Seminar 61 - Outliers Detection Techniques and their Benefits in Data-Driven Modeling

Machine Learning for Anomaly Detection in Subjective Thermal Comfort Votes

Zhe Wang

Lawrence Berkeley National Laboratory

zwang5@lbl.gov

ORLANDO
2020 WINTER CONFERENCE
AND AHR EXPO

Learning Objectives

- Define data outliers and their different types, along with different approaches for their detection and removal.
- Understand the outlier detection applicability in simulated and monitored data as relevant to data-driven modeling, fault detection, and operational diagnostics.
- Apply the techniques used in practical cases, shown in the session, of outlier detection in building energy performance data-driven models, thermal comfort modeling and controls, and whole building energy data quality assurance.
- Conclude that the proper outlier detection and removal is crucial in data analytics in order to avoid data manipulation and biased results.

Acknowledgements

LBNL team: Tianzhen Hong

UC Berkeley team: Thomas Parkinson, Peixian Li

This research used ASHRAE Global Thermal Comfort Database II. The authors appreciate the efforts to develop and open source this dataset for public research.

Outline/Agenda

- Motivation
- Method
- Result
- Discussion

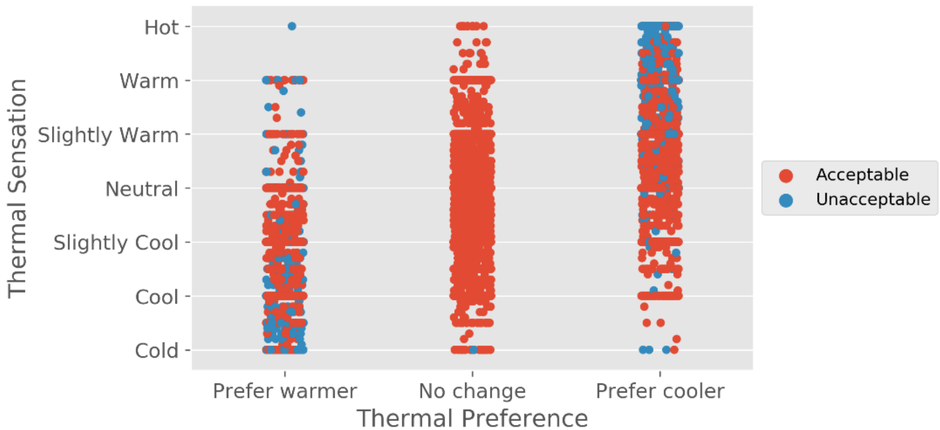
Motivation

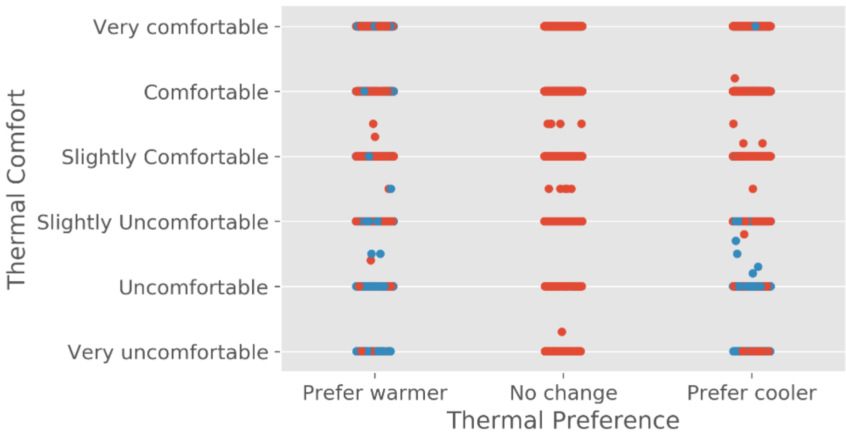
- Building thermal environment: High energy consumption, low satisfaction
  - In US, 50% of building energy are consumed for thermal environment management
  - The satisfaction level on thermal environment is low
- To better manage the thermal environment, we need to accurately measure it first
  - You cannot manage what you cannot measure -- Peter Drucker

