STACK DSA

Noise Surrogate Model for Urban Planning

Spacemaker AI

Head of Operations

Karoline Skatteboe

- Authors: Deepika Dittakavi, Lois Dankwa, Tyler Gmerek

- Date: June 10, 2020
- Industry Project - Data Science & Analytics Certificate Bootcamp with Stack Education

Noise Surrogate Model

GOAL

To analyze statistical models for predicting the fraction yellow zone based on different building and noise source configurations.

Design Methodology

- Introduction
- Data Exploration and Visualization
- Feature Engineering
- Model Analysis and Performance
- Results – Test Site
- Conclusion and recommendations

Introduction

At the core of Spacemaker AI's mission is the ability to generate urban building design proposals that minimizes the effect of noise, as excessive amount of noise is deemed to interfere with the natural biorhythms of everyday life.

An alternative to the high computational conventional method for noise calculation is surrogate models, which closely mimics the behavior of simulation models but computationally cheaper to evaluate.

Hence, the goal of this project.

Data Exploration - Data

Data	Number of Scenarios	Source of noise simulations	Number of noise simulations
Non-Specific - training	9	different sites scenarios	4500
Specific - training	1	test site scenario	250
Test - testing	1	test site scenario	250

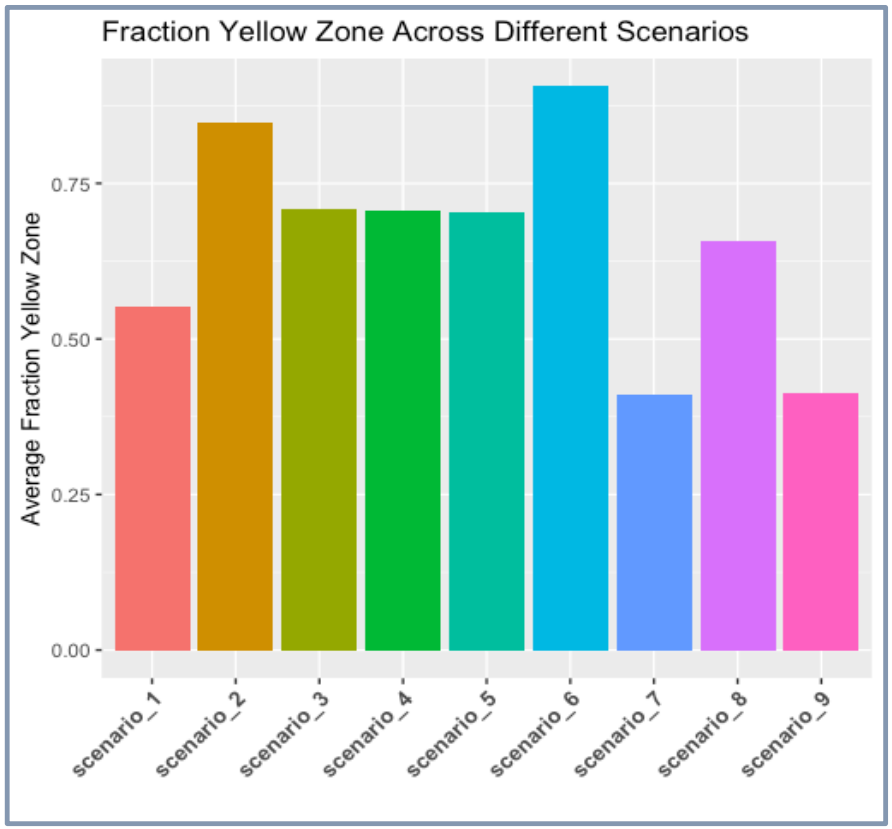
*Data source: Spacemaker AI

Data Dictionary	
Fraction Yellow Zone	Fraction of outdoor ground area in yellow zone
Building Grid	Numpy(npy) matrix file with locations and heights
Noise Source Grid	Numpy(npy) matrix file with locations
Scenario	Noise source configuration

