.00	96.57	0.00	64.46	-6.28

Some sensors' data (temperature, humidity, illumination, CO₂, sound and pressure) are continuous data, whereas the door open/close and web-based app data are events-based discrete data. We used the timestamp to integrate the data and generate the missing data using the interpolation method.

Each sensor datum was normalised between 0-1 to ensure consistency. The normalisation process was conducted for each type of sensor. All indoor human occupancy data from the web-based app input was converted to binary occupancy and was integrated with normalised sensor data using timestamp as the joint key.

2.5 Experiments and Results

In this chapter, we present three experiments with the data. The first experiment used the normalised raw data with the pre-processing method that was explained in the previous section. We used each of the data mining algorithms mentioned previously to determine the most accurate algorithm. In the second experiment, the normalised raw data were aggregated at ten-minute intervals. Several statistical features, such as maximum value (max), minimum value (min), average value (μ) and variance (σ), were used for analysis. We adopted the same data mining algorithms for each feature to see if the result differ from the first experiment. In the third experiment, the data were segmented into four temporal segments (morning, afternoon, evening, night) before using each data mining algorithm. This was done because every temporal segment has a different context that may indicate the difference in occupancy number (i.e. night time has fewer people), so this is a reasonable segmentation. We compared all three results and determined which condition generates the highest level of accuracy. Table 2.1 shows the coefficient results for Equation 2.1.

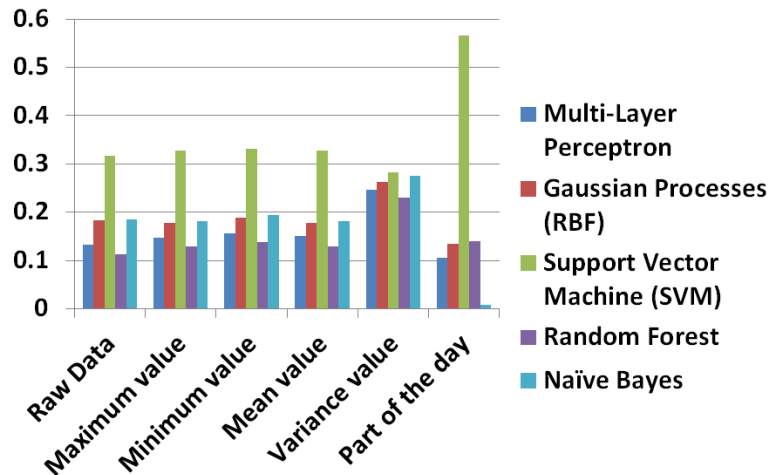


Figure 2.3: Root-mean-square Error Results of Various Machine Learning Algorithms.

2.5.1 Experiment 1 (Normalised Raw Data)

This experiment utilised a total of 32,152 normalised raw data points that were collected from one single-person office room. Table 2.1 shows that CO₂, light and sound are three dominant features to be considered for human occupancy recognition. Random forest has the highest accuracy result and has a more stable performance across different feature sets compared to other machine learning algorithms. Figure 2.3 shows that most machine learning algorithms have low RMSE, with the exception of SVM. With this finding, the accuracy result can be trusted.

2.5.2 Experiment 2 (Ten-Minute Time Segments)

For the ten-minute time segments, four sets of features were extracted: data analysis with the maximum value (max), the minimum value (min), the average value (μ) and variance (σ). For the first three analyses (max, min and μ), the results were similar to Experiment 1 (refer to Table 2.1 and Figure 2.3). However, for the last analysis (σ), the value was different. Table 2.1 shows that the relative humidity is ignored. The top three dominant features were the same (light, sound and CO₂), albeit in a different order to the previous ones. Figure 2.3 shows that the RMSE value is between 0.2 and 0.3. As the degree of error is quite high, the accuracy result for each machine learning algorithm could not be trusted.

Table 2.2: List of coefficients for each parameter for regression line polynomial order of 4.

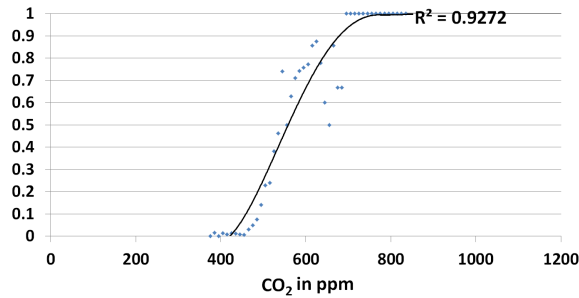
Algorithm	α	β	γ	δ	ϵ	R^2
Sound rate	8E-06	-0.0018	0.1345	-4.3749	51.352	0.932
Illumination	-2E-07	5E-05	0.0037	0.118	-0.32	0.767
CO ₂ rate	6E-11	-2E-07	0.0002	-0.0611	8.1978	0.856

2.5.3 Experiment 3 (Part of the Day Segments)

The level of accuracy for the majority of the machine learning algorithms is generally higher than the raw data. This result is acceptable, as aggregate data by part-of-the-day make the data more congested. There are two outputs that need to be highlighted. First, the accuracy of the SVM classifier dropped significantly (<70%). However, the naïve Bayes classifier achieves a very high accuracy level (>99%) with very low RMSE.

2.5.4 Most Dominant Sensor in Determining Human Presence

Based on the results of all experiments in Table 2.1, we conclude that the top three dominant sensors for recognising indoor human occupancy are the CO₂ rate, illumination level and sound rate sensors. Table 2.2 contains both R^2 value and Pearson correlation r-value for these top three sensors.

Figure 2.4: Occupancy Recognition Probability over CO₂ Level.

2.5.4.1 CO₂ Rate vs Indoor Human Occupancy

Figure 2.4 shows that CO₂ correlated well with human occupancy. Once the carbon dioxide exceeded 600 ppm, there was a high change (>0.7) that there was at least one person inside the room. The logistic regression model was good and had a very strong R^2 value (0.9272). In

Table 2.3: R^2 value and Pearson correlation r-value for the three most dominant sensors for recognising indoor human occupancy.

	R^2	Pearson correlation r
CO ₂ rate	0.9272	0.882469
Illumination level	0.7665	0.288449
Sound rate	0.9317	0.897066

Table 2.3, Pearson's r-value is high and shows a very strong correlation between CO₂ rate and indoor human occupation recognition.

2.5.4.2 Illumination Level vs Indoor Human Occupancy

Figure 2.5 shows that illumination level is not highly correlated with the number of occupants. The R^2 value (0.7665) is also not high. Our conclusion is that once the value of the illumination level is above 10 lux, there is a high chance (>0.8) that there is at least one person inside the room. Regardless of whether the room gets brighter or not afterwards, it has a low correlation to the occupancy recognition.

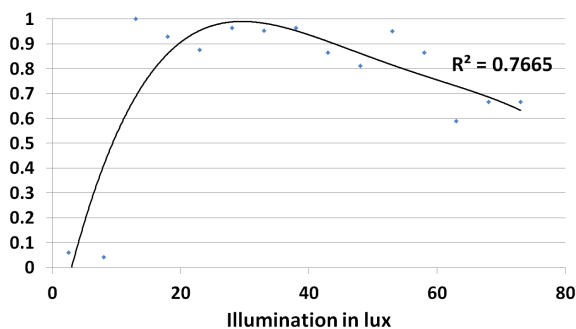


Figure 2.5: Occupancy Recognition Probability over Illumination Level.

The Pearson's r-value in Table 2.3 shows a weak correlation. An external factor such as window placement in the room can complicate this, as the sun will make the room brighter during midday compared to night time. This is the primary cause of why there is lux >70 but occupancy recognition probability is only close to 0.7.

2.5.4.3 Sound Rate vs Indoor Human Occupancy

In Figure 2.6, there is a strong correlation between sound rate and the number of occupants. Below 40 dB, the probability of a human being inside a room is 0. Once the value reaches

53 dB and above, we discover that there is at least one person inside the room. The logistic regression model also fit perfectly with a high goodness of fit value ($R^2 = 0.9317$), so we can confidently assume that this regression fitting line has a significant correlation with the sound rate aggregate data.

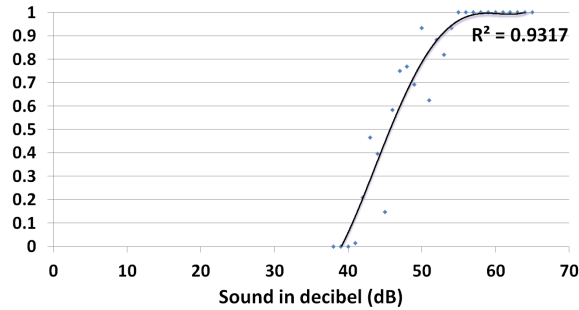


Figure 2.6: Occupancy Recognition Probability over Sound Sensor Rate.

Based on Pearson's r -value in Table 2.3, there was a very strong correlation between sound rate and indoor human occupation recognition.

2.6 Conclusion

We have shown that non-intrusive ambient sensor data can be used to accurately recognise indoor human occupancy. With a random forest data mining classifier on normalised raw data and aggregated data using various statistical methods, we managed to achieve an accuracy level as high as 98%. With the right treatment of data aggregation, naïve Bayes can achieve 99% accuracy. This level of accuracy is higher than any similar previous studies. We also proved that CO_2 rate, illumination level and sound rate are the top three most dominant features to detect indoor human occupancy, and that the most dominant is CO_2 rate. This result is encouraging, because when there are more people inside a room, the rate of CO_2 in that room gradually increases.

Chapter 3

SD-HOC: Seasonal Decomposition Algorithm for Mining Lagged Time Series

3.1 Introduction

Data mining technology is blended into human life and helps solve many problems that could not previously be solved. The problem we will consider in this chapter relates to the operational costs of buildings. From the US Department of Energy, 35-45% of the total operational costs of a building are spent on heating, ventilation, and air conditioning (HVAC) [DOE., 2010]. Substantial investment in energy usage research is needed to reduce HVAC costs in buildings according to their occupancy patterns. Reducing HVAC usage helps to reduce overall energy consumption. Furthermore, a building management system (BMS) can then intelligently adjust the HVAC based on the occupancy pattern so that the comfort of the occupants is not sacrificed.

Using sensor data to detect human presence is the current trend in ambient sensing research [Candanedo and Feldheim, 2016, Ekwevugbe et al., 2013a, Hailemariam et al., 2011, Khan et al., 2014, Leephakpreeda, 2005]. Yan highlighted the importance of occupant-related research [Yan et al., 2015]. In Chapter 2, it was highlighted that carbon dioxide (CO₂) is the best ambient sensor for detecting human presence. By using only CO₂, 91% accuracy was achieved for binary prediction of whether a room is occupied or vacant [Basu et al., 2015] and 15% accuracy was achieved for recognising the number of occupants. A hidden Markov model (HMM) was used for the CO₂ dataset to predict human occupancy, and a 65-80% range of accuracy was achieved

for predicting up to four occupants [Lam et al., 2009].

In this chapter, we develop a new algorithm for decomposing large datasets to extract the relevant features to be used for prediction and identification of seasonal trends. We can then apply the computations of these trends to various incomplete datasets, matching the time series to predict the relevant future features in the new dataset. We identify relevant seasonal trends in the data over time, apply these to the new dataset and use them to predict future trends in the data.

We have found this to be particularly useful for sensor data where we can extrapolate the CO₂ data to indoor human occupancy prediction with promising accuracy. We can match the sensor measurements for zero occupancy, at various times, possibly overnight, and tune our predictions to optimise individual comfort and the overall carbon footprint of the building.

We apply our new feature transformation algorithm to the prediction of the number of people in a room at a particular time through the measurement of the CO₂. Human occupancy prediction is of significant interest to the building industry, because it enables the automation of heating, cooling and lighting systems. If it is known that certain rooms are empty or under-utilised during certain times, then operational costs and carbon footprint can be reduced with better planning and scheduling. When the rooms are not occupied, the building system can also adjust these facilities to keep the inhabitants comfortable. This framework is called seasonal decomposition for human occupancy counting (SD-HOC).

SD-HOC preprocesses the data and integrates various machine learning algorithms. The experiment is conducted at two different locations and in two stages. In the first stage, the SD-HOC result is compared with a variety of other data mining prediction algorithms, such as decision tree, multi-layer perceptron, Gaussian processes (radial basis function), support vector machine and random forest. The second stage of our experiment compares SD-HOC with one of the best data mining prediction accuracies to predict the human occupancy number on a different number of prediction days. There are three advantages of using SD-HOC:

1. It has a low equipment cost due to pre-installation.
2. It ensures that users' privacy is protected.
3. It only uses CO₂ data, reducing the chance of errors caused by data integration.

3.2 Background and Related Work

When using image processing techniques [Erickson et al., 2009, Lee et al., 2011], the levels of accuracy for human occupancy detection can reach up to 80%. Unfortunately, these methods raise privacy concerns. Research communities have been investigating alternative methods to detect human occupancy without using cameras or image processing.

We focus on utilising only CO₂ sensors to estimate the indoor human occupancy number. The main reason is because CO₂ sensors are already integrated with the BMS and ventilation infrastructure and are commonly installed in buildings.

Machine learning algorithms, including a HMM, neural networks (NN) and support vector machine latent (SVM latent), were used in [Lam et al., 2009] by using CO₂ data with sensors deployed both inside and outside room. By feature engineering CO₂ data with first order and second order differences of CO₂, the accuracy achieved was between 65% and 80%.

[Dedesko et al., 2015] used a mass balance approach to predict both human occupancy and occupant activity using CO₂ and door sensors. The authors mentioned that sources of error and uncertainty are part of the limitations of this approach. CO₂ based occupancy detection in office and residential buildings was implemented in [Cali et al., 2015]. The binary occupancy prediction accuracy was 95.8% and the people counting accuracy was 80.6% for two or three people in each room.

PerCCS is a model with a non-negative matrix factorisation method to count people [Basu et al., 2015] using CO₂ as the only predictor. In predicting vacant occupancy, they achieved up to 91% accuracy, but achieved only 15% accuracy in predicting the number of occupants.

Overall, sensor-based detections have higher accuracy compared to radio-based detections. For example, Wi-Fi and received signal strength indication (RSSI) signals achieved 63% accuracy for indoor detection [Debatla et al., 2015] with nine occupants. For occupancy counting, CO₂ sensors alone have been investigated, with a maximum of 42 occupants and accuracy limit of 15% [Basu et al., 2015].

3.3 Problem Definition

Given the significant motivations for our research, this chapter addresses the problem of how we can use data mining techniques and feature selection to predict the number of people by using a single CO₂ sensor. We expect the results to have similar accuracy to state-of-the-art techniques in the occupancy-detection field. In Figure 3.1, the data show that there is a

dependency between CO₂ and occupancy data.

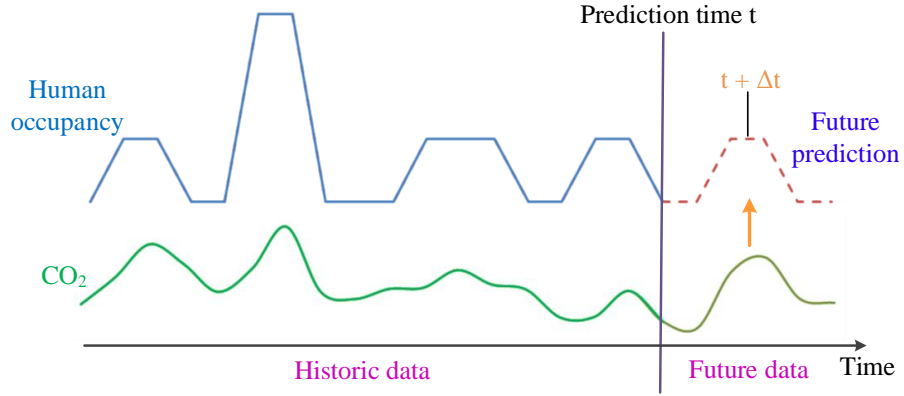


Figure 3.1: Real-time Prediction Scenario Showing Human Occupancy and CO₂ Fluctuation. The Fundamental Task is to Predict the Number of Occupants at Time $t + \Delta t$.

3.3.1 Scenario Assumption

Assume $|TS|$ represents the length of a time series, $TS = \{ts_1, ts_2, \dots, ts_q\}$, where q means the number of sample points. In our time series datasets, we have two aspects:

- Carbon dioxide (CO₂) concentration C , defined as

$$C = \{C_1, C_2, \dots, C_q\}$$
- Indoor human occupancy O , defined as

$$O = \{O_1, O_2, \dots, O_q\}$$

Our framework only depends on the CO₂ dataset to calculate the prediction. This is where the challenge lies, as the model needs to extract more features from a time series: this may seem simple, but it contains hidden trends. We introduce a term ‘lagged time series’ as a set of data in a regression time series, where each value relates to a situation in a surrounding context but belongs to a different time frame.

3.3.2 Time Series

In time series prediction, analysing one-step-ahead prediction is different from analysing multi-step-ahead prediction. Predicting multi-step-ahead needs a more complex method due to

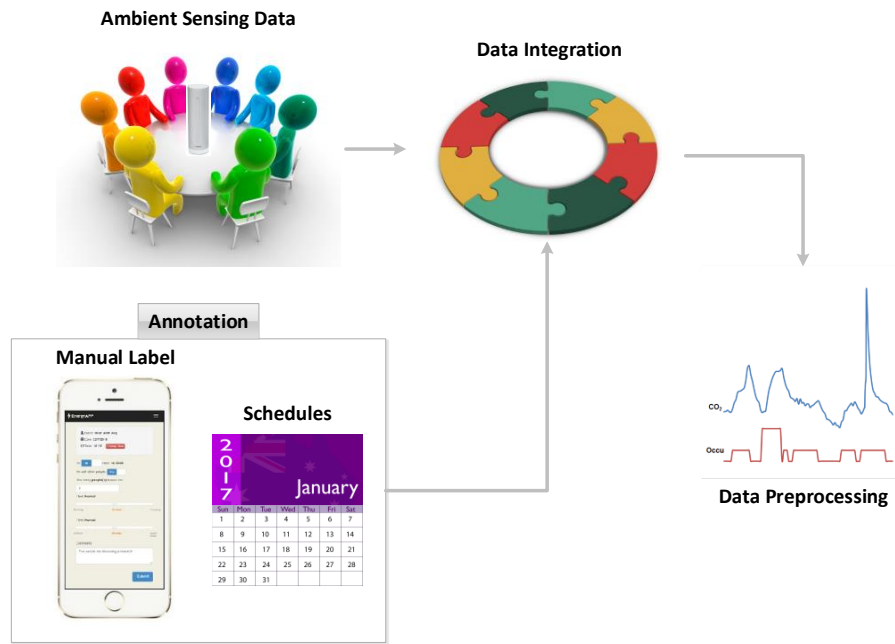


Figure 3.2: Data Collection and Analysis Framework.

the accumulation of errors and the number of uncertainties increasing with time. We focus on multi-step-ahead prediction with the support of one dependent variable to reduce uncertainties.

We have two different types of datasets: CO₂ concentration C and indoor human occupancy O . To explore the relationship between these factors, we need to identify the relevant features by analysing the correlations between CO₂ concentration, indoor human occupancy and all of their decomposed components.

3.4 The Features

This section explains our data preprocessing time series components, cross-correlation and line of best fit. Data preprocessing is crucial for our model, as it further increases the prediction analysis with various machine learning algorithms that we implemented in the experiment section. We collected data on the CO₂ concentration from the sensor data and the number of humans in the room, as shown in Figure 3.2. Both datasets are preprocessed and integrated using our novel method described in subsection 3.4.2. We transformed each dataset using feature engineering into more features, and applied our prediction model SD-HOC (described

in section 3.5) to predict the indoor human occupancy.

3.4.1 Time Delay Components

Time delay is a problem because it takes time for the concentration of CO₂ to build up enough to measure the presence of a person. To model a real time delay, we need the value of a time series regression function obtained after a specific time lag. This issue occurs in the majority of sensor data analyses, as data obtained from sensor readers need to travel to those sensor readers before being captured. In our study, when one person enters a room, it will take some time before the CO₂ level in the air increases proportionally. For this reason, we must preprocess the data to fix the time delay between CO₂ data and the indoor human occupancy number.

3.4.2 Cross-Correlation and the Line of Best Fit

Before analysing the data on CO₂ and the number of occupants, the data lagging issue needs to be considered. Data lagging means that it will take a certain time for CO₂ to populate the room, as there is a delay between the time of people exiting (or entering) the room and the decrement (or increment) of the CO₂ value in the air. To find out how much data lagging needs to be implemented, we first need to find the upper bound value (UB). UB is a maximum value that is calculated based on the room volume. UB will be used to calculate the time lag value, and is defined by the formula in Equation 3.1.

$$UB = |(RL * RW * RH)/C| \quad (3.1)$$

<i>UB</i>	upper bound value
<i>RL</i>	room length
<i>RW</i>	room width
<i>RH</i>	room height
<i>C</i>	constant value (100)

For each dataset from 0 minutes time lag to UB minutes time lag, the correlation of CO₂ data with the number of occupants is measured. If the room size is small, the UB value will be 1. The larger is room is, the bigger the UB value is. In our case study, for the small room A the UB value is 1, and for the big room B the upper bound of UB is 60. This value is aligned with the explanation above due to the large size of room B.

To calculate a line of best fit, we need to calculate the slope value between CO₂ and occupancy data, defined by Equation 3.2.

$$SL = \frac{\sum(O_t - \bar{O}_t)(C_t - \bar{C}_t)}{\sum(O_t - \bar{O}_t)^2} \quad (3.2)$$

SL slope of the linear regression line
O_t occupancy value
 \bar{O}_t sample means of the known occupancy value
C_t CO₂ value
 \bar{C}_t sample means of the known CO₂ value

Next, the intercept value between both datasets needs to be calculated, using the formula in Equation 3.3.

$$IC = \bar{C}_t - SL * \bar{O}_t \quad (3.3)$$

IC intercept of the linear regression line
 \bar{C}_t sample means of the known CO₂ value
 \bar{O}_t sample means of the known occupancy value

The main formula for the line of best fit (LBF) is shown in Equation 3.4.

$$LBF = (O_t - (SL * C_t + IC))^2 \quad (3.4)$$

O_t occupancy value
SL slope of the linear regression line
C_t CO₂ value
IC intercept of the linear regression line

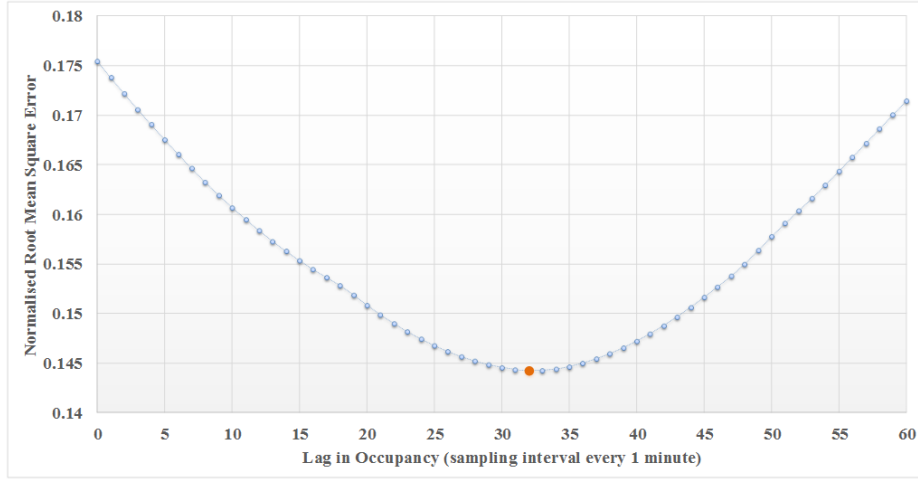


Figure 3.3: Ordinary Least Square Regression Normalised Root Mean Square Error Between CO₂ Data and Actual Occupancy for 60 Minutes Time Lag.

$$NRMSE = \frac{\sqrt{\frac{1}{n} \sum_{t=1}^n (C_t - \bar{C}_i)^2}}{O_{max} - O_{min}} \quad (3.5)$$

$NRMSE$	normalised root mean square error
t	total number of dataset
C_t	CO ₂ value
\bar{C}_t	sample means of the known CO ₂ value
O_{max}	maximum occupancy value
O_{min}	minimum occupancy value

3.4.3 Time Lag

For each line of best fit from subsection 3.4.2, we calculated the mean square error (MSE), root-mean-square deviation (RMSD) and the normalised root mean square error (NRMSE). The formula for calculating NRMSE is shown in Equation 3.5.

$$TL = \min(NRMSE) \quad (3.6)$$

This step is repeated UB times for each time lag. For the time lag analysis, we use least square regression to compare each NRMSE from time lag 0 until time lag UB. We pick the

lowest value of NRMSE as our time lag value (TL). The TL value formula is shown in Equation 3.6 and performs as our baseline time lag for the data analysis.

For the academic staff room, the TL value is 0. This value means that no time lag is needed for this analysis. For the cinema theatre, the lowest NRMSE value was at time lag $TL = 32$, as shown in Figure 3.3. This TL value is our base for the entire cinema theatre data analysis.

3.5 The Framework

There is no linear relationship between CO_2 and indoor human occupancy. To address this, we introduce a new SD-HOC analysis framework to decompose both CO_2 and occupancy data as shown in Figure 3.4. In this chapter, the main decomposition method we use is seasonal trend decomposition (STD).

The core feature transformation prediction model will be explained in the following subsections. The first subsection discusses STD in detail. The next subsection explains the correlation model for trend, seasonal and irregular features. The last subsection presents zero pattern adjustment (ZPA), a new method for analysing conditions when the room is vacant. The ZPA method can increase overall accuracy. This model needs to be re-trained for different locations to obtain the most optimal accuracy.

3.5.1 Seasonal-Trend Decomposition (STD)

STD is a decomposition technique in time series analysis. The X-11 method with moving average is one of the most famous variants [Shiskin et al., 1965], and X12-ARIMA is the most recent variant [Findley et al., 1998]. STD is an integral part of our framework.

To understand each time series dataset, we utilise STD to decompose the data into four main features: trend, cyclical, seasonal and irregular. The trend feature (T_t) represents the long-term progression of the time series during its secular variation. The cyclical feature (C_t) reflects a repeated but non-periodic fluctuation during a long period of time. The seasonal feature (S_t) is a systematic and regularly repeated event during a short period of time. The irregular feature (e_t , also known as error or residual) is a short term fluctuation from the time series and is the remainder after the trend, cyclical and seasonal features have been removed. In this chapter, we decided to combine the cyclical feature into the trend feature due to their similarity, to make the model simpler without sacrificing accuracy.

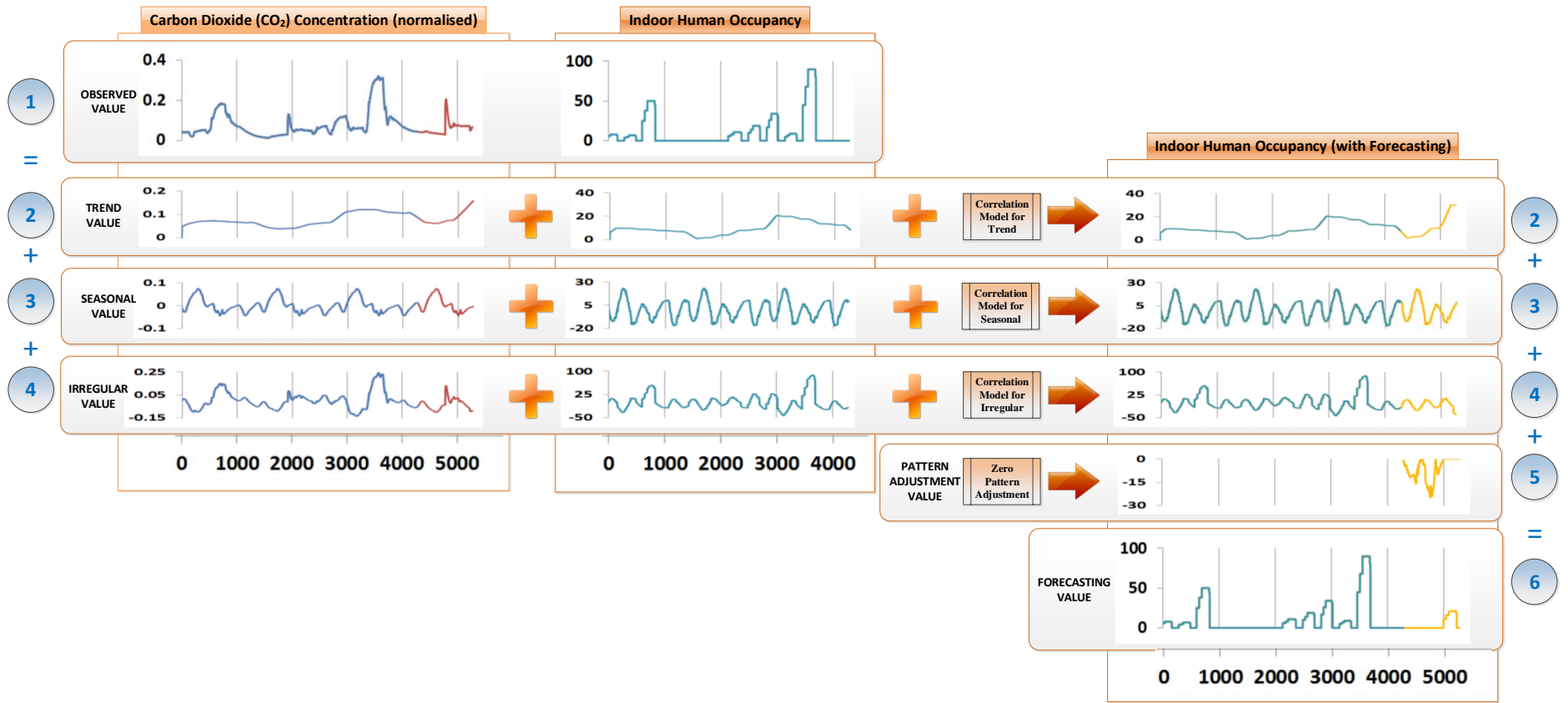


Figure 3.4: Seasonal Decomposition for Human Occupancy Counting (SD-HOC) Analysis Framework.

Below is the core logic for seasonal trend decomposition:

1. Calculate 2 x 12 moving average in both the CO₂ and occupancy raw datasets to obtain a rough trend feature data T_t for all periods (12 is the default because there are 12 months in a year).
2. Calculate ratios of the data to trend, named “centred ratios” (STD_t/T_t).
3. To form a rough seasonal feature (S_t) data estimation, apply a separate 2 x 2 moving average to each month of the centred ratios.
4. To obtain the irregular feature (e_t), divide the centred ratios by S_t .
5. Multiply the modified e_t by S_t to get modified centred ratios.
6. Repeat step 3 to obtain a revised S_t .
7. Divide the raw data by the new estimate of S_t to give the preliminary seasonal adjusted series, y_t/S_t .
8. Estimate the trend feature (T_t) by applying a weighted Henderson moving average [Hyndman, 2011] to the preliminary seasonally adjusted values.
9. Repeat step 2 to get new ratios by dividing the raw data by the new estimate of T_t .
10. Repeat steps 3 to 5 using the new ratios and applying a 3 x 5 moving average instead of a 3 x 3 moving average.
11. Repeat step 6 but using a 3 x 5 moving average instead of a 3 x 3 moving average.
12. Repeat step 7.
13. Finally, calculate the remainder feature by dividing the seasonally adjusted data from step 12 by the trend feature from step 8.

Our customised STD formulation is:

In this chapter, we decided to use additive decomposition, because it is the simplest way to give the first approximation. Our overall STD formula becomes:

$$STD_t = T_t + S_t + e_t \quad (3.8)$$

$$STD_t = f(T_t, S_t, e_t) \quad (3.7)$$

t	time
STD_t	actual value of a time series at time t
T_t	trend feature at t
S_t	seasonal feature at t
e_t	irregular feature at t

This general STD formula will be applied to both the CO₂ time series and human occupancy time series datasets:

$$C_t = T_t^C + S_t^C + e_t^C \quad (3.9)$$

$$O_t = T_t^O + S_t^O + e_t^O \quad (3.10)$$

To predict O_{t+1} up to O_{t+n} , we need to create a model to systematically predict each of T_{t+1}^O , S_{t+1}^O and e_{t+1}^O up to T_{t+n}^O , S_{t+n}^O and e_{t+n}^O , and then reconstruct the new prediction dataset using the additive method.

3.5.2 Correlation Models

There are three correlation models for the trend, seasonal and irregular features, as described in the following subsections.

3.5.2.1 Correlation Model for Trend Feature (T_t)

The trend feature (T_t) is defined as the long-term non-periodic progression of the time series during its secular variation. We assume that the trend feature for the CO₂ dataset (T_t^C) will be similar to the trend feature for indoor human occupancy (T_t^O), because there is dependency between both datasets.

The correlation model for the trend feature starts by checking the similarity between both trend features. We use the Pearson product-moment correlation coefficient (PCC) [Stigler, 1989] to validate it, as shown below:

Pearson's r-value ranges from -1 to $+1$. If the value is >0.7 , the correlation between both datasets is strongly positive. If the correlation is less than 0.7 , data preprocessing needs to be redone to find the new TL value (Equation 3.6).

$$PCC_r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \quad (3.11)$$

PCC_r	correlation coefficient for PCC
x	dataset x
y	dataset y
n	number of sample points

Once past the validation step, polynomial M5 linear regression is implemented. We chose the M5 method because it builds trees whose leaves are associated with multivariate linear models; the nodes of the tree are chosen over attributes that maximise the expected error reduction, given by the Akaike Information Criterion (AIC) [Akaike, 1974]. AIC is a measure for checking the relative goodness of fit of a statistical model. The purpose of using AIC is to evaluate the model. The value for each of the trend features needs to be positive, so we put the absolute value on both the CO₂ ($|T_t^C|$) and human occupancy trend features ($|T_t^O|$). The main formula for trend feature correlation is shown below:

$$|T_t^O| = |\alpha_0 + \alpha_1(T_t^C) + \alpha_2(T_t^C)^2 + \dots + \alpha_n(T_t^C)^n + \epsilon| \quad (3.12)$$

Linear regression with M5 will output each α_n and ϵ value. With these parameters, the future trend for T_{t+n}^O can be obtained.

3.5.2.2 Correlation Model for Seasonal Feature (S_t)

The seasonal feature (S_t) is a systematic and regularly repeated event during a short period of time. Due to this characteristic, every seasonal feature can be fitted by a finite Fourier series. To correlate S_t^C and S_t^O , we use dynamic time warping (DTW), a pattern matching technique to score the similarity between the shape of a specific signal within a certain duration [Petitjean et al., 2011]. The full correlation algorithm (Algorithm 2) is implemented to find regularly repeated events within each S_t .

Once we find repeated events in s_t^{fin} for both the CO₂ and occupancy seasonal features, we compare the length of $s_t^{fin(O)}$ and $s_t^{fin(C)}$. If the length of $s_t^{fin(O)} < s_t^{fin(C)}$, we apply an interpolation method inside $s_t^{fin(O)}$ so both have the same length. If the length of $s_t^{fin(O)} > s_t^{fin(C)}$, we apply a data reduction method so that both have the same length. The final

Algorithm 2 Finding a repeated event inside a seasonal feature

```

1: procedure REPEATED_EVENT( $S_t$ )
2:    $s_t^{temp}, s_t^{fin} \subset S_t$ 
3:    $len \leftarrow 0$ 
4:    $a \leftarrow S_t[len]$ 
5:   for each node  $i \in S_t$  do
6:      $len++$ 
7:      $s_t^{temp} \leftarrow s_t^{temp} + S_t[i]$ 
8:     if  $a = S_t[i]$  then
9:       if  $DTW(s_t^{temp}, S_t[i + 1..i + len]) > 95$  then
10:         $s_t^{fin} \leftarrow s_t^{temp}$ 
11:        break
12:      end if
13:    end if
14:  end for
15:  return  $s_t^{fin}$ 
16: end procedure

```

$\triangleright len$: Length for s_t^{temp}
 $\triangleright a$: Start Point

regression equation for seasonal feature correlation is shown in equation 3.13 where the future trend for $S_{t+n}^{fin(O)}$ can be obtained.

$$s_t^{fin(O)} = f(s_t^{fin(C)}) \quad (3.13)$$

3.5.2.3 Correlation Model for Irregular Feature (e_t)

Due to similar characteristics between the trend and irregular features, we apply the same correlation method from the trend feature:

$$|e_t^O| = |\beta_0 + \beta_1(e_t^C) + \beta_2(e_t^C)^2 + \dots + \beta_n(e_t^C)^n + \gamma| \quad (3.14)$$

The only difference from the trend feature is that we do not need to validate it using PCC, because the shape of the irregular feature will depend more on its trend and seasonal features.

3.5.3 Zero Pattern Adjustment

In human occupancy prediction research, inferring knowledge about when a room is vacant is paramount. By minimising false positives, the accuracy prediction can be improved. The zero pattern adjustment (ZPA) method learns behaviour from historical data and makes smart

adjustments for a vacant room when the normal algorithm returns an incorrect prediction. The ZPA technique overlays all previous datasets and places them on a single 24-hour x-axis chart, to determine the earliest start and end points when the room is vacant each day during the night to dawn period. We symbolise ZPA as zpa_t^O .

For our main occupancy model, we integrate each feature to get the occupancy prediction value.

3.6 Experiments and Results

In this section, our model is assessed for two different locations with distinct contexts to ensure the model's adaptability to various conditions. The first location, small room A, belongs to a staff member at RMIT University, Australia. This room is chosen for human occupancy prediction, since a controlled experiment can be conducted for an extended period of data collection.

The second dataset was collected inside a cinema theatre in Mainz, Germany [Wicker et al., 2015]. The cinema theatre was chosen because it has fluctuating numbers of people throughout the day. The number of people in the audiences can reach hundreds and can decrease to zero within a few hours. We refer to this room as big room B.

3.6.1 Experiment Setting

3.6.1.1 Small Room A

We used a commercial off-the-shelf Netatmo urban weather station (range: 0 – 5000 ppm; accuracy: ± 50 ppm) to read and collect ambient CO₂ data. The experiment took place between May and June 2015. The dataset were uploaded to a cloud service for integration purposes. We selected two weeks data from the whole dataset and used them in the further analysis. The room size is 3 x 4 m.

3.6.1.2 Big Room B

The cinema dataset was collected between December 2013 and January 2014 [Wicker et al., 2015], using mass spectrometry machinery installed on the air ventilation system. The air flows from the screening room and through the ventilation system to the mass spectrometer for data analysis.

3.6.1.3 Experimental Tool

We utilised Waikato environment for knowledge analysis (WEKA), matrix laboratory (MATLAB) and R to help us perform this experiment. WEKA is used for polynomial linear regression, using the M5 method for both correlation models for the trend (subsection 3.5.2.1) and irregular features (subsection 3.5.2.3). We also used WEKA for the majority of the data mining algorithms, such as multi-layer perceptron, Gaussian processes (with kernel RBF), support vector machine, random forest, naïve Bayes, decision tree (with random tree) and decision tree (with M5P). MATLAB code was run for the baseline method, support vector regression (SVR) and its prediction result. We used R to integrate all the data, including decomposition of STD and the majority of data preprocessing.

3.6.2 Experimental Parameters

The SD-HOC model predicts each future value for the whole period of time, based on a specific time window. To better understand this model and how well it performs compared with the baseline, we define x , the accuracy error tolerance parameter. Zero units of error tolerance means that only the exact number recognised is considered as a true positive. For example, with ten units of error tolerance, if the real indoor human occupancy is 150 people, predictions as low as 140 or as high as 160 are considered correct, as they are within ± 10 units of error tolerance. The value of parameter x will be different based on the size of the room.

Each machine learning algorithm's data was preprocessed using the same method as in section 3.4 to ensure that the comparison is valid.

3.6.2.1 Experiment for Small Room A Dataset

For the academic staff room dataset, we used a five-minute time window. We gathered 4,019 data from this room over 14 days. Due to the small room size, we decided not to use time lag for data analysis, as there is a negligible period between exhalation and sensor reading. For this room, we have seven pairs of the training-test dataset. It starts with seven days of the training dataset and seven days of the test dataset, and ends with 13 days of the training dataset to predict a one-day test dataset.

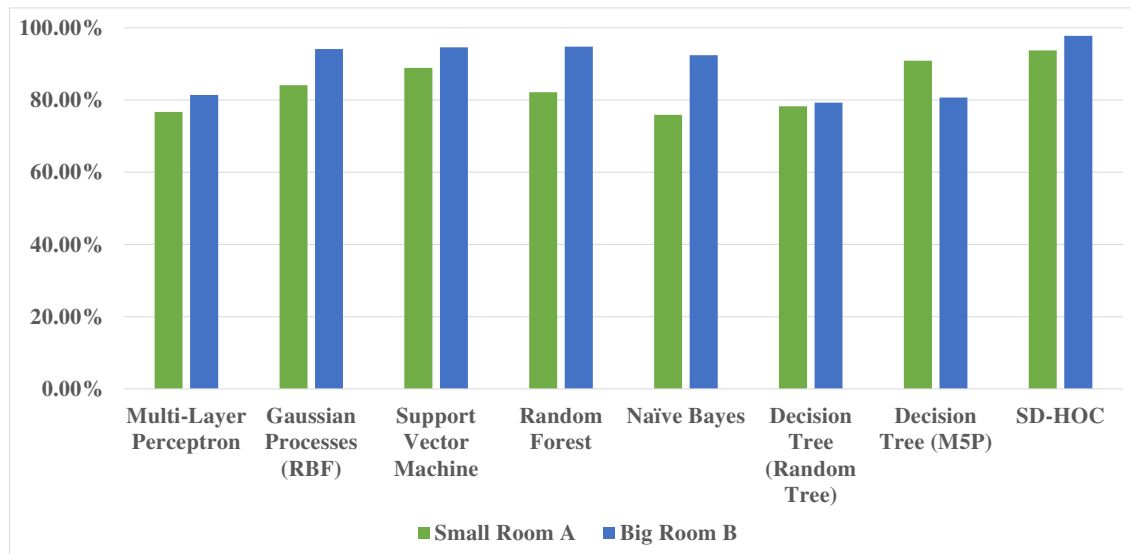


Figure 3.5: Accuracy Results of Various Machine Learning Algorithms.

3.6.2.2 Experiment for Big Room B Dataset

For the cinema theatre dataset, we use a three-minute time window for data analysis. We gathered 68,640 data from this cinema theatre over 23 days. The theatre capacity is 300 people and, for this experiment, we ran the line of best fit for time lag 0 to time lag 60. The lowest NRMSE was at time lag 32, therefore we used time lag 32 as the time lag baseline. This time lag is appropriate as the bigger room needs a longer time lag for the model to have a better accuracy. For this room, we used December 2013 data for training and January 2014 data for testing. We then replicated it using a similar method, by giving one day from the testing dataset to the training dataset and running the model again. This method was repeated until the test dataset consisted of only one day of data.

3.6.3 Experimental Results with Other Data Mining Algorithms

From Table 3.1 and Figure 3.5, we ran each data mining algorithm and compared the result with our novel model, SD-HOC. SD-HOC had the highest prediction accuracy, with 93.71% accuracy for the staff room and 97.73% for the cinema theatre.

Table 3.1: Accuracy results of various machine learning algorithms.

Machine Learning	Small Room A	Big Room B
Multi-Layer Perceptron	76.69%	81.39%
Gaussian Processes (RBF)	84.09%	94.09%
Support Vector Machine	88.86%	94.55%
Random Forest	82.16%	94.75%
Naïve Bayes	75.89%	92.40%
Decision Tree (Random Tree)	78.23%	79.26%
Decision Tree (M5P)	90.87%	80.68%
SD-HOC	93.71%	97.73%

3.6.4 Experimental Results with Support Vector Regression (SVR) on Different Number of Prediction Days

SVR had the highest prediction accuracy of the data mining algorithms. For this reason, we ran the experiment with different training and testing, to compare SD-HOC and the state-of-the-art machine learning algorithm baseline, SVR.