Motivation

- Two approaches to measure building thermal environment

	Physical parameters	Subjective responses
Metrics	air temperature, relative humidity, radiant temperature, air speed, and etc.	Thermal comfort, sensation, preference, satisfaction, and etc.
Problems	Lack of explanatory power due to inter-individual differences	Subject to concerns of reliability and precision





- Occupant-in-the-loop or occupant responsive control becomes a new trend
- Outlier exists, but there is no way to detect and correct it – research gap

Motivation

- Definition
  - Outliers: refer to those thermal comfort votes that are substantially and illegitimately different from their peers
- Why we need to detect them
  - Thought it might be a valid response
  - But it introduces noise and uncertainty to thermal comfort modelling and building control

Method

- Outlier: an occupant's vote is significantly different from its peers under similar conditions
- A two-step statically-based approach

Step	Method	Metrics
find its peers under similar conditions	K nearest neighbors using Euclidean distance	• Thermal comfort • Thermal sensation
measure the dissimilarity	Quantify the probability using Gaussian Regression	• Thermal preference • Thermal acceptability

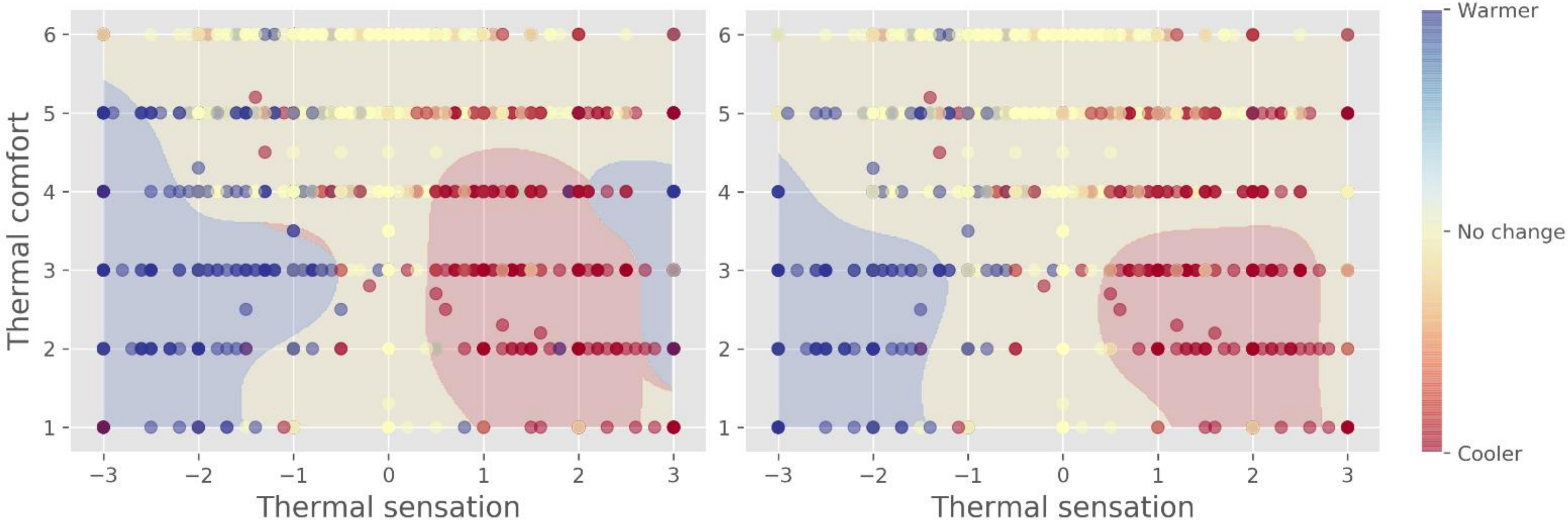
Method

Pseudocode

```
----------------------------------------------------------------------------------------------------
For each observation in the database:
    Rescaling each dimension to the same range of 0 to 1                    Step1: rescaling
----------------------------------------------------------------------------------------------------
    Find its nearest neighbors based on thermal sensation and thermal comfort by calculating
      its Euclidean Distance with the remaining observations in the database
                                                                            Step2: defining similar conditions
----------------------------------------------------------------------------------------------------
    Fit the simple multivariate Gaussian distribution on thermal acceptability and thermal
      preference with its neighbors                                         Step3: quantifying dissimilarities
    Calculate the p-value of the specific observation
----------------------------------------------------------------------------------------------------
    if the p-value is no less than the threshold:                           Step4: making decisions
        Flagged as a normal observation
    else:
        Flagged as a potential outlier
    end
end
```