* Source: Spacemaker AI

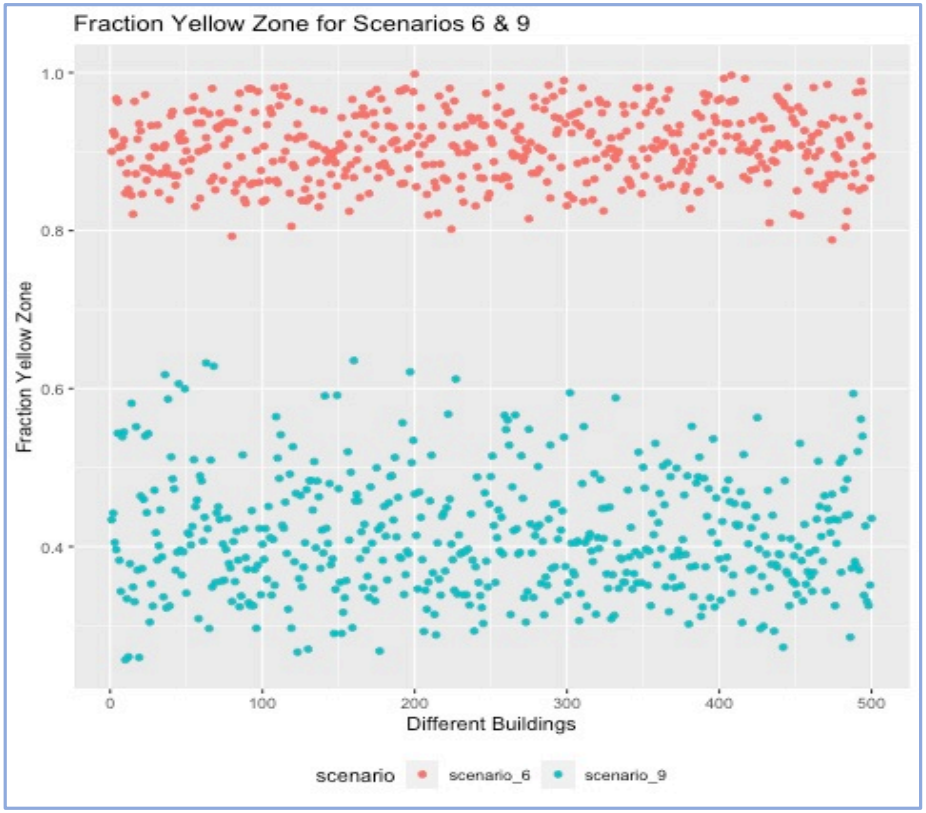
Data Exploration - scenarios

Fraction yellow zone(the dependent variable) varies with different scenarios.



Data Exploration - scenarios

- Variation is seen across different building configurations.
- Variation can be dense or sparse.