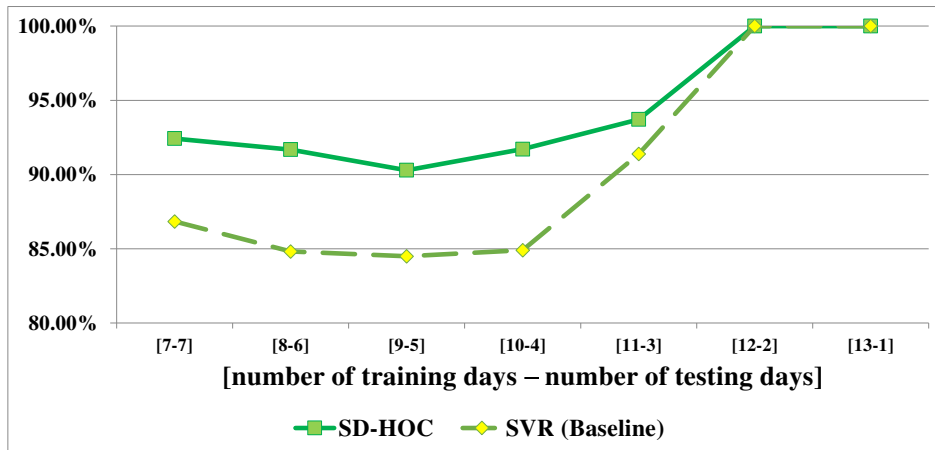


Figure 3.6: Small Room A Dataset - Comparison for Indoor Human Occupancy.

3.6.4.1 Evaluation and Baseline

To evaluate the result, we divided the data into two equal parts: the training dataset and the test dataset. To understand how well the model fits for a longer duration, we repeated the division of training and test datasets by adding one day of data from the test dataset to the

training dataset. This replication was repeated until the test dataset had only one day and the rest belonged to the training dataset. This evaluation method of incremental change to the days of training and reduction in days of testing ensured the robustness of the model.

3.6.4.2 Experimental Results for Small Room A Dataset

From Figure 3.6, SD-HOC performed better than the baseline by an average of 4.33%. As the last two days are Saturday and Sunday, both SD-HOC and the baseline models correctly predicted zero occupancy for those days.

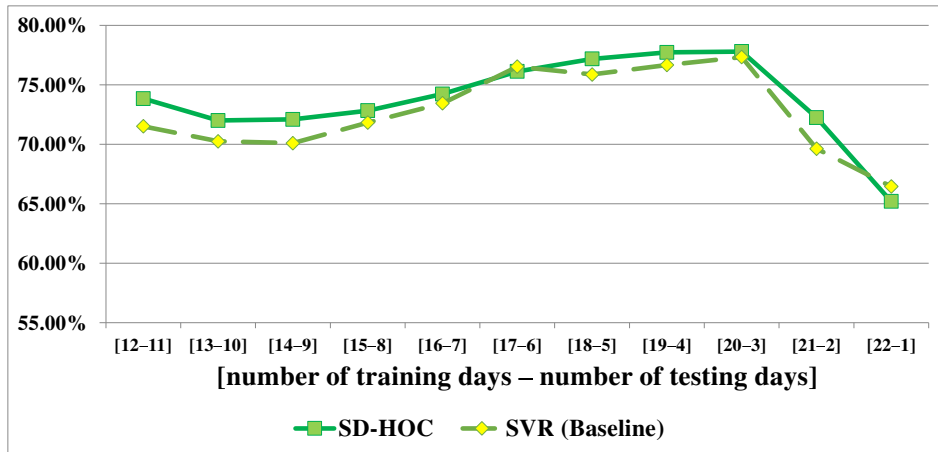


Figure 3.7: Big Room B Dataset - Comparison for Indoor Human Occupancy.

3.6.4.3 Experimental Results for Big Room B Dataset

For the cinema dataset, the accuracy comparison is shown in Figure 3.7. The SD-HOC method performed better than the baseline method; on average, SD-HOC method had 8.5% higher accuracy in predicting indoor human occupancy. The highest prediction accuracy was found when we used 22 days of data for training to predict the number of human occupants the next day.

The results from Figure 3.7 show that the SD-HOC method is more accurate in predicting indoor human occupancy. This result is encouraging. Furthermore, we can observe that the accuracy is higher for fewer days of prediction (bigger number of days of training data) than for more days of prediction (smaller number of days of training data), which is aligned with the results from the academic staff room experiment.

3.7 Discussion

Our experiment shows that our new framework has higher accuracy than most data mining algorithms for both small and large rooms, as shown in Figure 3.5. Furthermore, this SD-HOC model is robust enough to handle different scales of data, which was proven by evaluating our proposed model in two environments with different contexts (room size and maximum number of occupants).

Compared with the baseline (SVR), our framework showed a better prediction accuracy over differing numbers of days, for both the training and testing periods. This result demonstrates that seasonal decomposition can be utilised for predicting indoor human occupancy. The SD-HOC model can be used in many applications and is not limited to human occupancy prediction as it is based on seasonal decomposition methods.

SD-HOC performs well in comparison to other machine learning algorithms, due to the feature transformation step, where, for each transformed feature, a set of relevant algorithms is run. For the small room A, SVM and the decision tree method were the next best in prediction accuracy after SD-HOC. This is due to the fact that the number of people in this room fluctuated less from hour to hour, and there was a stable CO₂ concentration for an extended period.

For the big room B, SVR and random forest were the next best in prediction accuracy after SD-HOC. Random forest performs well when irrelevant features are present or when these features have skewed distributions. The number of people in the big room could fluctuate from zero to hundreds within 10 – 15 minutes. The SVM technique enables accurate discrete categorical labels to be predicted. This is why SVR was chosen as a baseline in section 3.6.4.

3.8 Conclusion

The roles of data mining algorithms in human life are becoming more important and the technology is being assimilated into human daily life. SD-HOC utilises several data mining algorithms and contributes to building and room occupancy counting. By understanding and knowing the numbers of people within a building, the heating, cooling, lighting, building energy consumption, emergency evacuation, security monitoring, and room utilisation can be made more efficient.

Although research in the area of human occupancy has been undertaken using various methods, including the use of ambient sensors, the occupancy models studied in previous work

required the use of many sensors. In the experiment in this chapter, we used a single sensor that is commonly available in BMSs to reduce the cost and complexity, as more sensors can mean less reliability.

There are many possibilities that can be explored by using this technique. SD-HOC can be used for any time series dataset to predict another time series dataset, as long as there is some dependency between those two datasets. SD-HOC is more than a simple correlation model, as it can solve many problems that a simple correlation model will not be able to solve. The further evolution of SD-HOC is discussed in Chapter 4.

Chapter 4

RUP: Large Room Utilisation Prediction with Carbon Dioxide Sensor

4.1 Introduction

Data obtained from the Australian Department of the Environment and Energy indicate that approximately 40% of total maintenance costs within a building are for heating, ventilation, and air conditioning (HVAC) [DOEE., 2013]. Hence, there is substantial investment in energy usage research to automate HVAC control in buildings based on occupancy patterns. Reducing HVAC usage will massively lessen overall energy consumption. However, this may compromise the comfort of the dwellers. A building management system (BMS) can intelligently adjust the HVAC based on the occupancy pattern.

The majority of buildings, especially older ones, do not have adequate infrastructure to accurately sense people and where they are within a building. Hence it is challenging to determine the precise ground truth value for analysis purposes. Some researchers use simulation models [Goldstein et al., 2010, Page et al., 2008] to reduce energy consumption. Methods of simulating the occupants' behaviours with the aim of reducing energy consumption were proposed in [Richardson et al., 2008, Saelens et al., 2011]. Unfortunately, simulations with agent-based models do not reflect the behaviour and uncertainty from the actual environment and are not adaptable to different types of rooms and buildings. Therefore, a better approach is to analyse data collected from the real world.

Using sensor data for indoor human occupancy detection is the current trend in ambient sensing research [Candanedo and Feldheim, 2016, Ekwevugbe et al., 2013a, Hailemariam et al., 2011, Khan et al., 2014, Mamidi et al., 2012, Yang et al., 2012]. In Chapter 2, it was highlighted that carbon dioxide (CO₂) is the best ambient predictor for detecting human presence. By using only CO₂, 91% accuracy was achieved for binary prediction of whether the room is occupied or vacant [Basu et al., 2015] and 15% accuracy was achieved for recognising the number of occupants. A hidden Markov model (HMM) was implemented for a CO₂ dataset to predict human occupancy, and 65-80% accuracy was achieved for predicting up to four occupants [Lam et al., 2009].

In this chapter, we propose a large Room Utilisation Prediction with CO₂ sensor (RUP), a new method to count indoor human occupancy based on the amount of CO₂ in the air. There are already reasons to monitor CO₂ concentration; for example, the Green Building Council of Australia (GBCA) gives buildings a score of 1 or 2 green points if CO₂ levels are maintained below 800 ppm or 700 ppm respectively ¹. One sample of a CO₂ data regression chart is shown in Figure 4.1.

The main reason we used CO₂-based occupancy counting is that CO₂ sensors are an integral part of infrastructure for demand-controlled ventilation [Basu et al., 2015]. There is a need to monitor CO₂ concentration in public indoor places in real time. A higher level of CO₂ concentration can have adverse effects on occupants' health [Zhang et al., 2017]. In addition, CO₂-based occupancy counting is not prone to accumulated counting errors, unlike sensors that detect transition between vacant and occupied states [Basu et al., 2015]. With common sensors such as temperature and CO₂, there is usually no need to equip a building with any additional occupancy counting sensors [Sangogboye et al., 2017].

Because people exhale CO₂ while they breathe, there is a correlation between CO₂ and occupancy. Machine learning model experiments were performed on two different sized rooms with a capacity of up to 300 occupants. We believe that if the algorithm works well with low and high occupancy, then this algorithm is robust. Three advantages of this method are:

- (a) RUP ensures that users' privacy is protected
- (b) Equipment cost is low due to pre-installation
- (c) Only CO₂ data are used, reducing the chance of errors caused by data integration

¹https://www.gbca.org.au/uploads/147/35475/Quality%20of%20Indoor_Draft_D1_distributed.pdf

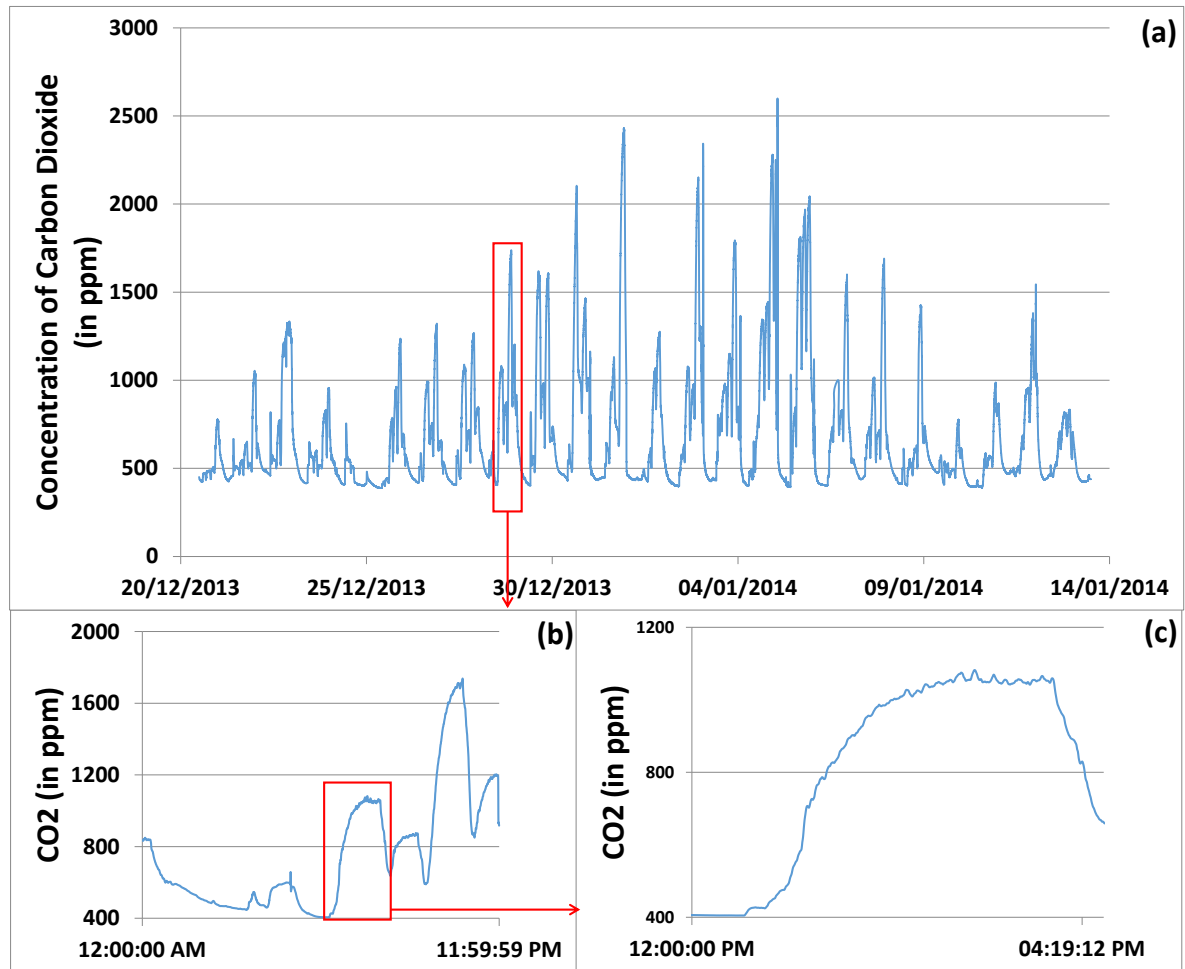


Figure 4.1: CO₂ concentration in the movie theatre over time. (a) shows an overview of all measurements; days, times and movies screening can be easily identified. (b) shows the concentration on December 28; movie screenings can be seen and the shape of each screening can be identified differently. (c) shows the CO₂ concentration during the movie “The Hunger Games 2: Catching Fire” on December 28, 2013 at 13:15.

4.1.1 Research Motivation

There are four primary motivations for this research. The first is to help space and room utilisation. By knowing the number of people in each room at a given time, the building manager can monitor which rooms are under-utilised and which rooms are over-utilised. For example, after integrating the RUP algorithm with the BMS, the building manager can observe the indoor human occupancy for each room and detect which rooms have low utilisation. With

this knowledge, he/she can adjust room allocation and utilisation accordingly, and can combine two meeting rooms if their usage numbers are both small or convert one of them into a storage chamber.

The second motivation is to support the BMS so that it can reduce power consumption when there is nobody in the room. Knowing the number of people at a given time in each room becomes crucial to achieving this motivation. Further development can be done to automate the BMS to adjust the HVAC higher or lower based on the number of occupants, or to automatically turn it off or on to low power mode when vacant.

The third motivation is for security purposes. With RUP and its human occupancy counting mechanism, an alarm can be set if the counter returns with value 1 or above in a room that should be vacant. There might be a burglar or uninvited guest inside the particular room. This functionality can be extended to monitor a location that should not be occupied for a very long time, such as a restroom. If this happens, there might have been an accident or the person inside might be breathing but unconscious.

The fourth and last motivation is for personal indoor comfort. By integrating the RUP algorithm into the BMS, the temperature can be adjusted based on the number of occupants in each room as determined by RUP. A room with many people inside may need a slightly lower temperature to compensate for the higher level of CO₂, and vice versa.

4.1.2 Research Contribution

The main contributions of this chapter are as follows:

- We propose a novel feature engineering method to process both CO₂ and human occupancy data into four main features: trend, seasonal, irregular and zero pattern adjustment.
- We develop a time-series model to predict the number of humans occupying a closed space. The accuracy of the proposed method is higher than the current state-of-the-art machine learning algorithm method, known as support vector regression (SVR) and PerCCS [Basu et al., 2015] technique, non-negative matrix factorisation with ensemble least square regression (NMF-ELSR).
- We compare our proposed two customised seasonal decomposition methods, to solve the human occupancy prediction problem and highlight the advantages of each method.

- We generalise our method to be implemented for any prediction problem Y_n if X_n is known and Y_n is dependent on X_n , where n is the number of sample points.

4.2 Background and Related Work

Occupancy detection has been explored extensively in the last decade and one of the biggest challenges is to achieve this without using image processing from cameras. When using image processing techniques [Erickson et al., 2009, Lee et al., 2011, Barandiaran et al., 2008], the levels of accuracy for human occupancy detection can reach 95%. Unfortunately, using cameras raises privacy concerns as people do not want to be identified. Research communities have been proposing various methods to detect human occupancy without using cameras or image processing. Occupancy prediction using depth sensors [Munir et al., 2017, Liu et al., 2017] can result in very accurate prediction, up to 99%. However, this chapter does not consider any research with vision-based sensing devices, because we are researching alternatives to these systems. Vision-based sensing devices require a clear line of sight and occlusion is always an issue [Teizer, 2015]. Given the need for clear line of sight, vision-based systems will not work well in counting crowds or a large number of people.

In this section, we divide human occupancy research into four subsections, based on their method: simulation in section 4.2.2, radio-based in section 4.2.3, sensor-based in section 4.2.4, and CO₂ sensor in section 4.2.5. In section 4.2.6, we present a summary of the current state of human occupancy research. Due to the extensive range of methods of machine learning algorithms that have been used for human occupancy recognition, a machine learning term list is presented in Table C.1 in the Appendix.

4.2.1 Background Study of Human Occupancy Calculation

This subsection discusses human occupancy calculation with flow rate of CO₂, using metabolic rate. A person exhales CO₂ as the natural process of breathing, but the amount varies from person to person. Equation 1 gives the exhalation rate equation of CO₂, derived from the metabolic rates formula [Arora, 2010].

We note a non-linear relationship between the flow rate of CO₂ and the average value of a person's height (H) and weight (W). This non-linear relationship needs to be considered if we wish to correlate human features (height and weight) with CO₂ data only. Our approach to addressing this non-linear gap will be explained further in section 4.4.3.

$$V_{\text{CO}_2} = \frac{M \cdot RQ \sqrt{H \cdot W}}{21132 \cdot (0.23RQ + 0.77)} \quad (4.1)$$

V_{CO_2} flow rate of CO_2
 M metabolic rate (in W/m^2)
 RQ respiratory quotient (CO_2 eliminated / O_2 consumed)
 H height (in cm)
 W weight (in kg)

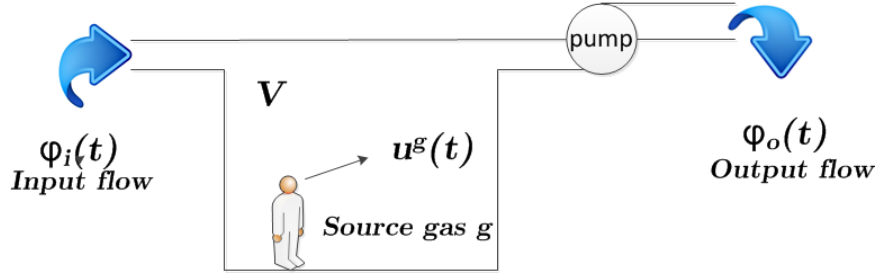


Figure 4.2: Simplified presentation of gas exchange in the respiratory chamber where $\varphi_i(t)$ is the flow rate of input air at time t , $\varphi_o(t)$ is the flow rate of output air at time t , V is the volume at time t , and $u^g(t)$ is the gas production rate at time t .

The size of the room is the next aspect that we need to consider to ensure our data are useful for indoor human occupancy prediction. Figure 4.2 explains the flow rate of inhalation [Granato et al., 2004]. For any given room that has standard ventilation, there will be an input flow $\varphi_i(t)$ and pumping out by a specific output flow $\varphi_o(t)$. For a specific room size V at the time t , the gas production rate is $u^g(t)$. From this finding, the size of the room is crucial for understanding the gas production rate.

From Equation 4.1 and Figure 4.2, we can see that many parameters are required to accurately measure indoor human occupancy by using the metabolic rates formula and flow rate of CO_2 . Parameters include the occupant's weight, height and rate of breath and also the size of the room. The need for occupant's weight and height can lead to privacy issues. In addition, it is very difficult to obtain accurate readings of room size.

4.2.2 Simulation-Based Indoor Human Occupancy Detection

The common method for detecting human occupancy is to construct a simulation model for a room or building [Goldstein et al., 2010, Page et al., 2008, Richardson et al., 2008, Saelens et al., 2011]. Simulation models will not be accurate for multiple buildings with different features

and characteristics, as there are many variables, including the ones mentioned in subsection 4.2.1, that need to be considered for indoor human occupancy.

Table 4.1: Models, parameters and reported accuracies for radio-based occupancy detection research.

Ref.	Occupancy Detection Algorithms	Detailed Devices	Max. # of people	Duration	Location	Accuracy (Occupancy)
[Depatla et al., 2015]	Multipath fading (MP) with line of sight, Kullback-Leibler (KL) divergence	Wi-Fi transmitter and receiver	9	not stated	indoor and outdoor	96% for outdoor and, 63% for indoor
[Dutta et al., 2006]	Doppler equation, Neyman-Person detector (noise detector), Exponentially-weighted moving average (EWMA)	Ultrawideband radar	binary occupancy	not stated	outdoor	not stated
[He and Arora, 2014]	Support Vector Regression (SVR)	A low-power pulser radar	40	750 minutes	four rooms	not stated
[Leephakpreeda, 2005]	Pareto distribution of occupancy's inactivity	Microwave motion sensor	not stated	10 days	staff office room	not stated
[Li et al., 2015]	Neural Network (NN), SVM and Sequential Counting (SC)	Wi-Fi access point, Samsung Galaxy S3	50	5 hours	two classrooms	93%
[Sigg et al., 2014]	K-nearest neighbours (KNN)	Radio frequency (wireless signal)	3	not stated	corridor	not stated

4.2.3 Radio-Based Indoor Human Occupancy Detection

Radio-based devices include but are not limited to Wi-Fi, bluetooth, any electromagnetic waves, and gamma rays. A summary of the radio-based research is in Table 4.1.

A model using only Wi-Fi power based on its RSSI is used for occupancy detection [Depatla et al., 2015]. By using the line of sight analysis, outdoor human occupancy accuracy can be up to 96%.

A low-power pulsed radar was utilised for people counting in [He and Arora, 2014]. The counting was modelled with support vector regression (SVR) for up to 40 occupants. A correlation coefficient of 0.97 was achieved with 2.17 mean absolute error between the estimated count and the ground truth.

Using wireless signals, device-free occupancy recognition was achieved by [Sigg et al., 2014]. They managed to distinguish the state of the wireless signal between an empty environment and an occupied one. Furthermore, this study successfully detected the occupants' activities, such as walking, lying, crawling or standing.

In [Leephakpreeda, 2005], a microwave motion sensor is used to detect motion and then the lighting is controlled with time delays to reduce electricity consumption. [Dutta et al., 2006] fused ultra-wideband radar that can function beyond the line of sight to detect people. However, such equipment is expensive and rarely justified for the purpose of occupancy detection alone.

4.2.4 Indoor Human Occupancy Detection with Sensors

Indoor human occupancy detection with sensors refers to research using any type of sensors for counting people. A summary of related sensor-based work can be found in Table 4.2.

Table 4.2: Models, parameters and reported accuracies for sensor-based occupancy detection research.

Ref.	Occupancy Detection Algorithms	Detailed Devices	Max. # of people	Duration	Location	Accuracy (Occupancy)
Chapter 2	MLP, GP with RBF, SVM, RF and NB	Temperature, humidity, CO ₂ , sound, pressure, and illumination (light) sensors	binary occupancy	2 weeks	1 single person office room	96% – 99%
[Agarwal et al., 2010]	Using hardware (CC2530 micro controller) built-in capability	Reed switches and PIR sensors	binary occupancy	2 weeks	10 offices	not stated
[Basu et al., 2015]	Non-negative matrix factorisation (NMF), ensemble least square regression (ELSR) and SVR	CO ₂ sensors	15 (lab) 42 (classroom)	13 days	1 lab and 1 classroom	91% for vacant prediction and 15% for occupied prediction
[Beltran et al., 2013]	KNN, linear regression (LR) and artificial neural networks (ANN)	PIR and thermal array sensors	not stated	3 weeks	a 17-node deployment covering 10 building areas	not stated
[Candanedo and Feldheim, 2016]	Random forest (RF), gradient boosting machines (GBM), linear discriminant (LD) analysis and classification and regression trees (CART)	Raspberry Pi with light, CO ₂ , DHT22 (temp/humid) sensors, Zigbee radio and digital video camera	2	not stated	1 room	95% – 99% with all predictors and 83% – 85% using temperature only
[Castanedo et al., 2011]	Latent dirichlet allocation (LDA)	PIR sensors	not stated	24 weeks	3 floor Inmotek building	not stated
[Dodier et al., 2006]	Bayesian probability theory and Belief Network (BN)	3 PIR sensors and telephone sensor	binary occupancy	2 days	2 offices	1 PIR: 20.1% 2 PIR: 97.8% 3 PIR: 99.9%
[Ekwevugbe et al., 2013a]	ANN with MATLAB and WEKA	CO ₂ , sound, relative humidity, temperature (air and computer), and PIR sensors	39	7 days	open-plan office	84.59%
[Ekwevugbe et al., 2013b]	ANN with MATLAB and WEKA	sound, case temperature, humidity, light, CO ₂ , motion, volatile organic compounds (VOCs) and PIR sensors	6	30 days	open-plan office	75%
[Gao and Keshav, 2013]	KNN and learning-based model predictive control (LBMPC)	Microsoft Kinect, infrared sensor and laser pointer	binary occupancy	more than 3 months	office room	not stated
[Garg and Bansal, 2000]	Motion detection	PIR sensors	binary occupancy	not stated	office room	not stated
[Hailemariam et al., 2011]	Decision trees (DT)	CO ₂ , computer current, light, PIR and sound sensors	1	7 days	a single cubicle	CO ₂ : 94.68% current: 96.27% light: 81.02% motion: 98.44% sound: 90.79%
[Howard and Hoff, 2013]	Modified Bayesian combined forecasting approach with seasonal ARIMA model, historic average, time delay NN and SVR	50 PIR sensors	not stated	2 years	2 large buildings	not stated

Table 4.2: Models, parameters and reported accuracies for sensor-based occupancy detection research. (cont'd)

Ref.	Occupancy Detection Algorithms	Detailed Devices	Max. # of people	Duration	Location	Accuracy (Occupancy)
[Jazizadeh and Becerik-Gerber, 2012]	not stated	64 wired BMS sensors, 50 moveable sensor boxes and several cameras	8	2 weeks	7 rooms	not stated
[Khan et al., 2014]	KNN and SVM	motion PIR, acoustic noise, temperature, light and humidity sensors	20	14 days for low traffic area and 10 days for high traffic area	large commercial buildings	not stated
[Khan et al., 2015]	RF	acoustic (microphone), locomotive (accelerometer) and location (magnetometer) sensors	8	not stated	not stated	not stated
[Lam et al., 2009]	Hidden Markov models (HMM), NN and support vector machines (SVM) latent	CO ₂ for both inside and outside room	4	58 days	open plan office with 16 rooms and 1 conference room	65% – 80%
[Liao and Barooah, 2010]	Classical linear minimum variance (LMV) estimator	binary motion sensors	1	not stated	1 room	not stated
[Mamidi et al., 2012]	Rule-based heuristic, Multi-layer perceptron (MLP), Gaussian processes (GP), LR, v-SVM-R and ensemble voting (EV)	BLEMS sensors	16	several weeks	2 shared lab spaces	46% – 95%
[Mohammadmoradi et al., 2017]	Otsu's thresholding and modelling thermal noise distribution	GirdEYE IR array sensor and a Raspberry Pi	3	fixed scenario experiment	university lab, classrooms, computer labs and conference rooms	93%
[Sangogboye et al., 2017]	Random forest, decision tree	temperature, CO ₂ and PIR	36 (study zone) 85 (classroom) 10 (conference room)	1 month	1 study zone and 1 classroom	not stated
[Shih and Rowe, 2015]	Principal component analysis (PCA), density-based spatial clustering of applications with noise (DBSCAN) and Sabine acoustic model	Transducer with microphone and tweeter	24(classroom) 150 (auditorium)	quick scenario	a conference room, 1 classroom and 1 auditorium	90%
[Srinivasan et al., 2010]	Density-based spatial clustering of applications with noise (DB-SCAN) and maximum likelihood estimate (MLE)	ultrasonic distance sensor for height on the doorway, motion sensors and magnetic reed switch sensors	20 (lab) 4 (home residents)	5 days	1 lab and 3 homes	95% (identification accuracy)
[Wang et al., 2005]	non-homogeneous Poisson model with two different exponential distributions	Infrared sensor behind a Fresnel lens	not stated	1 year	35 single person offices	not stated
[Yang et al., 2017]	Regularised regression	LED	20 volunteers	6 months	5×6m lab	≥ 90%
[Yang et al., 2012]	SVM, ANN with radial basis function (RBF) and HMM	light, sound, motion, CO ₂ , temperature, relative humidity and PIR sensors	9	20 days	2 shared lab spaces	87.62% for self estimation and 64.83% for cross-estimation
[Yang et al., 2014]	ANN, DT, KNN, naïve Bayesian (NB), tree augmented, naïve Bayes network (TAN) and SVM	light, sound, PIR, CO ₂ , Reed door sensor, relative humidity and temperature sensors	3 (single occupancy rooms) 9 (multi occupancy rooms)	1 month (single occupancy rooms) 20 days (multi occupancy rooms)	2 single-occupancy rooms and 2 multi-occupancy rooms	ANN: 92.5% – 97.1% DT: 96.0% – 98.2% KNN: 95.4% – 97.5% NB: 88.9% – 94.3% TAN: 95.3% – 98.0% SVM: 95.1% – 97.5%

Algorithms to detect indoor human occupancy for the purpose of reducing energy cost are proposed in [Agarwal et al., 2010, Gao and Keshav, 2013, Garg and Bansal, 2000, Lu et al., 2010, Beltran et al., 2013]. Energy saving results vary from 5% to 60%.

The most common device that is used for counting people is a passive infra-red (PIR) sensor, which is used as a motion sensor with a reed switch as the door sensor [Agarwal et al., 2010, Castanedo et al., 2011, Dodier et al., 2006, Gao and Keshav, 2013, Garg and Bansal, 2000, Howard and Hoff, 2013, Liao and Barooah, 2010, Lu et al., 2010, Wang et al., 2005]. In [Dodier et al., 2006], single, double and triple PIR sensors were compared to detect the presence of people, and the accuracy was 20.1%, 97.8% and 99.9% respectively.

Using sensor network data, 50 PIR sensors were deployed and the data were utilised to build a modified Bayesian forecasting method [Howard and Hoff, 2013]. They compared the accuracy with several other machine learning methods, such as seasonal autoregressive integrated moving average (ARIMA), neural networks (NN) and SVR, and found that their techniques have the lowest error. Their accuracy results were not presented.

ThermoSense, a system for estimating occupancy by using a thermal array sensor combined with passive infrared sensor was presented in [Beltran et al., 2013]. Using k-nearest neighbour (KNN), linear regression (LR) and artificial neural networks (ANN), ThermoSense could predict the room occupancy and reduce energy use by up to 25%.

In [Kleiminger et al., 2013], electricity consumption data from electricity meters was used as the feature in a occupancy analysis. The article presents four models of the regression in SVM, KNN, thresholding (THR) and HMM.

The Building Level Energy Management Systems (BLEMS) project from the University of Southern California uses a combination of sensors (light, sound, motion, CO₂, temperature and humidity sensor) to create a model to estimate human occupancy. A radial basis function (RBF) method showed 87.62% for self-estimation, and when the model was implemented in another room, the cross-examination result showed 64.83% occupancy accuracy [Yang et al., 2012]. Various machine learning algorithms such as multi-layer perceptron (MLP), Gaussian processes (GP), LR, SVM and ensemble voting (EV) were implemented, resulting in occupancy accuracy ranging from 46% to 95% [Mamidi et al., 2012].

Utilising density-based spatial clustering of applications with noise (DB-SCAN), [Srinivasan et al., 2010] produced 95% accuracy with ultrasonic distance sensors for height sensors, motion sensors and magnetic reed switch sensors. [Candanedo and Feldheim, 2016], using temperature sensors only, achieved between 83% and 85% accuracy in predicting two persons in a single room. Their accuracy was up to 99% when all predictors were used, including light, CO₂ and

humidity sensors.

An estimation algorithm based on unsupervised clustering of both overlapped and non-overlapped conversational data, with a change point detection algorithm for locomotive motion of the users to infer occupancy, was proposed as a mobile app by [Khan et al., 2015]. Users installed the app and provided consent for the app to access the smartphone sensor's data. Using the random forest (RF) algorithm, they applied occupancy detection, and the accuracy was 76% for counting a maximum of eight people.

A framework was created to produce occupancy estimates at different levels of granularity and provide confidence measures for effective building management in [Khan et al., 2014]. By using KNN and SVM, their accuracy varied from 64.6% to 94.7%.

Real-time occupancy detection by using decision trees (DT) with multiple types of sensors, such as light, sound, CO₂, motion and computer power, was conducted by [Hailemariam et al., 2011]. The lowest accuracy was obtained from sound sensors (90.79%) and the highest was from motion sensors (98.44%).

A model for real-time estimation of building occupancy sensing was presented by [Ekwevugbe et al., 2013a,b]. Both papers utilised ANN and ran in the Waikato Environment for Knowledge Analysis (WEKA) and MATLAB for an open plan office occupancy detection. The accuracy was between 75.00% and 84.59%. With ambient sensors such as temperature, relative humidity, sound, light and CO₂, they used volatile organic compounds (VOCs) data as one of their features.

A systematic approach to occupancy modelling by using various ambient sensor data for both single occupancy and multi-occupancy room was proposed by [Yang et al., 2014]. With ANN, DT, KNN, naïve Bayes (NB), tree augmented naïve Bayes network (TAN) and SVM, human occupancy was predicted with up to 98.2% accuracy in DT.

By utilising a smart phone sensor, such as the microphone, bluetooth or Wi-Fi, one study [Guo et al., 2016] obtained a precision and recall on group-aware recognition of over 80% with 45 participants using mobile crowd sensing.

A model that utilises temperature, humidity, CO₂, sound, pressure and illumination sensors was used in Chapter 2 to detect whether a room is occupied or vacant. The occupancy detection was improved by using feature engineering and various machine learning algorithms, such MLP, GP with RBF, SVM, RF and NB. The human occupancy was detected above 95% with multiple machine learning algorithms such as MLP, GP-RBF and RF.

Table 4.3: Algorithms, devices and reported accuracies for CO₂ sensor-based occupancy detection research.

Ref.	Occupancy Detection Algorithms	Detailed Devices	Max. # of People	Accuracy (Occupancy)
[Basu et al., 2015]	Non-negative matrix factorisation (NMF) ensemble least square regression support vector regression (SVR)	CO ₂ sensors (BACNet server)	15 (lab) 42 (classroom)	Binary prediction: 91% Counting prediction: 15%
[Cali et al., 2015]	Mass balance equation	CO ₂ sensors	12	Binary prediction: 95.8% Counting prediction: 80.6%
[Dedesko et al., 2015]	Time-lagged mass balance approach	PP Systems SBA-5 CO ₂ gas analysers	3	not stated
[Hailemariam et al., 2011]	Decision trees (DT) Hidden Markov models	CO ₂ sensors	1	94.68%
[Lam et al., 2009]	Neural networks Support vector machines (SVM) latent	Gas detection CO ₂ sensor network	4	65% – 80%

4.2.5 CO₂-Based Indoor Human Occupancy Detection

Since CO₂ sensors are already integrated within Australian BMS and ventilation infrastructure, we focus on utilising only CO₂ sensor data to estimate indoor human occupancy. CO₂ sensors typically cost more than PIR sensors. By taking advantage of BMS sensors, operational cost can be reduced by not purchasing or installing extra PIR or motion sensors. Table 4.3 presents a summary of related work on indoor occupancy detection using CO₂ sensors.

Machine learning algorithms, including HMM, NN and support vector machine (SVM) latent were used in [Lam et al., 2009], using CO₂ data with sensors deployed both inside and outside a room. By feature engineering CO₂ data with the first order and second order difference of CO₂, the accuracy achieved was between 65% and 80% for binary occupancy prediction.

A method to assess human occupancy and estimate occupant activity in ten hospital rooms was conducted by [Dedesko et al., 2015] as part of the Hospital Microbiome Project. Using a time-lagged mass balance approach, Dedesko et al. aimed to determine occupant characteristics and understand the interactions between humans and microbial communities. The accuracy of results was not provided in the paper.

CO₂-based occupancy detection in office and residential buildings was conducted by [Cali et al., 2015]. Testing and validation was done for both residential and non-residential buildings. A mass balance equation was used, and the vacant/occupied prediction accuracy was 95.8% and the human counting prediction accuracy was 80.6%.

in another study, a heterogeneous sensor array was deployed in an office workspace for the purpose of providing a real-time occupancy detector [Hailemariam et al., 2011]. Decision trees was used to perform the classification and to explore the relationship between different types of sensors, features derived from sensor data, and occupancy. The prediction accuracy for human occupancy using a CO₂ sensor was 94.68%, which was lower than both a motion sensor

(98.44%) and a current sensor (96.27%), but higher than a light sensor (81.02%) and a sound sensor (90.79%).

PerCCS is a model with a non-negative matrix factorisation method for counting people [Basu et al., 2015], with CO₂ as the only predictor. In predicting vacant/occupied, Basu et al. achieved up to 91% accuracy, but only 15% accuracy in predicting the number of occupants.

4.2.6 Summary of Related Work

After reviewing all the literature in occupancy detection, a few key points are derived as follows:

- PIR is the most commonly used research sensor in the literature for indoor human occupancy. PIR sensors are cheap, but not scalable; i.e. new PIR sensors must be bought for every new room. This means that, to be able to predict occupancy for 100 rooms, at least 100 PIR sensors must be provided.
- Not every research paper gives their accuracy. This makes direct comparison between each model difficult.
- Poor sensor calibration and lower frequency reduce the accuracy of prediction results.
- CO₂ is one of the best types of sensors to measure indoor human occupancy, but the majority of papers on indoor human occupancy using CO₂ provide binary occupancy analysis, not real people-counting analysis.
- Using too many types of sensors as features can result in lower accuracy.
- There is no single formula available to calculate the number of occupants. In subsection 4.2.1, we conclude that collecting all the parameters related to indoor human occupancy is not practical and is subject to privacy breaches.
- Datasets from most previous experiments are not publicly available.
- Ambient sensor research is done without user consent (which is one of the cons of smartphone-based apps).

Overall, sensor-based detections have higher accuracy than radio-based detections. For example, Wi-Fi and RSSI signals achieved 63% accuracy for indoor detection [DePATLA et al., 2015] with nine occupants. For occupancy counting, using only CO₂ sensors has been experimented with, using a maximum of 42 occupants and with an accuracy limit of 15% [Basu et al., 2015].

Table 4.4: Example of Indoor Human Occupancy History Data.

Timestamp (t)	CO ₂ (ppm)	Occupancy (person)	Activity
18/05/2015 09:36:53 AM	457	0	(room empty)
18/05/2015 10:01:55 AM	550	1	(arrived)
18/05/2015 10:41:59 AM	596	0	
18/05/2015 11:37:05 AM	580	3	(group meeting)
18/05/2015 12:07:07 PM	725	3	
18/05/2015 12:32:10 PM	500	0	(lunch break)
18/05/2015 12:42:11 PM	510	0	
18/05/2015 12:47:12 PM	514	1	(back from lunch)
18/05/2015 02:02:23 PM	508	0	(went to seminar)
18/05/2015 03:52:35 PM	503	1	
18/05/2015 03:02:30 PM	397	0	(external meeting)
18/05/2015 04:26:55 PM	570	1	
18/05/2015 05:38:22 PM	475	0	(went home)
...	

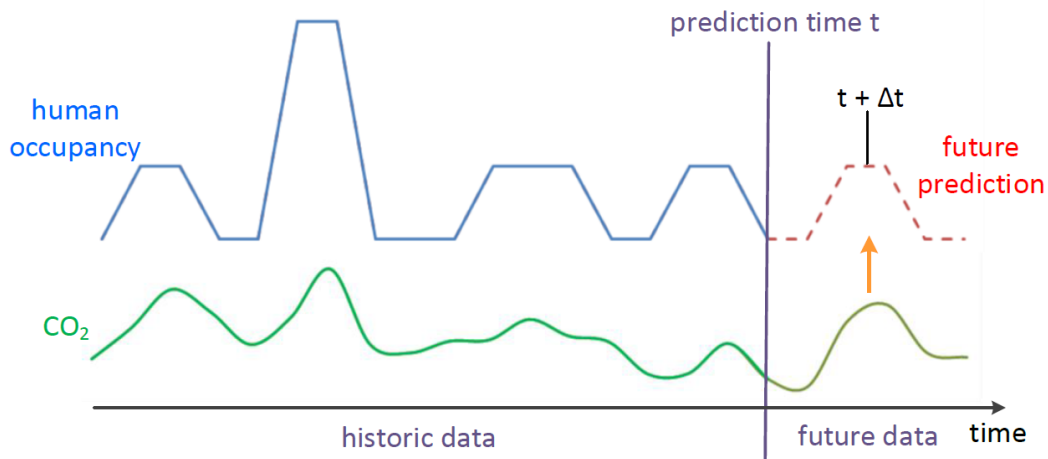


Figure 4.3: Real-time prediction scenario showing human occupancy and CO₂ fluctuations. The fundamental task is to predict the number of occupants at time $t + \Delta t$.

4.3 Problem Definition

Table 4.4 and Figure 4.3 show that there is a dependency between CO₂ and occupancy data. Our research question is: how can we predict the number of people by using a single CO₂ sensor, with an accuracy similar to the state-of-the-art techniques in the occupancy detection field?

4.3.1 Scenario Assumption

Assume TS represents the length of a time series and is expressed as $TS = \{ts_1, ts_2, \dots, ts_q\}$, where q is the number of sample points. In our time series datasets, we have two aspects:

- Carbon dioxide (CO_2) concentration C , defined as $C = \{C_1, C_2, \dots, C_q\}$
- Indoor human occupancy O , the number of people in the room at TS defined as $O = \{O_1, O_2, \dots, O_q\}$

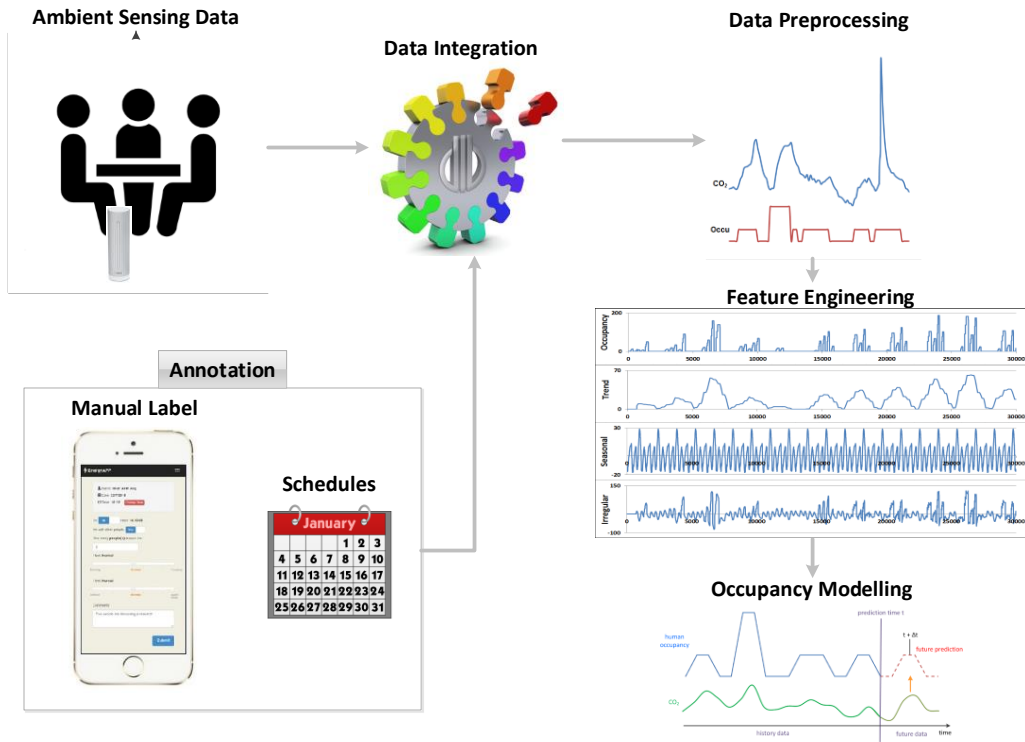


Figure 4.4: Data Collection and Analysis Framework.