Result: thermal preference

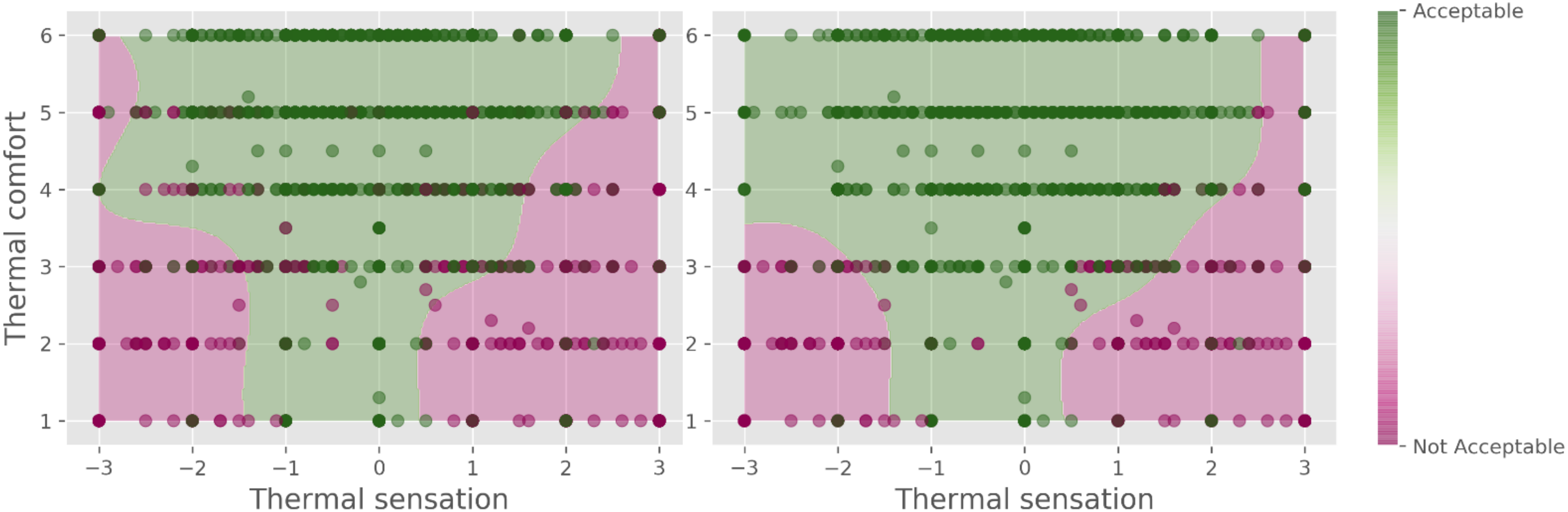
- Test our approach using ASHRAE Global Thermal Comfort Database II
- Strange voting behaviors have been removed
- Smoother boundary