Data Exploration - grid configurations

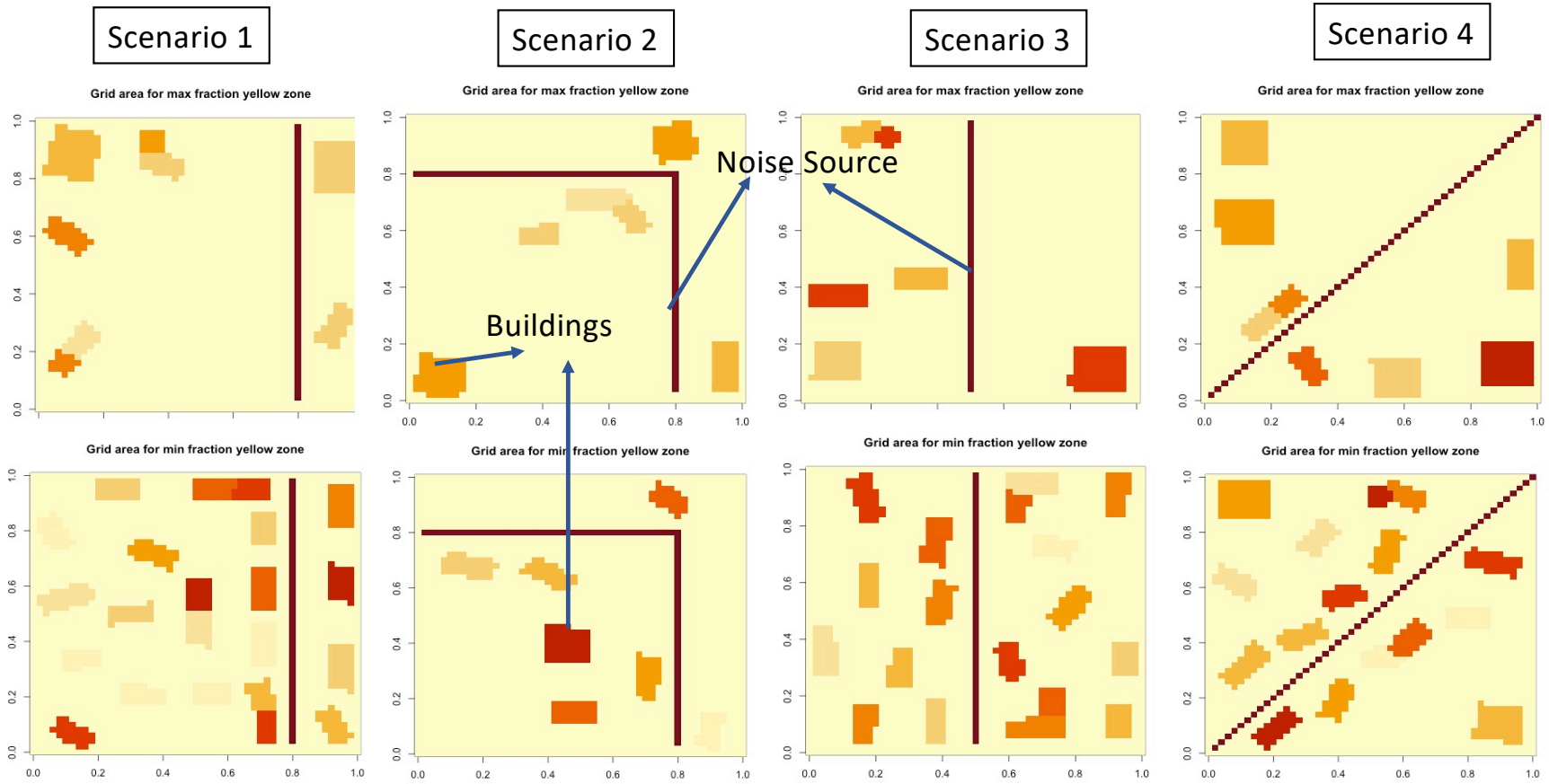
Maximum Fraction Yellow Zone:

- ❖ Visualizations for the various grid configurations for the combined building and noise source for:
  - Non-Specific data - scenario 1 – 9
  - Specific site data
  - Test site data

Minimum Fraction Yellow Zone:

- ❖ Visualizations for the various grid configurations for the combined building and noise source for:
  - Non-Specific data - scenario 1 – 9
  - Specific site data
  - Test site data

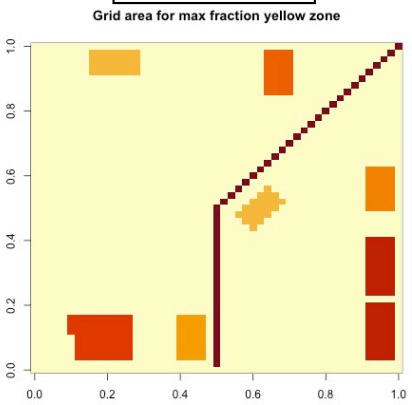
Data Exploration - grid configurations non-specific