4.3.2 Problem Definition

In time series prediction, analysing one-step-ahead prediction is different from analysing multi-step-ahead prediction. Predicting multiple steps ahead needs a more complicated method due to the accumulation of errors and the number of uncertainties increasing with time. We focus on multi-step-ahead prediction with the support of one dependent variable to reduce uncertainties.

We have two different types of datasets: CO₂ concentration C and indoor human occupancy O . To explore the relationship between the above two factors, two problems must be solved:

- Decompose both CO₂ concentration and indoor human occupancy to reduce the level of complexity.
- Explore the correlations between CO₂ concentration, indoor human occupancy and all of their decomposed components.

4.4 Method

We assess our model in two different locations with very different contexts to ensure this model works under various conditions. The first location is an academic office belonging to a staff member at RMIT University, Australia. This room is chosen for human occupancy prediction because a controlled experiment can be conducted for an extended period of data collection. A picture of the room is shown in Figure 4.5.

The second dataset was collected inside a cinema theatre in Mainz, Germany [Wicker et al., 2015]. The cinema theatre was chosen because it has fluctuating numbers of people throughout the day. The number of people can reach hundreds and then decrease to zero within a few hours.

The ventilation systems refresh the air once the CO₂ concentration reaches a limit where it is unhealthy for humans to breathe. Within the RUP algorithm, we assume that the reason the CO₂ level reaches this peak is because the number of occupants reaches a peak.

4.4.1 Data Collection Experimental Setup

4.4.1.1 Academic Staff Room

We used a commercial off-the-shelf Netatmo urban weather station (range: 0 – 5000 ppm; accuracy: ± 50 ppm) to read and collect ambient CO₂ data (Figure 4.5). The experiment took place between May and June 2015 and the data collection frequency was five minutes. The dataset was uploaded to a cloud service for integration purposes. Due to the characteristics of the small room (3 x 4 x 5 m), N for time lag is 0 as we assume that there is a negligible time period between exhalation and the sensor reading. We selected two weeks of data from the whole dataset and used this in the further analysis. To obtain actual occupancy data for this room, one staff member volunteered to manually record the occupancy for the whole duration

of the experiment. Once the records were obtained, the annotation was crosschecked with the online calendar to mitigate any errors.

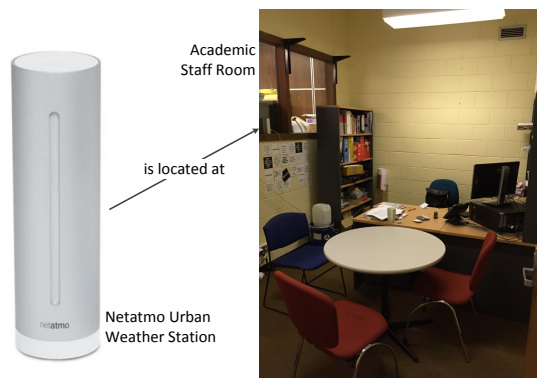


Figure 4.5: A Netatmo urban weather station (left), a sensor device to gather ambient CO_2 data that was set up near the window in the academic staff room (right).

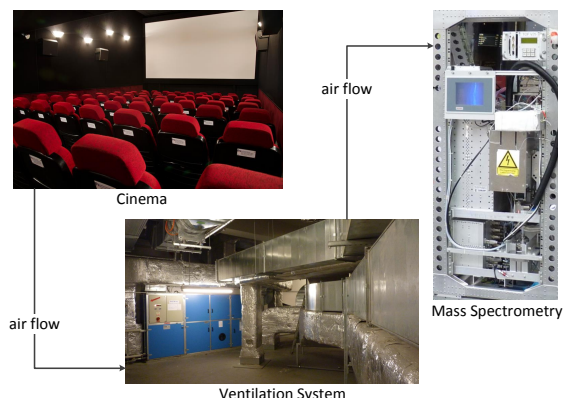


Figure 4.6: Measurement in the cinema theatre. Air is drawn out from the screen room via the ventilation system and is transported to the mass spectrometer. [Wicker et al., 2015]

4.4.1.2 Cinema Theatre

The cinema dataset was collected between December 2013 and January 2014 [Wicker et al., 2015]. The air-flow measurement and device arrangement for the cinema dataset is shown in Figure 4.6. The dataset was collected using mass spectrometry machinery installed on the air ventilation system. The air flows from the screening room via the ventilation system to the mass spectrometer (range: $10 - 500 \text{ m/z}$ with a time-of-flight (TOF) acquisition sampling time per channel of 0.1 ns , resolution: $\pm 3700 \text{ m}/\Delta\text{m}$). To obtain actual occupancy data for this cinema theatre, we collaborated with the cinema theatre and obtained the total number of tickets that were sold for each movie session. We applied a smoothing method to reduce the number by 20% during the first and last five minutes of the movie duration to model people entering and leaving the theatre.

4.4.1.3 Experimental Tool

We utilised Waikato Environment for Knowledge Analysis (WEKA), MATrix LABoratory (MATLAB) and programming language R in this experiment. WEKA was used for polynomial linear regression, with the M5 method for both correlation models for trend (subsection 4.4.3.2.1) and irregular features (subsection 4.4.3.2.3). MATLAB code was run for the baseline

method, SVR and its prediction result. We used R to integrate all the data, including decomposition of STD and STL and the majority of data preprocessing, and to compute another baseline algorithm, NMF-ELSR. We imported the data from R into Microsoft Excel for data analysis and visual output.

4.4.2 Data Preprocessing

This section explains our data preprocessing and why it is important for our model. We gathered both the CO₂ concentration from the sensor data and the indoor human occupancy from our annotations, as shown in Figure 4.4. Both datasets were integrated and preprocessed using our preprocessing method described in subsection 4.4.2.1 (autocorrelation and the line of best fit) and subsection 4.4.2.2 (time lag). With feature engineering, we factorised the data into different features and applied prediction models described in section 4.4.3 to predict indoor human occupancy.

To solve a time delay issue between the CO₂ data and our predicted indoor human occupancy number, the data required preprocessing. There is a time delay issue because when a person enters a room it takes some time before the CO₂ level in the air increases proportionally.

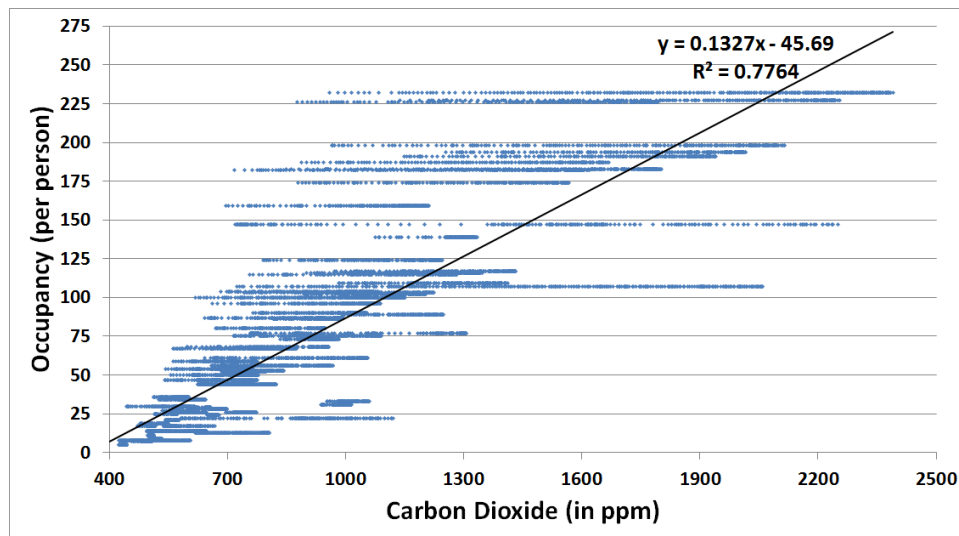


Figure 4.7: Correlation between the number of occupants and CO₂ readings (ppm) with time lag 0.

4.4.2.1 Autocorrelation and the Line of Best Fit

In order to analyse the data for CO₂ and the number of occupants, the data lagging issue needs to be considered. The data lagging means that it will take some time for CO₂ to populate the room, as there is a delay between the time of people entering (or exiting) the room and the increment (or decrement) of the CO₂ reading. For each dataset from 0 minutes time lag to the upper bound (UB) time lag (in minutes), the correlation between CO₂ data and the number of occupancies was calculated. The UB value is the maximum value that can be used to calculate the time lag value in subsection 4.4.2.2, and is defined by the formula in Equation 4.2.

$$UB = \lceil (roomlength * roomwidth * roomheight) / 100 \rceil \quad (4.2)$$

For the academic staff room, the UB value is 1. This small value means that the usefulness of time lag for this room's analysis is minimal. For the cinema data [Wicker et al., 2015], the UB value is 60, due to the large size of the theatre. A line of best fit was drawn, as shown in Figure 4.7.

To calculate the line of best fit, we first calculated the slope value (SL) between CO₂ and occupancy data, using Equation 3.2. After calculating the SL value, the intercept value between both datasets needs to be calculated, using Equation 3.3. The intercept value is the value at the intersection of the y axis by the linear regression line. The main formula for the line of best fit (LBF) is shown in Equation 3.4.

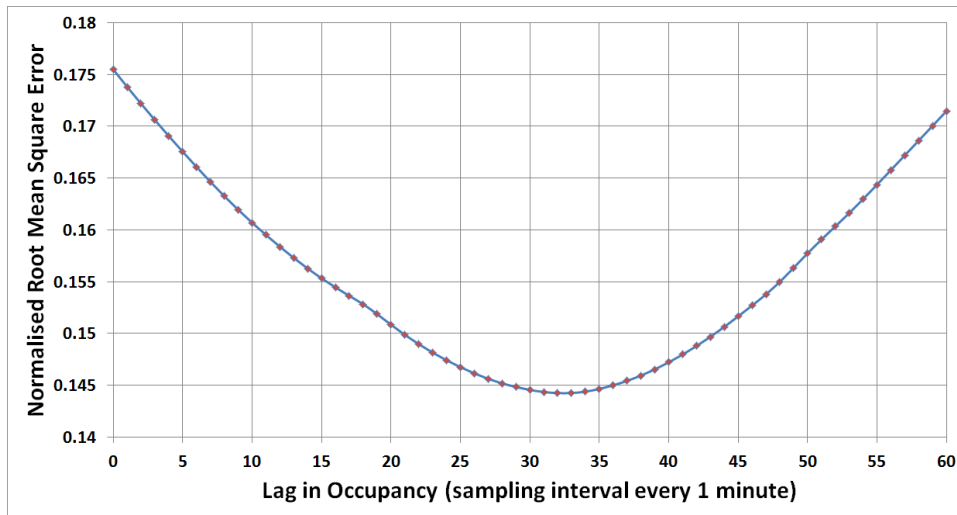


Figure 4.8: Ordinary Least Square Regression Normalised Root Mean Square Error (NRMSE) between CO₂ data and actual occupancy, for 60 minutes time lag.

4.4.2.2 Time Lag

For each line of best fit, the mean squared error (MSE), root-mean-square deviation (RMSD) and the normalised root mean square error (NRMSE) were calculated. The formula for calculating NRMSE is shown in Equation 4.3.

$$NRMSE = \frac{\sqrt{\frac{1}{n} \sum_{t=1}^n (C_t - \bar{C}_i)^2}}{O_{max} - O_{min}} \quad (4.3)$$

$NRMSE$	normalised root mean square error
t	total number of dataset
C_t	CO ₂ value
\bar{C}_i	sample means of the known CO ₂ value
O_{max}	maximum occupancy value
O_{min}	minimum occupancy value

This step was repeated up to UB times for each different time lag. For the time lag analysis, we used least square regression to compare each NRMSE from time lag 0 until time lag UB. We used the lowest value of NRMSE, time lag value (TL), as our baseline time lag for the data analysis, shown in Equation 4.4.

$$TL = \min(NRMSE) \quad (4.4)$$

By calculating NRMSE beyond the limit of UB, the TL result would not be significantly better. On the other hand, setting a lower UB value than the suggested output of Equation 4.2 may result in an incorrect TL and potentially reduce the overall prediction accuracy. For the academic staff room, the TL value was 0, which means that no time lag was needed for this analysis. For the cinema theatre, the lowest error value occurs at time lag $TL = 32$, as shown in Figure 4.8. This TL value is our base for the cinema theatre data and is used for the entire cinema data analysis process. The TL value needs to be calculated only once for every domain.

4.4.3 Room Utilisation Prediction Algorithm

Since there is no linear relationship between CO₂ and indoor human occupancy, we introduce a new prediction model that addresses this non-linear correlation by decomposing both CO₂ and occupancy data. There are two variants of this model, RUP seasonal trend decomposition



Figure 4.9: Large Room Utilisation Prediction (RUP) with Carbon Dioxide Sensor.

(RUP-STD) and RUP seasonal trend decomposition based on Loess (RUP-STL), which will be discussed in the following subsections. Each decomposition extracts a feature and a correlation can be developed from each feature.

The following subsections explain the core prediction model for this chapter, shown in Figure 4.9. In the first subsection, we discuss two data decomposition methodologies. The next subsection explains the correlation model for trend, seasonal and irregular features. The last subsection presents a new method to increase the overall accuracy by analysing conditions when the room is vacant, which we term zero pattern adjustment. This model needs to be re-trained for each location to obtain the best accuracy.

4.4.3.1 Decomposition Methodologies

There are two variants of decomposition methodologies that are utilised for RUP: seasonal-trend decomposition (STD), described in subsection 4.4.3.1.1; and seasonal-trend decomposition based on Loess (STL), described in subsection 4.4.3.1.2.

4.4.3.1.1 Seasonal-Trend Decomposition

STD is a mature technique in time series analysis. One of the most popular variants is the version X-11 method for using the moving average [Shiskin et al., 1965], and the most recent variant is version X12-ARIMA [Findley et al., 1998]. STD is an integral part of our framework.

To understand the time series data better, we use STD to decompose the model into four main features: trend, cyclical, seasonal and irregular. The trend feature (T_t) reflects the long-term progression of the time series during its secular variation. The cyclical feature (C_t) is a repeated but non-periodic fluctuation during an extended period of time. The seasonal feature (S_t) is a systematic and regularly repeated sequence during a short period of time. The irregular feature (e_t , also known as error or residual) is a short-term fluctuation from the time series and is the remainder after the trend, cyclical and seasonal features have been removed. For this chapter, as our experiment is within a short period of time (one month for each case), we combine the cyclical feature into the trend feature to make the model simpler without sacrificing accuracy.

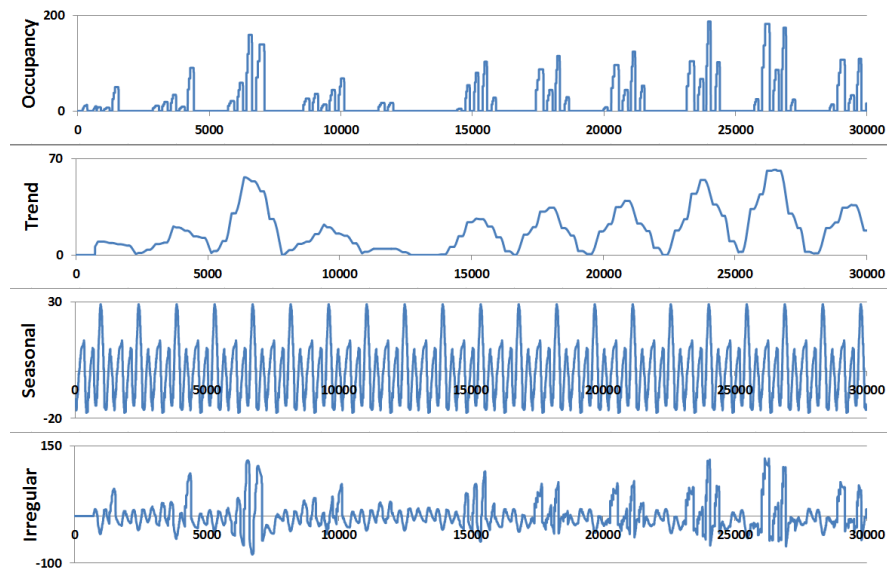


Figure 4.10: Examples of Time Series Decomposition.

Our customised STD formulation is:

$$STD_t = f(T_t, S_t, e_t) \quad (4.5)$$

t	time
STD_t	actual value of a time series at time t
T_t	trend feature at t
S_t	seasonal feature at t
e_t	irregular feature at t

The function $f()$ can be additive or multiplicative, yielding an additive decomposition or a multiplicative decomposition. An additive decomposition model is a data model in which each factor is added to model the data. Multiplicative decomposition occurs when the seasonal feature pattern is increased as the number of data increases. For multiplicative decomposition, the trend and seasonal features are multiplied and then added to the irregular feature. In our case, because the magnitude of the seasonal feature pattern in the data does not depend on the magnitude of the overall dataset, we decided to use additive decomposition as shown in Figure 4.10. From Equation 4.5, our general STD formula becomes:

$$STD_t = T_t + S_t + e_t \quad (4.6)$$

This general STD formula is applied to the time series for the CO₂ dataset and the human occupancy dataset:

$$C_t = T_t^C + S_t^C + e_t^C \quad (4.7)$$

$$O_t = T_t^O + S_t^O + e_t^O \quad (4.8)$$

To predict O_{t+1} to O_{t+n} , we need to create a model to systematically predict each of T_{t+1}^O , S_{t+1}^O and e_{t+1}^O up to T_{t+n}^O , S_{t+n}^O and e_{t+n}^O respectively, and then reconstruct the new prediction dataset using the additive method. In this chapter, we explore two variants of STD for our model and compare the accuracy result with the baseline. The first variant is standard STD that is implemented using a moving average, and the second is seasonal-trend decomposition based on Loess (STL) [Cleveland et al., 1990].

4.4.3.1.2 Seasonal-Trend Decomposition Based on Loess

Loess stands for locally weighted scatterplot smoothing, a non-parametric local regression method that is built on least square regression. This method is designed to estimate a non-linear relationship in a dataset. STL is a filtering procedure for decomposing a time series into trend, seasonal and irregular components with a Loess smoother. When compared with other STD variants, such as X-12 ARIMA, STL has several advantages, namely:

- Is very versatile and robust, can handle any type of seasonality, and will not be limited to a monthly or quarterly dataset.

- The seasonal feature and the smoothness of the trend-cycle can both be controlled by the user.
- It is robust on outliers, so occasional unusual data will not affect the estimation of trend-cycle and seasonal features. However, they will affect the irregular feature.

4.4.3.2 Correlation Models

There are three correlation models used: trend features, described in subsection 4.4.3.2.1; seasonal features, described in subsection 4.4.3.2.2; and irregular features, described in subsection 4.4.3.2.3.

4.4.3.2.1 Correlation Model for Trend Feature (T_t)

As the definition of the trend feature (T_t) is the long-term non-periodic progression of the time series during its secular variation, we assume that the trend feature for the CO₂ dataset (T_t^C) will be similar to the trend feature for indoor human occupancy (T_t^O).

To check the similarity between both trend features, we used the Pearson product-moment correlation coefficient (r) as shown below:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \quad (4.9)$$

r correlation coefficient

x dataset x

y dataset y

n number of sample points

The range of Pearson's r-value is from -1 to $+1$. If the value is >0.7 , the correlation between the datasets is strongly positive.

Once the validation step is done, we implement polynomial M5 linear regression. The M5 method will build trees whose leaves are associated with multivariate linear models, and the nodes of the trees are chosen over attributes that maximise the expected error reduction, given by the Akaike Information Criterion (AIC, a measure to check the relative goodness of fit of a statistical model) [Akaike, 1974]. The purpose of using AIC is to evaluate the model. The value for each trend feature needs to be a positive value, so we put the absolute value on both the CO₂ ($|T_t^C|$) and the human occupancy trend features ($|T_t^O|$). The main formula for trend feature correlation is:

$$T_t^O = \alpha_0 + \alpha_1(T_t^C) + \alpha_2(T_t^C)^2 + \dots + \alpha_n(T_t^C)^n + \epsilon \quad (4.10)$$

Linear regression with M5 will output each α_n and ϵ value. With these parameters, the future trend for T_{t+n}^O can be obtained.

Algorithm 3 Finding a repeated pattern sequence inside the seasonal feature.

```

1: procedure REPEATED_SEQUENCE( $S_t$ )
2:    $s_t^{temp}, s_t^{fin} \subset S_t$ 
3:    $len \leftarrow 0$  ▷  $len$ : Length for  $s_t^{temp}$ 
4:    $a \leftarrow S_t[len]$  ▷  $a$ : Start Point
5:   for each node  $i \in S_t$  do
6:      $len++$ 
7:      $s_t^{temp} \leftarrow s_t^{temp} + S_t[i]$ 
8:     if  $a = S_t[i]$  then
9:       if  $DTW(s_t^{temp}, S_t[i + 1..i + len]) > 95$  then
10:         $s_t^{fin} \leftarrow s_t^{temp}$ 
11:        break
12:      end if
13:    end if
14:  end for
15:  return  $s_t^{fin}$ 
16: end procedure

```

4.4.3.2.2 Correlation Model for Seasonal Feature (S_t)

The seasonal feature (S_t) is a systematic and regularly repeated sequence during a short period of time. Due to this characteristic, every seasonal feature can be fitted by a finite Fourier series. To correlate S_t^C and S_t^O , we use dynamic time warping (DTW), a pattern matching technique for scoring the similarity between the shape of particular signals within a certain duration [Petitjean et al., 2011]. The full correlation algorithm to find regularly repeated sequences within each S_t is shown in Algorithm 3.

Once we find a sequence that is repeating in s_t^{fin} for both the CO₂ and occupancy seasonal features, we compared the length of $s_t^{fin(O)}$ and $s_t^{fin(C)}$. If the length of $s_t^{fin(O)} < s_t^{fin(C)}$, we applied an interpolation method inside $s_t^{fin(O)}$, so that both had the same length. If the length of $s_t^{fin(O)} > s_t^{fin(C)}$, we applied a data reduction method, so that both had the same length. The final regression equation for seasonal feature correlation is:

$$s_t^{fin(O)} = f(s_t^{fin(C)}) \quad (4.11)$$

With this equation, the future trend for $S_{t+n}^{fin(O)}$ can be obtained.

4.4.3.2.3 Correlation Model for Irregular Feature (e_t)

Due to similar characteristics existing between trend and irregular features, we apply the same correlation method from the trend feature:

$$e_t^O = \beta_0 + \beta_1(e_t^C) + \beta_2(e_t^C)^2 + \dots + \beta_n(e_t^C)^n + \gamma \quad (4.12)$$

The only difference from the trend feature is that we do not need to validate it using PCC, because the shape of the irregular feature will depend more on its trend and seasonal features.

4.4.3.2.4 Zero Pattern Adjustment

In human occupancy prediction research, inferring knowledge to detect when a room is vacant is paramount. By minimising false positives, the accuracy prediction can be improved. The zero pattern adjustment (ZPA) method learns behaviour from historical data and makes smart adjustments for a vacant room when the normal algorithm returns an incorrect prediction. The ZPA technique overlays all previous datasets and puts them on a single 24-hour x-axis to determine the earliest start and end points when the room is vacant each day, from night to dawn. We symbolise ZPA as zpa_t^O .

Algorithm 4 Finding the indoor human occupancy.

```

1: procedure NUMBER_OCCUPANCY( $T_t^O, S_t^O, e_t^O, zpa_t^O$ )
2:    $O_t^{temp} \leftarrow T_t^O + S_t^O + e_t^O + zpa_t^O$   $\triangleright O_t^{temp}$ : temporary occupancy
3:   if  $O_t^{temp} \geq 0$  then
4:      $O_t = O_t^{temp}$ 
5:   else
6:      $O_t = 0$ 
7:   end if
8:   return  $O_t$ 
9: end procedure

```

4.4.3.3 Occupancy Model

Algorithm 4 contains the occupancy calculation model, where we integrate each feature to compute the occupancy prediction value. This model can be trained to have a high accuracy, using training data for as little as two weeks. With more training and ground truth data, the accuracy prediction can be improved.

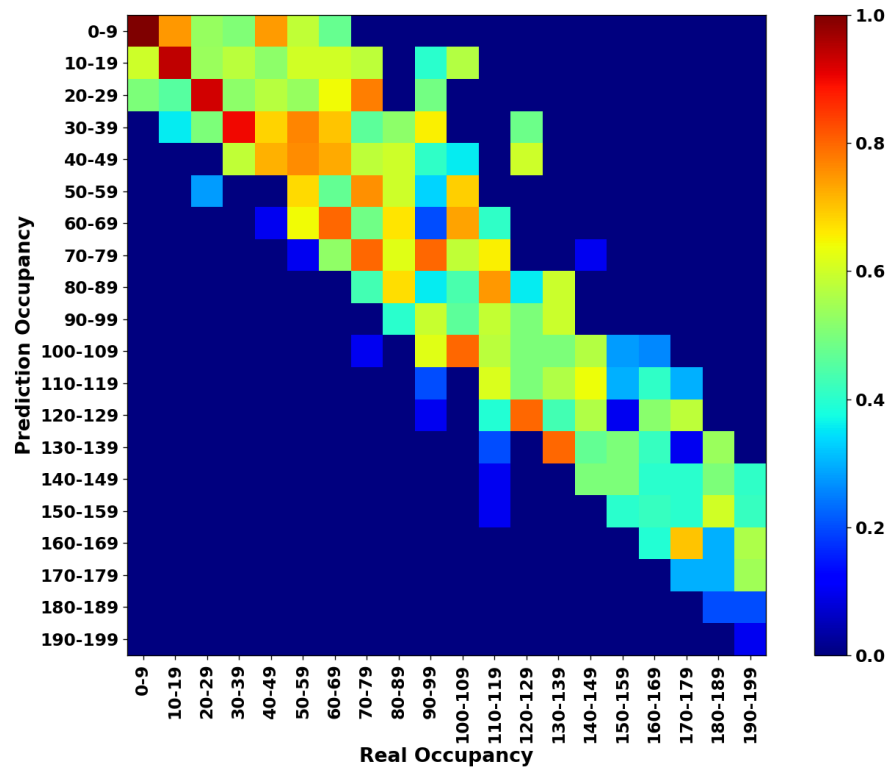


Figure 4.11: Confusion Matrix between Real Occupancy and Prediction Occupancy for Cinema Theatre.

4.5 Experiments, Results and Discussion

4.5.1 Experiments

There are two types of feature decomposition: RUP-STD and RUP-STL. We implemented both RUP models for each different combination of training and test dataset. Finally, we compared each RUP accuracy result with both SVR and NMF-ELSR.

The RUP model predicts each future value for the whole period of time based on specific time window. To better understand this model and how well it performs compared with the baseline, we define x , the accuracy error tolerance parameter. Zero units of error tolerance means that only the exact number recognised is considered as a true positive. For example, with ten units of error tolerance, if the real indoor human occupancy is 150 people, predictions as low as 140 or as high as 160 are considered correct, as they are within ± 10 units of error tolerance. The value of parameter x will be different based on the size of the room.

4.5.1.1 Experimental Parameters for Academic Staff Room Dataset

For the academic staff room dataset, we used five-minute time windows. We gathered 4019 data from this room over 14 days. Due to the small room size, we decided not to use time lag for data analysis. For this room, we have seven pairs of training-test datasets. It starts with seven days of training data and seven days of test data, and ends with 13 days of training data to predict one day of test data.

4.5.1.2 Experimental Parameters for Cinema Dataset

For the cinema theatre dataset, we used a three-minute time window for data analysis. We gathered 68,640 data from this cinema over 23 days. The cinema theatre capacity is 300 people, and we ran the line of best fit for time lag 0 to time lag 60. The lowest normal root mean square error point is at time lag 32, so we used time lag 32 as the time lag baseline. This time lag is appropriate, because a bigger room needs a larger time lag for the model to have a better accuracy. For this room, we used December 2013 data for training and January 2014 data for testing. We then replicated this in the similar method, by giving one day from the testing dataset to the training dataset and running the model again. This method was repeated until the test dataset consisted of only one day of data. Finally, we ran the same training-test dataset using two baseline methods, SVR and NMF-ELSR. The general confusion matrix for cinema occupancy prediction performance is shown in Figure 4.11.

4.5.1.3 Evaluation and Baselines

To evaluate the result, we divided the data into two equal parts: the training dataset and the test dataset. To understand how well the model fits for a longer duration, we repeated the division of the training and test datasets by adding one day of data from the test dataset to the training dataset. This replication was repeated until the test dataset had only one day

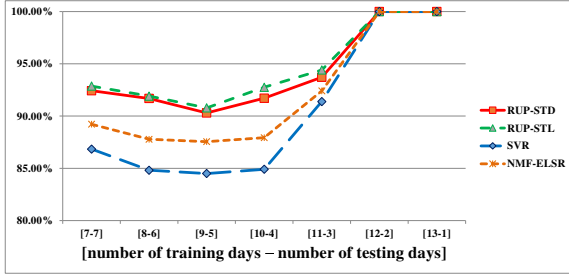


Figure 4.12: Academic staff room dataset: comparison for indoor human occupancy with zero unit of error tolerance.

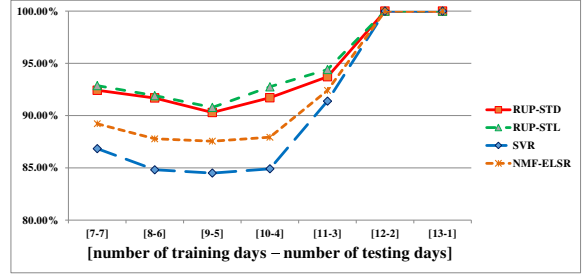


Figure 4.13: Academic staff room dataset: comparison for indoor human occupancy with one unit of error tolerance.

Table 4.5: Academic staff room indoor human accuracy result.

# of training days	# of testing days	zero units error tolerance				one unit error tolerance			
		SVR	NMF-ELSR	RUP-STD	RUP-STL	SVR	NMF-ELSR	RUP-STD	RUP-STL
7	7	86.84%	89.22%	92.42%	92.87%	97.81%	98.31%	99.35%	99.35%
8	6	84.82%	87.78%	91.68%	91.91%	97.44%	98.16%	99.30%	99.19%
9	5	84.50%	87.55%	90.29%	90.78%	97.14%	98.02%	99.16%	99.16%
10	4	84.90%	87.93%	91.71%	92.76%	97.29%	98.15%	98.95%	98.95%
11	3	91.39%	92.41%	93.71%	94.41%	98.72%	99.12%	99.88%	100.00%
12	2	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
13	1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Average accuracy		90.35%	92.13%	94.26%	94.68%	98.34%	98.82%	99.52%	99.52%

and the rest were in the training dataset. This evaluation method of incremental change to the days of training and reduction in days of testing ensured the robustness of the model.

We used two state-of-the-art algorithm baselines to compare our technique. The first one is SVR, chosen because the latest human occupancy counting method with CO₂ data, PerCCS [Basu et al., 2015], used SVR as their baseline. The second baseline we used was the PerCCS technique, a non-negative matrix factorisation with ensemble least square regression (NMF-ELSR).

4.5.2 Experimental Result

4.5.2.1 Experimental Result for Academic Staff Room Dataset

The average accuracy for indoor human occupancy with RUP-STD was 94.26%, with RUP-STL it was 94.68%, and with SVR it was 90.35%. From Figure 4.12, both RUP-STD and RUP-STL performed better than the baseline by an average of 4.33%. The last two days were Saturday and Sunday, and both RUP and the baseline model correctly predicted zero occupancies for each day.

To understand the model, we ran the experiment and checked the level of accuracy with

one unit of error tolerance. The accuracy result is shown in Figure 4.13. For this experiment, the maximum number of people is four at any one time. The average accuracy with one unit of error tolerance for indoor human occupancy was 99.52% with RUP-STD, 99.52% with RUP-STL, and 98.34% with SVR.

4.5.2.2 Experimental Result for Cinema Dataset

For the cinema dataset, the comparison accuracy result is shown in Figure 4.14. Both RUP methods performed better than the baseline method by an average of 8.5%. The highest prediction accuracy was found when we used 22 days of data for training to predict the number of human occupants the next day.

Because the maximum number of people allowed in the cinema is 300, predicting the exact number of people at one particular time is challenging. The accuracy of the baseline methods were on average 39.4% for SVR and 47.8% for RUP. We decided to calculate the accuracy number with ten units of error tolerance, so that a difference of up to ten occupants is counted as a true positive. Ten units of error tolerance is acceptable for the cinema theatre, because it will not make a major difference for controlling the HVAC and BMS whether there are 280 or 290 people. The result for ten units of error tolerance was shown in Figure 4.15. RUP-STL performed worse than the baseline, but RUP-STD is the most accurate for almost every test case, with an average accuracy of 73.76%. The average accuracy for SVR was 72.7%.

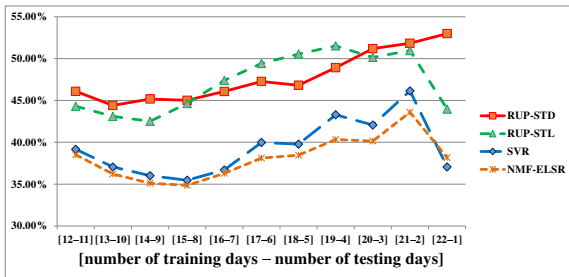


Figure 4.14: Cinema dataset: comparison for indoor human occupancy with zero units of error tolerance.

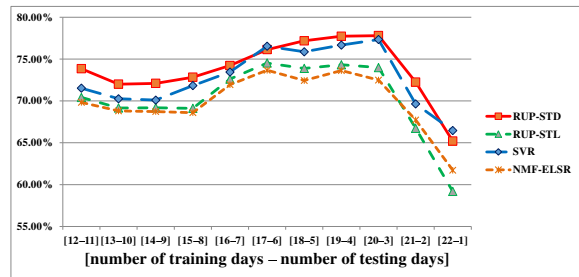


Figure 4.15: Cinema dataset: comparison for indoor human occupancy with ten units of error tolerance.

The results from Figures 4.14 and 4.15 show that, overall, the RUP method was more accurate in predicting indoor human occupancy. RUP-STL was slightly more accurate on zero units of error tolerance, and RUP-STD was more accurate on average. This result is encouraging, as RUP-STD uses a moving average and a smoothing method usually works best

Table 4.6: Cinema theatre indoor human occupancy accuracy result.

# of training days	# of testing days	zero units error tolerance				one unit error tolerance			
		SVR	NMF-ELSR	RUP-STD	RUP-STL	SVR	NMF-ELSR	RUP-STD	RUP-STL
12	11	39.16%	38.51%	46.09%	44.31%	71.52%	69.83%	73.85%	70.42%
13	10	37.05%	36.22%	44.41%	43.10%	70.26%	68.81%	72.00%	69.16%
14	9	36.01%	35.11%	45.17%	42.51%	70.10%	68.73%	72.10%	69.18%
15	8	35.47%	34.86%	45.02%	44.63%	71.83%	68.60%	72.84%	69.12%
16	7	36.70%	36.29%	46.08%	47.40%	73.46%	71.94%	74.23%	72.62%
17	6	39.98%	38.12%	47.27%	49.42%	76.54%	73.67%	76.13%	74.53%
18	5	39.78%	38.46%	46.82%	50.54%	75.87%	72.44%	77.18%	73.89%
19	4	43.29%	40.33%	48.92%	51.53%	76.68%	73.63%	77.73%	74.32%
20	3	42.06%	40.15%	51.19%	50.15%	77.34%	72.49%	77.80%	73.99%
21	2	46.14%	43.59%	51.84%	50.97%	69.64%	67.72%	72.25%	66.73%
22	1	37.06%	38.18%	52.99%	43.98%	66.46%	61.72%	65.21%	59.19%
Average accuracy		39.34%	38.17%	47.80%	47.14%	72.70%	69.96%	73.76%	70.29%

with some error tolerance. Furthermore, we can observe that the accuracy is higher for fewer days of prediction than for more days of prediction, which is aligned with the results from the academic staff room experiment.

4.5.3 Discussion

Our new framework had a high accuracy (94.4% as shown in Table 4.5) for a small room with up to four residents, and it performed better than the baseline method for a large room with up to 300 occupants. This RUP model is robust enough to handle different scales of data, which was proven by doing research in two different environment contexts (room size and the maximum number of occupants).

When comparing two baselines, NMF-ELSR performed strongly in an environment with a small number of occupants, whereas SVR was more accurate in a room with a large number of people. With a room containing five people, RUP-STD was the most accurate method, followed by RUP-STL, then NMF-ELSR, and SVR was the least accurate. With a large room, both of our RUP models performed better than the baselines by approximately 8%. On average, our RUP algorithms were more accurate than both baselines.

From the cinema theatre result, we can observe that the RUP model performed better than the better baseline method (SVR), and that with ten units of error tolerance for up to 300 occupants, the occupancy prediction accuracy was 77.8%. Estimating the number of occupants in a large room is a challenging problem and, to the best of the author’s knowledge, there has been no research that has achieved this level of accuracy for occupancy prediction for this many number occupants. The best state-of-the-art performance using CO₂ data is 15% accuracy for predicting up to 40 occupants, reported by [Basu et al., 2015]. This has shown NMF-ELSR to

perform worse than the RUP and SVR techniques, as shown in Table 4.6.

From the experimental results, the RUP model's accuracy in predicting up to 10 days ahead was reduced to only 6.9% on average, compared to predicting only one day ahead. This result is also encouraging, because it shows that the RUP model is significantly stable for long-term predictions.

The RUP-STL method performed better than SVR for precise accuracy, but the standard accuracy decreased when we considered ten units of error tolerance. This means that the value estimated can be very accurate at one time and inaccurate in other cases. The ranges fluctuated significantly. RUP-STD, on the other hand, was more stable and had a stable range of prediction. This is because STD is based on the moving average smoothing method.

4.6 Conclusion

Research on building and room occupancy counting is becoming more important. By understanding and knowing the numbers of people within a building, the heating, cooling, lighting, building energy consumption, emergency evacuation, security monitoring, and room utilisation can all be made more efficient. Thorough research in this area has been implemented with various methods, including the use of ambient sensors. However, occupancy models in previous work require the use of many sensors, which is expensive for installation and ongoing operation. In this experiment, we used a single sensor that is commonly available in Australian BMSs. This reduces the cost and complexity, as more sensors mean less reliability. The research into using a single type of information, such as CO₂, and using it to predict human occupancy is novel. Hence, many possibilities can be explored by using this technique. Furthermore, our model can be trained to have a high accuracy with training data gathered over only two weeks. This algorithm is robust in handling both low and high occupancy, up to three hundred occupants. Furthermore, with CO₂ data, the privacy of every individual is protected as no personal information is required. This method is device-free, in that no device was attached to the body throughout the experiment phases. Our method produced on average 8.46% better accuracy compared to the baseline method. In addition, our model reduces the accuracy by 6.9% when predicting more than 10 days ahead.

Chapter 5

DA-HOC++: A Scalable Room Occupancy Prediction with Transferable Time Series Decomposition of CO₂ Sensor Data

5.1 Introduction

Predicting human occupancy in a building is crucial for maximising building utilisation and improving building management and operations. However, to predict the number of occupants accurately, there needs to be sufficient labelled data to train the predictors and validate the model.

Many time-series prediction problems suffer from missing labelled data and are unable to achieve acceptable accuracy. Unfortunately, there is an abundance of unlabelled time series data that exist for either training or test data for machine learning model input. To utilise an unlabelled dataset, domain adaptation techniques need to be implemented. There are three types of domain adaptation based on the availability of the labels in the test dataset:

1. Unsupervised domain adaptation
2. Semi-supervised domain adaptation
3. Supervised domain adaptation

To address the problem of unlabelled data, the unsupervised domain adaptation is the best option, because supervised and semi-supervised domain adaptation techniques will not work in the absence of labelled data. For semi-supervised domain adaptation, the learning model needs to have a set of labelled source samples, a set of unlabelled source samples and an unlabelled set of target samples.

Many research projects have used domain adaptation to solve real-world problems [Pan and Yang, 2010, Taylor and Stone, 2009, Csurka, 2017]. In this chapter, we focus on implementing a domain adaptation technique for indoor human occupancy prediction. Data obtained from the Australian Department of the Environment and Energy indicate that approximately 40% of total maintenance costs within a building are spent on heating, ventilation, and air conditioning (HVAC) [DOEE., 2013]. Reducing HVAC usage will massively lessen overall energy consumption. A building management system (BMS) can intelligently adjust the HVAC based on the occupancy pattern.

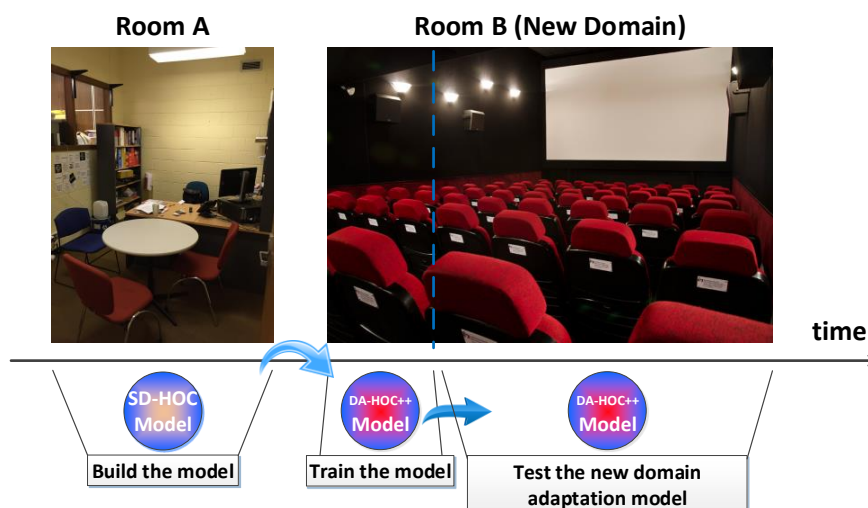


Figure 5.1: Illustration for the Main Algorithm DA-HOC++.

The research community has explored a variety of ways to develop an accurate method for determining how many occupants are in one particular room at one specific time. The current trend is utilising sensor data for indoor human occupancy detection [Candanedo and Feldheim, 2016, Ekwevugbe et al., 2013a, Khan et al., 2014, Yang et al., 2012]. Our previous research compared different types of ambient sensors, and highlighted that CO₂ is the best ambient sensor predictor for detecting human presence, as shown in Chapter 2. By using only

CO₂, 93.71% accuracy was achieved for the binary prediction of whether the room is vacant or occupied, as shown in Chapter 3. Non-negative matrix factorisation with ensemble least square regression was implemented for a CO₂ dataset to predict the occupancy state of a room, and 91% accuracy was obtained [Basu et al., 2015]. For each research outcome, a reliable historical dataset is required for training data. Without adequate past datasets, higher accuracy is difficult to achieve.

In this chapter, we propose a domain adaptation model for carbon dioxide - human occupancy counter double plus (DA-HOC++). The domain adaptation concept is shown in Figure 5.1. The human occupancy counting baseline model accurately predicts the number of people when the training and labelled data are available. We have designed a novel, semi-supervised domain adaptation for a human occupancy counting model, for implementation in any room without adequate labelled data. We compare our results with two baseline methods: support vector regression (SVR) and seasonal decomposition for human occupancy counting (SD-HOC) from Chapter 3.

Domain adaptation is useful as we usually have few or no historical data. With domain adaptation, we leverage the previous working model and adapt the model to the new environment. We used the latest baseline model for indoor human occupancy and integrated our domain adaptation method on top of the baseline model to make the new DA-HOC++.