The boundary was predicted by SVM

Result: thermal acceptability

- Cannot remove all strange voting behaviors

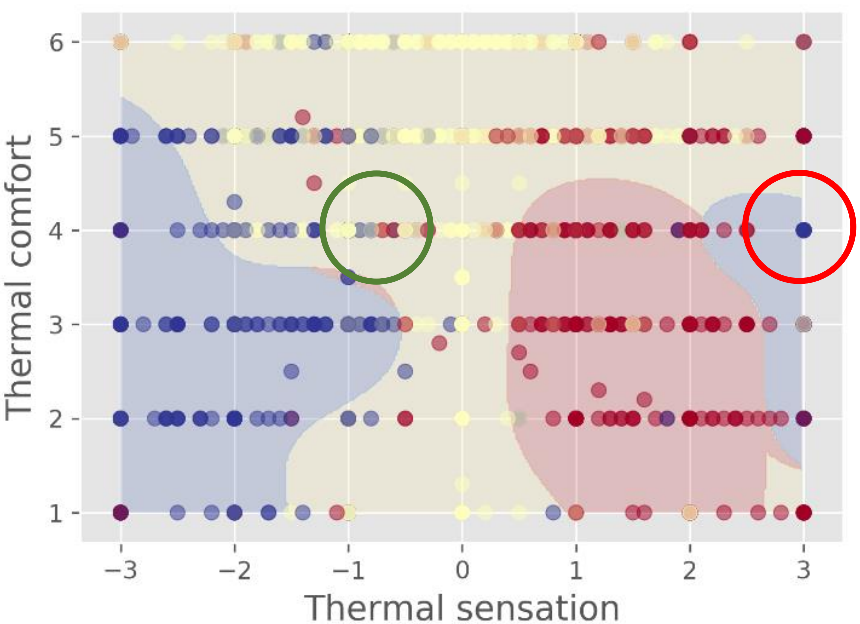


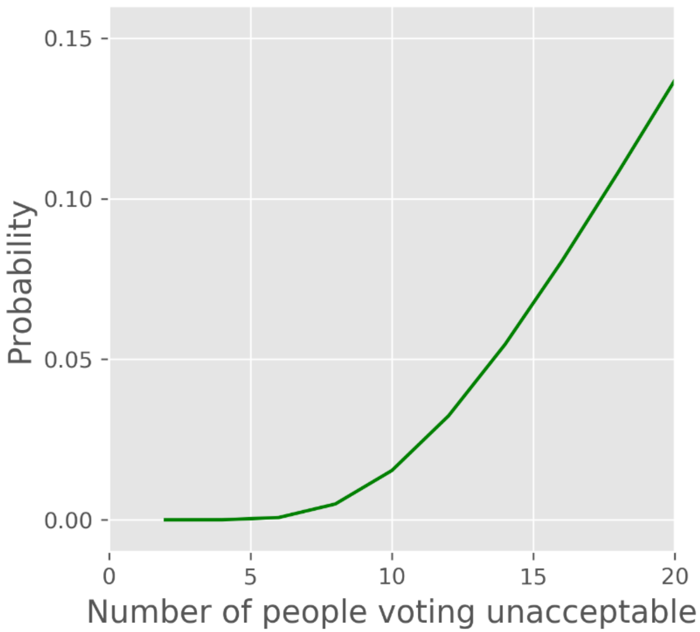
The boundary was predicted by SVM

Discussion

- How to distinguish individual differences from outliers

- Could be handled by Gaussian Regression
  - Diversified opinions → Large std. → High probability to vote differently





Assume

- *100 voters*
- *You vote unacceptable*

Discussion: Contribution

- Filled in the research gap of outlier detection for subjective thermal comfort votes
- Proposed a two-step statistical-based framework for outlier detection
  - Tune ***hyper-parameters***: the number of neighbors, the p-threshold to determine whether outlier or not
  - Use different ***metrics*** (e.g. indoor temperature) to define similar conditions
  - Use different ***approach*** (e.g. density based clustering) to define similar conditions
  - Use different ***approach*** (e.g. distance based dissimilarity) to quantify dissimilarities

Discussion: Limitation

- Just an approach to flag potential outliers, from the statistical point of view
- What is the best approach to provide comfort for occupants with unusual or significantly different thermal preferences remains an open question

Questions?

Thanks for your time and attention!

Zhe Wang

zwang5@lbl.gov

Tianzhen Hong

thong@lbl.gov

Wang, Z., Parkinson, T., Li, P., Lin, B. and Hong, T.*, 2019. The Squeaky wheel: Machine learning for anomaly detection in subjective thermal comfort votes. Building and Environment, 151, pp.219-227.*