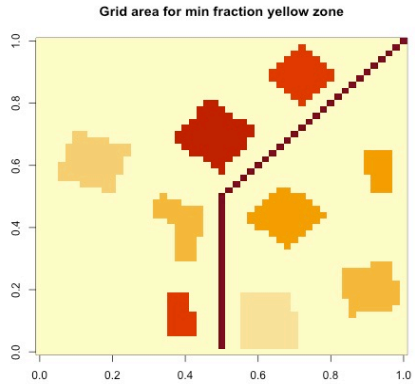


Building darker color=taller

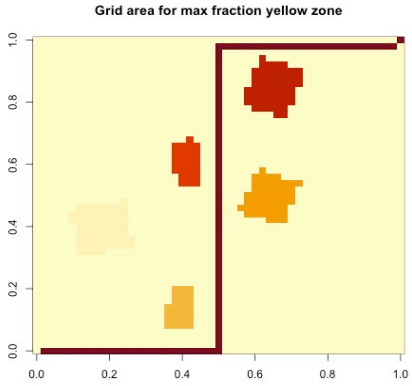
Data Exploration - grid configurations non-specific

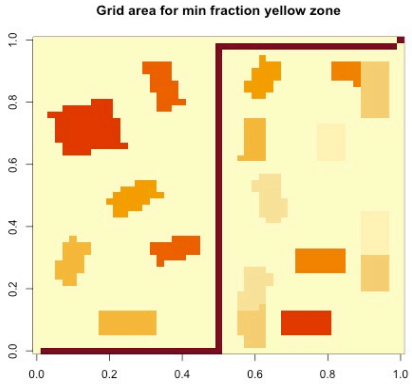
Scenario 5



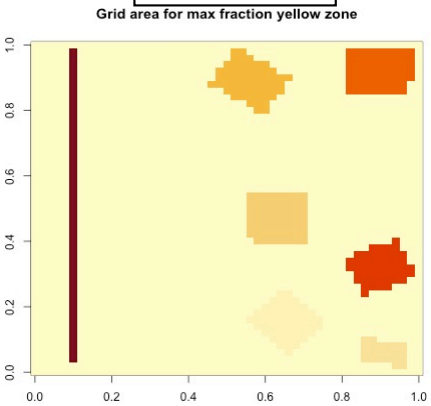


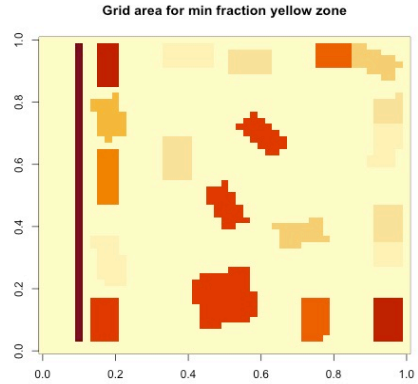
Scenario 6



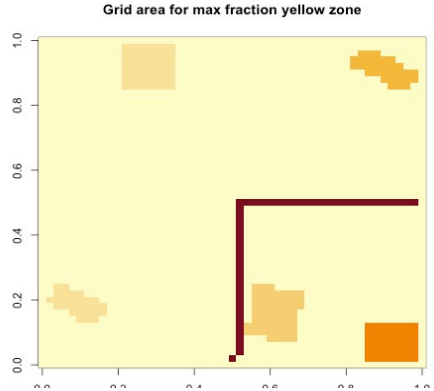


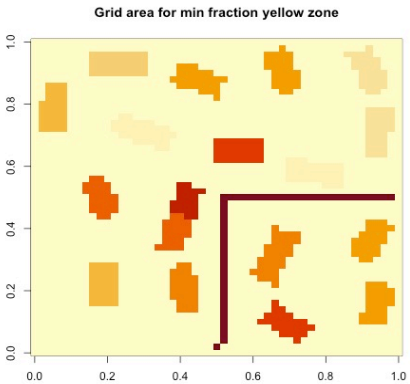
Scenario 7





Scenario 8



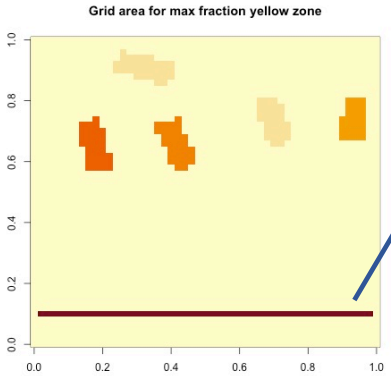


Data Exploration - grid configurations

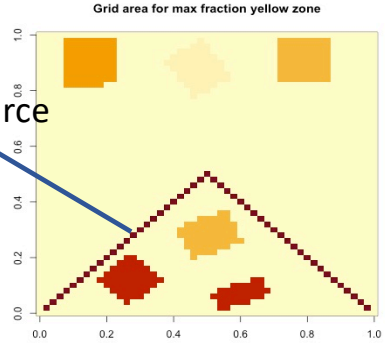
Scenario 9
(non- specific site data)

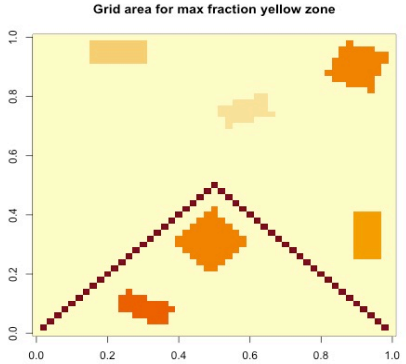
Specific site
(specific site data)

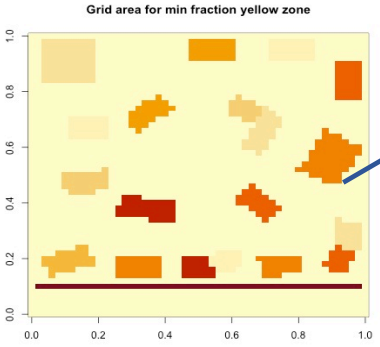
Specific site
(Test site data)