5.1.1 Research Motivation

Many practical, real-life advantages can be obtained by knowing the number of occupants residing in one building or one room at any given time. The benefits include reducing energy consumption, human indoor comfort, and improved security of the area. Knowing the number of people in advance can be used to adjust the HVAC to reduce power cost if there is nobody inside a room for a given period. Comfort can be improved by increasing or decreasing the temperature based on the crowdedness of a room. For security reasons, if the owner of a building knows that there should be no one inside the building at a particular time, detecting a person could indicate a security breach.

Understanding the occupancy pattern is also necessary for space utilisation. If the usage pattern of two similar rooms is below 50% for both rooms, it may be beneficial to move the occupants from one room to another and utilise the newly vacant room for other more useful purposes. Maximising space and room utilisation creates greater space efficiency and could increase both individual and group productivity.

Implementation of the domain adaptation model on indoor human occupancy is a complex new research area. There are several conditions that can cause a building to not have enough historical data. For example, a newly developed structure does not have any historical data. also, highly secured buildings such as presidential and other government buildings may not have many sensors installed due to security concerns, and would therefore lack past data. Finally, old buildings that do not have any ambient sensors installed before also lack of prior data. For each scenario, the domain adaptation technique can be very beneficial and much desired.

The DA-HOC++ technique is an improvement of the DA-HOC technique with a more robust preprocessing method and advancement for the post-adjustment DA-HOC. Furthermore, the power of domain adaptation lies in how transferable the method is. In this chapter, we rigorously test DA-HOC++ with multiple different locations in different countries, in up to five locations. In addition, DA-HOC++ is a more generalised technique for time series and can be used in other contexts with similar features.

5.1.2 Research Contribution

The main contributions of this chapter are twofold. First, we propose a new semi-supervised domain adaptation data preprocessing method for human occupancy counting. Second, by making use of this preprocessing method, we present a novel framework to address semi-supervised domain adaptation problems. The performance is evaluated using a real-world dataset and is compared with state-of-the-art techniques to measure the robustness of our new framework.

This model is not only applicable for indoor human occupancy, but also for other contexts with similar condition. The model could be generalised to be implemented for any domain adaptation prediction problem Y^{target}_n if X^{source}_n , Y^{source}_n and X^{target}_n are known and $Y^{\text{source/target}}_n$ is dependent on $X^{\text{source/target}}_n$, where n is the number of sample points.

5.2 Background and Related Works

5.2.1 Related Work

Data mining and machine learning technologies have enabled great advances in many knowledge engineering areas covering classification, clustering and regression [Wu et al., 2008, Yang and Wu, 2006, Spiegel, 2016]. Unfortunately, most machine learning techniques are designed to work with the best results under one condition, and the training and test data are extracted

from the same distribution and feature space. Every time the distribution changes, the majority of the statistical models need to be re-created with new gathered training data. For this reason, domain adaptation research is emerging in recent years [Pan and Yang, 2010, Taylor and Stone, 2009] as a way to leverage a dataset containing minimal labelled data for prediction purposes.

There are a variety of knowledge engineering methods that can fully utilise the benefit of the domain adaptation technique, for example in the fields of image recognition [Kulis et al., 2011] and web-document classification [Fung et al., 2006, Al-Mubaid and Umair, 2006, Sarinnapakorn and Kubat, 2007, Dai et al., 2007]. A more prominent example where domain adaptation can shine is to classify product categories and their product reviews automatically. In [Blitzer et al., 2007], a domain adaptation technique is used to save a significant amount of labelling effort and be more cost-effective. A classification model is adapted and trained on some products, to help classification models adapt to the other products and reduce the effort of annotating and labelling reviews for various products.

Domain adaptation can also be implemented with deep learning techniques [Lu et al., 2017]. The author performed the domain adaptation technique for fault analysis. Domain adaptation is also suitable to analyse regression series [Cortes and Mohri, 2011]. Other research in semi-supervised domain adaptation includes [Daume III and Marcu, 2006], [Lopez-Paz et al., 2012] and [Yan et al., 2016]. A pipeline for unsupervised domain adaptation has been previously proposed in [Farajidavar et al., 2014]. An active domain adaptation framework is proposed by [Kale et al., 2015] for the purpose of adding a label for unlabelled target data and to generate effective label queries during active learning. Both supervised and unsupervised domain adaptation experiments are applied for activity recognition using simple in-home sensors by [Inoue and Pan, 2016].

The other case for domain adaptation is when data can be easily outdated. Data gathered during a given period might have a different distribution when they are collected in a later period. This problem usually appears on indoor Wi-Fi localisation, where a system may be required to locate a user's position based on Wi-Fi data. The value of Wi-Fi signal varies throughout the time and the recalibration process can thus be expensive. The domain adaptation technique can adopt the localisation model from the previous period and adjust it for a later period [Pan et al., 2008b].

A new dimensionality reduction method is proposed to find a latent space that minimises the distance between distributions of data in different domains [Pan et al., 2008a]. The experiment was supported by two real-world applications, including indoor Wi-Fi localisation and binary text classification. In the activity recognition field, a transfer learning framework is introduced

that is based on automatically learning a bridge between different sets of sensors using transfer learning technology [Hu and Yang, 2011]. Research in transfer learning across different feature spaces was studied by [Dai et al., 2009], and they came up with the new term ‘translated learning’. The newly proposed translated learning claims to significantly outperform many state-of-the-art baseline methods.

5.2.2 Semi-Supervised Domain Adaptation

Within pattern classification, methods for semi-supervised domain adaptation are designed to handle cases in which one cannot assume that training and test sets are sampled from the same distribution, because they were collected from different domains. However, some unlabelled samples that belong to the same domain as the test set are available, enabling the learner to adapt their parameters.

By definition, semi-supervised domain adaptation can be explained as follows: given a source domain D_S and a corresponding learning task L_S , and a target domain D_T and a corresponding learning task L_T , semi-supervised domain adaptation aims to improve the learning of the target predictive function $f_T(\cdot)$ in D_T using the knowledge in D_S and L_S , where $D_S \neq D_T$ and $L_S = L_T$. In addition, some unlabelled target-domain data must be available at training time.

In the semi-supervised domain adaptation setting, the source and target tasks are the same, while the source and target domains are different. In this situation, a small amount of labelled data in the target domain is available while labelled data from the source domain exist. Besides, according to different situations between the source and target domains, we can further categorise the semi-supervised domain adaptation setting into two cases:

1. The feature spaces (χ) between the source and target domains are different, $\chi_S \neq \chi_T$.
2. The feature spaces between domains are the same, $\chi_S = \chi_T$, but the marginal probability distributions of the input data are different, $P(\chi_S) \neq P(\chi_T)$.

In this chapter, we focus on the second case, where the feature spaces between source and domain is similar, but there is a difference in their marginal probability distributions.

5.2.3 Carbon Dioxide - Human Occupancy Counter Model

Many techniques can be utilised to predict human occupancy prediction. Dutta P.K. et al. [Dutta et al., 2006] used a radar to predict human presence. Other research has used multiple

Table 5.1: Algorithms, devices and reported accuracies for CO₂-sensor-based occupancy detection research.

Ref.	Occupancy Detection Algorithms	Detailed Devices	Max. # of People	Accuracy (Occupancy)
Chapter 4	Large Room Utilisation Prediction (RUP)	CO ₂ sensors	300	Small room: 94.68% Larger room: 73.76%
[Arief-Ang et al., 2017]	DA-HOC (Domain Adaptation)	CO ₂ sensors	230	Binary prediction: 67.85% – 75.42% Counting prediction: 59.45% – 63.75%
Chapter 3	Seasonal decomposition - human occupancy counting (SD-HOC)	CO ₂ sensors	not stated	93.71% – 97.73%
[Basu et al., 2015]	Non-negative matrix factorisation (NMF), Ensemble least square regression, support vector regression (SVR)	CO ₂ sensors (BACNet server)	15 (lab) 42 (classroom)	Binary prediction: 91% Counting prediction: 15%
[Cali et al., 2015]	Mass balance equation	CO ₂ sensors	12	Binary prediction: 95.8% Counting prediction: 80.6%
[Dedesko et al., 2015]	Time-lagged mass balance approach	PP Systems SBA-5 CO ₂ gas analysers	3	not stated
[Hailemariam et al., 2011]	Decision trees (DT) Hidden Markov models	CO ₂ sensors	1	94.68%
[Lam et al., 2009]	Neural networks Support vector machines (SVM) latent	Gas detection CO ₂ sensor network	4	65% – 80%

ambient sensors, including temperature, humidity, illumination, sound, and CO₂, as covered in Chapter 2. Wi-Fi power measurement can also be used to estimate occupancy [DePATLA et al., 2015].

Human occupancy prediction using CO₂ is gaining in popularity as a model, and PerCCS is a model with a non-negative matrix factorisation method for counting people [Basu et al., 2015] using CO₂ as the only predictor. They achieved up to 91% accuracy for predicting binary occupancy (whether the room is vacant or occupied), but suffered from 15% accuracy in predicting the number of occupants. PerCCS used SVR as their baseline method. Table 5.1 presents a summary of related work on indoor occupancy detection using CO₂ sensors.

The seasonal decomposition for human occupancy counting (SD-HOC) model is used to solve the non-linear correlation issue between CO₂ and indoor human occupancy, by decomposing both the CO₂ and occupancy data, as shown in Chapter 3. In seasonal trend decomposition (STD), there are three main components. The trend feature (TF_t) reflects the long-term progression of the time series during its secular variation. The seasonal feature (SF_t) is a systematic and regularly repeated event during a short period of time. The irregular feature (IF_t , also known as error or residual) is a short-term fluctuation from the time series and is the remainder after the trend and season features have been removed.

This general SD-HOC formula will be applied to the time series for both the CO₂ and human occupancy datasets. The main formula is:

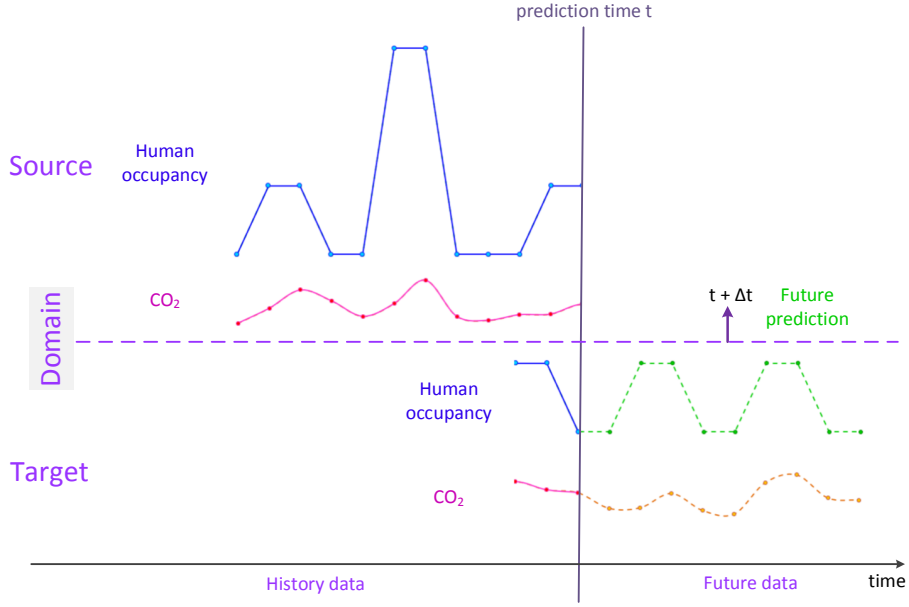


Figure 5.2: Domain adaptation prediction scenario in target domain showing human occupancy and CO₂ fluctuations. The fundamental task is to predict the number of occupants at time $t + \Delta t$.

$$O(t) = TF_O(t) + SF_O(t) + IF_O(t) + ZF_O(t) \quad (5.1)$$

- $O(t)$ Indoor human occupancy
- $TF_O(t)$ Trend feature for occupancy
- $SF_O(t)$ Seasonal feature for occupancy
- $IF_O(t)$ Irregular feature for occupancy
- $ZF_O(t)$ Zero pattern adjustment for occupancy

5.3 Problem Definition

We proposed a solution to build a transferable model to be used for indoor human occupancy prediction by using only carbon dioxide (CO₂) in the other domain. From Figure 5.2, there is a fully labelled dataset from the source domain. As occupancy data is dependent on CO₂, a model can be built to predict indoor human occupancy from both source and target domain CO₂ sensors. We then develop a semi-supervised domain adaptation method from the SD-HOC model (the source model) and integrate this with the limited labelled data from the target domain to predict future human occupancy.

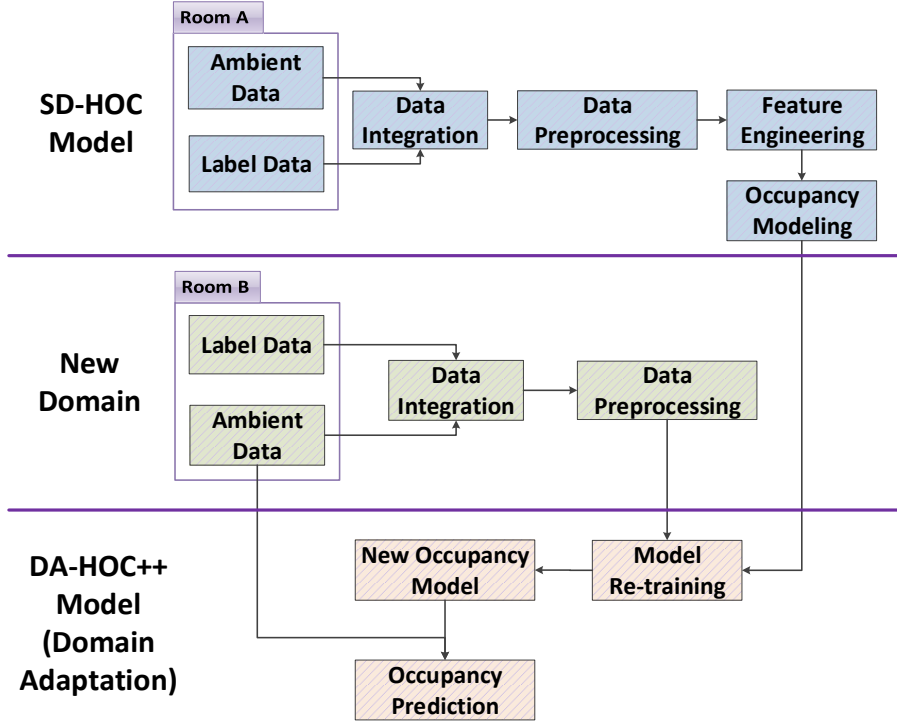


Figure 5.3: Data Collection and Analysis Framework.

5.3.1 Scenario

Assume that χ represents the length of a time series and is expressed as $\chi = \{\chi_1, \chi_2, \dots, \chi_q\}$, where q means the number of sample points. There are two types of time series: source domain time series (χ_S) and target domain time series (χ_T). Each time series dataset has two aspects: CO₂ concentration C and indoor human occupancy O . To summarise, in total we have four aspects:

- CO₂ concentration from source domain C_S , defined as $C_S = \{C_{S1}, C_{S2}, \dots, C_{Sq}\}$
- CO₂ concentration from target domain C_T , defined as $C_T = \{C_{T1}, C_{T2}, \dots, C_{Tq}\}$
- Indoor human occupancy from source domain O_S , defined as $O_S = \{O_{S1}, O_{S2}, \dots, O_{Sq}\}$
- Indoor human occupancy from target domain O_T , defined as $O_T = \{O_{T1}, O_{T2}, \dots, O_{Tq}\}$

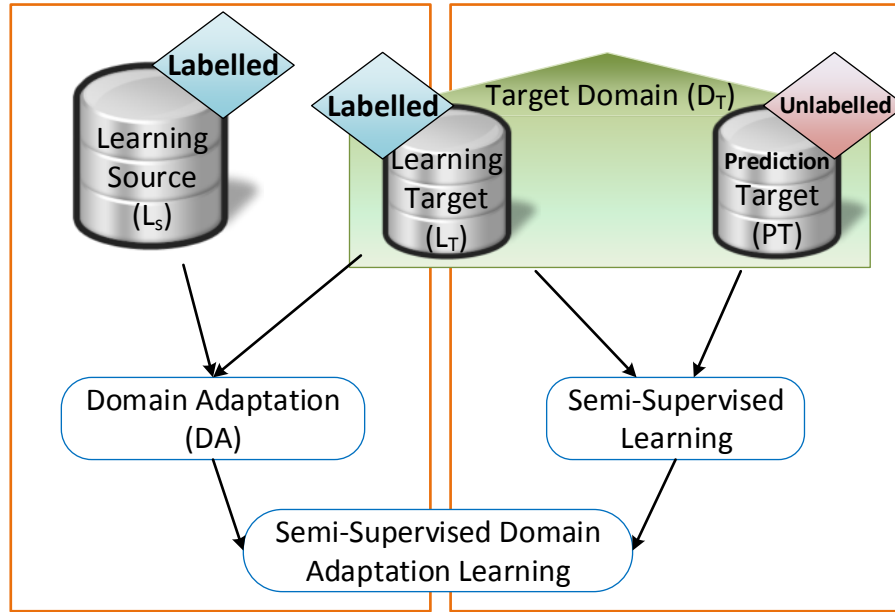


Figure 5.4: Overview of the Semi-supervised Domain Adaptation Learning Method.

5.3.2 Domain Adaptation

In the non-domain adaptation problem, χ_S and χ_T are both assumed to have been drawn from the same distribution, χ . In the domain adaptation setting, however, we would like to apply our trained classifier to the examples drawn from a distribution different from the one upon which it was trained. Therefore, we assume there are two separate distributions, χ_S and χ_T , from which data may be drawn.

It is important to note that one can have a semi-supervised algorithm that does or does not make the domain adaptation assumption, and vice versa. However, much of the work in this chapter was inspired by the belief that, although distinct, these problems are nevertheless intimately related. More specifically, when trying to solve a transfer problem between two domains, it seems intuitive that looking at the data of the target domain during training will improve performance over ignoring this source of information. Similarly, even if one believes that a transfer problem is not being solved, it may still be beneficial to model one's training and test data as if they were not identically distributed.

For the domain adaptation prediction model, two problems need to be focused on:

- A data preprocessing method for DA-HOC++ to ensure source and target domains have

the same granularity.

- Developing a semi-supervised domain adaptation for each SD-HOC component (TF_t, SF_t, IF_t and ZF_t).

For the domain adaptation method, we build a model using one dataset from a single location as source domain (χ_S), and then construct a semi-supervised domain adaptation framework so that the previous model can be utilised to predict in another environment as target domain (O_T).

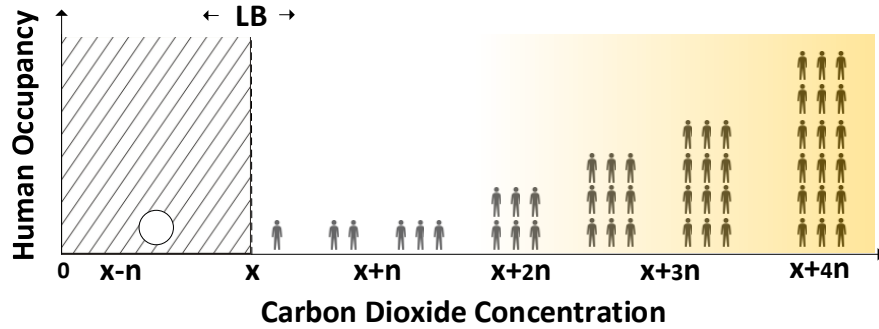


Figure 5.5: Correlation between human occupancy number and carbon dioxide concentration.

5.4 Methodology

There are three methodology steps that are implemented for the data collection and analysis framework shown in Figure 5.3. The first step is to build the SD-HOC model to the source model that has fully labelled data. The second step is to train the model with new labelled data from the target domain. The final step is to implement the DA-HOC++ model to predict the human occupancy. The semi-supervised domain adaptation concept is shown in Figure 5.4.

The DA-HOC++ model consists of three main phases: the preprocessing of CO₂ data, the main algorithm and the post adjustment model.

5.4.1 Preprocessing DA-HOC++

In this section, detailed explanations are provided of the data preprocessing steps and the reason why data preprocessing is crucial for our model. During the preprocessing phase, data from both source and target are merged. The autocorrelation, line of best fit and time lag

analysis from subsection 3.4 are an integral part of the preprocessing phase. The purpose of this step is to adjust the time taken for CO₂ gas to permeate and populate the whole room. The time lag value differs for each case and the larger the room is, the greater the time lag value.

$$LB_{CO_2} = \frac{\sum_{n=1}^{N_{max}-1} C_{min}}{N_{max} - 1} \quad (5.2)$$

LB_{CO_2}	Lower bound for CO ₂ value
n	Counter number
N_{max}	Total number of local maximum points
C_{min}	Local minimum CO ₂ concentration value located after local maximum n

In Equation 5.2, we calculate the lower bound value (LB) to adjust the time series and find a lower threshold for the CO₂ concentration when the room is supposed to be vacant. This value reduces the calculation complexity for the main algorithm DA-HOC++. Figure 5.5 shows that the LB value fluctuated based on a room's characteristics and conditions and on the demographics of humans inside it. The complexity reduction algorithm is shown in Algorithm 5.

Algorithm 5 Infused LB_{CO_2} with the dataset for complexity reduction calculation

```

1: procedure VACANT_LOWER_BOUND( $LB_{CO_2}$ )
2:    $len \leftarrow 0$  ▷  $len$ : Length for  $C_q$ 
3:    $temp \leftarrow 0$ 
4:   for each node  $c \in C_q$  do
5:      $len++$ 
6:     if  $c \leq LB_{CO_2}$  then
7:        $temp \leftarrow len$ 
8:        $C_q[temp] \leftarrow LB_{CO_2}$ 
9:     end if
10:  end for
11: end procedure

```

The value of LB_{CO_2} is used for the input parameter in DA-HOC++ to modify any concentration value that is lower than the threshold value. Binary prediction accuracy is boosted with this uniform threshold value of carbon dioxide concentration for the vacant room.

Temporal frequencies and time intervals from both source and target domains with each time lag value dataset are then correlated and adjusted based on the differences in temporal

$$A = \left| \frac{\alpha \times F_S^{Temp} - (1 - \alpha)F_T^{Temp}}{T_i} \right| \quad (5.3)$$

A	Preprocessing value
α	Weight value for temporal frequency for source
F_S^{Temp}	Temporal frequency for source value
F_T^{Temp}	Temporal frequency for target value
T_i	Time interval value

frequencies. Preprocessing values can then be extracted using the formula in Equation 5.3.

With the preprocessing value, the dataset with the higher temporal frequency can be adjusted by reducing the period to equalise the lower temporal frequency dataset. By reducing the period, excess data issues can be mitigated and the higher temporal frequency dataset undergoes a data reduction process. Once both datasets have similar temporal frequencies, the preprocessing phase is complete.

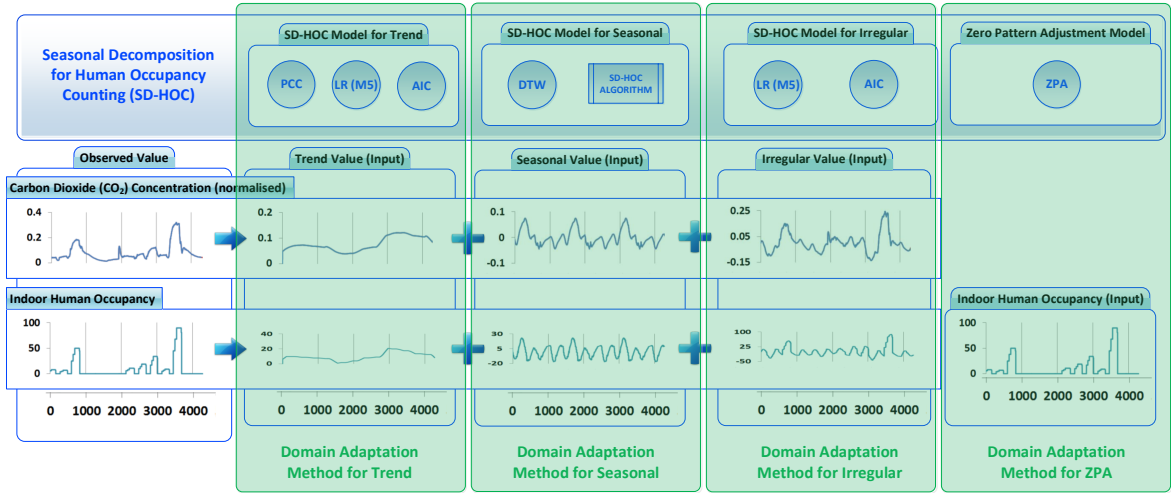


Figure 5.6: Semi-supervised domain adaptation method for Seasonal Decomposition for Human Occupancy Counter Double Plus (DA-HOC++).

5.4.2 Main Algorithm DA-HOC++

The main algorithm DA-HOC++ divides the datasets into three portions: the source dataset (D_S), the target dataset (D_T) and the learning target dataset (L_T). The source dataset contains all the labelled datasets from the source domain, the target dataset contains all the

unlabelled datasets from the target domain, and the learning target dataset contains all the labelled datasets from the target domain. The learning target dataset quantity is exiguous compared to both source and target datasets. A proportional comparison between learning target and target datasets is 0.1 or less. Figure 5.2 gives an illustration of this division.

The second data partition is divided between the vacant room prediction and occupied room prediction. The original SD-HOC algorithm does not differentiate between these two, because by retraining the model for every domain, the new model can capture the condition and the essence from each domain. However, because the majority of the target domain dataset is unlabelled, differentiating the vacant and occupied rooms can be designed by the model in a more proportional way.

As DA-HOC++ is adapted from SD-HOC, the dataset is factorised to four different features: trend feature (TF_t^{DA}), seasonal feature (SF_t^{DA}), irregular feature (IF_t^{DA}) and zero pattern adjustment feature (ZF_t^{DA}) as shown in Figure 5.6. The general formula for DA-HOC++ shown in Equation 5.4.

$$O_t^{DA} = TF_t^{DA} + SF_t^{DA} + IF_t^{DA} + ZF_t^{DA} \quad (5.4)$$

t	Time
O_t^{DA}	Occupancy number with domain adaptation at t
TF_t^{DA}	Trend feature with domain adaptation at t
SF_t^{DA}	Seasonal feature with domain adaptation at t
IF_t^{DA}	Irregular feature with domain adaptation at t
ZF_t^{DA}	Zero Pattern Adjustment with domain adaptation at t

5.4.2.1 Domain Adaptation Method for the Trend Feature (TF_t^{DA})

The trend feature behaves similarly in CO₂ concentration and human occupancy due to their linear correlation. The first algorithm for the trend feature model is the maximum mean discrepancy (MMD) [Fortet and Mourier, 1953]. The purpose of MMD is to bring the average of the two distributions closer to each other, while projecting the data into the principal components directions of the full dataset including the source and target domain. We want to find a function that assumes different expectations on two separate distributions.

Equation 5.5 definition: Let \mathcal{F} be a class of functions $f: X \rightarrow \mathbb{R}$. Let p and q be Borel probability distributions and let $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_m)$ be samples composed of independent and identically distributed observations drawn from p and q , respectively. We

define the maximum mean discrepancy (MMD) and its empirical estimate as:

$$MMD[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} \left(\mathbf{E}_p[f(x)] - \mathbf{E}_q[f(y)] \right) \quad (5.5)$$

Putting trend feature source domain (TF_S^{DA}) as p and trend feature target domain (TF_T^{DA}) as q , we can elaborate the MMD formula as shown in Equation 5.6.

$$MMD[\mathcal{F}, TF_S^{DA}, TF_T^{DA}] := \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(TF_{S_i}^{DA}) - \frac{1}{n} \sum_{i=1}^n f(TF_{T_i}^{DA}) \right) \quad (5.6)$$

With DA-HOC++, understanding the trend feature distribution's distance for each domain is necessary, due to the lack of training data from the related environment dataset. A trend component pseudo model needs to be created to mimic the model from the other domain and compare the boundary between the top and the bottom values for human occupancy. The CO₂ concentration behaves similarly due to the normalisation process during data preprocessing. Finding the mapping between CO₂ level and the number of occupants is the essential contribution from the trend feature model. The general function with Taylor's expansion for trend feature analysis is shown in Equation 5.7.

$$TF_t^{DA} = \alpha_0 + \alpha_1(TF_t^{DA}) + \alpha_2(TF_t^{DA})^2 + \dots + \alpha_n(TF_t^{DA})^n + \epsilon \quad (5.7)$$

5.4.2.2 Domain Adaptation Method for the Seasonal Feature (SF_t^{DA})

The seasonal feature is the repeating part of data during a short period. Due to its nature, once the DA-HOC++ model can obtain the repeating pattern for the seasonal feature, the prediction is more accurate. The SD-HOC seasonal feature focuses on the regularly repeated event during a short period. The domain adaptation method for the seasonal feature will learn from the short amount of new testing data and then correlate them with the original model to find the similarity. The duration of each repetition is essential and will be translated to the new model. The full algorithm for the seasonal feature is shown in Algorithm 6. With the DA-HOC++ model, the data pattern from the seasonal feature from the source domain is extracted, and from this the seasonal feature from the target domain can be deciphered.

Algorithm 6 Finding a repeated pattern sequence inside seasonal feature

```

1: procedure REPEATED_SEQUENCE( $SF_t^{CD}, SF_t^{DA}$ )
2:    $len \leftarrow 0$   $\triangleright len$ : Length for  $SF_t^{CD}$ 
3:    $a \leftarrow SF_t^{DA}[len]$   $\triangleright a$ : Start Point
4:   for each node  $i \in SF_t^{DA}$  do
5:      $len++$ 
6:      $SF_t^{CD} \leftarrow SF_t^{CD} + SF_t^{DA}[i]$ 
7:     if  $a = SF_t^{DA}[i]$  then
8:       if  $DTW(SF_t^{CD}, SF_t^{DA}[i + 1..i + len]) > 95$  then
9:          $SF\_Fin_t^{DA} \leftarrow SF_t^{CD}$ 
10:        break
11:      end if
12:    end if
13:  end for
14:  return  $SF\_Fin_t^{DA}$ 
15: end procedure

```

5.4.2.3 Domain Adaptation Method for the Irregular Feature (IF_t^{DA})

The irregular feature is the residual component from the raw data minus both the trend and seasonal features. The DA-HOC++ irregular feature is similar to the trend feature without the Pearson product-moment correlation coefficient. We implemented the similar domain adaptation method from section 5.4.2.1.

$$MMD[\mathcal{F}, IF_S^{DA}, IF_T^{DA}] := \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(IF_{S_i}^{DA}) - \frac{1}{n} \sum_{i=1}^n f(IF_{T_i}^{DA}) \right) \quad (5.8)$$

For the DA-HOC++ irregular feature model, we correlated it, like the trend feature, with weight value, due to the possibility of differences in scale between the source and the target components.

$$IF_t^{DA} = \beta_0 + \beta_1(IF_t^{DA}) + \beta_2(IF_t^{DA})^2 + \dots + \beta_n(IF_t^{DA})^n + \gamma \quad (5.9)$$

5.4.2.4 Domain Adaptation Method for the Zero Pattern Adjustment Feature (ZF_t^{DA})

The SD-HOC ZPA feature was invented to adjust the condition where the room is vacant and also aims to minimise false positives. The domain adaptation method for ZPA includes an adapted ZPA border adjustment for the start point and end point that has been declared in the original model with a new weighting mechanism. This weighting mechanism is needed because the new testing data will not be enough to understand the complete structure of the new environment, and the original model for ZPA can only provide a minimal contribution. The weighting calculation is shown in Equation 5.10.

$$w = \frac{1}{2} \left[\frac{\sum_{i=1}^n ZFStart_t^{CD}}{n} + \frac{\sum_{i=1}^n ZFEnd_t^{CD}}{n} \right] \quad (5.10)$$

w	Weight value
$ZFStart_t^{CD}$	ZPA starting point for SD-HOC
$ZFEnd_t^{CD}$	ZPA end point for SD-HOC
n	Total number in dataset
i	Counter value

Once the weight value has been obtained, the new $ZFStart_t^{DA}$ and $ZFEnd_t^{DA}$ can be obtained, using equation 5.11.

$$ZF[Start/End]_t^{DA} = w \cdot ZF[Start/End]_t^{CD} \quad (5.11)$$

$ZF[Start/End]_t^{DA}$	ZPA starting/end point for DA-HOC++
$ZF[Start/End]_t^{CD}$	ZPA starting/end point for SD-HOC
w	Weight value

5.4.3 Post Adjustment DA-HOC++

The post adjustment phase contains the final calibration and model evaluation. The final calibration includes plotting each feature from the previous subsection and correlating each value with its vacant and non-vacant condition using Gaussian mixture models (GMM), as shown in Figure 5.7.

From Figure 5.7, we can conclude that the correlation between seasonal and irregular factors resulted in clear separation between vacant and non-vacant, so we can group the prediction

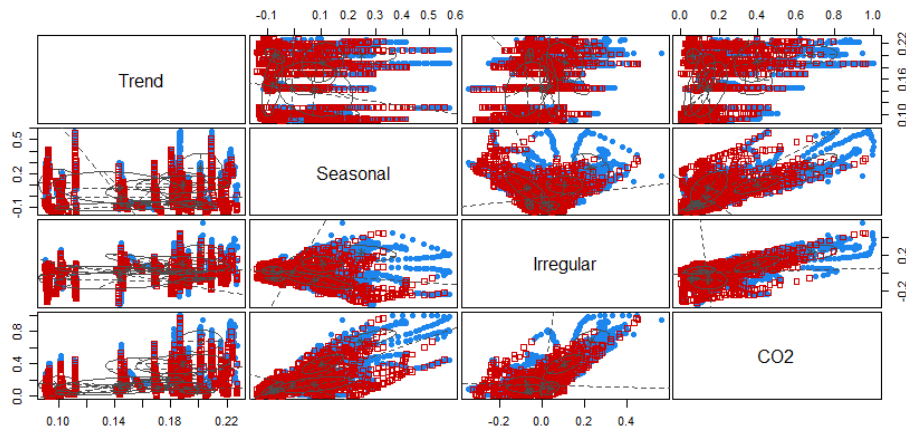


Figure 5.7: The correlation plot between trend, seasonal, irregular and CO₂ value between vacant (red) and non-vacant (blue) room.

accordingly. If each correlation mixed vacant and non-vacant together as in the trend and the seasonal factors above, the final calibration step is finished.

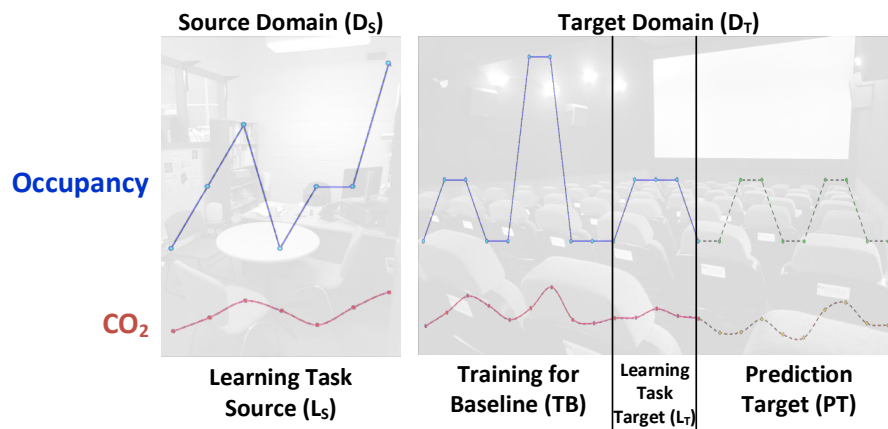


Figure 5.8: Source Domain and Target Domain.

For model evaluation, we decided to use two baselines, SVR and SD-HOC. Figure 5.8 shows data from both the source domain (D_S) and target domain (D_T). The target domain is divided into three different parts. The first part is training data for the baseline (TB), which is only used for baselines. The second part is the learning task from the target domain (L_T), and the duration is short. The third part is the prediction target (PT), which is the duration in which we predict the number of people for every minute.

For each baseline, they are trained both with and without TB , as in the real world scenario it is not easy to have a long duration of training dataset. We have $SVR(-TB)$ and $SD-HOC(-TB)$ for baselines that are trained using L_T dataset only, $SVR(+TB)$ and $SD-HOC(+TB)$ for baselines that are trained using both TB and L_T datasets, and our proposed DA-HOC++ model that is trained with another domain dataset (L_S) and thus improved using a short duration of L_T .

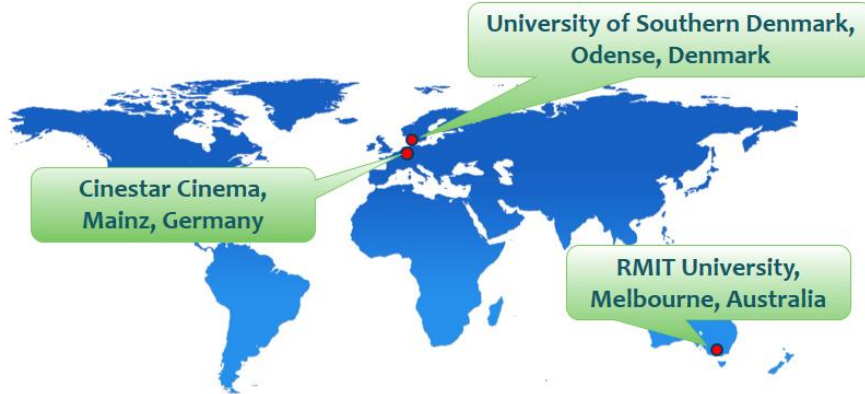


Figure 5.9: The DA-HOC++ experiment covers multiple locations in three different countries, in six different rooms with a variety of sizes, surrounding environments and characteristics.

We used two baseline methods to ensure that the new proposed domain adaptation algorithm performs. Domain adaptation research is hard to compare and it is almost impossible to compare the prediction accuracy based on the numbers only. For this reason, we decided to add another dataset before L_T so that the baseline could have a proper training dataset. We called this new dataset as ‘training for baseline’ (TB). Because of this, it is expected that any baseline with TB will have a better accuracy prediction compared to any algorithm with domain adaptation. The cross-domain analysis will perform worse than the same domain analysis. The question that we want to solve is: how much worse is the result compared to the same domain algorithm? Furthermore, we subsequently ran the same baseline algorithm without TB . If our domain adaptation algorithm can perform better than the baseline algorithm without TB result, it provides a promising research direction, as producing a new model from another domain is propitious to increase prediction accuracy given the lack of data from the target domain.

Table 5.2: Detailed statistical information on the datasets for small and large rooms.

Information	Single office room (Melbourne)	Cinema theatre (Mainz)	Study Zone 1 (Odense)	Study Zone 2 (Odense)	Classroom 1 (Odense)	Classroom 2 (Odense)
Min. # of occupants	0 people	0 people	0 people	0 people	0 people	0 people
Max. # of occupants	5 people	230 people	29 people	35 people	67 people	39 people
Avg. # of occupants	0.16 people	20.69 people	2.79 people	2.55 people	7.46 people	3.69 people
Mode # of occupants	0 people	0 people	0 people	0 people	0 people	0 people
Min. CO ₂	378 ppm	388.59 ppm	268 ppm	256 ppm	304 ppm	370.88 ppm
Max. CO ₂	1002 ppm	3518.06 ppm	688 ppm	907.52 ppm	1384 ppm	844.8 ppm
Avg. CO ₂	464.82 ppm	690.95ppm	380.73 ppm	386.38 ppm	471.64 ppm	463.02 ppm
Mode. CO ₂	453 ppm	424.26 ppm	312 ppm	296.96 ppm	384 ppm	403.84 ppm
Room size	12 m ²	not stated	125 m ²	125 m ²	139 m ²	139 m ²
Number of entrances	1	3	2	2	2	2
Door condition	Default state is closed	Closed	Default state is closed	Default state is closed	Default state is closed	Default state is closed
Windows condition	Closed at all times	Closed at all times	Closed at all times	Closed at all times	Closed at all times	Closed at all times
Sensor used	NetAtmo weather station	Mass spectrometry	Indoor CO ₂ sensor	Indoor CO ₂ sensor	Indoor CO ₂ sensor	Indoor CO ₂ sensor

5.5 Location Detail, Settings and Parameters

For this chapter, we utilised six different datasets from six different rooms. For the first room, we gathered ambient data from an academic office belonging to a staff member at RMIT University, Australia. This room was chosen for DA-HOC++ source domain because a controlled experiment could be conducted for an extended period of data collection.

The other five rooms are our target domain. For the second room, we utilised a large dataset of volatile organic compounds collected with a mass spectrometer in a cinema theatre in Mainz, Germany [Wicker et al., 2015]. We extracted and used the CO₂ channel from the dataset for our study. This cinema theatre dataset is used because of its nature of having

fluctuating numbers of people throughout the day. The numbers of people in the audiences can reach three hundred and can decrease to zero within two hours.

The last four rooms' datasets are utilised from a building in the University of Southern Denmark, Odense, Denmark [Sangogboye et al., 2017]. Two of these are study zones and the other two are classrooms. The value of domain adaptation research lies in how transferable the model is. With a total of six rooms of data scattered in three different countries across two continents (Figure 5.9), achieving acceptable prediction accuracy for multiple buildings means the knowledge transfer techniques are applicable in this domain.

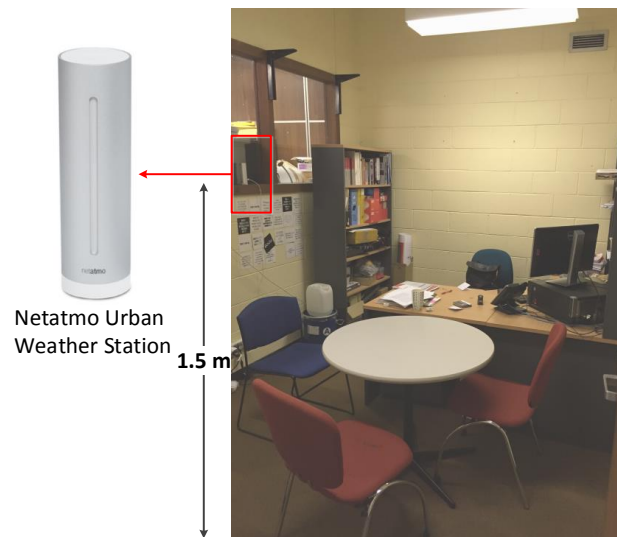


Figure 5.10: A Netatmo urban weather station (left), a sensor device to gather ambient CO₂, data which was set up near the window in the academic staff room (right).

5.5.1 Location Settings and Parameters

In this subsection, we explain the location of source and target domains.

5.5.1.1 Source Domain Location

A commercial off-the-shelf Netatmo urban weather station (range: 0-5000 ppm; accuracy: ± 50 ppm) was used to collect ambient CO₂ data, as shown in Figure 5.10. The duration of data gathering was two months, from May to June 2015. The dataset was uploaded to a cloud service for integration purposes. Due to the small room dimensions (3 x 4 x 3.5 m), the time

lag is 0 as we assume that there is a negligible period between exhalation and sensor reading. The time window that we used for this dataset is five minutes.

We conducted manual labelling for the number of people, because we were doing a controlled experiment. The Netatmo urban weather station was chosen because it has a wide range of CO₂ sensor reader. The device was put on a window at the height of a human occupant's nose, because this height at which CO₂ concentration changes first. Further detail about this room's dataset is in Table 5.2.

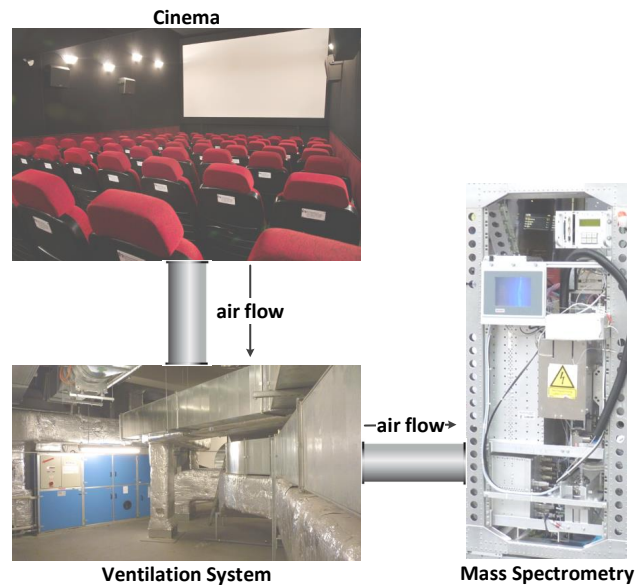


Figure 5.11: Measurement in the cinema theatre. Air is drawn out from the screen room via the ventilation system and is transported to the mass spectrometer [Wicker et al., 2015].

5.5.1.2 Target Domain Location

Five locations were used as target domain. The first one is the Cinestar cinema theatre in Germany. The cinema dataset was collected between December 2013 and January 2014 [Wicker et al., 2015]. Due to the size of this room, the dataset was collected using a mass spectrometry machinery installed on the air ventilation system. The air flowed from the screening room via the ventilation system to the mass spectrometer for data analysis. The air-flow measurement and device arrangement for the cinema dataset are shown in Figure 5.11.

For the cinema theatre dataset, we used a 30-second time window for data analysis. The capacity of the cinema theatre is 230 people, and for this experiment we ran the line of best fit

for time lag 0 to time lag 60. The lowest NRMSE is at time lag 32, and we used time lag 32 as the time lag baseline. Further detailed information about this cinema dataset is in Table 5.2.



Figure 5.12: One of the classrooms located in OU44 building (shown top-right) at the University of Southern Denmark [Sangogboye et al., 2017].

The other four datasets are obtained from two study zones (125 m²) and two classrooms (139 m²) located in OU44 building, University of Southern Denmark. The datasets were collected from March to April 2017. Project work was the typical activity that occurred in the study zones, and the classrooms were generally used for teaching purposes. The building is equipped with a centralised BMS that controls ventilation, heating and lighting systems. For all four datasets, the time windows were 60 seconds. The study zones can cater for up to 36 occupants and the classrooms can hold up to 85. Images of the OU44 building image and one of the classrooms are in Figure 5.12. Further detailed information about the datasets for these four rooms is in Table 5.2.

5.6 Experiments and Results

Our experiment was divided into two stages, binary occupancy prediction and occupancy counting prediction. Binary occupancy predicts whether the room is vacant or occupied. For the single office room, 89.38% of the dataset showed when the room was unoccupied. For the cinema theatre dataset, it showed 67.89%. Study zones 1 and 2 had unoccupied percentage values of 38.98% and 49.55%, respectively. Classrooms 1 and 2 had unoccupied percentage

values of 64.96% and 62.98% respectively. The reason why the unoccupied percentage is much lower for study zones is that many students stay in the room during evenings and weekends. Occupancy counting prediction focuses on predicting the number of people in the room.

5.6.1 Experiment Tools

Waikato Environment for Knowledge Analysis (Weka), MATLAB and R were used to help perform this experiment. Weka is used for machine learning algorithms and data analysis. MATLAB is used in building some models and R is used for data integration, analysis and visualisation. We imported the data from R into Microsoft Excel for data analysis and visual output.

All experiments were run in a 64 bit Windows 7 Enterprise service pack 1 operating system machine. The model was HP EliteOne 800 G2 23-in NT GPU AiO. The processor was Intel Core i7-6700 CPU 3.40 GHz with 16 GB installed memory (RAM). The total capacity of our internal hard drive was 512 GB.

5.6.2 Baselines

For the evaluation of time lag, we implemented the SD-HOC preprocessing method as shown in subsection 3.4.3. The time lag for the cinema theatre was 32 minutes. This value was derived from the size of the room and the location of the mass spectrometer. Time lag means that there is a time delay between when a person enters a room, the measurement of the CO₂, and the reaction from our model. CO₂ needs to travel from the occupant's exhalation to the mass spectrometer's sensor.

There are a total of four baselines that we used for this experiment: the SVR algorithm that was trained with the TB dataset and without the TB dataset, and the SD-HOC algorithm that was trained with and without the TB dataset. We called these SVR($-TB$), SVR($+TB$), SD-HOC($-TB$) and SD-HOC($+TB$) respectively. We ran each experiment for both binary occupancy prediction and occupancy counting. Binary occupancy prediction means that each algorithm only predicts whether the room is empty or occupied. Occupancy counting means that each algorithm needs to predict the exact number of occupants for each PT time frame. The binary occupancy prediction accuracy should be higher than the occupancy counting prediction accuracy. We ran the experiment multiple times for a different number of days for both L_T and PT . A longer duration of L_T results in a higher accuracy for each algorithm.

Table 5.3: SVR (+TB), SD-HOC (+TB), SVR (-TB), SD-HOC (-TB) and DA-HOC++ Human Binary Occupancy and Occupancy Counting Prediction Accuracy Results for the Cinema Dataset.

Cinema	Non Domain Adaptation		Domain Adaptation		
Binary Occupancy Prediction					
$L_T - PT \left(\frac{L_T}{L_T + PT} \right)$	SVR (+TB) [TB + $L_T \rightarrow PT$]	SD-HOC (+TB)	SVR (-TB) [$L_T \rightarrow PT$]	SD-HOC (-TB)	DA-HOC++ [$L_S + L_T \rightarrow PT$]
1 - 13 (7.14%)	81.62%	88.58%	56.85%	59.85%	67.85%
2 - 12 (14.29%)	80.93%	88.30%	56.84%	59.84%	68.84%
3 - 11 (21.43%)	80.44%	89.51%	52.58%	54.58%	67.58%
4 - 10 (28.57%)	84.19%	90.38%	56.79%	58.79%	67.79%
5 - 9 (35.71%)	87.52%	91.04%	59.19%	61.19%	71.19%
6 - 8 (42.86%)	90.00%	93.89%	60.24%	62.24%	73.24%
7 - 7 (50.00%)	91.28%	95.84%	63.42%	64.42%	75.42%
Occupancy Counting Prediction					
1 - 13 (7.14%)	71.52%	73.85%	50.45%	51.45%	59.45%
2 - 12 (14.29%)	70.26%	72.00%	51.84%	52.84%	59.84%
3 - 11 (21.43%)	70.10%	72.10%	50.19%	51.19%	59.19%
4 - 10 (28.57%)	71.83%	72.84%	51.29%	53.29%	59.29%
5 - 9 (35.71%)	73.46%	74.23%	50.02%	51.02%	62.02%
6 - 8 (42.86%)	74.54%	76.13%	52.51%	53.51%	62.51%
7 - 7 (50.00%)	75.87%	77.18%	53.75%	55.75%	63.75%

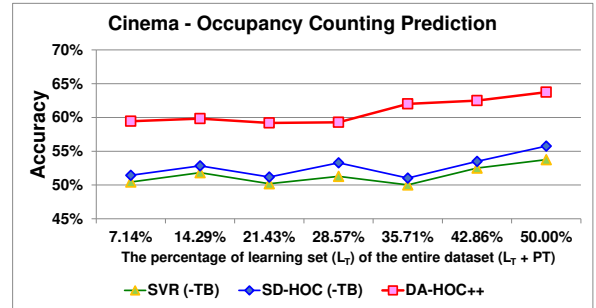
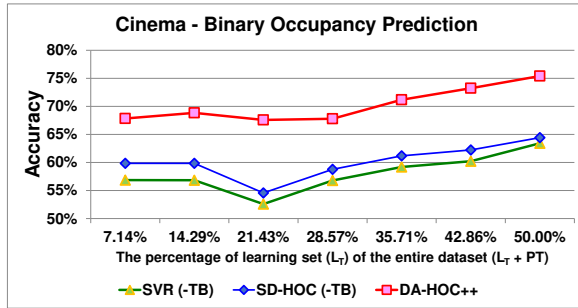


Figure 5.13: Binary Occupancy Prediction Result for the Cinema Dataset.

Figure 5.14: Occupancy Counting Accuracy Result for the Cinema Dataset.

If the prediction value is within ± 5 people from the ground truth, we counted this as a true positive. If the room was vacant and the algorithm predicted correctly, we counted this as a true negative. We calculated the prediction accuracy as the number of true positive values divided by the number of PT records in the dataset, which includes true positives, true negatives, false positives and false negatives as shown Equation 5.12.

Table 5.4: SVR (+TB), SD-HOC (+TB), SVR (-TB), SD-HOC (-TB) and DA-HOC++ Human Binary Occupancy and Occupancy Counting Prediction Accuracy Results for the Study Zone 1 Dataset.

Study Zone 1	Non Domain Adaptation		Domain Adaptation		
Binary Occupancy Prediction					
$L_T - PT \left(\frac{L_T}{L_T + PT} \right)$	SVR (+TB) [TB + $L_T \rightarrow PT$]	SD-HOC (+TB) [TB + $L_T \rightarrow PT$]	SVR (-TB) [$L_T \rightarrow PT$]	SD-HOC (-TB) [$L_T \rightarrow PT$]	DA-HOC++ [$L_S + L_T \rightarrow PT$]
1 - 9 (10.00%)	82.44%	85.42%	58.33%	60.12%	70.11%
2 - 8 (20.00%)	83.31%	86.12%	59.27%	61.53%	71.64%
3 - 7 (30.00%)	83.44%	87.21%	60.18%	62.49%	71.68%
4 - 6 (40.00%)	84.50%	87.98%	61.44%	63.77%	72.45%
5 - 5 (50.00%)	85.67%	88.05%	61.91%	64.18%	73.86%
Occupancy Counting Prediction					
1 - 9 (10.00%)	71.56%	73.86%	46.14%	49.61%	58.71%
2 - 8 (20.00%)	72.13%	74.21%	47.24%	50.10%	58.96%
3 - 7 (30.00%)	72.86%	74.56%	47.98%	50.76%	59.32%
4 - 6 (40.00%)	74.05%	76.72%	48.76%	51.68%	59.81%
5 - 5 (50.00%)	74.96%	77.13%	49.14%	52.13%	60.15%

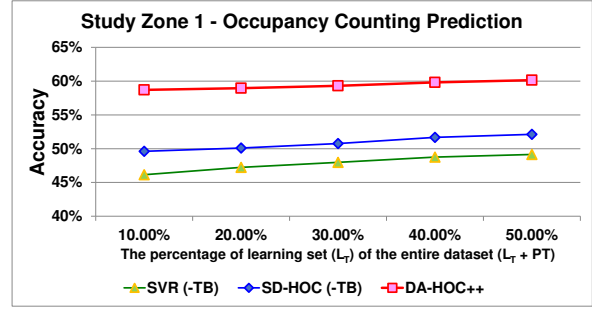
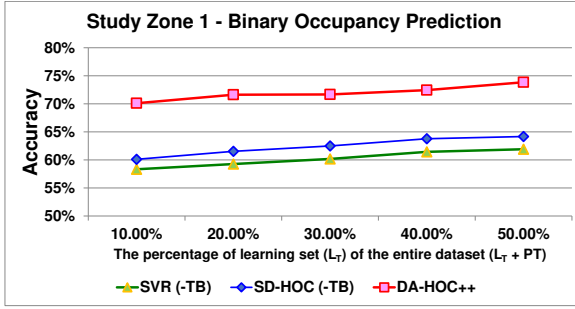


Figure 5.15: Binary Occupancy Prediction Results for the Study Zone 1 Dataset.

Figure 5.16: Occupancy Counting Accuracy Results for the Study Zone 1 Dataset.

$$Accu = \frac{TP + TN}{TP + FP + TN + FN} \quad (5.12)$$

- $Accu$ Prediction accuracy (%)
- TP Total number of true positive prediction records
- FP Total number of false positive prediction records
- TN Total number of true negative prediction records
- FN Total number of false negative prediction records

Table 5.5: SVR (+TB), SD-HOC (+TB), SVR (-TB), SD-HOC (-TB) and DA-HOC++ Human Binary Occupancy and Occupancy Counting Prediction Accuracy Results for the Study Zone 2 Dataset.

Study Zone 2	Non Domain Adaptation		Domain Adaptation		
Binary Occupancy Prediction					
$L_T - PT \left(\frac{L_T}{L_T + PT} \right)$	SVR (+TB) [TB + $L_T \rightarrow PT$]	SD-HOC (+TB)	SVR (-TB) [$L_T \rightarrow PT$]	SD-HOC (-TB)	DA-HOC++ [$L_S + L_T \rightarrow PT$]
1 - 9 (10.00%)	83.47%	86.23%	61.51%	62.51%	72.51%
2 - 8 (20.00%)	85.32%	87.44%	62.64%	63.64%	72.64%
3 - 7 (30.00%)	86.72%	88.71%	63.15%	64.15%	73.15%
4 - 6 (40.00%)	87.31%	89.11%	63.47%	64.47%	73.47%
5 - 5 (50.00%)	88.55%	90.08%	64.12%	64.82%	73.82%
Occupancy Counting Prediction					
1 - 9 (10.00%)	72.71%	74.93%	47.81%	50.81%	59.81%
2 - 8 (20.00%)	73.82%	75.66%	47.99%	51.99%	60.99%
3 - 7 (30.00%)	74.65%	76.72%	48.27%	52.27%	61.27%
4 - 6 (40.00%)	75.07%	77.32%	48.63%	52.63%	62.63%
5 - 5 (50.00%)	76.22%	78.37%	49.94%	53.94%	62.94%

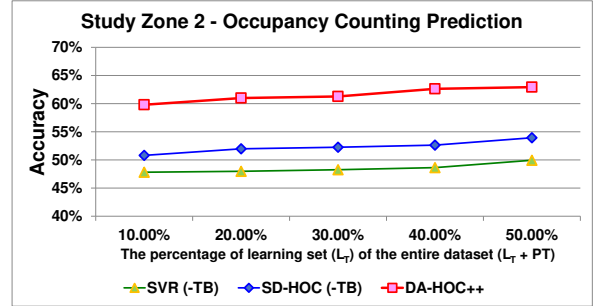
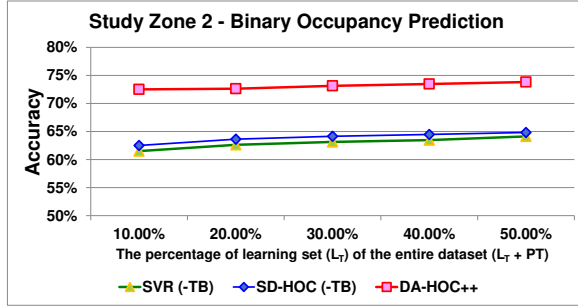


Figure 5.17: Binary Occupancy Prediction Results for the Study Zone 2 Dataset.

Figure 5.18: Occupancy Counting Accuracy Results for the Study Zone 2 Dataset.

5.6.3 Experimental Results

5.6.3.1 Domain Adaptation for Cinema Dataset

Starting from day one in the learning task target data (L_T), we split the data into two sections. We used the L_T data from day 1 to predict the next 13 days (PT), and compared this baseline result with our DA-HOC++ method. This result is line one of Table 5.3. We then used L_T data from two days to predict 12 days (PT), and compared this in line two of Table 5.3. This step was repeated until learning and prediction data had a 50/50 split.

For Figures 5.13 and 5.14, we calculated $\frac{L_T}{L_T + PT}$ and presented this as a percentage, to visualise the comparison of binary occupancy and occupancy counting prediction for the varying dataset predictions from $\frac{1}{1+13}$ (equal to 7.14%) to $\frac{7}{7+7}$, which corresponds to a

Table 5.6: SVR (+TB), SD-HOC (+TB), SVR (-TB), SD-HOC (-TB) and DA-HOC++ Human Binary Occupancy and Occupancy Counting Prediction Accuracy Results for the Classroom 1 Dataset.

Classroom 1	Non Domain Adaptation		Domain Adaptation		
	Binary Occupancy Prediction				
$L_T - PT \left(\frac{L_T}{L_T + PT} \right)$	SVR (+TB) [TB + L_T ->PT]	SD-HOC (+TB)	SVR (-TB) [L_T ->PT]	SD-HOC (-TB)	DA-HOC++ [L_S + L_T ->PT]
1 - 9 (10.00%)	81.51%	82.33%	59.14%	61.77%	68.16%
2 - 8 (20.00%)	82.63%	84.17%	59.54%	62.93%	69.55%
3 - 7 (30.00%)	83.44%	84.81%	60.22%	63.67%	70.18%
4 - 6 (40.00%)	84.50%	85.08%	61.72%	64.04%	71.48%
5 - 5 (50.00%)	84.67%	85.69%	62.15%	64.88%	72.06%
Occupancy Counting Prediction					
1 - 9 (10.00%)	68.44%	70.11%	43.73%	46.75%	55.74%
2 - 8 (20.00%)	69.03%	71.13%	44.88%	47.91%	56.88%
3 - 7 (30.00%)	69.81%	71.95%	45.07%	48.12%	56.92%
4 - 6 (40.00%)	70.45%	72.46%	45.45%	48.51%	57.46%
5 - 5 (50.00%)	71.39%	73.45%	46.21%	49.18%	59.19%

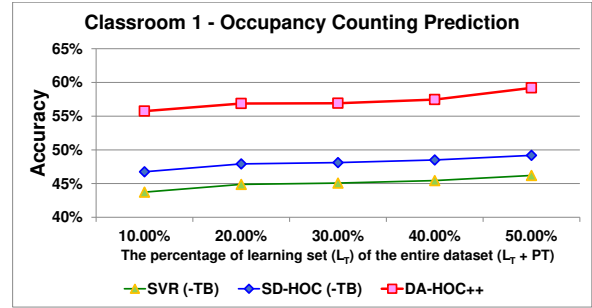
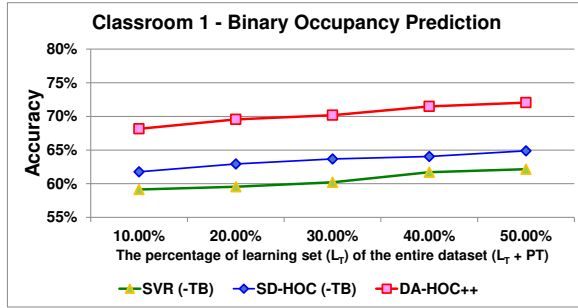


Figure 5.19: Binary Occupancy Prediction Results for the Classroom 1 Dataset.

Figure 5.20: Occupancy Counting Accuracy Results for the Classroom 1 Dataset.

50/50 split between the learning task target and prediction target data.

The binary occupancy prediction accuracy results are shown in Table 5.3 and Figure 5.13. The accuracy of SD-HOC(+TB) ranges from 88.58% - 95.84% (see Table 5.3 column 2) and is the highest compared to other algorithms. The accuracy of SVR(+TB) is the second best (see Table 5.3 column 1). DA-HOC++ is the most accurate algorithm (compare Table 5.3 column 5 with column 3 and 4) if the TB dataset is not available. DA-HOC++'s performance is better by 12.29% in comparison to SVR and 10.14% in comparison to SD-HOC. Overall, SD-HOC is more accurate than SVR and this result agrees with previous research as shown in Chapter 3. The accuracy increases for the larger learning dataset, higher L_T .

The occupancy counting prediction accuracy results are shown in Table 5.3 and Figure 5.14. Any algorithms with TB as part of their training dataset result in higher accuracy. Without

Table 5.7: SVR (+TB), SD-HOC (+TB), SVR (-TB), SD-HOC (-TB) and DA-HOC++ Human Binary Occupancy and Occupancy Counting Prediction Accuracy Results for the Classroom 2 Dataset.

Classroom 2	Non Domain Adaptation		Domain Adaptation		
	Binary Occupancy Prediction				
$L_T - PT \left(\frac{L_T}{L_T + PT} \right)$	SVR (+TB) [TB + $L_T \rightarrow PT$]	SD-HOC (+TB)	SVR (-TB) [$L_T \rightarrow PT$]	SD-HOC (-TB)	DA-HOC++ [$L_S + L_T \rightarrow PT$]
1 - 9 (10.00%)	87.46%	89.58%	62.11%	65.11%	72.70%
2 - 8 (20.00%)	87.56%	90.10%	62.64%	66.64%	73.57%
3 - 7 (30.00%)	88.30%	90.41%	63.68%	66.68%	74.43%
4 - 6 (40.00%)	88.41%	91.47%	64.45%	67.45%	74.86%
5 - 5 (50.00%)	89.09%	91.86%	65.86%	67.86%	75.25%
Occupancy Counting Prediction					
1 - 9 (10.00%)	72.16%	75.58%	48.74%	50.74%	60.35%
2 - 8 (20.00%)	72.43%	76.08%	48.84%	50.88%	60.72%
3 - 7 (30.00%)	72.46%	76.28%	49.92%	50.92%	60.90%
4 - 6 (40.00%)	73.05%	76.88%	49.96%	51.46%	61.20%
5 - 5 (50.00%)	73.16%	77.85%	50.19%	51.59%	61.30%

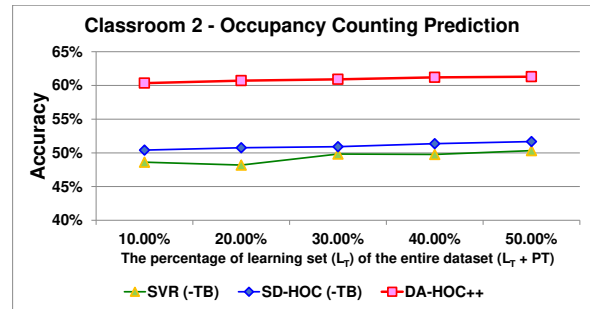
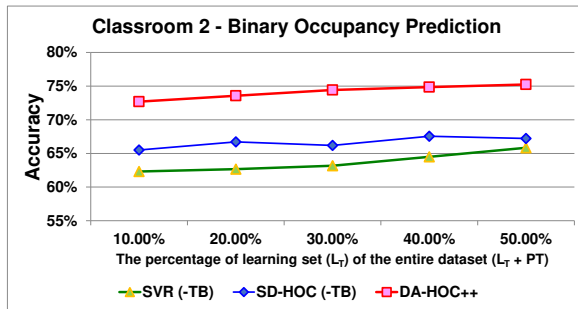


Figure 5.21: Binary Occupancy Prediction Results for the Classroom 2 Dataset.

Figure 5.22: Occupancy Counting Accuracy Results for the Classroom 2 Dataset.

using the TB dataset, DA-HOC++ had the highest accuracy, averaging between 59.45% and 63.75% with ± 5 people error tolerance (see Table 5.3 column 5). Even though the accuracy seems low, predicting the number of occupants between 0 and 230 with more than 60% accuracy is acceptable.

There were two periods for the empty room prediction in the cinema theatre. The first was from midnight to morning, when the cinema theatre is closed. Our ZPA method covers this interval. The second period was between cinema sessions, when the audience for the previous screening leaves the cinema before the next audience enters. Some of our predictions returned negative occupancy values (up to -19 persons). Due to this inaccuracy, we forced our model to convert each negative prediction into zero occupancy. The accuracy of the binary occupancy prediction was higher than that of the occupancy counting prediction. This result is expected as

Table 5.8: The Accuracy, Precision, Recall and F-Score for Cinema Dataset.

$L_T - PT \left(\frac{L_T}{L_T + PT} \right)$	Binary Occupancy Prediction			
	Accuracy	Precision	Recall	F-score
1 - 13 (7.14%)	67.85%	70.02%	59.33%	64.23%
2 - 12 (14.29%)	68.84%	70.38%	61.46%	65.62%
3 - 11 (21.43%)	67.58%	70.91%	63.45%	66.97%
4 - 10 (28.57%)	67.79%	70.96%	64.64%	67.65%
5 - 9 (35.71%)	71.19%	71.48%	66.56%	68.93%
6 - 8 (42.86%)	73.24%	71.87%	68.77%	70.29%
7 - 7 (50.00%)	75.42%	73.93%	70.21%	72.02%
Occupancy Counting Prediction				
1 - 13 (7.14%)	59.45%	53.04%	50.91%	51.95%
2 - 12 (14.29%)	59.84%	55.13%	53.41%	54.26%
3 - 11 (21.43%)	59.19%	56.29%	54.12%	55.18%
4 - 10 (28.57%)	59.29%	58.56%	56.14%	57.32%
5 - 9 (35.71%)	62.02%	59.55%	56.30%	57.88%
6 - 8 (42.86%)	62.51%	59.96%	56.57%	58.22%
7 - 7 (50.00%)	63.75%	61.84%	58.14%	59.93%

predicting whether the room is vacant or not is more straightforward than predicting the exact number of people inside. The significance of this research is that any large room utilisation prediction can be conducted with a negligible amount of training data.

5.6.3.2 Domain Adaptation for Study Zones Dataset

Both of the study rooms were the smallest target domain in our experiment, with a size of 125 m² per room. Both SVR(+*TB*) and SD-HOC(+*TB*) returned the highest accuracy, a result from the cinema theatre dataset experiment that only arose because there was enough historical data. Without adequate historical data, DA-HOC++ is the most accurate domain adaptation technique. On average, the DA-HOC++ resulted in the highest binary prediction accuracy, with 71.95% for study zone 1 and 73.12% for study zone 2 (Tables 5.4 and 5.5). Similar information can also be seen in Figures 5.15 and 5.17. It is interesting to note that in study zone 2, the baselines SVR(-*TB*) and SD-HOC(-*TB*) had a similar accuracy, whereas in study zone 1, SD-HOC(-*TB*) had a better accuracy than SVR(-*TB*). For the occupancy counting, the overall trend between study zones 1 and 2 is similar, as shown in Figures 5.16 and 5.18.

5.6.3.3 Domain Adaptation for Classrooms Dataset

For the classrooms dataset experiment, the overall result was aligned with the other cinema and study zones datasets. The non domain adaptation technique had the highest accuracy

Table 5.9: The Accuracy, Precision, Recall and the F-Score for Study Zones 1 and 2.

$L_T - PT \left(\frac{L_T}{L_T + PT} \right)$	Study Zone 1				Study Zone 2			
	Binary Occupancy Prediction				Occupancy Counting Prediction			
	Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
1 - 9 (10.00%)	70.11%	64.59%	61.08%	62.79%	72.51%	67.20%	65.12%	66.14%
2 - 8 (20.00%)	71.64%	65.96%	61.23%	63.51%	72.64%	69.57%	65.37%	67.40%
3 - 7 (30.00%)	71.68%	66.67%	62.25%	64.38%	73.15%	69.59%	65.46%	67.46%
4 - 6 (40.00%)	72.45%	69.07%	62.26%	65.49%	73.47%	70.00%	66.58%	68.25%
5 - 5 (50.00%)	73.86%	70.41%	63.25%	66.64%	73.82%	70.12%	66.71%	68.37%
1 - 9 (10.00%)	58.71%	55.07%	53.08%	54.06%	59.81%	56.47%	50.63%	53.39%
2 - 8 (20.00%)	58.96%	55.43%	54.82%	55.12%	60.99%	56.48%	53.11%	54.74%
3 - 7 (30.00%)	59.32%	57.74%	55.52%	56.61%	61.27%	57.07%	55.29%	56.17%
4 - 6 (40.00%)	59.81%	59.01%	57.06%	58.02%	62.63%	59.53%	57.73%	58.62%
5 - 5 (50.00%)	60.15%	60.49%	59.55%	60.02%	62.94%	61.93%	59.10%	60.48%

prediction in SVR(+ TB) and SD-HOC(+ TB), as shown in Tables 5.6 and 5.7. Other than the cinema dataset, classroom 1 had the most occupants, up to 67 people. This is the main reason why the prediction accuracy in classroom 1 suffered compared to other locations like classroom 2 and both study zones (Figures 5.19 and 5.20). On the other hand, classroom 2 had the best accuracy prediction, as shown in Figures 5.21 and 5.22. A possible reason for this is that the dataset between CO₂ and occupants showed a strong positive correlation. Another reason is that, even though the room size was bigger than the study zone (139 m² compared to 125 m²), the number of occupants in classroom 2 was relative low throughout the experiment period, reaching a maximum of only 39 people; classroom 1 was more populated, with a maximum of 67 occupants.

5.6.4 Evaluation Metrics

From the results in section 5.6.3, our new model, DA-HOC++ has a better overall accuracy of domain adaptation prediction in multiple domains. We added three evaluation metrics to ensure that this high prediction accuracy means a better model.

$$Precision = \frac{TP}{TP + FP} \quad (5.13)$$

The first evaluation metric is precision. Precision is the ratio of correctly predicted positive observations (TP) to the total predicted positive observations ($TP+FP$), as shown in Equa-

Table 5.10: The Accuracy, Precision, Recall and F-Score for Classrooms 1 and 2.

$L_T - PT \left(\frac{L_T}{L_T + PT} \right)$	Classroom 1				Classroom 2			
	Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
	Binary Occupancy Prediction							
1 - 9 (10.00%)	68.16%	68.31%	66.73%	67.51%	72.70%	67.93%	66.69%	67.30%
2 - 8 (20.00%)	69.55%	69.32%	67.15%	68.22%	73.57%	70.34%	68.19%	69.25%
3 - 7 (30.00%)	70.18%	70.25%	67.83%	69.02%	74.43%	71.36%	69.23%	70.28%
4 - 6 (40.00%)	71.48%	70.30%	70.17%	70.23%	74.86%	73.49%	71.31%	72.38%
5 - 5 (50.00%)	72.06%	71.54%	70.27%	70.90%	75.25%	74.00%	73.67%	73.83%
	Occupancy Counting Prediction							
1 - 9 (10.00%)	55.74%	55.34%	53.77%	54.54%	60.35%	57.86%	56.30%	57.07%
2 - 8 (20.00%)	56.88%	56.69%	56.04%	56.36%	60.72%	59.72%	56.49%	58.06%
3 - 7 (30.00%)	56.92%	58.69%	58.44%	58.56%	60.90%	60.55%	56.57%	58.49%
4 - 6 (40.00%)	57.46%	59.71%	60.24%	59.97%	61.20%	60.67%	57.69%	59.14%
5 - 5 (50.00%)	59.19%	61.78%	60.76%	61.27%	61.30%	60.83%	59.06%	59.93%

tion 5.13. The results from the cinema dataset prediction in Table 5.8 show that, for binary occupancy prediction, the precision was relatively high, ranging from 70.02% to 73.93%.

The evaluation metrics for study zones 1 and 2 are shown in Table 5.9. Binary occupancy prediction showed a high precision, ranging from 65% to 70%. A similar trend was apparent for classrooms 1 and 2 (Table 5.10). Although the precision for occupancy counting prediction was lower than for binary occupancy prediction, it was still greater than 55%, which is encouraging.

The second evaluation metric was recall, which is the ratio of correctly predicted positive observations (TP) to all the observations in the actual class ($TP+FN$), as shown in Equation 5.14.

$$Recall = \frac{TP}{TP + FN} \quad (5.14)$$

Recall values were slightly weaker than precision (Tables 5.8, 5.9 and 5.10). However, the value above 50% is still acceptable. The binary occupancy prediction of study zone 1 showed the lowest overall recall value.

The final evaluation metric is the F-score which is the weighted average of precision and recall, as shown in Equation 5.15. This score takes both false positives (FP) and false negatives (FN) into account.

$$F - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (5.15)$$

Overall, the F-score was stable and similar to the respective prediction accuracies (Tables 5.8, 5.9 and 5.10). Based on the three evaluation metrics, we can confidently conclude that this model is highly accurate.

5.7 Conclusion

Human occupancy counting research is useful in many areas, such as space and room utilization, energy consumption reduction, human comfort, and security. In some cases, a proper training dataset to build a robust human occupancy counting model could not be obtained. DA-HOC++ is the latest state-of-the-art domain adaptation technique to predict indoor human occupancy, using a minimum amount of training data and leveraging the knowledge of prediction models from other content. Domain adaptation allows us to transfer classification knowledge into a new domain.

DA-HOC++'s binary prediction accuracy of up to 75.34% with minimal training data is encouraging. Even though DA-HOC++ performed on average 20% less accurately for binary prediction and 15% less accurately for occupancy counting, it was the best model when the right amount of historical data for the target domain could not be obtained. DA-HOC++ is also the optimal option for making occupancy predictions for a newly constructed building, as the building will have no historical data to draw on.

Chapter 6

THERMO: Thermal Comfort Prediction and Adjustment in Shared Office Environments

6.1 Introduction

Understanding indoor comfort for the occupants of a building is paramount. By ensuring that their comfort is maintained at a satisfactory level results in a high level of work productivity [Akimoto et al., 2010]. Some studies suggest that the degree of productivity has a direct relationship to a workers' level of thermal comfort [Wyon, 2004, Feige et al., 2013]. Studying and working in an uncomfortable environment reduces not only productivity but also affects the overall health of the worker [Jiang et al., 2011]. The American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE) defines the thermal comfort of an occupant as a mental state of satisfaction with regards to the surrounding thermal environment [A.S.H.R.A.E., 2013].

Reports in the literature in ubiquitous computing fields addressing thermal comfort is limited to:

- **Cost reductions:** energy savings and reduction in costs while maintaining user comfort level [Giaccone et al., 2012, Lam et al., 2014, Sarkar et al., 2016, Winkler et al., 2016]. Most research in this area focuses on monitoring and reducing energy consumption without compromising the occupants' thermal comfort.

- **Comfort booster:** increasing the overall occupants' thermal comfort [Fanger and Toffum, 2002, Homod et al., 2012, Sunnam et al., 2015]. The most popular method predicts the mean consensus and percentage of dissatisfied occupants (Subsection 6.2.1).
- **Health and productivity:** maintain health, wellbeing and productivity [Wyon, 2004, Matzarakis and Amelung, 2008, Laforteza et al., 2009, Feige et al., 2013, Ortiz et al., 2017]. A good thermal comfort building correlates positively with the occupants' health and wellbeing, which also improves their quality of work.
- **Feedback mechanism:** feedback mechanism technique from the users to fine-tune the HVAC (Heating, ventilation and air conditioning) system [Winkler et al., 2016, Shin et al., 2017]. This is a direct approach with every system adjustment translating back to the occupants, who give feedback via a survey. However, this method requires many users' personal information and continuous engagement of users.

Adaptive thermal comfort research has been done for the last 20 years [Brager and De Dear, 1998, De Dear et al., 1998] but there is little research reported in the literature focused on autonomous prediction of thermal comfort using the help of artificial intelligence and machine learning technology. If general thermal comfort and sensation could be predicted accurately, improving overall comfort would be feasible with minimal effort. Currently, it is possible to determine occupants' perceived thermal comfort and automate thermal predictions and adjustments using machine learning techniques.

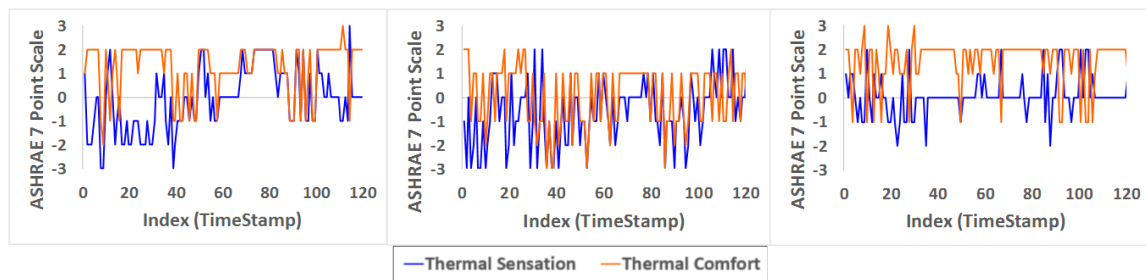


Figure 6.1: ASHRAE 7 Point Scale plots of Friend Center for three participants survey results regarding their **general thermal comforts** and **thermal sensations**.

The advances in ambient sensor technology enable the improvement of overall thermal comfort for occupants. A variety of ambient sensors are becoming affordable. Smart homes and buildings install different types of sensors based on their purposes. Thermometer (temperature

sensor) measures the ambient temperature. Hygrometer (humidity sensor) measures the level of water vapour in the air. Data from ambient sensors is vital in understanding the occupants' overall thermal comfort. For example, if the room is too hot decreasing the occupants' thermal comfort, a thermometer can be used to detect this. Actuator technology used in smart buildings are designed and programmed to recognise and react by decreasing the temperature automatically, thereby maintaining a stable overall thermal comfort.

Today, temperature control of the majority of public buildings (e.g., university classrooms, offices) is based on a standard of an average of people's comfort level. The U.S.A. Federal Occupational Safety and Health Administration (OSHA) recommends that the temperature should be maintained between 20-24.5°C, which is considered to be comfortable¹. In reality, people have diverse comfort preferences depending upon different factors, such as, age, gender, weight and nationality [Yamtraipat et al., 2005, Kingma and van Marken Lichtenbelt, 2015]. Furthermore, a predetermined temperature control system is not necessarily a better solution, as there is no guarantee that the majority of occupants are comfortable [Lam et al., 2014]. A snapshot of standard general thermal comfort and sensation for three different individuals is shown in Figure 6.1. Extreme thermal sensation (too hot or cold) causes the general thermal comfort to decrease, from comfortable to uncomfortable. However, sometimes this is not the case, as some occupants prefer cooler or hotter temperatures.

It is better for the system to predict the most suitable thermal comfort for its occupants based on their voting history compared to adopting an HVAC system from user feedback afterwards. If the system can pre-empt how users would react with regards to their thermal comfort and adjust the temperature accordingly in advance, the users will experience thermal comfort without requiring action. In this chapter, a THERMO, thermal comfort and sensation prediction in shared office environments is proposed. THERMO model was based on both ambient sensor data (such as indoor and outdoor temperatures, humidity and air velocity), occupants' background and daily surveys. THERMO implements sparse non-negative matrix factorisation as part of a task driven approach for de-noising ambient sensor data. A dual-layer model structure is designed to build a model. The first layer is a pre-model that uses the technique from chemometrics, an ensemble rotation forest. The second layer focuses on the continuity of ambient sensors and uses partial least squares for classification to predict the occupant's thermal comfort and sensation. Using sparse non-negative matrix factorisation as the de-noising method, ensemble rotation forest and partial least squares for classification,

¹https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_table=INTERPRETATIONS&p_id=24602

as double predictive models, have not been explored in the study of thermal comfort in the reported literature, to the best of the author's knowledge.

Furthermore, THERMO also has adjustment functionality. Once set features are optimised, THERMO can determine the optimal setting to maintain the comfort level of the occupants. This predictive method was tested against multiple machine learning techniques from the literature, including random forest, SARIMA, decision tree, logistic regression, multilayer perceptron, naïve Bayes and support vector machine.

In this study, a model is proposed which predicts the thermal comfort of occupants in a shared office environment. The thermal analysis category is divided into two groups, general thermal comfort and thermal sensation. General thermal comfort is a mental state of satisfaction with regards to the surrounding thermal environment [A.S.H.R.A.E., 2013]. General thermal comfort ranges from 1 (very uncomfortable) to 6 (very comfortable). Thermal sensation is the apparent sensation that people feel regarding the thermal environment in their surroundings. Thermal sensation ranges from -3 (very hot) to +3 (very cold), based on the ASHRAE 7 point scale. There is little correlation between general thermal comfort and thermal sensation, as individuals have different preferences in conditions of comfort.

6.1.1 Research Motivation

There are three aims for this study. The first aim is to maintain the health and wellbeing of each occupant in a shared office environment [Matzarakis and Amelung, 2008, Ortiz et al., 2017]. The better the perception of thermal comfort of each occupant, the higher overall health quality of the individual [Laforteza et al., 2009]. This is important due to the psychological benefits that often results in a healthier physical body [Grossman et al., 2004, Tugade et al., 2004]. Hence, maintaining health and wellbeing from both a physical aspect of thermal comfort as well as their psychological level of comfort is crucial.

The second is to improve the occupants' productivity via a comfortable and well balanced ambient thermal working environment. It is generally accepted that thermal comfort can boost individual productivity [McCartney and Humphreys, 2002, Akimoto et al., 2010]. Knowing the occupants' thermal comfort level and the ability to adjust it accordingly in BMS is beneficial. Further development to automate the HVAC would result in occupants not needing to do anything to maintain thermal comfort, thus increasing and maintaining productivity.

The third aim is to reduce cost. Many buildings and offices set their HVAC system at a fixed temperature, based on the time of day/year and outdoor temperature [Zhuang et al.,

2017]. Sometimes, this setting is not a suitable temperature for the occupants. Knowing the occupants' thermal comfort preferences could lead to a reduction in energy consumption, as the HVAC system does not need to make the building as cold or hot as the defined, standard temperature [Sarkar et al., 2016, Winkler et al., 2016]. This optimisation is not only for temperature, but also for setting the relative humidity and air velocity. This is only possible if the system considers the occupants' thermal comfort.

6.1.2 Research Contribution

The main contributions of this chapter are two-folds. The first contribution is a novel technique to predict occupants' general thermal comfort and thermal sensation using THERMO. The second contribution is a one-step-ahead further adjustment technique to optimise the HVAC system to maximise the occupants' general thermal comfort. Together, the user comfort is predicted and maintained.

6.2 Background and Related Works

There are many reports in the literature which study the thermal comfort building occupants [Saelens et al., 2011, Giaccone et al., 2012, Sunnam et al., 2015]. To date, there is no single standard that can be used to satisfy the preference for all occupants. The level of comfort varies between individuals, which can change based on multiple factors that influence the occupants' standard of comfort level. The gender of an occupant can affect whether an individual perceives a room as too cold or comfortable [Karjalainen, 2012]. Human sensations of comfort are more complex than other perceptions, such as acoustical, visual stimuli or air quality environments [Ortiz et al., 2017]. Comfort is a reaction to the environment that is strongly influenced by cognitive and behavioural processes. Several main features related to human thermal comfort are shown in Figure 6.2.

Despite half a decade of thermal comfort research and analysis [Fanger et al., 1970], adapting acceptable temperatures to the desired comfort level for each occupant of a building still presents multiple challenges. Many systems enable occupants to give their feedback with regards to the temperature of a room. The complexity comes in integrating the users' feedback into the HVAC system of the building. In the literature, there are many approaches to address this challenge. For example, systems have been proposed to obtain feedback from people automatically by determining their thermal profiles using information about their gender, age,



Figure 6.2: Some of the Factors influencing Thermal Comfort.

height and weight [Byrne et al., 2005], or manually by enabling them to vote using their smartphones [Hang-yat and Wang, 2013, Lam et al., 2014]. Different mechanisms have also been proposed to translate user preference into the appropriate HVAC settings [Feldmeier and Paradiso, 2010, Erickson and Cerpa, 2010, 2012]. All of the solutions require continuous feedback and communication between the occupants and the building. Using THERMO and the advantages of machine learning analysis and process automation, a technique was developed to improve indoor thermal comfort that does not require user feedback.

In this section, we divide it into two subsections. The background study of indoor thermal comfort is presented in subsection 6.2.1. Multiple research in indoor thermal comfort-related research is covered extensively in subsection 6.2.2.

6.2.1 Background Study of Indoor Thermal Comfort

For some time the study of indoor thermal comfort focused on calculating both the predicted mean vote (PMV) and predicted percentage of dissatisfied (PPD) [Fanger et al., 1970]. PMV considers both environmental and physiological factors using an equation. The general formula for PMV is shown in Equation 6.1.

$$\begin{aligned}
 PMV = & [0.303e^{-0.036M} + 0.028] \{ (M - W) - 3.96E^{-8} f_{cl} [t_{cl} + 273]^4 - (t_r + 273)^4 \\
 & - f_{cl} h_c (t_{cl} - t_a) - 3.05 [5.73 - 0.007(M - W) - p_a] - 0.42 [(M - W) \\
 & - 58.15] - 0.0173M(5.87 - p_a) - 0.0014M(34 - t_a) \}
 \end{aligned} \tag{6.1}$$

PMV	The predicted mean vote
e	Euler's number (2.718)
f_{cl}	clothing factor
h_c	convective heat transfer coefficient
l_{cl}	clothing insulation [clo]
M	metabolic rate [W/m^2] 115 for all scenarios
p_a	vapour pressure of air [kPa]
R_{cl}	clothing thermal insulation
t_a	air temperature [$^{\circ}C$]
t_{cl}	surface temperature of clothing [$^{\circ}C$]
t_r	mean radiant temperature [$^{\circ}C$]
V	air velocity [m/s]
W	external work (assumed = 0)

Individual data of the predicted mean vote (PMV) [Fanger and Toftum, 2002] was calculated from a large group of parameters. The parameters included a combination of air temperature, mean radiant temperature, relative humidity, air velocity, metabolic rate and clothing insulation. Zero was the ideal value, as it represented thermal neutrality. The comfort zone was defined by the combination of the six parameters for which the PMV was within the recommended limits, in between negative and positive half points ($-0.5 < PMV < +0.5$).