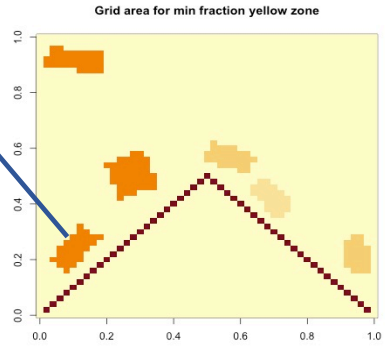
Noise Source

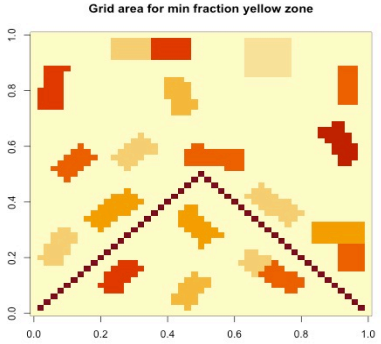






Buildings





Building darker color=taller

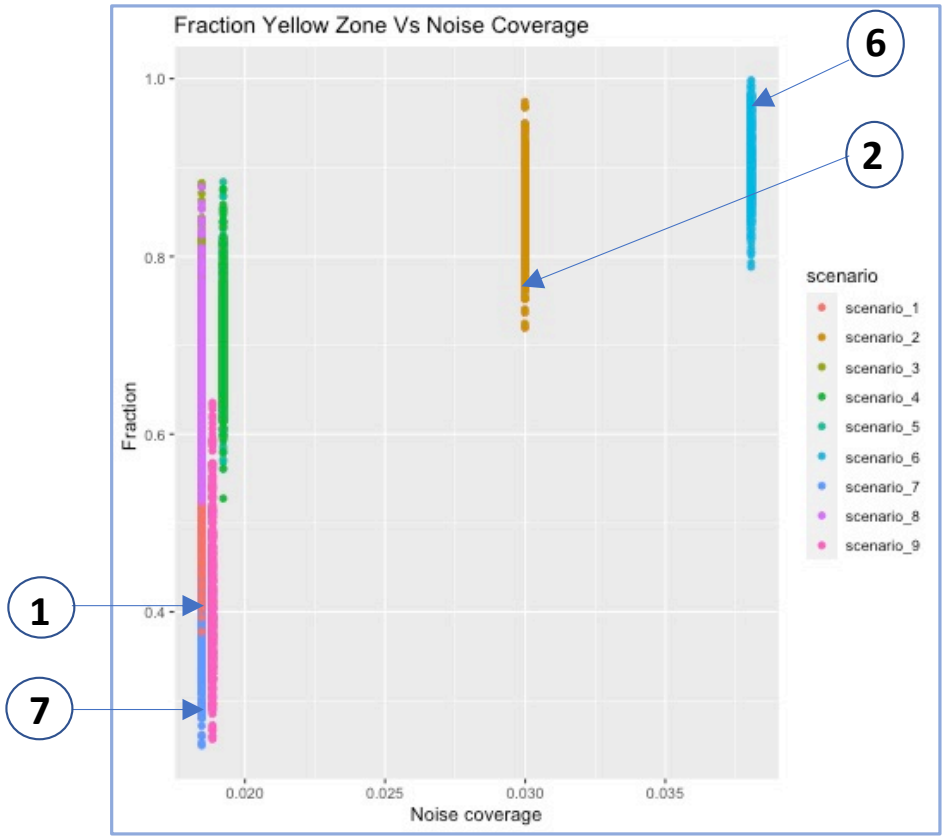
Feature Engineering - creating features

Based on the outcome of the trends gathered from the data exploration, all features used to analyze the statistical models fall under one of the following categories:

- Building coverage