PPD could only be calculated after PMV as it was derived from PMV. PPD predicted the percentage of occupants that would be dissatisfied with the thermal conditions. As the PMV value moves further from 0 or neutral, the PPD value increases. The maximum number of people dissatisfied with their comfort conditions was 100%, and the recommended acceptable PPD range for thermal comfort from ASHRAE 55 is less than 10% of people dissatisfied for an interior space ². The general formula for PPD is shown in Equation 6.2.

$$PPD = 100 - 95_e^{[-(0.3353PMV^4 + 0.2179MPV^2)]} \tag{6.2}$$

PPD The predicted percentage of dissatisfied

²[https://osr.ashrae.org/Public%20Review%20Draft%20Standards%20Lib/Add-55-2004-d-PPR1-Draft%20\(chair-approved\).pdf](https://osr.ashrae.org/Public%20Review%20Draft%20Standards%20Lib/Add-55-2004-d-PPR1-Draft%20(chair-approved).pdf)

PMV and PPD are good indicators of general human indoor thermal comfort. To obtain an accurate value of PMV, multiple parameters were required, including clothing factor, clothing insulation, personal metabolic rate, vapour pressure of air, clothing thermal insulation, air temperature, surface temperature of clothing, mean radiant temperature, air velocity and external work value. In most cases, obtaining all of these parameters was impossible due to a variety of reasons. To address this and with the advancement in technology, the possibility of predicting the general thermal comfort and thermal sensation using only ambient sensor data was explored.

6.2.2 Related Works

Some thermal comfort-related studies focus more on energy savings while attempting to improve occupants' comfort level [Sarkar et al., 2016, Winkler et al., 2016].

PreHeat system aims to more efficiently heat homes using occupancy sensing and prediction to control home heating automatically [Scott et al., 2011]. MissTime (the time that the house was occupied but not warm) is decreased by a factor of 6-12 using Preheat. The Nest is a smart thermostat that utilises machine learning, sensing, and networking technology, as well as eco-feedback features [Yang and Newman, 2013]. The Nest study tested 23 participants, mostly tech-savvy, affluent males, and therefore the participants were biased.

An intermediary communication platform between occupants and the building management system (BMS) has been proposed in the literature [Jazizadeh and Becerik-Gerber, 2012], with the objective to create adaptive end-user comfort management to compensate for the high rate of discomfort in office buildings. The results suggest there is a weak to moderate correlation between ambient temperature, humidity and occupants' preferences. Furthermore, the variation in correlation between different occupants is high.

A smart system called indoor Lighting and Temperature Controller (iLTC), which eliminates the fixed set-points and requirement of additional light sensors, is introduced in the literature [Sarkar et al., 2016]. iLTC provides maximal user comfort to all co-occupants in a shared space by operating set-points more aggressively. Additionally, using the proposed distributed system, the energy consumption of HVAC is reduced by 39%.

A Model Predictive Control (MPC) framework has been developed for optimal HVAC control that minimises energy consumption while maintaining the comfort boundaries of the occupants [Beltran and Cerpa, 2014]. The Blended Markov Chain (BMC) occupancy prediction model predicts thermal load and occupancy of each zone in a building. MPC/BMC control

framework results indicate that an energy savings of 15.5% in summer and 9.4% in winter are achievable.

Genie, a novel software thermostat, is designed and used at a university for 21 months [Balaji et al., 2016]. Due to the clarity of information and wider thermal control provided by Genie, users felt more comfortable in their office environments. Furthermore, the overall energy consumption did not increase or lead to misuse of the HVAC controls.

Analysis of personal thermal comfort systems and SPOT* Personal Environmental Control (PEC) systems designed for rapid and scalable deployment is described in the literature [Rabani and Keshav, 2016]. Intuitive web-based interfaces for user controls enable SPOT* to be installed in approximately 15 minutes. SPOT* The average absolute discomfort reduction by SPOT* is 67% by a typical user compared to the same HVAC system in the absence of SPOT*. Shin introduced fairness in a participatory model to adjust thermal comfort control in smart buildings [Shin et al., 2017]. An aggregation method that ensures fairness is implemented based on voting points of each user.

A field study across 30 UK households over a month used three different smart thermostats that automate heating based on the users' heating preferences and real-time price variations [Alan et al., 2016]. Thematic analysis of the data indicates that the participants had different understandings and expectations of the smart thermostat, and used it in different ways to effectively respond in real-time while maintaining their thermal comfort.

Human-building interaction (HBI), as a new research domain within Human-computer interaction (HCI), exposes the fundamental characteristics of HBI, such as user immersion in the "machine" and extensive space and time scales, and proposes an operational definition of the domain [Nembrini and Lalanne, 2017]. HCI studies include a focus on energy efficiency, comfort and user awareness of building dynamics improvements.

FORCES presents several methods of feedback that use data and environmental interaction in comfort voting applications [Winkler et al., 2016]. Winkler stated that although the use of feedback using comfort voting applications has improved the quality of service, the effects of the application feedback and user interface design has not been investigated previously. FORCES reduces the energy consumption up to 18.99% and increases user satisfaction of thermal conditions.

There are various ways to understand an occupant' thermal comfort. Some study design and use surveys which focus on a combination of thermal comfort and sensation [Lai et al., 2014, Langevin et al., 2015, Martinez-Molina et al., 2017], while others combine the default metrics of thermal comfort and sensation with two other metrics, thermal preference and acceptability

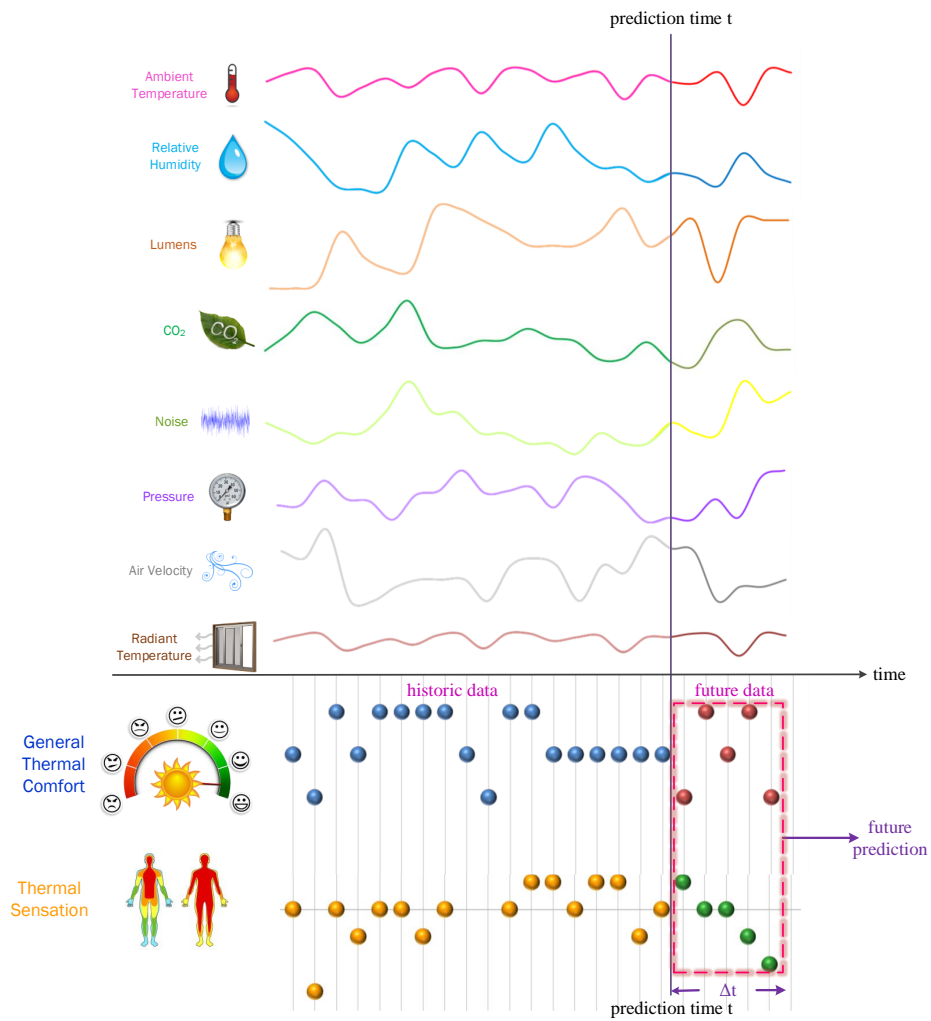


Figure 6.3: Real-time prediction scenario for continuous t showing multiple ambient sensor fluctuations. The fundamental task is to predict both general thermal comfort and thermal sensation at time $t + \Delta t$.

[Luo et al., 2015, Rupp et al., 2018]. In this study, the target prediction was simplified by focusing on general thermal comfort and thermal sensation prediction only.

6.3 Problem Definition

6.3.1 Scenario Assumption

Every human has an individual comfort zone. If C_n is the individual comfort zone where n represents each individual. There is C_1 for individual 1, C_2 for individual 2, C_3 for individual 3 until C_n for individual n .

Assume TS represents the length of a time series and is expressed as $TS = \{ts_1, ts_2, \dots, ts_q\}$, where q is the number of sample points. General thermal comfort for each set of data is represented by TC and thermal sensation by TS . For each feature, we have F^a whereas a is the order of feature. To summarise, in total the three main aspects were:

- General thermal comfort TC , defined as $TC = \{TC_1, TC_2, \dots, TC_q\}$
- Thermal sensation TS , defined as $TS = \{TS_1, TS_2, \dots, TS_q\}$
- Each of the different features based on order number a , F^a , defined as $F^1 = \{F_1^1, F_2^1, \dots, F_q^1\}$ for the first feature, $F^2 = \{F_1^2, F_2^2, \dots, F_q^2\}$ for the second feature, and so on.

6.3.2 Problem Definition

In the prediction of thermal comfort, the predicted mean vote (PMV) and the predicted percentage of dissatisfied (PPD) are the main standards to calculate each occupant's personal thermal comfort. Multiple thermal comfort-related information is required each time these are calculated. This is costly and contains large errors due to the extensive duration of the survey required each time the building occupants' thermal comfort is assessed.

The main problem that is proposed in this chapter is whether the system can predict when the occupants feel uncomfortable with the features surrounding them even before the persons themselves feel uncomfortable. The model builds on multiple features from all ambient sensors, background survey and daily survey data. We predict both occupants' general thermal comfort and thermal sensation by using only ambient sensor data that can be gathered continuously without troubles and annoys the users by another time-consuming survey as shown in Figure 6.3.

This study addresses the question: whether a system can predict the occupants feeling uncomfortable using features surrounding them, before the individuals feel uncomfortable. The model was built on multiple features from all ambient sensors, background surveys and daily survey data. The occupants' general thermal comfort and thermal sensation were predicted

using only ambient sensor data that could be stored continuously without consuming the users time (Figure 6.3).

There were multiple features affecting general thermal comfort (Figure 6.2). To develop a model that predicts both the occupants' general thermal comfort and thermal sensation, there are two problems that need to be addressed:



Figure 6.4: The data flow in THERMO and the detailed source of groups of features covering ambient sensors, daily survey data and background survey with the distribution of different weighted levels.

- Building a pre-model with all ambient sensors, daily surveys and background surveys data as part of THERMO.
- Explore the correlation between each feature, particularly from model feature group and both general thermal comfort and thermal sensation for the main THERMO model.

6.4 Methodology

Using a machine learning technique from the artificial intelligence domain, THERMO was built, which was a prediction and adjustment model for general thermal comfort and thermal sensation in shared office environments. The prediction model included data pre-processing and a dual-layer model design for THERMO. The first layer was a pre-model where multiple different features from numerous sources were integrated together (Figure 6.4). The second model used ambient sensors only as predictor features for the main model. This double layer model was used to address the phenomenon that people do not always behave in a rational way, therefore a linear model would not make accurate predictions.

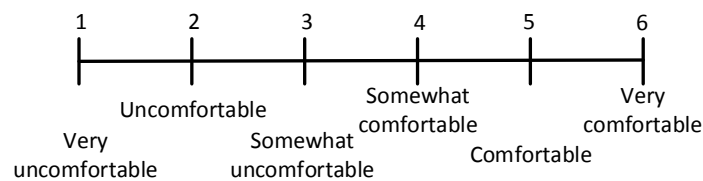


Figure 6.5: The measurement of General Thermal Comfort (From very uncomfortable [1] to very comfortable [6]).

There were two primary prediction targets used to understand occupants' individual thermal comfort. Firstly, general thermal comfort was the level of comfort of each individual with regards to their thermal environment, ranging from 1 to 6. The standard survey used to understand the occupants' thermal comfort is shown in Figure 6.5 [Langevin et al., 2015].

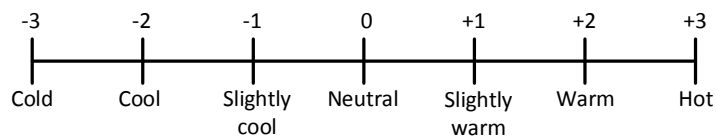


Figure 6.6: The measurement of Thermal Sensation (From cold [-3] to hot [+3]).

For most building managers, their aim is to get as many 5 (comfortable) or 6 (very comfortable) survey results as possible. There are a variety of reasons why users gave a low score, such as an extreme high-temperature day and it became too hot, or an extremely low-temperature day and it became too cold. Other reasons could be humidity issues and wearing cloth that was too bulky or thin. The food and drink that the occupants consumed also affected the

user's general thermal comfort.

The second prediction target was the occupants' thermal sensation. Thermal sensation is the thermal feeling of the user at that moment. It could be too cold or hot, or a pleasant temperature. The standard survey used to identify the occupants' thermal sensation is shown in Figure 6.6.

The ideal value for thermal sensation is close to zero. Unfortunately, some individuals prefer a temperature that is cooler or warmer. Therefore, collecting and understanding a background survey from each individual is crucial. The model develops from the background survey and makes predictions tailored to each user.

Individuals have a relative comfort zone, rather than one absolute value for comfort. A relative comfort zone is a range of pleasant temperatures in which an individual can enjoy and is comfortable in. The outside temperature also contributes to the individuals indoor thermal comfort. Seasonal ranges in temperature causes unconscious awareness and individuals will adjust their indoor thermal comfort closer to this range. As an example, an individual can be satisfied with lower indoor temperature during winter than summer.

6.4.1 The Main Group of Features in THERMO

The THERMO technique combines quantitative data from ambient sensor and daily survey data. To obtain the best prediction results for system automation, additional background surveys to understand each individuals' preference was necessary.

6.4.1.1 Ambient Sensor

Ambient intelligence is an idea where technology could naturally blend with everyday life as shown in Chapter 2. There are many types of ambient data that could be gathered by sensors, such as temperature, relative humidity, illumination, carbon dioxide, sound, pressure, air velocity et cetera. Data from each sensor was differently weighted to the relative thermal comfort. Ambient information differed from indoor and outdoor, so this information is crucial in the prediction of user thermal comfort reaction. Ambient Sensor data was part of features in both the pre-model and main model.

6.4.1.2 Daily Survey Data

To build an accurate model, daily survey data was essential. The technique needs to understand its occupants and daily feedback is the best method of communication between the user and

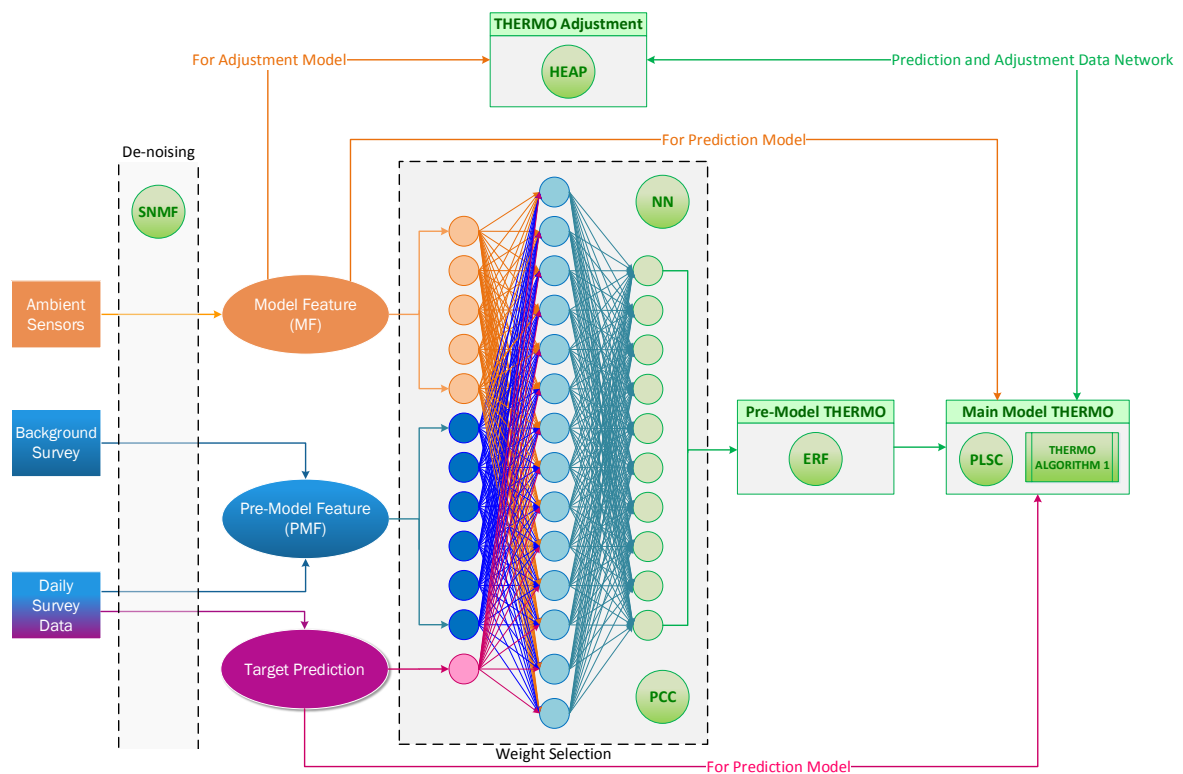


Figure 6.7: The Complete Framework Structure of THERMO.

machine. The content of the survey was also relevant. Every feature was given different weight depending on how directly it affected the current thermal comfort of each user. The required survey data contained current general thermal comfort, current general thermal sensation and other thermal comfort-related data. Daily survey data was one part of the pre-model feature.

6.4.1.3 Background Survey

Background survey for each individual is optional to have for this technique but can improve the model prediction accuracy. Background survey can include information about user tolerance of hot and cold, could be a range of minimum and maximum temperature that is generally acceptable for each user. In a country that has four seasons a year, sometimes these values differ between summer and winter and should be captured separately. General thermal comfort preference and personal clothing are usually part of the background survey. Moreover, anything around each user that can affect individual thermal comfort must be recorded such as fans, windows and doors availability nearby. Each information needs to be weight differently as well.

Table 6.1: Five different level of weighting in THERMO.

Weight level	Weight level category	Weighting value	Sample of features of this weight level
1	Zero contribution	0	Indoor pressure.
2	Minor contribution	0.2	Gender, survey time, floor number, office type, indoor CO2, indoor lumen.
3	Medium contribution	0.5	Door availability, window availability, blinds availability, food availability, drink availability, survey season period, body discomfort (depend on location).
4	Major contribution	0.8	Outdoor ambient temperature, outdoor relative humidity, outdoor air velocity, clothing availability, central air conditioner availability, personal acceptance of thermal sensation.
5	Critical contribution	1.0	Indoor ambient temperature, indoor relative humidity, indoor air velocity, indoor mean radiant temperature, local heater availability, local fan availability, metabolic rates.

Background survey data is only part of the pre-model feature.

Background survey for each individual was optional for this technique but improved the model prediction accuracy. Background survey included information about user tolerance of hot and cold, a range of minimum and maximum temperatures that were generally acceptable by each user. Sometimes these values differed between summer and winter and should be captured separately. General thermal comfort preference and personal clothing were usually part of the background survey. Moreover, anything located in the vicinity of each user that can affect individual thermal comfort was recorded, such as fans, windows and doors. The information was weighted. Background survey data was only used as feature in the pre-model.

6.4.2 Data Pre-processing

The data pre-processing process was divided into two steps: The first step was data de-noising and disaggregation using sparse non-negative matrix factorisation (SNMF). SNMF is one approach for source separation in classification analysis. One of its applications was speaker identification in the signal processing. A detailed description of data cleansing given in the subsection 6.4.2.1. The second assigned the THERMO weight level to the data. Every feature was assigned to one weight level and borrowing the first layer of neural network to find the optimal weighting value. The weight level rules are discussed in subsection 6.4.2.2. The complete framework structure of THERMO is shown in Figure 6.7.

6.4.2.1 Data Cleansing

Data cleansing and de-noising were done using SNMF. The equation used for SNMF in THERMO data cleansing is shown in Equation 6.3. The second and third terms represent the smoothness constraints which aims to penalise large values of the elements in basis and coefficient matrices.

The fourth term represents the sparseness constraint that is imposed on the coefficient matrix using L1 norm minimisation.

$$\min_{B, X} \frac{1}{2} \|D - BX\|_F^2 + \frac{\alpha_1}{2} \|B\|_F^2 + \frac{\alpha_2}{2} \|X\|_F^2 + \frac{\beta}{2} \sum_{i=1}^n \|X(:, i)\|_1^2 \quad (6.3)$$

D	data matrix, $D \approx B \times X$
B	basis matrix, $B \geq 0$
X	coefficient matrix, $X \geq 0$
$\alpha_1, \alpha_2, \beta$	positive constants

6.4.2.2 THERMO Weight Level

Once data cleansing and the de-noising process was performed, the weight assignment was done using neural network. The weighting process is crucial in THERMO, as some features were not relevant and the relevancy of each feature differed based on its contribution to user's thermal comfort. The first layer of neural network technique was used to approximate the weight [Hagan et al., 1996]. Furthermore, THERMO categorises the results into five different levels of weighting for each feature, as shown in Table 6.1. An assign features to categories, the Pearson product-moment Correlation Coefficient (PCC) was used (Equation 6.4).

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \quad (6.4)$$

r	correlation coefficient
x	dataset x
y	dataset y
n	number of sample points

The range of Pearson's r value was from -1 to +1. If the value was >0.7, the correlation between both datasets was strongly positive. A PCC test for each feature was run for general thermal comfort and thermal sensation data. The r value was averaged to obtain the final PCC value for the feature. A detail description of each weight is described below:

6.4.2.2.1 Zero contribution:

the feature was unrelated to human thermal comfort and was assigned the weight value of zero. A feature that had a final average PCC value <0 was assigned to this category.

6.4.2.2.2 Minor contribution:

two different types were assigned to this level. The first one was the feature had a small contribution to thermal comfort. The second type was due to the data being corrupted and data cleansing not being possible. Every feature in the minor contribution level was given a weight value of 0.2. Any feature with a final average PCC value between 0 and 0.2 was assigned to this category.

6.4.2.2.3 Medium contribution:

the feature had an indirect effect on thermal comfort. Surrounding environment information usually was assigned to this group, such as doors or windows nearby. Every feature in the medium contribution level is given a weight value of 0.5. A feature that had a final average PCC value between 0.2 and 0.5 was assigned to this category.

6.4.2.2.4 Major contribution:

the feature had a direct effect on thermal comfort. Some examples for this level were clothing insulation and outdoor temperature. Every feature in the major contribution level was given a weight value of 0.8. A feature that had a final average PCC value between 0.5 and 0.8 was assigned to this category.

6.4.2.2.5 Critical contribution:

the feature had a direct effect on thermal comfort and the data from this level was considered important by each occupant. One example for this level was indoor temperature. Every feature of the critical contribution level was given a weight of 1. A feature that had a final average PCC value between 0.8 and 1 was assigned to this category.

6.4.3 THERMO Prediction Algorithm

The THERMO algorithm was divided into two processes. The first process was to build a pre-model, which was the raw, preliminary model built from the pre-model feature (PMF) as a base model. From this base model, the THERMO main model was implemented. The main model was built using only the model feature (MF) as a learning source and target prediction data, dataset as shown in Figure 6.8.

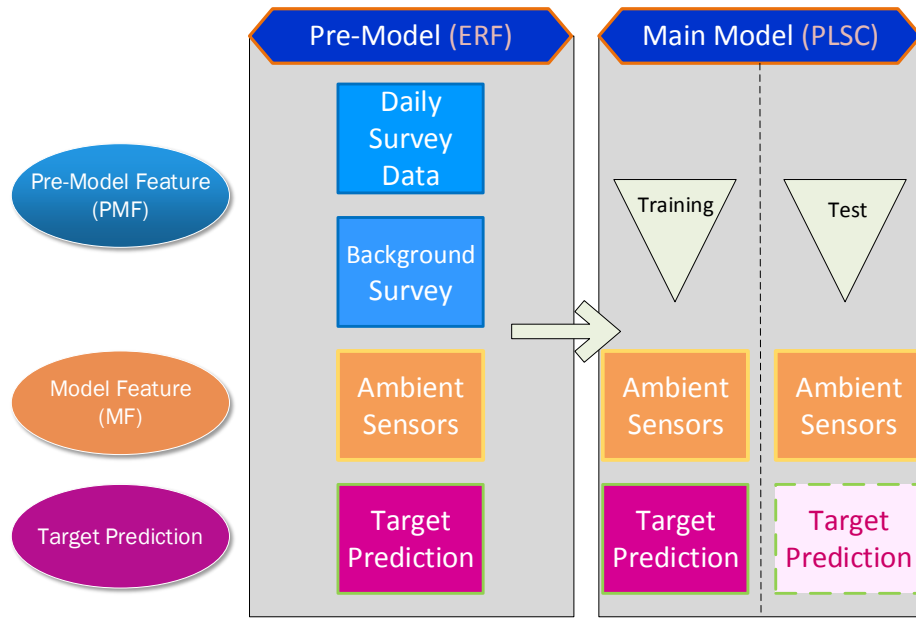


Figure 6.8: Complete THERMO Prediction Algorithm Model.

6.4.3.1 Pre-model

In the THERMO pre-model, all data from different sources were used. PMF group included data from background surveys and daily survey data. Data from MF contained ambient sensor data. Each data set was given corresponding weighting based on its relation to thermal comfort. Data from background survey acted as a foundation. A minimum, median and maximum acceptance of sensation in the ASHRAE 7 point scale was surveyed during different seasons. Clothing, drinks, heater locality, fan locality, thermostat, window, door, blinds and effectiveness were recorded.

Due to the amount of features used as predictors in the thermal comfort domain, THERMO requires a prediction model that can identify the subset of factors that influenced the prediction target and discard factors that were irrelevant but may skew the predictors. Using this definition, the ensemble rotation forest (ERF) technique was used as shown in Equation 6.5, which is fast to compute [Rodriguez et al., 2006].

6.4.3.2 THERMO Main Model

Once the pre-model was established, preparation for the THERMO Main Model was completed. Partial Least Squares was used for classification (PLSC) [35]. The advantage of PLSC was that

$$\mu_j(x) = \frac{1}{L} \sum_{i=1}^L d_{i,j}(xR_i^a), j = 1, \dots, c \quad (6.5)$$

μ	ERF confidence
x	objects in the training dataset (a matrix)
d	probability value
R	rotation matrix
L	number of classifiers in the ensemble

this technique reduced the predictor number to a smaller set of uncorrelated components. It performed partial least squares technique on these components, instead of the original data. The partial least squares technique is widely used in chemometrics, chemical and food areas [Wold et al., 2001]. Use of this method in the thermal comfort area is novel, to the best of the author's knowledge. The logic behind the THERMO Main Model is shown in Algorithm 7.

6.4.3.2.1 THERMO Season

To further understand the model, a seasonal experiment was performed due to thermal comfort analysis being different based on seasons. The model focused on a country that had four distinct seasons in spring, summer, autumn and winter. The model for each season was different due to the way people perceive thermal comfort and thermal sensation differently during warm and cool seasons [Cao et al., 2011]. The average temperature for both indoor and outdoor also differed between individuals.

There was some similarity between the trend in temperature during spring and autumn. Due to this, the model for spring and autumn were identical. It is important to note that there was a significant difference in temperatures during these seasons, due to both being transition periods that could experience extreme temperatures. THERMO does not need any seasonal data to build a prediction model.

6.4.4 THERMO Adjustment Algorithm

THERMO adjustment algorithm model was based on the THERMO prediction algorithm model. The relationship between THERMO adjustment and THERMO prediction can be observed from Figure 6.7. Once the prediction model was complete, THERMO identified the ambient sensors that could be adjusted in the HVAC system. Usually, ranges from ambient temperature, relative humidity, lumens, air velocity and radiant temperature were used. For

Algorithm 7 Partial least squares for classification pseudocode for THERMO (capital letters are matrices, lower case letters are vectors when superscripted and scalars when subscripted).

```

1: procedure PLS1( $X, y, l$ )
2:    $X^{(0)} \leftarrow X$  ▷  $X, Y$ : matrices,  $l$ : length of matrices
3:    $w^{(0)} \leftarrow X^T y / \|X^T y\|$  ▷  $T$ : an orthonormal matrix
4:   for  $k=0$  to  $l-1$  do ▷  $B$ : slope,  $B_0$ : intercept
5:      $t^{(k)} \leftarrow X^{(k)} w^{(k)}$  ▷  $w^{(0)}$ : an initial estimate of  $W$ 
6:      $t_k \leftarrow t^{(k)T} t^{(k)}$ 
7:      $t^{(k)} \leftarrow t^{(k)} / t_k$ 
8:      $p^{(k)} \leftarrow X^{(k)T} t^{(k)}$ 
9:      $q_k \leftarrow y^T t^{(k)}$ 
10:    if  $q_k = 0$  then
11:       $l \leftarrow k$ 
12:      break
13:    end if
14:    if  $k < (l - 1)$  then
15:       $X^{(k+1)} \leftarrow X^{(k)} - t_k t^{(k)} p^{(k)T}$ 
16:       $w^{(k+1)} \leftarrow X^{(k+1)T} y$ 
17:    end if
18:  end for
19:  define  $W$  to be the matrix with column  $w^{(0)}, w^{(1)}, \dots, w^{(l-1)}$ 
20:  Do the same to form the  $P$  matrix and  $q$  vector.
21:   $B \leftarrow W(P^T W)^{-1} q$ 
22:   $B_0 \leftarrow q_0 - P^{(0)T} B$ 
23:  return  $B, B_0$ 
24: end procedure

```

temperature, a range of ± 5 °C, for relative humidity $\pm 10\%$ was used. Lightning adjustment limit was approximately 10 lumens and air velocity 3 m/s. Once the list had been defined, THERMO adjustment algorithm calculated and permuted each combination using heap's algorithm [Berry et al., 2014]. Every time a better result for either general thermal comfort or thermal sensation was obtained, it was recorded in the THERMO's thermal list. Once the permutation was performed, the best combination of ambient sensor data was sent to the HVAC system so the occupants could experience thermal comfort. The higher the general thermal comfort value, the better. Unfortunately, this did not achieve good thermal sensation for every individual as every perceived comfort differently. The adjustment algorithm only achieved improvement in general thermal comfort.

6.5 Evaluation

In this section, we divide it into four different subsections. The first one, subsection 6.5.1 discusses the dataset in depth. Subsection 6.5.2 covers the software tools and hardware machine that we used to execute all the experiments. The third subsection 6.5.3 presents the evaluation matrix and baseline.



Figure 6.9: The Friend Center in Philadelphia, USA.

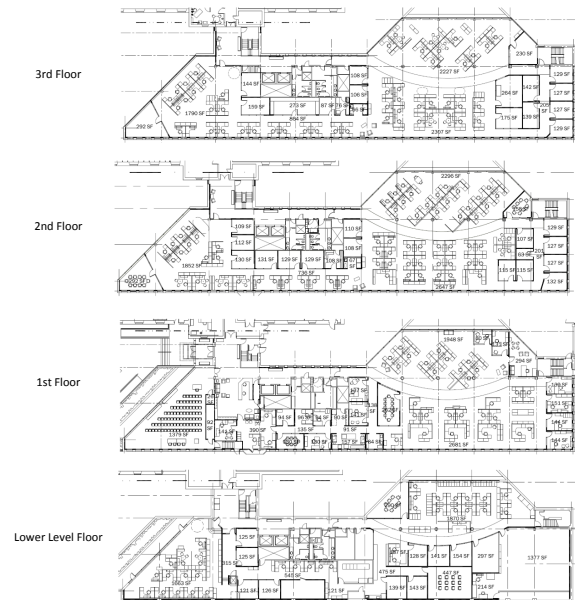


Figure 6.10: The Friend Center Floor Map [Langevin et al., 2015].

6.5.1 Datasets

Comfort related data gathered across one year from the Friend Center in Philadelphia, USA was used (Figure 6.9) [Langevin et al., 2015]. This building has been home to diverse groups working for peace and justice. There is a total of 840,984 data points across 118 features from 24 occupants, collected from four different floors of the building, as shown in Figure 6.10. The distribution was three occupants on the lower level floor, four occupants on the first floor, nine occupants on the second floors and eight occupants on the third floor. Data types vary from sensors and daily surveys across a period of one year. Sensor data, including but not limited to, indoor temperature, relative humidity, air velocity and the concentration of carbon dioxide (CO₂) were used. Daily survey data included personal thermal sensation (too cold or too hot)

and general thermal comfort level. The features of the dataset were categorised into five types of variables, based on how the data was gathered:

1. Background survey (demographic information, office characteristics, thermal comfort and preferences, control options, personal values and typical work schedules);
2. Daily surveys (thermal comfort, thermal sensation, clothing, activity and controls);
3. HOBO data logger (indoor ambient temperature, relative humidity, air velocity, lumens, CO₂ concentration and mean radiant temperature);
4. Friends center BMS (current thermostat cooling and heating setpoint and base thermostat cooling and heating setpoint) and
5. Weather analytics (outdoor ambient temperature, outdoor relative humidity and outdoor air velocity).

Background survey and daily survey data were used in the model without any modification. HOBO data logger, Friends center BMS and outdoor weather analytics were integrated into the ambient sensor category. Each feature was categorised into a corresponding weight category and given a weight accordingly.

6.5.2 Experiment Tool

Waikato Environment for Knowledge Analysis (Weka), Anaconda navigator, Spyder (Python IDE) and R were used to perform the experiment. Weka was used for machine learning algorithms and data analysis. Anaconda navigator and Spyder were used to build models and R were for data integration, analysis and visualisation. The data from R and Spyder was imported into Microsoft Excel for data analysis and visual output enhancement. For image visualisation, Microsoft Visio was also used.

All experiments were run in 64 bit Windows 7 Enterprise service pack 1 operating system, model HP EliteOne 800 G2 23-in NT GPU AiO. The processor was Intel Core i7-6700 CPU 3.40 GHz with an installed memory (RAM) of 16 GB. The total internal hard drive capacity was 512 GB.

Table 6.2: General Thermal Comfort and Thermal Sensation Prediction Accuracy Results from THERMO and other Machine Learning Algorithms.

Machine Learning Technique	General Thermal Comfort	Thermal Sensation
Random Forest	0.581497797	0.506112469
SARIMA	0.455907089	0.327120545
Decision Tree	0.488586304	0.436924309
Support Vector Machine	0.426111334	0.334801762
Multilayer Perceptron	0.542651181	0.391670004
Naïve Bayes	0.498197837	0.439727673
Logistic Regression	0.444933921	0.467761314
THERMO	0.604235489	0.523454465

6.5.3 Evaluation and Baseline

Multiple machine learning techniques were chosen as baselines. Random forest was selected as representative from ensemble learning for classification. Decision tree belongs to standard prediction techniques representing the basic prediction method. Logistic regression is a linear approach modelling between a scalar dependent variable and respective independent variable. The multilayer perceptron is the representative from the artificial neural network family. Naïve Bayes from Bayes family technique was chosen due to it being fast and works well with high dimensions. Support vector machine is an efficient technique for handling high dimensional spaces. Finally, we picked seasonal autoregressive integrated moving average (SARIMA) as the final baseline, representing the ARIMA algorithm family.

For the evaluation, the receiver operating characteristic (ROC) curves were plotted for general thermal comfort and thermal sensation for comparison. ROC curve was created by plotting the true positive rate against the false positive rate at different threshold settings. ROC curve demonstrated multiple findings:

- It demonstrated the trade-off between sensitivity and specificity.
- The closer the curve started from the left-hand border, and continued to the top edge of the ROC, the more accurate the model.
- The closer the curve to the middle 45° diagonal of the ROC space, the less accurate the model.

- The area under the curve was the measure of accuracy of the model. The larger the area, the more accurate the model.

6.6 Experimental Results

This section is divided into four subsections. The first one discusses the result and analysis for general thermal comfort prediction (Subsection 6.6). The second subsection covers thermal sensation prediction result and analysis in Subsection 6.6.1. Subsection 6.6.2 presents an analysis about how season affects human judgement based on the data point of view. The last part, Subsection 6.6.3 covers how THERMO model adjusts indoor ambient sensors to an improved environment surrounding, proven by an overall better general thermal comfort for the occupants.

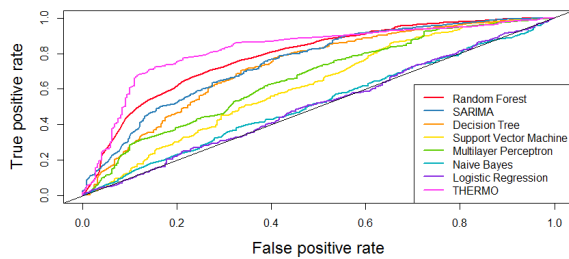


Figure 6.11: ROC Curve Plot of General Thermal Comfort for Various Machine Learning Algorithms.

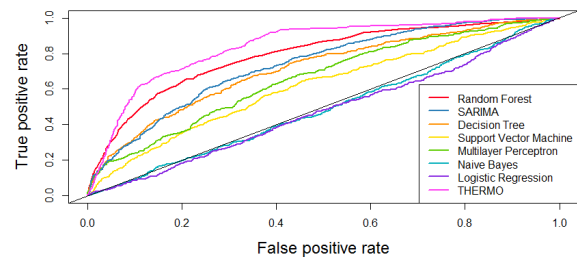


Figure 6.12: ROC Curve Plot of Thermal Sensation for Various Machine Learning Algorithms.

subsectionGeneral Thermal Comfort Results and Analysis Table 6.2 summarises the prediction accuracy results for general thermal comfort and thermal sensation. THERMO had the highest prediction accuracy (60.42%), followed by random forest and multilayer perceptron. The rest of algorithms prediction accuracy were below 50%. From Figure 6.11, the ROC curve indicated that THERMO was the best technique, having the greatest area under the curve. THERMO can predict the occupants' general thermal comfort more accurately than the majority of machine learning techniques. THERMO can detect when individuals were comfortable in their surroundings or when they were uncomfortable.

6.6.1 Thermal Sensation Results and Analysis

Detailed results for the thermal sensation prediction is shown in Table 2. THERMO had a 52.23% prediction accuracy and was the best technique to predict the occupants' thermal sensation in a building. Random forest and logistic regression techniques followed after. Other than THERMO, random forest technique is the only technique that had a prediction accuracy greater than 50%. From ROC curve, THERMO was better compared to other machine learning techniques and had the greatest area under the curve (Figure 6.12). Having an understanding of thermal sensation was laborious compared to general thermal comfort as the algorithm only predicted how comfortable or uncomfortable each occupant was. For thermal sensation, there were three categories. Negative values when occupants felt cold and how cold it was, zero when it was neutral and a positive value when the occupants felt hot and how hot it was.

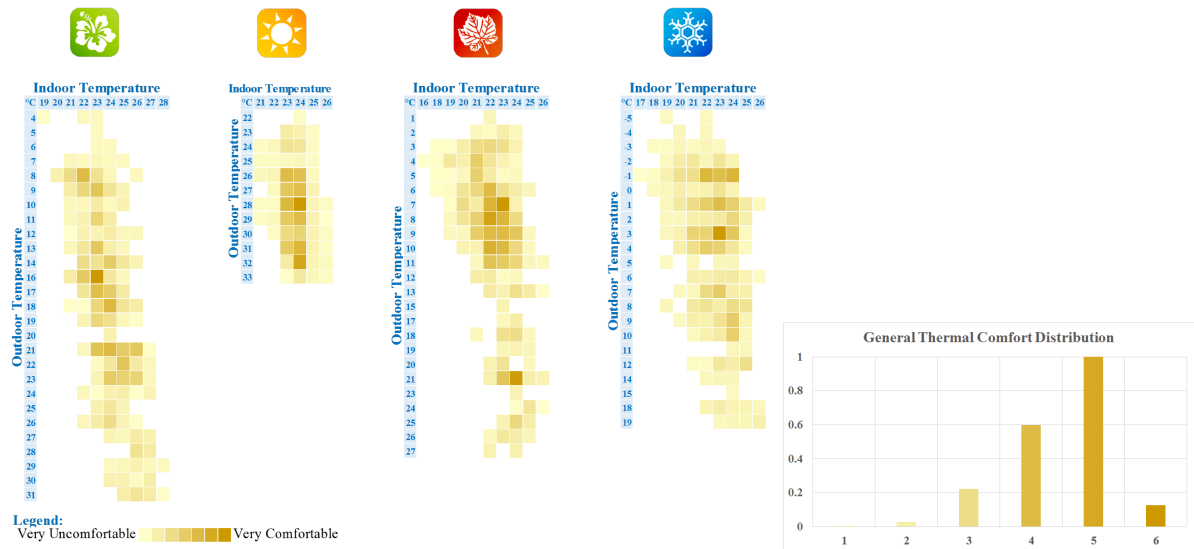


Figure 6.13: The distribution of general thermal comfort by Figure 6.14: General Thermal each season (From the left to the right: Spring, Summer, Comfort Distribution. Autumn and Winter).

6.6.2 Analysis related to Season

To better understand the relationship to seasons, an experiment comparing occupants' thermal comfort and thermal sensation voting relative to outdoor and indoor temperatures. From Figures 6.13 and Figure 6.15, each thermal comfort's preference and their reaction in relation to indoor and outdoor temperature was determined. In general, during Spring, the occupants were

more comfortable with lower temperatures when it was cold outside. The same temperature, for example, 25°C gave different results when outside was 7°C (a bit uncomfortable) or 21°C (comfortable). Figure 6.14 indicates that although the majority of occupants were comfortable with the general thermal comfort, the number of uncomfortable occupants was significant. Figure 6.16 shows that majority of occupants who felt neutral, with fewer in the extreme left (colder) and extreme right (hotter).

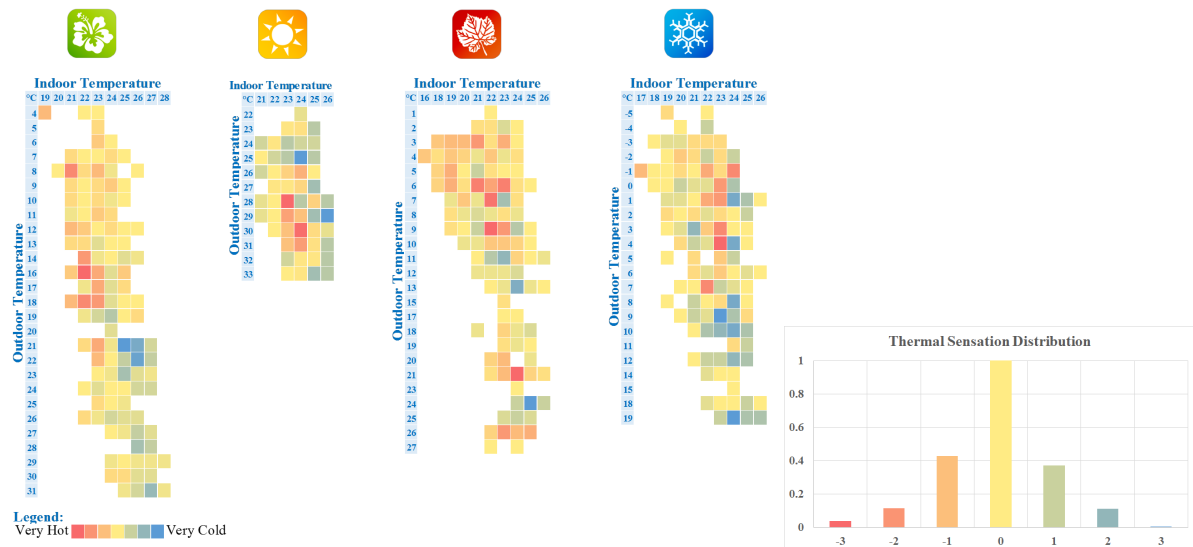


Figure 6.15: The distribution of thermal sensation by each season (From the left to the right: Spring, Summer, Autumn and Winter). Figure 6.16: Thermal Sensation Distribution.