- Noise coverage
- Building coverage to noise coverage ratio
- Average distance to noise
- Building coverage in the 4 zones
- Building heights in the 4 zones

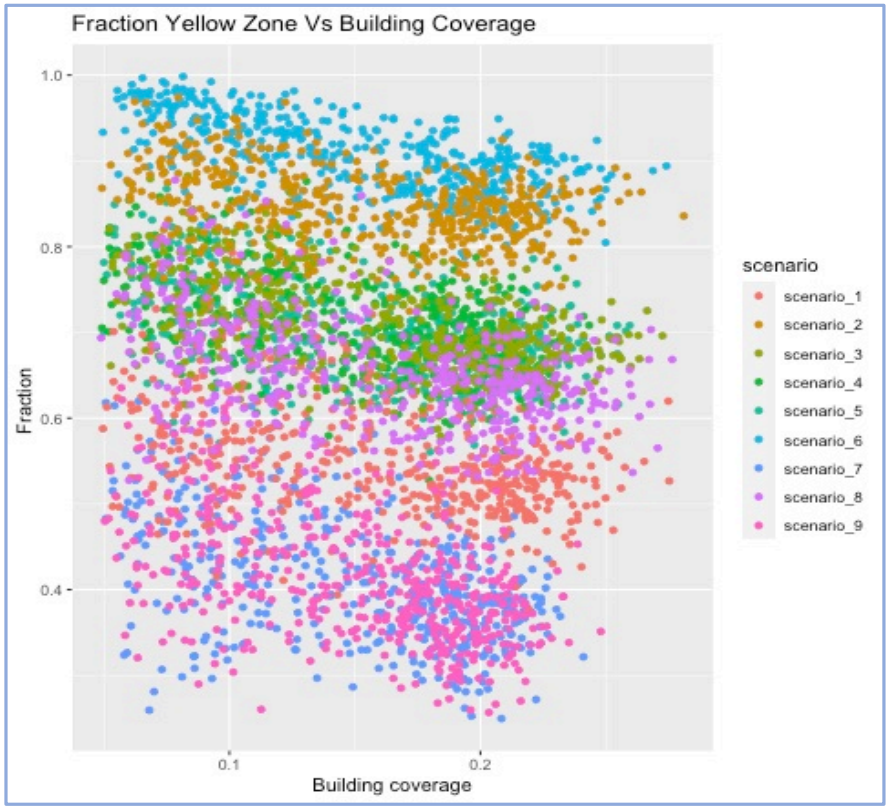
Feature Engineering - exploration

- Fraction yellow zone increases with increasing noise coverage.
- Scenario 1 and 7 have lesser noise coverage whilst Scenario 2 and 6 have a greater noise coverage.