Interestingly, during summer, colder did not always equate to increased comfort. Indoor temperatures between 23 to 24°C was comfortable for most occupants. One or two degrees lower and higher made occupants feel less comfortable. The data suggested that 21°C was too cold for the occupants during summer. THERMO, can make accurate predictions without using seasonal data as an input parameter.

The autumn result was similar to that of the spring results, with a significant temperature gap. During autumn, the occupants preferred an indoor temperature between 22 to 23°C. During winter, the result was different. Fewer high thermal comfort levels were recorded. One significant finding was that during winter on days when the outside temperature was high (>10°C), the majority of occupants were not comfortable. Any adjustment to the HVAC did not increase occupant comfort.

Table 6.3: A General Thermal Comfort Comparison Snapshot between the Original and after Adjustment.

Timestamp (t)	General Thermal Comfort	General Thermal Comfort (Adjusted)	Improvement
18/05/2015 09:36:53 AM	3	3	0
18/05/2015 0:01:55 AM	4	4	0
18/05/2015 10:31:58 AM	4	5	1
18/05/2015 10:41:59 AM	5	5	0
18/05/2015 11:02:01 AM	4	5	1
18/05/2015 11:37:05 AM	4	5	1
18/05/2015 12:07:07 PM	5	5.5	0.5
18/05/2015 12:32:10 PM	5	5	0
18/05/2015 12:37:10 PM	4	5	1
18/05/2015 12:42:11 PM	3	4	1
18/05/2015 12:47:12 PM	2	3	1
18/05/2015 12:52:13 PM	2	2.5	0.5
18/05/2015 01:17:16 PM	1	2	1
18/05/2015 02:02:23 PM	1	2	1
18/05/2015 02:42:28 PM	2	3	1
18/05/2015 03:12:31 PM	3	3	0

6.6.3 Adjustment Results

Even though the adjustment technique is simple, this was a promising improvement as the system could improve the overall experience of the occupants without any action being required by the occupants of the building. There are multiple external aspects that are not yet integrated when we run this experiment such as HVAC response times or environmental changes created by the users (open doors or windows). Furthermore, there are some unpredictable disruptions that could affect the occupants. As the technique is a simple heap algorithm, we implemented a simple evaluation method by comparing the original general thermal comfort value and the new adjusted general thermal comfort. The improvement matrix is shown in Table 6.3. On average, our THERMO adjustment method successfully increases the occupants' general thermal comfort by 0.625 from the original value.

6.7 Conclusion

Research into indoor comfort is growing due to its importance to the wellbeing of occupants, using machine learning techniques and historical data. This problem is challenging to address as individuals have different levels of acceptance of what is comfortable. In this study, dataset spanning across one year including ambient sensors and occupants' survey data was used to predict group general thermal comfort and thermal sensation with a high degree of accuracy in predicting results, compared to other prediction techniques previously reported in the literature. Furthermore, a one-step-ahead further adjustment technique was used to optimise

the HVAC system for the purpose of maximising the occupants' general thermal comfort and thermal sensation.

This study could be used to develop around the thermal comfort automation field, as a prediction model and an adjustment technique were developed. This model could be implemented in any building, as the compile list of features relate to thermal comfort were taken from multiple research sources, as shown in Figure 6.4. The model was designed with simplicity in mind so it could be integrated into HVAC systems for future automation. Moreover, there are multiple benefits in achieving group relative thermal comfort and thermal sensation predictions. Improving the occupants' health and wellbeing, boosting occupants' productivity and an overall reduction in power consumed by the building, could be achieved using this technique.

Chapter 7

Conclusion

To conclude, current research on human occupancy counting is growing and with the advancements of technology and processing power, millions of data can be stored, processed, mined and analysed within an acceptable period. This is revolutionary within the machine learning and data mining area.

We defined the context required to find out which features are best for understanding more about human occupancy. Research into counting of human occupancy is emerging and correlating each feature with it can result in interesting findings. Ambient sensors were the options that we decided to use, because they provided the least intrusive way to link something back to each person. We do not want to closely monitor a person until this feels too intimidating and breaches his/her privacy boundary. For these reasons, camera-assisted tools are not being utilised for these kinds of experiments.

The aim of the research was to build a scalable human occupancy counting framework with predictive capability for both single domain modelling and domain adaptation. From the outset, we identified that when CO₂ concentration data is continuously collected and accumulated for a specified period, the number of occupants can be recognised and predicted. The benefit of this study is extensive, ranging from the reduction of energy consumption to indoor security enhancement.

Several issues around building utilisation analytics that are tackled and elucidated in this thesis can be summarised as follows:

1. CO₂ is a gas and therefore needs some time to reach detectors. There is a time delay that needs to be integrated into the model.

2. Knowing the number of people present at a particular time could not happen automatically, hence an occupancy label needs a well-designed gathering and identifying method.
3. Detecting long-term trend changes using context histories and their inter-relationships.
4. Making the learning models adaptive and continuous so that they can adapt based on the changes in contexts and can still predict future events continuously with reasonable accuracy.
5. Predicting the future occurrence of an unknown domain, based on the knowledge learned about that event from available data for a previously modelled domain.

The research presented in this thesis has focused on the development of a human occupancy framework solution for any domain. It explored different machine-learning solutions for large-scale ambient sensor data. Some of the experiments were carried out in the cloud environment and some on a powerful academic server. The developed techniques reduced the rate of false predictions and were designed to assist building management system experts in remote monitoring centres with early recognition of human occupancy by using prediction output.

Another research area was thermal comfort and thermal sensation prediction, using sensor data and machine learning techniques to support and improve the wellbeing and productivity of each inhabitant. This research completed the exploration of building utilisation analytics.

7.1 Research Questions and Answers

The core chapters of this thesis addressed key research challenges related to building utilisation analytics, which include indoor human occupancy counting and thermal comfort prediction. This thesis only utilised data from ambient sensing and survey data. Four research questions were developed, as shown in subsection 1.3. To find the solution for each research question, we researched, designed, developed and analysed both fundamental and technical domains. Specifically, we developed multiple frameworks and intelligent machine-learning algorithms to provide a customised solution for each challenge. We backed up every experiment, not with generated simulation data but with real-world datasets to ensure the applicability of each solution. Each research question and its solution is summarised below.

RQ-1. *How to recognise indoor human occupancy using multivariate ambient sensor data?*

The first research question (RQ-1) was addressed in Chapter 2. We presented a simple human occupancy recognition framework with multivariate ambient sensors and experimented with a variety of different time segments. We deployed four different types of off-the-shelf sensors from two manufacturers, to ensure we could collect the following data reliably from within one staff office: illumination, temperature, humidity, levels of carbon dioxide, pressure, and sound. We also collected motion, power consumption, door opening and closing data, and annotations from a self-developed mobile app as ground truth. We presented methods to preprocess the data and compute the number of people in the room with different classifiers, and identified sensors with strong and weak correlations. We explained our methodology for integrating large amounts of sensor data in section 2.3, and discussed our experiments and findings in relation to the binary occupancy of a single person office, providing a baseline for recognising human occupancy and showing that random forest is the most suitable machine learning algorithm for indoor human occupancy in section 2.5. Every ambient sensor coefficient was compared in Table 2.1 and the most dominant sensor in determining human presence was chosen; this was carbon dioxide.

RQ-2. *How to perform room utilisation prediction using carbon dioxide data?*

To address the second research question (RQ-2), we developed several new techniques for indoor human occupancy counting with carbon dioxide datasets. Seasonal decomposition for human occupancy counting (SD-HOC) was introduced in Chapter 3 to preprocess both carbon dioxide and human occupancy datasets. In this experiment, we utilised a dataset from a building in Melbourne, Australia, and another from a cinema theatre in Mainz, Germany. Multiple machine learning techniques were implemented and the results were compared to SD-HOC. SD-HOC excelled in prediction accuracy and had the highest accuracy of the techniques. SD-HOC integrated time lag and a line of best fit model into the preprocessing algorithms. SD-HOC utilised seasonal-trend decomposition with moving average to transform the preprocessed data, and for each trend, seasonal and irregular component, different regression algorithms were modelled to predict each respective human occupancy component value. Utilising the M5 method of linear regression for trend, and irregular component and dynamic time warping for the seasonal component, a set of the prediction values for each component was obtained. A zero pattern adjustment model was used to increase the accuracy and, finally, additive decomposi-

tion was used to reconstruct the prediction value. The accuracy results were compared with other data mining algorithms, such as decision tree, multi-layer perceptron, Gaussian processes - radial basis function, support vector machine, random forest, naïve Bayes, and support vector regression; this was undertaken in two different locations that have different contexts.

In Chapter 4, we developed large Room Utilisation Prediction (RUP) with a carbon dioxide sensor (Figure 4.9). RUP is the extended version of SD-HOC and caters to large-scale rooms with more occupants. RUP de-noises and preprocesses the carbon dioxide and indoor human occupancy data. We utilised seasonal-trend decomposition based on local regression (RUP-STL) and seasonal-trend decomposition with moving average (RUP-STD) to factorise both datasets. For each trend, seasonal and irregular component, we modelled different regression algorithms to predict respective human occupancy component values. We ran our model in two different locations that have different contexts. The first location was an academic staff room (Figure 4.5) and the second was a cinema theatre with up to 300 people (Figure 4.6). Our results showed an average 4.33% increase in accuracy for the small room with 94.68% indoor human occupancy counting, and an 8.46% increase for the cinema theatre in comparison to the accuracy of the baseline method (support vector regression).

RQ-3. *How to perform transfer learning to use an existing occupancy prediction model to predict the utilisation of another room with limited training data?*

In Chapter 5, the third research question (RQ-3) was addressed. A domain adaptation technique was explored and a semi-supervised domain adaptation method for carbon dioxide - human occupancy counter plus plus (DA-HOC++) was proposed, which is a robust way to estimate the number of people within in one room by using data from a carbon dioxide sensor. DA-HOC++ is the domain adaptation expansion technique from SD-HOC (Chapter 3). The SD-HOC model can accurately predict the number of individuals when adequate training and labelled data are available (Figure 5.6). DA-HOC++ was able to predict the number of occupants with minimal training data, as little as one day of data. DA-HOC++ accurately predicted indoor human occupancy for five different rooms across different countries, using a model trained from a small room and adapted to the other rooms. We evaluated DA-HOC++ with two baseline methods: the support vector regression technique and the SD-HOC model (Figures 5.3, 5.4, 5.5, 5.6 and 5.7). The results demonstrated that the performance of DA-HOC++ was better by an average of 10.87% compared to SVR and 8.65% compared to SD-HOC.

RQ-4. *How to predict indoor comfort based on ambient sensor and users' survey data?*

The fourth research question (RQ-4) is covered in Chapter 6. We developed THERMO, a prediction and adjustment model for both general thermal comfort and thermal sensation in a shared office environment. We utilised one-year of comfort-related data from ambient sensors, background surveys and daily survey data from the Friend Center in Philadelphia, USA (Figure 6.9). THERMO implemented sparse non-negative matrix factorisation for de-noising the ambient sensor data and building a double layer model for prediction of thermal comfort and sensation (Figure 6.7). The first pre-model used an ensemble rotation forest (with most of the features in the survey and sensor data), and the second primary model implemented partial least squares for classification using ambient sensors as predictors. THERMO prediction accuracy was compared with multiple machine learning techniques such as random forest, decision tree, logistic regression, multilayer perceptron, naïve Bayes, and support vector machine; it had superior thermal comfort prediction accuracy. THERMO's adjustment technique improved the general thermal comfort of 24 building inhabitants by 0.625 on a 1-6 comfort scale.

7.2 Future Directions for Research

There are several recommended directions for future research on building utilisation analytics using extensive sensor data and machine learning techniques. The three most important aspects are data quality, context awareness, and effective predictive analytics. Data quality focuses on how to maintain the integrity of the data. In the era of big data, the temporal information from sensor data becomes more widely available and techniques to ensure and preserve data quality are paramount.

Context awareness relates to the content surrounding building utilisation and to environmental analysis. Indoor human occupancy prediction techniques that are aware of surrounding contexts could become a future norm for research analysis. Improving the accuracy of prediction analytics around building utilisation is always the primary goal in this area.

For domain adaptation research and transfer learning research, these are still in their infancy and more researchers are focusing on these areas. For future work, incremental learning techniques are a good research direction. Other future work may include unsupervised domain adaptation without any labelling. Because of this, the unsupervised technique removes the cost

of gathering the labels and finding the ground-truth. Progressive learning models can learn and will gradually improve their prediction accuracy over time.

In the domain of thermal comfort automation, there are many directions that can be pursued. As our model and technique are based on machine learning, implementing them in multiple locations with real-time thermal comfort datasets is a recommended area for future works. The model could be trained with people from different countries and ethnicities so that the model and technique can be used worldwide. This model is built based on the historical dataset. Incremental learning is another possible future area, where the model can adapt itself automatically by gradually learning new data using an adaptive learning model. Progressive learning models can be implemented and online learning models will be investigated in the future.

Bibliography

- Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, and T. Weng. Occupancy-driven energy management for smart building automation. In *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building*, pages 1–6. ACM, 2010. Cited on pages 3, 14, 54, and 56.
- Y. Agarwal, B. Balaji, S. Dutta, R. K. Gupta, and T. Weng. Duty-cycling buildings aggressively: The next frontier in hvac control. In *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, pages 246–257. IEEE, 2011. Cited on page 15.
- H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974. Cited on pages 38 and 70.
- T. Akimoto, S.-i. Tanabe, T. Yanai, and M. Sasaki. Thermal comfort and productivity-evaluation of workplace environment in a task conditioned office. *Building and environment*, 45(1):45–50, 2010. Cited on pages 9, 112, and 115.
- K. Akkaya, I. Guvenc, R. Aygun, N. Pala, and A. Kadri. Iot-based occupancy monitoring techniques for energy-efficient smart buildings. In *Wireless Communications and Networking Conference Workshops (WCNCW), 2015 IEEE*, pages 58–63. IEEE, 2015. Cited on page 3.
- H. Al-Mubaid and S. A. Umair. A new text categorization technique using distributional clustering and learning logic. *IEEE Transactions on Knowledge and Data Engineering*, 18(9):1156–1165, 2006. Cited on page 83.
- A. T. Alan, M. Shann, E. Costanza, S. D. Ramchurn, and S. Seuken. It is too hot: An in-situ study of three designs for heating. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5262–5273. ACM, 2016. Cited on page 120.

- I. B. A. Ang, F. D. Salim, and M. Hamilton. Human occupancy recognition with multivariate ambient sensors. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pages 1–6. IEEE, 2016. Cited on pages viii and 10.
- I. B. Arief-Ang, F. D. Salim, and M. Hamilton. DA-HOC: Semi-Supervised Domain Adaptation for Room Occupancy Prediction using CO2 Sensor Data. In *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments, BuildSys '17*, pages 1:1–1:10, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5544-5. doi: 10.1145/3137133.3137146. URL <http://doi.acm.org/10.1145/3137133.3137146>. Cited on pages viii, 11, and 85.
- I. B. Arief-Ang, M. Hamilton, and F. D. Salim. Rup: Large room utilisation prediction with carbon dioxide sensor. *Pervasive and Mobile Computing*, 46:49–72, 2018a. Cited on pages viii and 11.
- I. B. Arief-Ang, F. D. Salim, and M. Hamilton. *SD-HOC: Seasonal Decomposition Algorithm for Mining Lagged Time Series*, volume 845 of *Communications in Computer and Information Science*. Springer International Publishing, 2018b. ISBN 978-981-13-0292-3. Cited on pages viii and 10.
- R. C. Arora. Comfort - Physiological Principles, IAQ and Design Conditions. In *Refrigeration and Air Conditioning*, chapter 19, pages 819 – 871. PHI Learning Pvt. Ltd., eastern edition, 2010. ISBN 9788120339156. Cited on page 51.
- F. Ascione, N. Bianco, R. F. De Masi, F. de’Rossi, and G. P. Vanoli. Energy refurbishment of existing buildings through the use of phase change materials: Energy savings and indoor comfort in the cooling season. *Applied Energy*, 113:990–1007, 2014. Cited on page 3.
- A.S.H.R.A.E. Standard 55-2013, Thermal Environmental Conditions for Human Occupancy. Technical report, 2013. Cited on pages 112 and 115.
- B. Balaji, J. Koh, N. Weibel, and Y. Agarwal. Genie: a longitudinal study comparing physical and software thermostats in office buildings. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1200–1211. ACM, 2016. Cited on page 120.

- J. Barandiaran, B. Murguia, and F. Boto. Real-time people counting using multiple lines. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*, pages 159–162. IEEE, 2008. Cited on pages 7 and 51.
- A. Barbato, L. Borsani, A. Capone, and S. Melzi. Home energy saving through a user profiling system based on wireless sensors. In *Proceedings of the first ACM workshop on embedded sensing systems for energy-efficiency in buildings*, pages 49–54. ACM, 2009. Cited on page 15.
- T. Basten, L. Benini, A. Chandrakasan, M. Lindwer, J. Liu, R. Min, and F. Zhao. Scaling into ambient intelligence. In *Proceedings of the conference on Design, Automation and Test in Europe-Volume 1*, page 10076. IEEE Computer Society, 2003. Cited on page 13.
- C. Basu, C. Koehler, K. Das, and A. K. Dey. Percs: person-count from carbon dioxide using sparse non-negative matrix factorization. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 987–998. ACM, 2015. Cited on pages 26, 28, 48, 50, 54, 58, 59, 75, 77, 81, and 85.
- A. Beltran and A. E. Cerpa. Optimal hvac building control with occupancy prediction. In *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pages 168–171. ACM, 2014. Cited on page 119.
- A. Beltran, V. L. Erickson, and A. E. Cerpa. Thermosense: Occupancy thermal based sensing for hvac control. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8. ACM, 2013. Cited on pages 3, 54, and 56.
- K. J. Berry, J. E. Johnston, and P. W. Mielke Jr. A chronicle of permutation statistical methods. *Springer International Publishing, Cham. doi*, 10:978–3, 2014. Cited on page 132.
- J. Blitzer, M. Dredze, F. Pereira, et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447, 2007. Cited on page 83.
- G. S. Brager and R. J. De Dear. Thermal adaptation in the built environment: a literature review. *Energy and buildings*, 27(1):83–96, 1998. Cited on page 113.
- N. M. Byrne, A. P. Hills, G. R. Hunter, R. L. Weinsier, and Y. Schutz. Metabolic equivalent: one size does not fit all. *Journal of Applied physiology*, 99(3):1112–1119, 2005. Cited on page 117.

- D. Cali, P. Matthes, K. Huchtemann, R. Streblow, and D. Müller. CO2 based occupancy detection algorithm: experimental analysis and validation for office and residential buildings. *Building and Environment*, 86:39–49, 2015. Cited on pages 28, 58, and 85.
- L. M. Candanedo and V. Feldheim. Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. *Energy and Buildings*, 112:28–39, 2016. Cited on pages 26, 48, 54, 56, and 80.
- B. Cao, Y. Zhu, Q. Ouyang, X. Zhou, and L. Huang. Field study of human thermal comfort and thermal adaptability during the summer and winter in beijing. *Energy and Buildings*, 43(5):1051–1056, 2011. Cited on page 131.
- F. Castanedo, D. López-de Ipina, H. K. Aghajan, and R. P. Kleihorst. Building an occupancy model from sensor networks in office environments. *ICDSC*, 3:1–6, 2011. Cited on pages 54 and 56.
- A. K. Clear, J. Morley, M. Hazas, A. Friday, and O. Bates. Understanding adaptive thermal comfort: new directions for ubicomp. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 113–122. ACM, 2013. Cited on page 3.
- R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990. Cited on page 69.
- C. Cortes and M. Mohri. Domain adaptation in regression. In *International Conference on Algorithmic Learning Theory*, pages 308–323. Springer, 2011. Cited on page 83.
- G. Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017. Cited on page 80.
- W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200. ACM, 2007. Cited on page 83.
- W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across different feature spaces. In *Advances in neural information processing systems*, pages 353–360, 2009. Cited on page 84.

- H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006. Cited on page 83.
- R. J. De Dear, G. S. Brager, J. Reardon, F. Nicol, et al. Developing an adaptive model of thermal comfort and preference/discussion. *ASHRAE transactions*, 104:145, 1998. Cited on page 113.
- S. Dedesko, B. Stephens, J. A. Gilbert, and J. A. Siegel. Methods to assess human occupancy and occupant activity in hospital patient rooms. *Building and Environment*, 90:136–145, 2015. Cited on pages 28, 58, and 85.
- D. T. Delaney, G. M. O’Hare, and A. G. Ruzzelli. Evaluation of energy-efficiency in lighting systems using sensor networks. In *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pages 61–66. ACM, 2009. Cited on page 3.
- S. Depatla, A. Muralidharan, and Y. Mostofi. Occupancy estimation using only wifi power measurements. *IEEE Journal on Selected Areas in Communications*, 33(7):1381–1393, 2015. Cited on pages 15, 28, 53, 59, and 85.
- R. H. Dodier, G. P. Henze, D. K. Tiller, and X. Guo. Building occupancy detection through sensor belief networks. *Energy and buildings*, 38(9):1033–1043, 2006. Cited on pages 3, 14, 54, and 56.
- U. DOE. Building Energy Databook. Technical report, 2010. Cited on pages 13 and 26.
- A. DOEE. The Basics of HVAC Energy Efficiency Factsheet. Technical report, 2013. Cited on pages 47 and 80.
- L. Douglas. 3d data management: Controlling data volume, velocity and variety. *Gartner. Retrieved*, 6(2001):6, 2001. Cited on page 6.
- P. K. Dutta, A. K. Arora, and S. B. Bibyk. Towards radar-enabled sensor networks. In *Information Processing in Sensor Networks, 2006. IPSN 2006. The Fifth International Conference on*, pages 467–474. IEEE, 2006. Cited on pages 14, 53, and 84.
- T. Ekwevugbe, N. Brown, and V. Pakka. Real-time building occupancy sensing for supporting demand driven hvac operations. Energy Systems Laboratory, 2013a. Cited on pages 26, 48, 54, 57, and 80.

- T. Ekwevugbe, N. Brown, V. Pakka, and D. Fan. Real-time building occupancy sensing using neural-network based sensor network. In *2013 7th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, pages 114–119. IEEE, 2013b. Cited on pages 3, 54, and 57.
- V. L. Erickson and A. E. Cerpa. Occupancy based demand response hvac control strategy. In *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, pages 7–12. ACM, 2010. Cited on page 117.
- V. L. Erickson and A. E. Cerpa. Thermovote: participatory sensing for efficient building hvac conditioning. In *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pages 9–16. ACM, 2012. Cited on page 117.
- V. L. Erickson, Y. Lin, A. Kamthe, R. Brahme, A. Surana, A. E. Cerpa, M. D. Sohn, and S. Narayanan. Energy efficient building environment control strategies using real-time occupancy measurements. In *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pages 19–24. ACM, 2009. Cited on pages 3, 7, 14, 28, and 51.
- P. O. Fanger and J. Toftum. Extension of the pmv model to non-air-conditioned buildings in warm climates. *Energy and buildings*, 34(6):533–536, 2002. Cited on pages 113 and 118.
- P. O. Fanger et al. Thermal comfort. analysis and applications in environmental engineering. *Thermal comfort. Analysis and applications in environmental engineering.*, 1970. Cited on pages 116 and 117.
- N. Farajidavar, T. de Campos, and J. Kittler. Transductive transfer machine. In *Asian Conference on Computer Vision*, pages 623–639. Springer, 2014. Cited on page 83.
- A. Feige, H. Wallbaum, M. Janser, and L. Windlinger. Impact of sustainable office buildings on occupant’s comfort and productivity. *Journal of Corporate Real Estate*, 15(1):7–34, 2013. Cited on pages 3, 112, and 113.
- M. Feldmeier and J. A. Paradiso. Personalized hvac control system. In *Internet of Things (IOT), 2010*, pages 1–8. IEEE, 2010. Cited on page 117.
- D. F. Findley, B. C. Monsell, W. R. Bell, M. C. Otto, and B.-C. Chen. New capabilities and methods of the x-12-arima seasonal-adjustment program. *Journal of Business & Economic Statistics*, 16(2):127–152, 1998. Cited on pages 34 and 67.

- E. Fortet and E. Mourier. Convergence de la re'paration empirique vers la re'paration the'orique. *Ann. Scient. E'cole Norm. Sup.*, 70:266–285, 1953. Cited on page 92.
- G. P. C. Fung, J. X. Yu, H. Lu, and P. S. Yu. Text classification without negative examples revisit. *IEEE transactions on Knowledge and Data Engineering*, 18(1):6–20, 2006. Cited on page 83.
- P. X. Gao and S. Keshav. Optimal personal comfort management using spot+. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8. ACM, 2013. Cited on pages 14, 54, and 56.
- V. Garg and N. Bansal. Smart occupancy sensors to reduce energy consumption. *Energy and Buildings*, 32(1):81–87, 2000. Cited on pages 14, 54, and 56.
- L. Giaccone, A. Guerrisi, P. Lazzeroni, M. Martino, and M. Tartaglia. An effective monitoring of indoor comfort and evaluation of energy consumption in a complex urban energy system. In *Renewable Energy Research and Applications (ICRERA), 2012 International Conference on*, pages 1–6. IEEE, 2012. Cited on pages 112 and 116.
- R. Goldstein, A. Tessier, and A. Khan. Schedule-calibrated occupant behavior simulation. In *Proceedings of the 2010 Spring Simulation Multiconference*, page 180. Society for Computer Simulation International, 2010. Cited on pages 4, 47, and 52.
- L. Granato, A. Brandes, C. Bruni, A. V. Greco, and G. Mingrone. Vo2, vco2, and rq in a respiratory chamber: accurate estimation based on a new mathematical model using the kalman-bucy method. *Journal of Applied Physiology*, 96(3):1045–1054, 2004. Cited on page 52.
- P. Grossman, L. Niemann, S. Schmidt, and H. Walach. Mindfulness-based stress reduction and health benefits: A meta-analysis. *Journal of psychosomatic research*, 57(1):35–43, 2004. Cited on page 115.
- B. Guo, Z. Yu, L. Chen, X. Zhou, and X. Ma. Mobigroup: Enabling lifecycle support to social activity organization and suggestion with mobile crowd sensing. *IEEE Transactions on Human-Machine Systems*, 46(3):390–402, 2016. Cited on page 57.
- M. T. Hagan, H. B. Demuth, M. H. Beale, et al. *Neural network design*, volume 20. Pws Pub. Boston, 1996. Cited on page 128.

- E. Hailemariam, R. Goldstein, R. Attar, and A. Khan. Real-time occupancy detection using decision trees with multiple sensor types. In *Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design*, pages 141–148. Society for Computer Simulation International, 2011. Cited on pages 26, 48, 54, 57, 58, and 85.
- L. A. Hang-yat and D. Wang. Carrying my environment with me: A participatory-sensing approach to enhance thermal comfort. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8. ACM, 2013. Cited on page 117.
- J. He and A. Arora. A regression-based radar-mote system for people counting. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*, pages 95–102. IEEE, 2014. Cited on page 53.
- R. Z. Homod, K. S. M. Sahari, H. A. Almurib, and F. H. Nagi. Rlf and ts fuzzy model identification of indoor thermal comfort based on pmv/ppd. *Building and Environment*, 49: 141–153, 2012. Cited on page 113.
- J. Howard and W. Hoff. Forecasting building occupancy using sensor network data. In *Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, pages 87–94. ACM, 2013. Cited on pages 14, 54, and 56.
- D. H. Hu and Q. Yang. Transfer learning for activity recognition via sensor mapping. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1962, 2011. Cited on page 84.
- L. Huang, Y. Zhu, Q. Ouyang, and B. Cao. A study on the effects of thermal, luminous, and acoustic environments on indoor environmental comfort in offices. *Building and Environment*, 49:304–309, 2012. Cited on page 3.
- R. J. Hyndman. *Moving Averages*, pages 866–869. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. Cited on page 36.
- S. Inoue and X. Pan. Supervised and unsupervised transfer learning for activity recognition from simple in-home sensors. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 20–27. ACM, 2016. Cited on page 83.

- F. Jazizadeh and B. Becerik-Gerber. Toward adaptive comfort management in office buildings using participatory sensing for end user driven control. In *Proceedings of the Fourth ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, pages 1–8. ACM, 2012. Cited on pages 55 and 119.
- Y. Jiang, K. Li, L. Tian, R. Piedrahita, X. Yun, O. Mansata, Q. Lv, R. P. Dick, M. Hannigan, and L. Shang. Maqs: a personalized mobile sensing system for indoor air quality monitoring. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 271–280. ACM, 2011. Cited on pages 3 and 112.
- Y. Jiang, X. Pan, K. Li, Q. Lv, R. P. Dick, M. Hannigan, and L. Shang. Ariel: Automatic wi-fi based room fingerprinting for indoor localization. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 441–450. ACM, 2012. Cited on pages 3 and 14.
- M. Jradi, F. C. Sangogboye, C. G. Mattera, M. B. Kjærgaard, C. Veje, and B. N. Jørgensen. A world class energy efficient university building by danish 2020 standards. *Energy Procedia*, 132:21–26, 2017. Cited on page 3.
- D. Kale, M. Ghazvininejad, A. Ramakrishna, J. He, and Y. Liu. Hierarchical active transfer learning. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 514–522. SIAM, 2015. Cited on page 83.
- S. Karjalainen. Thermal comfort and gender: a literature review. *Indoor air*, 22(2):96–109, 2012. Cited on page 116.
- A. Khan, J. Nicholson, S. Mellor, D. Jackson, K. Ladha, C. Ladha, J. Hand, J. Clarke, P. Olivier, and T. Plötz. Occupancy monitoring using environmental & context sensors and a hierarchical analysis framework. In *BuildSys@ SenSys*, pages 90–99, 2014. Cited on pages 3, 26, 48, 55, 57, and 80.
- M. A. A. H. Khan, H. Hossain, and N. Roy. Infrastructure-less occupancy detection and semantic localization in smart environments. In *proceedings of the 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services on 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 51–60. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2015. Cited on pages 14, 55, and 57.

- B. Kingma and W. van Marken Lichtenbelt. Energy consumption in buildings and female thermal demand. *Nature climate change*, 5(12):1054–1056, 2015. Cited on page 114.
- W. Kleiminger, C. Beckel, T. Staake, and S. Santini. Occupancy detection from electricity consumption data. In *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, pages 1–8. ACM, 2013. Cited on page 56.
- B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1785–1792. IEEE, 2011. Cited on page 83.
- R. Laforteza, G. Carrus, G. Sanesi, and C. Davies. Benefits and well-being perceived by people visiting green spaces in periods of heat stress. *Urban Forestry & Urban Greening*, 8(2):97–108, 2009. Cited on pages 113 and 115.
- D. Lai, D. Guo, Y. Hou, C. Lin, and Q. Chen. Studies of outdoor thermal comfort in northern china. *Building and Environment*, 77:110–118, 2014. Cited on page 120.
- A. H.-y. Lam, Y. Yuan, and D. Wang. An occupant-participatory approach for thermal comfort enhancement and energy conservation in buildings. In *Proceedings of the 5th international conference on Future energy systems*, pages 133–143. ACM, 2014. Cited on pages 3, 112, 114, and 117.
- K. P. Lam, M. Höynck, B. Dong, B. Andrews, Y.-S. Chiou, R. Zhang, D. Benitez, J. Choi, et al. Occupancy detection through an extensive environmental sensor network in an open-plan office building. *IBPSA Building Simulation*, 145:1452–1459, 2009. Cited on pages 27, 28, 48, 55, 58, and 85.
- J. Langevin, P. L. Gurian, and J. Wen. Tracking the human-building interaction: a longitudinal field study of occupant behavior in air-conditioned offices. *Journal of Environmental Psychology*, 42:94–115, 2015. Cited on pages xx, 120, 124, and 133.
- H. Lee, C. Wu, and H. Aghajan. Vision-based user-centric light control for smart environments. *Pervasive and Mobile Computing*, 7(2):223–240, 2011. Cited on pages 7, 28, and 51.
- T. Leephakpreeda. Adaptive occupancy-based lighting control via grey prediction. *Building and environment*, 40(7):881–886, 2005. Cited on pages 14, 15, 26, and 53.

- H. Li, E. C. Chan, X. Guo, J. Xiao, K. Wu, and L. M. Ni. Wi-counter: smartphone-based people counter using crowdsourced wi-fi signal data. *IEEE Transactions on Human-Machine Systems*, 45(4):442–452, 2015. Cited on page 53.
- C. Liao and P. Barooah. An integrated approach to occupancy modeling and estimation in commercial buildings. In *American Control Conference (ACC), 2010*, pages 3130–3135. IEEE, 2010. Cited on pages 55 and 56.
- K. S. Liu, S. Munir, J. Francis, C. Shelton, and S. Lin. Long term occupancy estimation in a commercial space: an empirical study. In *IPSN*, pages 307–308, 2017. Cited on page 51.
- D. Lopez-Paz, J. M. Hernández-lobato, and B. Schölkopf. Semi-supervised domain adaptation with non-parametric copulas. In *Advances in neural information processing systems*, pages 665–673, 2012. Cited on page 83.
- J. Lu, T. Sookoor, V. Srinivasan, G. Gao, B. Holben, J. Stankovic, E. Field, and K. Whitehouse. The smart thermostat: using occupancy sensors to save energy in homes. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, pages 211–224. ACM, 2010. Cited on pages 3, 14, 15, and 56.
- W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, and T. Zhang. Deep model based domain adaptation for fault diagnosis. *IEEE Transactions on Industrial Electronics*, 64(3):2296–2305, 2017. Cited on page 83.
- M. Luo, B. Cao, J. Damians, B. Lin, and Y. Zhu. Evaluating thermal comfort in mixed-mode buildings: A field study in a subtropical climate. *Building and environment*, 88:46–54, 2015. Cited on page 121.
- R. C. Luo, S. Y. Lin, and K. L. Su. A multiagent multisensor based security system for intelligent building. In *Multisensor Fusion and Integration for Intelligent Systems, MFI2003. Proceedings of IEEE International Conference on*, pages 311–316. IEEE, 2003. Cited on page 3.
- S. Mamidi, Y.-H. Chang, and R. Maheswaran. Improving building energy efficiency with a network of sensing, learning and prediction agents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 45–52. International Foundation for Autonomous Agents and Multiagent Systems, 2012. Cited on pages 15, 48, 55, and 56.

- A. Martinez-Molina, P. Boarin, I. Tort-Ausina, and J.-L. Vivancos. Post-occupancy evaluation of a historic primary school in Spain: Comparing pmv, tsv and pd for teachers' and pupils' thermal comfort. *Building and Environment*, 117:248–259, 2017. Cited on page 120.
- A. Matzarakis and B. Amelung. Physiological equivalent temperature as indicator for impacts of climate change on thermal comfort of humans. In *Seasonal forecasts, climatic change and human health*, pages 161–172. Springer, 2008. Cited on pages 9, 113, and 115.
- K. McCartney and M. Humphreys. Thermal comfort and productivity. *Proceedings of Indoor Air*, 2002:822–827, 2002. Cited on pages 9 and 115.
- H. Mohammadmoradi, S. Munir, O. Gnawali, and C. Shelton. Measuring people-flow through doorways using easy-to-install ir array sensors. In *The annual International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2017. Cited on page 55.
- S. Munir, R. S. Arora, C. Hesling, J. Li, J. Francis, C. Shelton, C. Martin, A. Rowe, and M. Berges. Real-time fine grained occupancy estimation using depth sensors on arm embedded platforms. In *Real-Time and Embedded Technology and Applications Symposium (RTAS), 2017 IEEE*, pages 295–306. IEEE, 2017. Cited on page 51.
- J. Nembrini and D. Lalanne. Human-building interaction: When the machine becomes a building. In *IFIP Conference on Human-Computer Interaction*, pages 348–369. Springer, 2017. Cited on page 120.
- M. A. Ortiz, S. R. Kurvers, and P. M. Bluysen. A review of comfort, health, and energy use: Understanding daily energy use and wellbeing for the development of a new approach to study comfort. *Energy and Buildings*, 152:323–335, 2017. Cited on pages 9, 113, 115, and 116.
- J. Page, D. Robinson, N. Morel, and J.-L. Scartezzini. A generalised stochastic model for the simulation of occupant presence. *Energy and buildings*, 40(2):83–98, 2008. Cited on pages 4, 47, and 52.
- S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010. Cited on pages 80 and 83.
- S. J. Pan, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *AAAI*, volume 8, pages 677–682, 2008a. Cited on page 83.

- S. J. Pan, V. W. Zheng, Q. Yang, and D. H. Hu. Transfer learning for wifi-based indoor localization. In *Association for the advancement of artificial intelligence (AAAI) workshop*, page 6, 2008b. Cited on page 83.
- F. Petitjean, A. Ketterlin, and P. Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693, 2011. Cited on pages 38 and 71.
- A. Rabbani and S. Keshav. The spot* personal thermal comfort system. In *BuildSys@ SenSys*, pages 75–84, 2016. Cited on page 120.
- I. Richardson, M. Thomson, and D. Infield. A high-resolution domestic building occupancy model for energy demand simulations. *Energy and buildings*, 40(8):1560–1566, 2008. Cited on pages 4, 47, and 52.
- J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1619–1630, 2006. Cited on page 130.
- R. F. Rupp, R. de Dear, and E. Ghisi. Field study of mixed-mode office buildings in southern brazil using an adaptive thermal comfort framework. *Energy and Buildings*, 158:1475–1486, 2018. Cited on page 121.
- D. Saelens, W. Parys, and R. Baetens. Energy and comfort performance of thermally activated building systems including occupant behavior. *Building and Environment*, 46(4):835–848, 2011. Cited on pages 4, 47, 52, and 116.
- F. C. Sangogboye, K. Arendt, A. Singh, C. T. Veje, M. B. Kjærgaard, and B. N. Jørgensen. Performance comparison of occupancy count estimation and prediction with common versus dedicated sensors for building model predictive control. In *Building Simulation*, volume 10, pages 829–843. Springer, 2017. Cited on pages xix, 3, 48, 55, 99, and 101.
- K. Sarinapakorn and M. Kubat. Combining subclassifiers in text categorization: A dst-based solution and a case study. *IEEE Transactions on Knowledge and Data Engineering*, 19(12):1638–1651, 2007. Cited on page 83.
- C. Sarkar, S. A. U. Nambi, and R. V. Prasad. iltc: Achieving individual comfort in shared spaces. In *EWSN*, pages 65–76, 2016. Cited on pages 112, 116, and 119.

- J. Scott, A. Bernheim Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges, and N. Villar. Preheat: controlling home heating using occupancy prediction. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 281–290. ACM, 2011. Cited on page 119.
- O. Shih and A. Rowe. Occupancy estimation using ultrasonic chirps. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems*, pages 149–158. ACM, 2015. Cited on page 55.
- E.-J. Shin, R. Yus, S. Mehrotra, and N. Venkatasubramanian. Exploring fairness in participatory thermal comfort control in smart buildings. 2017. Cited on pages 3, 113, and 120.
- J. Shiskin, A. H. Young, and J. C. Musgrave. *The X-11 variant of the census method II seasonal adjustment program*. Number 15. US Department of Commerce, Bureau of the Census, 1965. Cited on pages 34 and 67.
- S. Sigg, M. Scholz, S. Shi, Y. Ji, and M. Beigl. Rf-sensing of activities from non-cooperative subjects in device-free recognition systems using ambient and local signals. *IEEE Transactions on Mobile Computing*, 13(4):907–920, 2014. Cited on page 53.
- S. Spiegel. Transfer learning for time series classification in dissimilarity spaces. In *Proceedings of AALTD 2016: Second ECML/PKDD International Workshop on Advanced Analytics and Learning on Temporal Data*, page 78, 2016. Cited on page 82.
- V. Srinivasan, J. Stankovic, and K. Whitehouse. Using height sensors for biometric identification in multi-resident homes. *Pervasive computing*, pages 337–354, 2010. Cited on pages 14, 15, 55, and 56.
- S. M. Stigler. Francis galton’s account of the invention of correlation. *Statistical Science*, 4(2): 73–79, 1989. ISSN 08834237. URL <http://www.jstor.org/stable/2245329>. Cited on page 37.
- R. Sunnam, A. Marston, Z. Fu, and O. Baumann. Analyzing indoor comfort conditions through simulation and on-site measurements. In *Proceedings of the Symposium on Simulation for Architecture & Urban Design*, pages 144–151. Society for Computer Simulation International, 2015. Cited on pages 113 and 116.

- M. Swan. Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator Networks*, 1(3):217–253, 2012. Cited on page 3.
- M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009. Cited on pages 80 and 83.
- J. Teizer. Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites. *Advanced Engineering Informatics*, 29(2):225–238, 2015. Cited on page 51.
- Y.-t. Tsao and J. Y.-j. Hsu. Demand-driven power saving by multiagent negotiation for hvac control. In *Joint Proceedings of the Workshop on AI Problems and Approaches for Intelligent Environments and Workshop on Semantic Cities*, pages 9–14. ACM, 2013. Cited on page 15.
- M. M. Tugade, B. L. Fredrickson, and L. Feldman Barrett. Psychological resilience and positive emotional granularity: Examining the benefits of positive emotions on coping and health. *Journal of personality*, 72(6):1161–1190, 2004. Cited on page 115.
- D. Wang, C. C. Federspiel, and F. Rubinstein. Modeling occupancy in single person offices. *Energy and buildings*, 37(2):121–126, 2005. Cited on pages 14, 15, 55, and 56.
- J. Wicker, N. Krauter, B. Derstorff, C. Stöner, E. Bourtsoukidis, T. Klüpfel, J. Williams, and S. Kramer. Cinema data mining: The smell of fear. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1295–1304. ACM, 2015. Cited on pages xviii, xix, 40, 62, 63, 65, 98, and 100.
- D. A. Winkler, A. Beltran, N. P. Esfahani, P. P. Maglio, and A. E. Cerpa. Forces: feedback and control for occupants to refine comfort and energy savings. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1188–1199. ACM, 2016. Cited on pages 112, 113, 116, 119, and 120.
- S. Wold, M. Sjöström, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001. Cited on page 131.
- X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008. Cited on page 82.

- D. P. Wyon. The effects of indoor air quality on performance and productivity. *Indoor air*, 14 (s7):92–101, 2004. Cited on pages 112 and 113.
- N. Yamtraipat, J. Khedari, and J. Hirunlabh. Thermal comfort standards for air conditioned buildings in hot and humid thailand considering additional factors of acclimatization and education level. *Solar Energy*, 78(4):504–517, 2005. Cited on page 114.
- D. Yan, W. O’Brien, T. Hong, X. Feng, H. B. Gunay, F. Tahmasebi, and A. Mahdavi. Occupant behavior modeling for building performance simulation: Current state and future challenges. *Energy and Buildings*, 107:264–278, 2015. Cited on page 26.
- K. Yan, L. Kou, and D. Zhang. Domain adaptation via maximum independence of domain features. *arXiv preprint arXiv:1603.04535*, 2016. Cited on page 83.
- Q. Yang and X. Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(04):597–604, 2006. Cited on page 82.
- R. Yang and M. W. Newman. Learning from a learning thermostat: lessons for intelligent systems for the home. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 93–102. ACM, 2013. Cited on page 119.
- Y. Yang, J. Hao, J. Luo, and S. J. Pan. Ceilingsee: Device-free occupancy inference through lighting infrastructure based led sensing. In *Pervasive Computing and Communications (PerCom), 2017 IEEE International Conference on*, pages 247–256. IEEE, 2017. Cited on page 55.
- Z. Yang, N. Li, B. Becerik-Gerber, and M. Orosz. A multi-sensor based occupancy estimation model for supporting demand driven hvac operations. In *Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design*, page 2. Society for Computer Simulation International, 2012. Cited on pages 48, 55, 56, and 80.
- Z. Yang, N. Li, B. Becerik-Gerber, and M. Orosz. A systematic approach to occupancy modeling in ambient sensor-rich buildings. *Simulation*, 90(8):960–977, 2014. Cited on pages 55 and 57.
- M. Youssef, M. Mah, and A. Agrawala. Challenges: device-free passive localization for wireless environments. In *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, pages 222–229. ACM, 2007. Cited on page 15.

- H. Zhang, E. Arens, and W. Pasut. Air temperature thresholds for indoor comfort and perceived air quality. *Building Research & Information*, 39(2):134–144, 2011. Cited on page 3.
- X. Zhang, P. Wargocki, Z. Lian, and C. Thyregod. Effects of exposure to carbon dioxide and bioeffluents on perceived air quality, self-assessed acute health symptoms, and cognitive performance. *Indoor air*, 27(1):47–64, 2017. Cited on page 48.
- L. Zhuang, X. Chen, and X. Guan. Parameters analysis of hierarchical evolutionary method with an application of hvac system. In *Automation Science and Engineering (CASE), 2017 13th IEEE Conference on*, pages 858–863. IEEE, 2017. Cited on page 115.

Appendix A

Sensor Devices

A.1 List of devices (sensors) and Their Capacities.

For this experiment, we purchased several devices that will be explained below:

- Z-Wave Aeon Multi Sensors


	Type of sensor	Range	Unit
	Illuminate (light)	> 0	Lux
	Temperature	0 to 50	Celsius degree (°C)
	Humidity	0 to 100	Percentage (%)
	Motion	0 or 1	Boolean value

Figure A.1: Z-Wave Aeon Multi Sensors

- SmartThings SmartSense Open/Closed Sensor


	Type of sensor	Range	Unit
	Door open/close	Open or close	Boolean value

Figure A.2: SmartThings SmartSense Open/Closed Sensor

- SmartThings SmartPower Outlet


	Type of sensor	Range	Unit
	Electric current flow	0 to 100	Watt

Figure A.3: SmartThings SmartPower Outlet

- Netatmo Urban Weather Station


	Type of sensor	Range	Unit	Accuracy
	Indoor temperature	0 to 50	Celsius degree (°C)	+ - 0.3°C
	Humidity	0 to 100	Percentage (%)	+ - 3%
	Barometer (pressure)	260 to 1160	mbar	+ -1 mbar
	CO2 meter	0 to 5000	ppm	+ -50 ppm
	Sound meter	35 to 120	dB	

Figure A.4: Netatmo Urban Weather Station

Appendix B

App System Architecture used in Chapter 2

Below is the architecture design for data gathering:

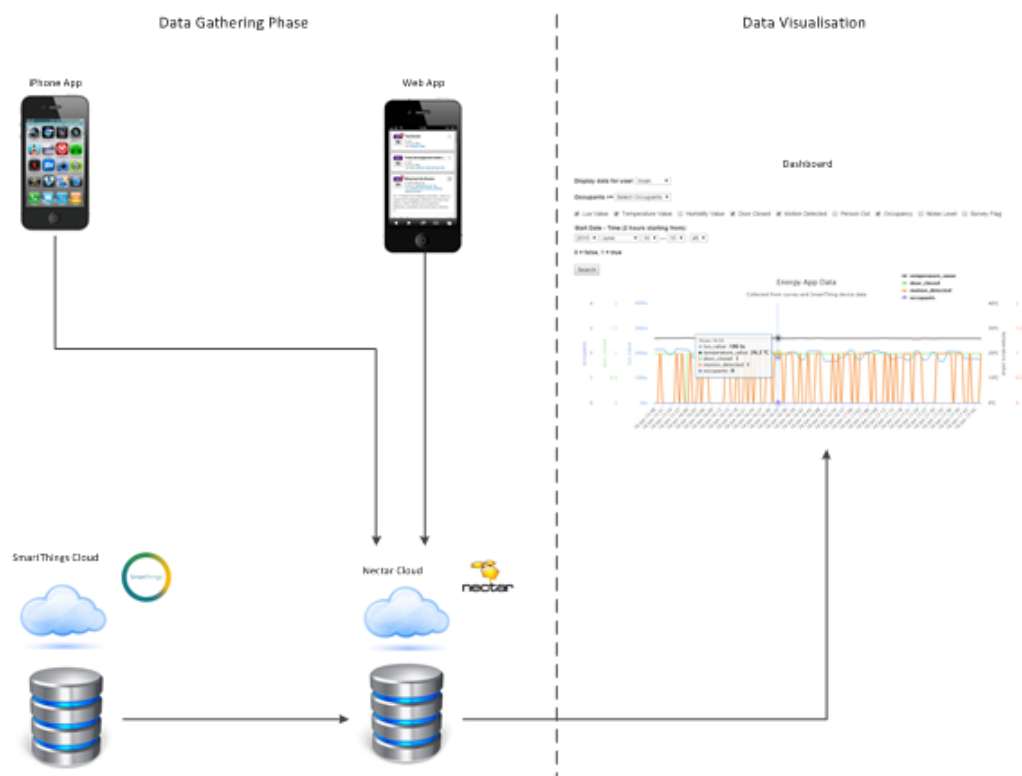


Figure B.1: App System architecture

Appendix C

Machine Learning Techniques and Their Abbreviations

Table C.1: Machine Learning Techniques and their abbreviations that are used in related works.

Abbreviation	Machine Learning Techniques
ARIMA	Autoregressive integrated moving average
ANN	Artificial Neural Network
BN	Belief Network
CART	Classification and Regression Trees
DB-SCAN	Density-Based Spatial Clustering of Applications with Noise
DT	Decision Trees
EV	Ensemble Voting
ELSR	Ensemble Least Square Regression
GBM	Gradient Boosting Machines
GP	Gaussian Processes
HMM	Hidden Markov Model
KNN	K-Nearest Neighbour
KL divergence	Kullback–Leibler Divergence
LBMPC	Learning-Based Model Predictive Control
LD	Linear Discriminant
LDA	Latent Dirichlet Allocation
LMV	Linear Minimum Variance
LP	Linear Regression
MLE	Maximum Likelihood Estimate
MLP	Multi-Layer Perceptron
NB	Naïve Bayes
NMF	Non-negative Matrix Factorization
NN	Neural Networks
MP	MultiPath fading
RF	Random Forest
RBF	Radial Basis Function
SVM	Support Vector Machine
SVR	Support Vector Regression
TAN	Tree Augmented Naïve Bayes network
THR	Thresholding

Appendix D

Completed Ethics Proposal, Ethics Approval and Survey Documents



RMIT Human Research Ethics Committee

Risk Assessment and Application Form

INSTRUCTIONS

Completion Instructions

This is a dynamic form designed to be completed and submitted electronically. It is comprised of two parts - **Part A: Risk Assessment Checklist** and **Part B: Human Research Ethics Committee Application Form** (More than Low Risk Research) or **CHEAN Application Form** (Negligible/Low Risk/Learning & Teaching Practice Research).

Please do not print this form. Save a blank copy to your computer and open it with the free Adobe Reader software (version 9.0 or later). If you do not have Adobe Reader installed, or are using an older version, you can download the latest free version for Mac and PC from the [Adobe website](#). Mac Users please note that the default program 'Preview' will not display this form correctly - you will need to install Adobe Reader to view and edit this form.

Please do not save and re-use this form. For each new ethics risk assessment and application download the current version from the HREC/CHEAN website. The risk assessment checklist and application form is updated regularly. 'Old' versions of the form will not be accepted by the HREC/CHEAN.

Submission Instructions

HREC Application Form: attach the 'more than low risk' application, along with any relevant documentation, to an email addressed to the Secretary of the HREC at human.ethics@rmit.edu.au

CHEAN Application Form: attach the 'negligible or low risk' application, along with any relevant documentation, to an email addressed to the Secretary of the appropriate CHEAN for your College as follows:

Science Engineering and Health CHEAN	SEH-Human-Ethics@ems.rmit.edu.au	(+ 61 3) 9926 7096
Business CHEAN	BCHEAN@rmit.edu.au	(+ 61 3) 9925 5596
Design and Social Context CHEAN	DSCethics@rmit.edu.au	(+61 3) 9925 3283

Applicants are reminded that the research must only commence after formal Human Research Ethics Committee or CHEAN final approval has been granted. Please allow a minimum of 30 working days from the submission of your application to receive a first response from the Human Research Ethics Committee or CHEAN.

PART A: RISK ASSESSMENT CHECKLIST

The following Risk Assessment Checklist will determine the level of risk associated with the research to be undertaken. Based on the outcome the appropriate application form will need to be submitted to either the CHEAN (negligible/low risk) or the Human Research Ethics Committee (more than low risk). The risks identified below may apply to the participants, the research team, the University or the wider community. These questions should assist you to identify risks that exist, and then to develop strategies to negate, minimise or manage these risk factors.

1. Assessment of Research Topics and Procedures

If you are unsure whether you have answered any of the following questions correctly seek further information via the secretary of the appropriate CHEAN (see above) and/or continue to complete the risk assessment.

A research project may be classified as 'exempt from ethics review' if certain criteria can be met.

1.1 Do any of the following criteria apply to your research project?

- | | | |
|---|--|------------------------------|
| 1.1.1 Are you using an existing data set and conducting a secondary analysis of the data? | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 1.1.2 Are the data of this existing data set non-identifiable to you or any of the investigators? | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 1.1.3 Have you received written permission to access and use information from this data set? | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |

If you responded **Yes** to all of the above three questions then your research may be classified as “exempt from review”. You do not need to seek formal ethics approval for your research project.

If you responded ‘No’ to any question, continue to complete the risk assessment.

The *National Statement on Ethical Conduct in Human Research* defines negligible risk as no foreseeable risks of harm or discomfort and any foreseeable risk is no more than inconvenience to the participants (Section 2.1.7). The *National Statement* describes inconvenience as the least form of harm that is possible for human participants in research. The most common example of inconvenience is participating in a non-identifiable survey and giving up time to do so. Does this research only involve negligible risk in your opinion? No Yes

Continue the risk assessment to confirm whether your proposed research is eligible for consideration for negligible level risk review by the College Human Ethics Advisory Network (CHEAN).

1.2 Are any of the following procedures or participants included?

- 1.2.1 Participants are identifiable or re-identifiable (i.e. codes are used) No Yes
- 1.2.2 Some form of deception is involved No Yes
- 1.2.3 Participants are aged less than 18 years No Yes
- 1.2.4 Participants are cognitively or emotionally impaired No Yes
- 1.2.5 Participants consider themselves to be Aboriginal or Torres Strait Islander people No Yes
- 1.2.6 Participants belong to a cultural/minority group No Yes
- 1.2.7 The procedure used in the research involves any experimental manipulation or includes the presentation of any stimulus other than question-asking No Yes
- 1.2.8 The questions asked include personally sensitive and/or culturally sensitive issues No Yes
- 1.2.9 There is a power-dependency relationship between researcher(s) and participant(s) e.g. the doctor/patient or teacher/student relationship No Yes

If ‘No’ has been answered to ALL questions, the project may be considered of negligible risk and is eligible for review by the CHEAN. Please proceed to complete the negligible risk application form. Go to the end of risk assessment form and select Negligible Risk.

If you responded ‘Yes’ to any question, continue to complete the risk assessment.

2.1 Are any of the following topics directly under investigation in part or in whole?

- 2.1.1 Any disease or health problem No Yes
- 2.1.2 Any psychological disorder, depression, mood states and/or anxiety No Yes
- 2.1.3 Cultural issues that may be sensitive to a particular community No Yes
- 2.1.4 Eating disorders No Yes
- 2.1.5 Fertility No Yes
- 2.1.6 Gambling No Yes
- 2.1.7 Gender identity No Yes
- 2.1.8 Grief, death or serious/traumatic loss No Yes
- 2.1.9 Illicit drug taking No Yes

- | | | |
|---|--|------------------------------|
| 2.1.10 Information or issues that may be sensitive to an individual | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.1.11 Parenting behaviour | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.1.12 Race or ethnic identity | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.1.13 Self report of criminal behaviour / illegal activity | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.1.14 Sexuality and sexual behaviour | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.1.15 Substance abuse | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.1.16 Suicide | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.1.17 Termination of pregnancy | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.1.18 Young people under the age of 18, except in a normal educational context involving standard procedures | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |

2.2 Are any of the following procedures to be employed?

- | | | |
|--|--|------------------------------|
| 2.2.1 Administration of drugs or placebos | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.2 Administration of ionising radiation | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.3 Administration of physical stimulation | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.4 Audio or visual recording | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.5 Collection of tissue / blood / body fluid / genetic material | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.6 Covert observation | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.7 Deception of participants | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.8 Invasive physical procedures / risk of physical injury | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.9 Physical exertion / risk of physical injury | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.10 Procedures inflicting pain | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.11 Psychological interventions or treatments | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.12 Substance abuse | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.13 Use of hazardous substance (e.g. carcinogens, teratogens, explosive materials) | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.14 Use of medical records where participants can be identified or linked | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.15 Use of microorganisms (e.g. bacteria, fungi) | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.16 Use of personal data obtained from Commonwealth or State Gov't Department/Agency | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 2.2.17 Withholding from one group specific treatments or methods of learning, from which they may "benefit" (e.g. in medicine or teaching) | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |

2.3 External Obligations

- | | | |
|--|--|------------------------------|
| 2.3.1 Is the research funded externally? | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
|--|--|------------------------------|

3. Participant Vulnerability Assessment

3.1 Does the research specifically target participants from any of the following groups?



- | | | |
|---|--|------------------------------|
| 3.1.1 Members of a socially identifiable group with special cultural or religious needs or political vulnerabilities | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 3.1.2 People able to be identified in any final report when specific consent for this has not been given | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 3.1.3 People highly dependent on medical care | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 3.1.4 People in a workplace setting with the potential for coercion or problems of confidentiality (e.g. employer/ employee) | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 3.1.5 People in a dependent or unequal relationship with the researchers (e.g. lecturer/student, doctor/patient, teacher/pupil, professional/client) | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 3.1.6 People not usually considered vulnerable but would be thought so in the context of the project | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 3.1.7 People unable to give free informed consent because of difficulties in understanding the Plain Language Statement or Information Sheet (e.g. language difficulties) | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 3.1.8 People whose ability to give consent is impaired | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 3.1.9 People with a physical disability or vulnerability | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 3.1.10 People with existing relationships with the researcher (e.g. relative, friend, co-worker) | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 3.1.11 Residents of a custodial institution | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 3.1.12 Aboriginal and Torres Islander individuals, communities or groups | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |

4. Research in Overseas Settings Assessment

4.1 Does the research involve any of the following?

- | | | |
|--|--|------------------------------|
| 4.1.1 Research being undertaken in a politically unstable area | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 4.1.2 Research in countries where criticism of government and institutions might put participants and/or researchers at risk | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |
| 4.1.3 Research involving sensitive social / cultural / political / ethnic / religious issues | <input checked="" type="checkbox"/> No | <input type="checkbox"/> Yes |

5. Risk Assessment Outcome

Please indicate the Level of Risk associated with your research, based on your responses to the Risk Assessment Checklist (using the information below as a guide). Once you have selected a checkbox below, proceed to Part B.

- Negligible Risk (CHEAN)**

 Low Risk (CHEAN)

 More than Low Risk (HREC)



Negligible Risk. If you answered 'No' to all of the questions in Section 1, select the Negligible Risk checkbox and proceed to Part B to complete an application for CHEAN review.

Low Risk. If you answered 'No' to all the above questions OR you answered 'Yes' to any of the above questions, but 'No' to all of the related sub-questions, your research would normally be deemed Low Risk and eligible for review by the CHEAN. Select the Low Risk checkbox and proceed to Part B to complete an Application for CHEAN review.

Research into learning and teaching practice. If you answered 'Yes' to question 3.1.5 and the research is into learning and teaching practice, your research would normally be deemed Low Risk and eligible for review by the CHEAN. Select the Low Risk checkbox and proceed to Part B to complete an Application for CHEAN review.

More than Low Risk. If you answered 'Yes' to any questions that did not have sub-questions OR you answered 'Yes' to several questions and their related sub-questions, your research would normally be deemed More than Low Risk. Select the More than Low Risk checkbox and proceed to Part B to complete a full HREC Application.

Exception: If you still believe that, due to the particular nature of the project or the participants, your proposal may still be eligible for review by CHEAN please provide details below. Then, select the Low Risk checkbox and proceed to Part B to complete an Application for CHEAN review.

PART B: CHEAN APPLICATION FORM

Please complete all of the following sections. The questions displayed in this Application form are specific to the Level of Risk associated with the project you are undertaking, as indicated by the checkbox selected in the previous section. Please note answer boxes expand.

1. General Details

1.1 Project Title

1.2 Chief Investigator

Title & Full Name	<input type="text" value="Dr. Flora Salim"/>		
Position	<input type="text" value="Lecturer"/>	Staff ID	<input type="text"/>
School/Institute	<input type="text" value="School of Computer Science and IT (CSIT)"/>	Phone	<input type="text"/>
Email	<input type="text"/>		
Position / Other Affiliations Relevant to this Application	<input type="text"/>		

Have the relevant online training modules (Human Research Ethics & Research Integrity) been completed? Yes No
If no, then see [here](#).

1.3 Principal Research Student

Title & Full Name	<input type="text" value="Mr. Irvan Bastian Arief Ang"/>	Student ID	<input type="text"/>
Course of Study	<input type="text" value="DR 221"/>	Staff ID	<input type="text"/>
School/Institute	<input type="text" value="School of Computer Science and IT (CSIT)"/>		

Have the relevant online training modules (Human Research Ethics & Research Integrity) been completed? Yes No
If no, then see [here](#).

1.4 Co-investigator(s) Working on Procedures that Involve Humans

Title & Full Name	<input type="text" value="A/Prof Margaret Hamilton"/>	Staff ID	<input type="text"/>
Position	<input type="text" value="Associate Professor"/>	Student ID	<input type="text"/>
School/Institute	<input type="text" value="School of Computer Science and IT (CSIT)"/>	Phone	<input type="text"/>
Email	<input type="text"/>		
Position / Other Affiliations Relevant to this Application If Student, provide Details on Level and Course of Study	<input type="text"/>		

Have the relevant online training modules (Human Research Ethics & Research Integrity) been completed? Yes No
If no, then see [here](#).

Add more Co-Investigators

Remove Co-Investigator

1.5 Proposed Project Duration

Proposed project commencement date Proposed project conclusion date

Project Summary
(50 words or less)

We aim to research and understand the factors that influence comfort and energy use within office environments. This data will be used by a PhD Candidate researching individual's and group's habits and behaviours related to energy use and their perception on thermal and environmental comfort.

Project aims & justification
(including reference to key literature) approx 500 words

This research will investigate how individuals use energy in their office, in relation to the ambient environmental conditions. Using controlled-group research, we can gather the main variables and set the ground-truth data for the future research.

The goal of this research is to encourage building users to be aware of their energy use and carbon footprint and help them to make an informed decision about reducing them.

This project will produce the following deliverables:
 (1) Tools for monitoring and analysing individual's and group's energy consumption in an office setting.
 (2) An accurate dataset obtained from real experiments.
 (3) Several papers that will be submitted to top conferences / journals in Computer Science and related areas.

The data gathered can be used in preparing a new course material on Green Office and Classrooms.

We will design, develop and create a set of tools (hardware and software) to gather quantitative measurements of light, heat, noise, humidity, occupancy, and energy use of each electrical device or appliance, possibly providing real time information to the participants. We will also develop a set of questionnaires to gather participant's qualitative feedback about their office environment.

Using sensors and power meters placed in participants' office, each individual data is collected. Those data will include but not limited to electric power consumption, heat data, lighting data, noise data, temperature data, humidity data, etc. Once the data is gathered, we can integrate those data with carbon footprint and energy use calculation.

The research can be conducted in three phases:

Phase 1 - Ground Truth Data Collection
 Designing and developing monitoring tools to be used by individuals in a controlled group experiment. We will fit the offices of 5-10 participants with sensors and power monitors. We will survey the participants to ensure the ground truth is acquired. This will be conducted over a period of 3 months.

Phase 2 - Data Analysis
 The data gathered from Phase 1 will be analysed to infer users' occupancy and space and appliance usage and correlate them with comfort and energy use.

Phase 3 - Further Data Collection
 Similar method with phase one with more participants to obtain more proper data with high accuracy rate. We will conduct a longer study with the original participants, and with new participants, particularly CS&IT PhD students, to observe changes over different seasons.

Research design/approach

2. Project Details

2.1 Project Description

Methodology/data collection techniques & analysis

Phase 4 - User Feedback and Behavioural Change Monitoring

In this phase, users will be exposed to their personal energy use, comfort, and carbon footprint data through a web-based visualisation. The changes in their habits and behaviours, if any, will be observed.

Each participant need to fill a survey at the end of the participation.

Based on the description above, phase 1 and phase 3 are related to data collection:

Phase 1: Ground Truth Data Collection

In this phase, we will pick small number of participant (5-10 people) on RMIT building 14 level 10. Power meter(s) device will be installed in participant electronic device and multiple sensors need to be planted in the office or participant's desk to measure heat, lightning, noise, temperature, humidity, etc to get the good data related to green variable.

Daily surveys are performed to ascertain the monitoring data collected using the sensors. A final survey at the end of Phase 1 will be given out to collect feedback from participants.

Phase 3: Further Data Collection

This phase is similar with phase 1 but will engage more participants, mainly PhD students, to monitor shared office environments.

2.2 Project Type

- | | | |
|---|---|--|
| <input checked="" type="checkbox"/> Research by Academic Staff Member | <input checked="" type="checkbox"/> Postgraduate Research | <input type="checkbox"/> Contract Research |
| <input type="checkbox"/> Masters by Coursework | <input type="checkbox"/> Undergraduate Research | <input type="checkbox"/> Honours Research |
| <input type="checkbox"/> Research into Teaching and Learning Practice | | <input type="checkbox"/> Other |

2.3 Is this project part of a larger project?

- No Yes

Please specify the title of the larger project, the name, title and affiliation of the Chief Investigator for the larger project, the name of the approving ethics committee(s) of the larger project, and the notice of approval.

1. iCo2mmunity funded by Sustainable Urban Precinct Project (SUPP)
2. Greener Office and Classrooms funded by SUPP

2.4 Does this project involve multi-centre research?

- No Yes

2.5 Is this research project specific to research into university learning and teaching practice?

- No Yes

3. Participant Details

Does your research project involve:

- Human participants *Complete Section 3 below*
- Use of data banks only
- Aboriginal and Torres Straight Islander people or communities

3.1 Number of Participants

Males Females Total

If a gender imbalance in the number of participants is apparent, please explain why:

3.2 Age Range

18 and above

3.3 Will any participants be ill or frail?

No Yes

3.4 Are there any criteria that will determine whether participants are included or excluded from the research?

No Yes

Provide details of all inclusion and exclusion criteria and explain why each criterion is important to the purpose of the research:

Participants must be a student or staff of RMIT with an office in the school of CSIT, can speak English fluently and not be ill or frail.

3.5 Recruitment Method

Please state how names and contact details of potential participants will be obtained, from where they will be recruited, how they will be invited to participate, and who will approach potential participants to seek their involvement. Sample copies of recruitment advertisements should be submitted with this application.

Note: Where participants are recruited from schools, hospitals, prisons or other institutions, written permission/approval from the institution or appropriate authority must be provided. See Question 7.

Personal emails to students and staff with an office in the School of Computer Science and IT, face-to-face recruitment.

3.6 Compensation to Participants

Not Applicable Applicable



3.7 Are any of the participants students of RMIT?

No Yes

Please explain the steps taken by the investigators to ensure that the students' participation is purely voluntary:

We will clearly explain that participation is purely voluntary and will not have any repercussions on their academic results. They will sign a consent letter to state their voluntary participation, and they may be able to remove themselves from the research at anytime during the experiment.

3.8 Does this project require the researchers to have a *Working with Children* Check? Information about this requirement is available on the Working with Children web site and in the Guidelines to Complete the Application Form

No Yes

4. Research Using Existing Databases

4.1 If the research involves access to existing database provided by an institution(s), please indicate:

Source(s) and number of records	N/A
Whether data to be used will be de-identified, potentially identifiable (e.g. coded), or identified	N/A
Whether permission has been granted by donors to use these data for research purposes	N/A
Whether formal permission/clearance has been sought or obtained from the relevant institution(s) (see also Section 7 below)	N/A

5. Description of Procedures

5.1 Describe in detail below exactly what participants will be asked to do and emphasise anything that may have adverse consequences.

Participants will be asked to fill a daily survey.
At the end of the experiment, participant will need to fill a completion survey.

5.2 Will questionnaire(s) (including those that are published or commercially available) be used in the project? Please attach a copy to this application.

No Yes (please attach a copy to this application)

5.3 If interviews or focus groups are to be held, please indicate the kinds of questions to be asked below or attach the interview schedule in the case of structured interviews.

N/A

5.4 Will participants at any time have pictures taken of them, either photographed or video recordings, or be audio recorded?

Not applicable Yes

6. Study Location

6.1 Please identify the precise location of the study

If permission is required to use the location, also indicate how permission will be obtained:

In RMIT. No permission required, except the consent from participants.

7. External Approvals

If a project requires approval from other institutions or ethics committees, next of kin or guardian, or representative or authority in the case of special groups, copies of such approvals must be provided to the RMIT HREC at the time of application or be made available as soon as possible thereafter.

7.1 Institutional

Name(s) of institution/ethics committee/authority:

Yes (details below) Yes, to follow (estimate when likely to be obtained below) No (please explain below) N/A

8. Informed Consent

8.1 How will consent be obtained?

Written consent form Informed consent implied by return of anonymous survey

Verbal Consent (explain below how consent will be recorded)

8.2 How will competence to give consent be determined and who will make this determination?

Please provide details below:

All participants are above 18 and have offices in CS&IT, therefore they have full competence to provide consent.



8.3 Will Participant Information Sheets and Consent Forms be used?

- Yes (copies attached) No (please explain below) An alternative method of obtaining consent will be used (please specify below)

8.4 Will Participant Information Sheets and Consent Forms be translated into the participants' first language?

- Yes (please provide copies of translations) No (please explain below) N/A

The information will be provided in English.

8.5 Will all participants have the capacity to give voluntary and informed consent?

- Yes No

9. Recording and Security of Project Documentation

9.1 How will data be recorded? (e.g. written questionnaires, interview notes, photographs, audio/video recording, direct electronic data entry).

Surveys, sensors data and power meters data.

9.2 Will confidentiality of results be maintained?

- Yes No

Please provide details:

Results will be known only by researchers. Users can register to view the statistical results on the web app, but they are unable to identify the contributors of data, as all users will not use their real name, a username will be used.

9.3 Indicate how the security of project documentation will be maintained and specify the precise location of the storage place(s):

9.3.1 During the study	Stored in a secured PC in RMIT and a cloud-based server with authenticated access, with password protection.
9.3.2 Following completion of the study	Stored in a secured PC in RMIT and a cloud-based server with authenticated access, with password protection.

Project documentation should be stored in secure, lockable locations, preferably on campus. Computer files should be password protected. Data, de-identified where appropriate, and consent forms should normally be kept for a period consistent with the Public Records Office of Victoria Standard (02/01) (normally 5 years for non clinical trial data and 15 years for clinical trial data following publication).



9.4 Will any data (including samples) be preserved for possible future use in another project either by yourself or another researcher?

Yes No

Please explain the nature of the data to be preserved, when the data might be used in another project, how that data might be used, for what purpose it might be used, and who might be given access to the data for another project:

In a secure server in RMIT, with password. The data may be used in another project for further data analytics and statistics. Future PhD students in RMIT may request data.

10. Dissemination of Results

10.1 Will participants be informed that results from the study may appear in publications, be included in a thesis or report, or be presented at conferences? (If relevant, this information should be included in the Participant Information Sheet and given to participants prior to obtaining informed consent).

Yes No

Please provide details:

If participants want to know about the outcomes of this study, they will provide their email address for correspondence purposes only, to be contacted later when the publications are complete.

10.2 Will participants be informed that results from the study will be available to them on request? (If relevant, this information should be included in the Participant Information Sheet and given to participants prior to obtaining informed consent).

Yes No

10.3 Will participants be informed that their personal data collected in the course of the research will be available to them on request? (If relevant, this information should be included in the Participant Information Sheet and given to participants prior to obtaining informed consent).

Yes No

Please explain:

The visualization of the aggregated data can be viewed directly on the mobile application.

11. Risk and Indemnity

11.1 Is there any risk of physical, psychological, social, legal or financial and/or community, employment and/or professional harm to the participant or organisation?

Yes No

ATTACHMENT CHECKLIST

Ensure the following attachments are included (where applicable)

- copy of recruitment advertisement(s) (Section 3.5)
- copy of questionnaire(s) and/or proposed interview/focus group outline (Section 5.3)
- debriefing documentation for participants (Section 5.5)
- evidence of permission to use places off-campus (Section 6)
- evidence of external approvals (Section 7)
- copy of the proposed Participant Information Sheet(s) & Consent Form(s) (Section 8.3)
- copy of translations of Participant Information Sheet(s) and Consent Form(s) (Section 8.4)
- copy of statement from medical practitioner or other relevant health professional accepting responsibility for procedures (Section 11.2 and/or 13.1)
- details of arrangements for first aid (Section 11.3)
- copy of funding application(s) (Section 16)

DECLARATION

By submitting this application, **we, the Chief Investigator and Co-Investigators**, declare that we:

- have read and agree to abide by the conditions and constraints of the *National Statement on Ethical Conduct in Human Research (2007)* and any other relevant University and / or statutory requirements;
- accept responsibility for the accuracy of the information provided in this application and for the conduct of this research, in accordance with the principles contained in the *National Statement* and any other conditions specified by the University Human Research Ethics Committee;
- abide by the terms and conditions set by the University Human Research Ethics Committee
- will ensure that the qualifications and / or experience of all RMIT personnel involved with the project are appropriate to their role and/or to the procedures performed;
- will ensure that appropriate permits from external organisations or agencies will be obtained and that any imposed conditions will be observed;
- certify that the information contained in this application is true and accurate;
- understand that the information contained in this application is given on the basis that it remains confidential in accordance with relevant University and statutory requirements;
- will seek approval for modifications to the research prior to their implementation.

By submitting this application⁽¹⁾, **I, the Chief Investigator**, declare that I :

- have ensured that the head of school has sighted this application and that s/he agrees that the required academic expertise and resources are available to complete this proposed research. Evidence of this assurance will be retained⁽²⁾.

¹ The application must be submitted electronically by the Chief Investigator from his/her RMIT staff email account.

² This evidence may consist of a hard-copy signed document or an email from the head of school acknowledging that they have sighted your application and have agreed that the required academic expertise and resources are available to complete this proposed research.

D.1 Ethics Approval Document



14th January 2015

Flora Salim
Building 14 Level 10, Room 3
School of Computer Science & IT
RMIT University

Dear Flora

BSEHAPP 42-14 SALIM-ANG iCO2mmunity: Individual and Group Energy Usage Measurement and Comfort Analysis in Office Environments

Thank you for submitting your amended application for review.

I am pleased to inform you that the CHEAN has approved your application for a period of **3 Years** from the date of this letter to **14th January 2018** and your research may now proceed.

The CHEAN would like to remind you that:

All data should be stored on University Network systems. These systems provide high levels of manageable security and data integrity, can provide secure remote access, are backed up on a regular basis and can provide Disaster Recover processes should a large scale incident occur. The use of portable devices such as CDs and memory sticks is valid for archiving; data transport where necessary and for some works in progress. The authoritative copy of all current data should reside on appropriate network systems; and the Principal Investigator is responsible for the retention and storage of the original data pertaining to the project for a minimum period of five years.

Please Note: Annual reports are due on the anniversary of the commencement date for all research projects that have been approved by the CHEAN. Ongoing approval is conditional upon the submission of annual reports failure to provide an annual report may result in Ethics approval being withdrawn.

Final reports are due within six months of the project expiring or as soon as possible after your research project has concluded.

The annual/final reports forms can be found at:
www.rmit.edu.au/staff/research/human-research-ethics

Yours faithfully,

Dr Falk Scholer
Deputy Chair, Science Engineering & Health
College Human Ethics Advisory Network

Cc Student Investigator/s: Irvan Bastian Arief Ang School of Computer Science & IT RMIT University
Other Investigator/s: Margaret Hamilton School of Computer Science & IT RMIT University

RMIT University

**Science Engineering
and Health**

**College Human Ethics
Advisory Network
(CHEAN)**

Plenty Road
Bundoora VIC 3083

PO Box 71
Bundoora VIC 3083
Australia

Tel. +61 3 9925 7096
Fax +61 3 9925 6506
• www.rmit.edu.au

D.2 Participant Information Document



School of Computer Science and
Information Technology

GPO Box 2476V
Melbourne VIC 3001
Australia

• www.rmit.edu.au

INVITATION TO PARTICIPATE IN A RESEARCH PROJECT

PARTICIPANT INFORMATION

iCO2mmunity: Individual and Group Energy Usage Measurement and Comfort Analysis in Office Environments

Chief Investigators:

Irvan Bastian Arief Ang, School of Computer Science & IT, RMIT University

Dr. Flora Salim, School of Computer Science & IT, RMIT University

A/Prof Margaret Hamilton, School of Computer Science & IT, RMIT University

You are invited to participate in research being conducted by people in the School of Computer Science and IT at RMIT University.

Please read this information carefully and be confident that you understand its contents before deciding whether to participate.

Participation is purely voluntary. Participation in this study does not impact you directly or your assessment in any courses being conducted at RMIT University. If you have any questions about the project, please ask any one of the investigators listed above, by emailing them.

Who is involved in this research? Why is it being conducted?

This research aims to determine what factors affect the energy consumption of individuals and groups in offices and calculate to what extent these factors contribute to global warming. We are interested to know how many watts of power that each individual uses and plan to measure the lighting levels, noise, heat and humidity. We would also like to gather information about comfort perception at various places and times which will enable us to have a better understanding and develop a strategy to reduce individual carbon footprints.

Why have you been approached?

You have been approached because you work, study and have an office in RMIT School of Computer Science and IT.

If I agree to participate, what will I be required to do?

If you agree, we will arrange a convenient time with you to install a power meter on your most commonly used electronic devices. Furthermore, we will also install sensors in your office to measure light, noise level, heat, temperature, and humidity data.

Participants will need to complete a short survey every day, and a longer survey at the end.

What are the possible risks and disadvantages?

There are no direct risks of physical or emotional harm.

What are the benefits associated with participation?

The anticipated benefits are a better knowledge or intelligence about carbon footprint and how individuals can measure and alter it.

What will happen to the information I provide?

Your response will be stored in a secured way with other responses to be analysed and the researchers will develop an optimal ways to reduce individual carbon footprints. The information will also be used for future related research.

What are my rights as a participant?

As a participant in this research you have the rights to:

- withdraw from participation at any time.
- have any unprocessed data withdrawn and destroyed provided it can be reliably identified.
- have any questions answered at any time.
- request access to any publications resulting from this study.

Whom should I contact if I have any questions?

Please contact any of the investigators listed at the beginning of this participant information.

Yours sincerely

Irvan Bastian Arief Ang
Flora Salim
Margaret Hamilton

Any complaints about your participation in this project may be directed to the Ethics Officer, RMIT Human Research Ethics Committee, Research & Innovation, RMIT, GPO Box 2476V, Melbourne, 3001.

The telephone number is (03) 9925 2251.
Details of the complaints procedure are available on the

[Complaints with respect to participation in research at RMIT](#) page

D.3 Consent Form Document

CONSENT FORM

- 1. I have had the project explained to me, and I have read the information sheet
- 2. I agree to participate in the research project as described
- 3. I agree:

to undertake the tests or procedures outlined above
to be interviewed and/or complete a questionnaire
that my participation will involve identifying my daily routine activities and I agree that researcher may use them for publication in journals and conferences and the resulting research publication will be also be published online in the RMIT University online repository: <http://researchbank.rmit.edu.au/>.

- 4. I acknowledge that:

- (a) I understand that my participation is voluntary and that I am free to withdraw from the project at any time and to withdraw any unprocessed data previously supplied (unless follow-up is needed for safety).
- (b) The project is for the purpose of research.
- (c) The privacy of the personal information I provide will be safeguarded and only disclosed where I have consented to the disclosure or as required by law.
- (d) The security of the research data will be protected during and after completion of the study. The data will be stored in secured PC in RMIT and a cloud-based server with authenticated access and password protection for five years. The data collected during the study may be published, and a report of the project outcomes will be provided to Irvan Bastian Arief Ang. Any information which will identify me will not be used.

- I agree the research data will be stored for future use
- I disagree the research data will be stored for future use

Participant's Consent

Participant: _____ Date: _____
(Signature)

Participants should be given a photocopy of this PICF after it has been signed.

D.4 Survey Form Document

iCO2mmunity: Individual and Group Energy Usage Measurement and Comfort Analysis in Office Environments

----- **Final Survey** -----

1. I consider myself as a green person – I never waste any resources.
 Strongly Disagree Disagree Neutral Agree Strongly Agree

2. I consider and think to reduce my individual carbon footprint.
 Strongly Disagree Disagree Neutral Agree Strongly Agree

3. I think this experiment will encourage the participant to be more aware about their carbon footprint.
 Strongly Disagree Disagree Neutral Agree Strongly Agree

4. Because of this experiment, I will share my carbon footprint related knowledge to my friends.
 Strongly Disagree Disagree Neutral Agree Strongly Agree

5. Because of this experiment, I will influence my friends to reduce their carbon foot print.
 Strongly Disagree Disagree Neutral Agree Strongly Agree

6. How have this experiment influence your behaviour in regard to energy saving?

7. Do you want to add something?

Thank you for participating in this experiment!

Appendix E

Full Experiment Accuracy Result for both Academic Staff Room and Cinema Theatre

Table E.1: Academic staff room indoor human occupancy accuracy result for 1 day and 2 days prediction with 7-13 days training data

training \ test	1 SVR	1 RUP-STD	1 RUP-STL	2 SVR	2 RUP-STD	2 RUP-STL
7	98.95%	99.30%	98.95%	96.34%	99.30%	98.61%
8	93.71%	100.00%	98.26%	89.01%	92.15%	91.80%
9	84.62%	83.92%	83.92%	76.79%	84.99%	84.12%
10	68.88%	86.06%	84.32%	69.81%	83.42%	85.51%
11	74.13%	81.12%	83.22%	87.09%	90.58%	91.62%
12	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
13	100.00%	100.00%	100.00%	-	-	-

Table E.2: Academic staff room indoor human occupancy accuracy result for 3 days and 4 days prediction with 7-11 days training data

training \ test	3 SVR	3 CDHOC-STD	3 CDHOC-STL	4 SVR	4 CDHOC-STD	4 CDHOC-STL
7	92.32%	94.42%	94.07%	85.17%	91.54%	91.19%
8	80.56%	89.42%	88.72%	77.23%	87.70%	87.87%
9	74.16%	83.82%	84.63%	80.63%	87.87%	88.48%
10	79.86%	88.95%	90.35%	84.90%	91.71%	92.76%
11	91.39%	93.71%	94.41%	-	-	-

Table E.3: Academic staff room indoor human occupancy accuracy result for 5 days and 6 days prediction with 7-9 days training data

training \ test	5 SVR	5 CDHOC-STD	5 CDHOC-STL	6 SVR	6 CDHOC-STD	6 CDHOC-STL
7	81.56%	89.67%	90.02%	84.64%	91.28%	91.69%
8	81.77%	90.16%	90.30%	84.82%	91.68%	91.91%
9	84.50%	90.29%	90.78%	-	-	-

Table E.4: Academic staff room indoor human occupancy accuracy result for 7 days prediction with 7 days training data

training \ test	7 SVR	7 CDHOC-STD	7 CDHOC-STL
7	86.84%	92.87%	92.42%

Table E.5: Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 1 day and 2 days prediction with 12-22 days training data

training \ test	1 SVR	1 CDHOC-STD	1 CDHOC-STL	2 SVR	2 CDHOC-STD	2 CDHOC-STL
12	83.85%	87.54%	83.43%	77.18%	81.07%	76.83%
13	70.31%	74.27%	71.27%	65.42%	68.90%	68.11%
14	60.49%	65.08%	66.31%	58.79%	61.69%	61.13%
15	59.97%	61.40%	56.59%	58.33%	57.35%	54.94%
16	56.04%	52.62%	46.62%	68.94%	66.18%	48.77%
17	82.56%	82.42%	80.09%	77.65%	78.81%	74.91%
18	72.71%	75.65%	71.92%	73.59%	77.72%	73.46%
19	74.59%	79.85%	75.73%	83.49%	84.01%	74.32%
20	92.41%	88.85%	87.24%	82.58%	84.41%	81.77%
21	72.61%	79.87%	74.64%	69.64%	72.25%	66.73%
22	66.46%	65.21%	59.19%	-	-	-

Table E.6: Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 3 days and 4 days prediction with 12-20 days training data

training \ test	3 SVR	3 CDHOC-STD	3 CDHOC-STL	4 SVR	4 CDHOC-STD	4 CDHOC-STL
12	71.57%	75.81%	72.81%	68.20%	71.72%	68.51%
13	62.95%	65.17%	64.02%	61.16%	61.95%	60.89%
14	57.61%	58.57%	58.45%	63.81%	63.79%	62.60%
15	66.02%	64.37%	61.82%	67.81%	66.82%	63.73%
16	70.24%	69.03%	45.85%	71.31%	71.41%	49.43%
17	76.61%	76.98%	74.98%	80.13%	78.72%	77.87%
18	79.92%	81.20%	77.52%	78.14%	80.33%	77.20%
19	79.99%	82.11%	79.45%	76.68%	77.73%	74.32%
20	77.34%	77.80%	73.99%	-	-	-

Table E.7: Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 5 days and 6 days prediction with 12-18 days training data

training \ test	5 SVR	5 CDHOC-STD	5 CDHOC-STL	6 SVR	6 CDHOC-STD	6 CDHOC-STL
12	65.64%	68.05%	65.44%	67.82%	70.15%	66.94%
13	65.34%	65.45%	63.56%	65.93%	67.05%	64.73%
14	65.56%	66.04%	64.16%	66.41%	67.99%	66.34%
15	69.21%	69.14%	66.08%	72.59%	72.58%	69.54%
16	75.25%	74.86%	55.11%	74.66%	75.65%	57.43%
17	78.66%	77.46%	77.43%	76.54%	76.13%	74.53%
18	75.87%	77.18%	73.89%	-	-	-

Table E.8: Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 7 days and 8 days prediction with 12-16 days training data

training \ test	7 SVR	7 CDHOC-STD	7 CDHOC-STL	8 SVR	8 CDHOC-STD	8 CDHOC-STL
12	68.54%	70.94%	67.48%	69.31%	71.65%	68.58%
13	67.19%	68.37%	66.51%	70.39%	71.19%	69.30%
14	70.19%	71.33%	69.45%	70.52%	72.70%	70.30%
15	72.64%	73.77%	70.35%	71.83%	72.84%	69.12%
16	73.46%	74.23%	57.62%	-	-	-

Table E.9: Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 9 days and 10 days prediction with 12-14 days training data

training \ test	9 SVR	9 CDHOC-STD	9 CDHOC-STL	10 SVR	10 CDHOC-STD	10 CDHOC-STL
12	71.95%	73.75%	70.73%	68.63%	74.52%	71.45%
13	70.67%	72.50%	70.18%	70.26%	72.00%	69.16%
14	70.10%	72.10%	69.18%	-	-	-

Table E.10: Cinema theatre indoor human occupancy accuracy result with ten unit error tolerance for 11 days prediction with 12 days training data

training \ test	11 SVR	11 CDHOC-STD	11 CDHOC-STL
12	71.52%	73.85%	70.42%