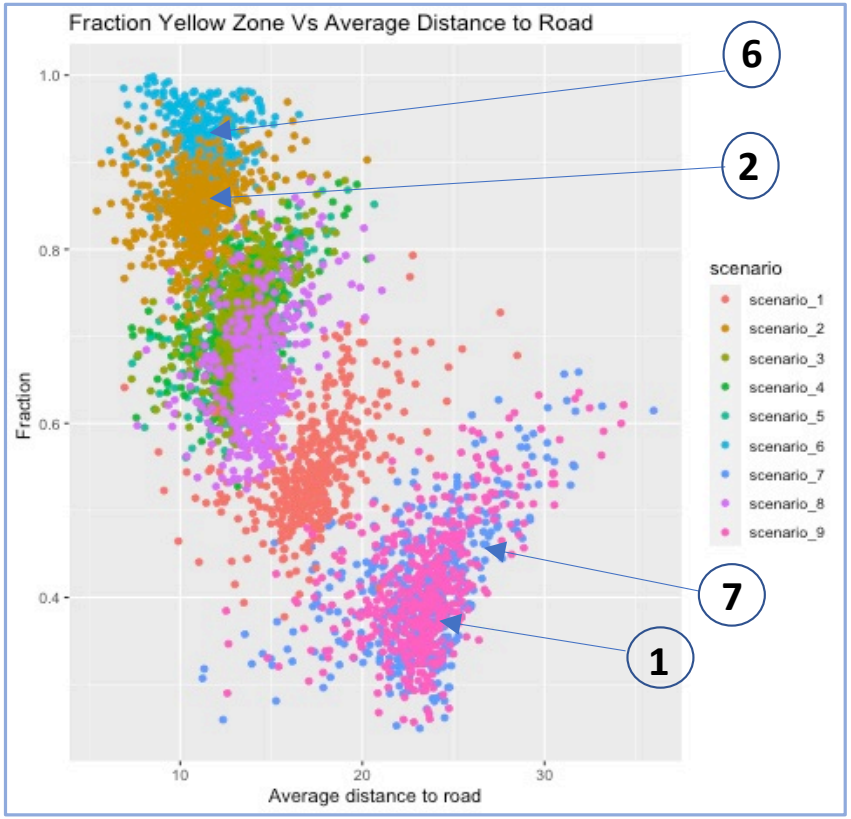
Feature Engineering - exploration

Fraction yellow zone decreases with increasing building coverage.



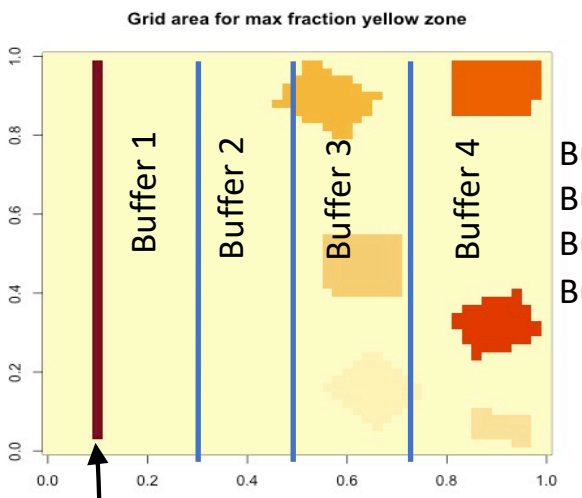
Feature Engineering - exploration

- Fraction yellow zone decreases with increasing average distance to noise source.
- Average distance to noise source for Scenario 1 and 7 have a broader range whilst that of Scenario 2 and 6 have a narrower range.

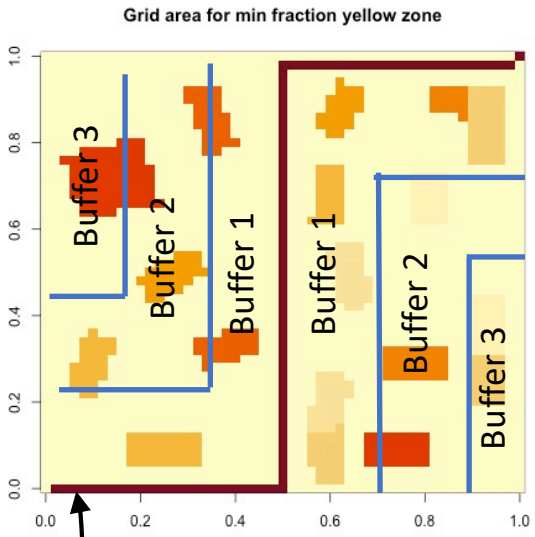


Feature Engineering - creating features

Definition of Buffer zones for creating features



Buffer zone 1: within 12 units
Buffer zone 2: 12 to 24 units
Buffer zone 3: 24 to 36 units
Buffer zone 4: beyond 36 units



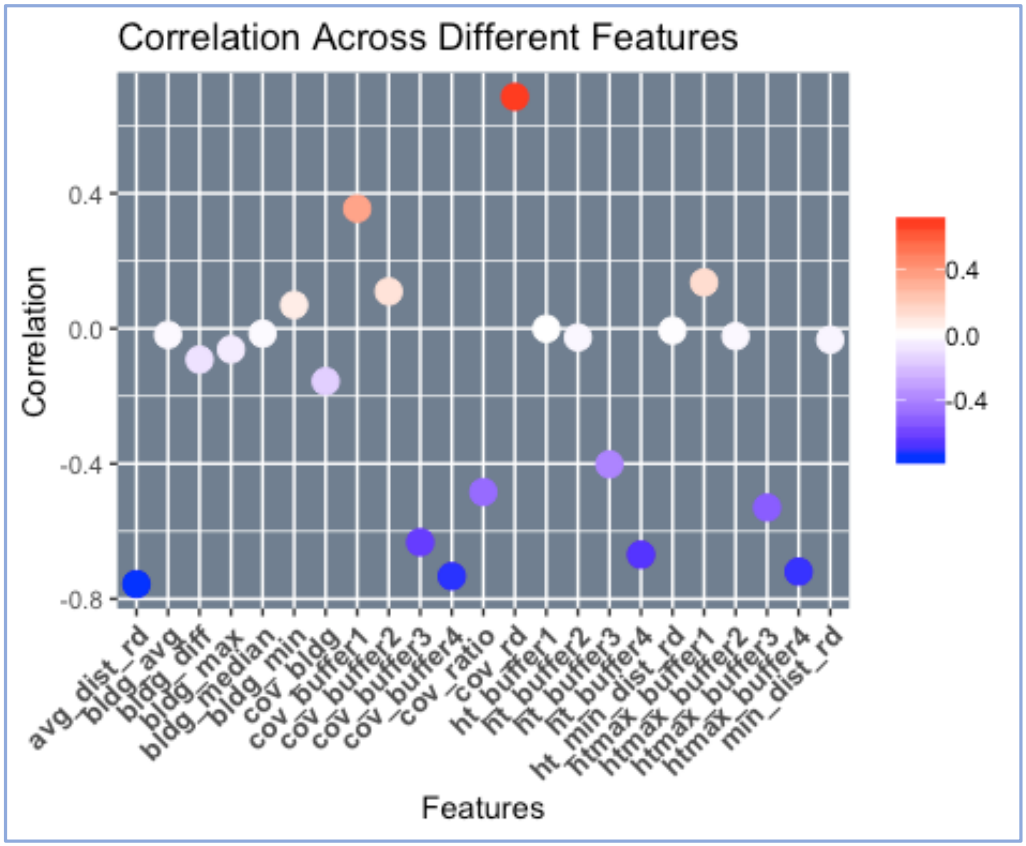
Not all scenarios have all the 4 buffer zones

Buffer 4 values will be zero for this scenario

Noise source

Feature Engineering - correlation

- Correlation was used to evaluate the important features.
- Features with zero or almost no correlation were omitted.

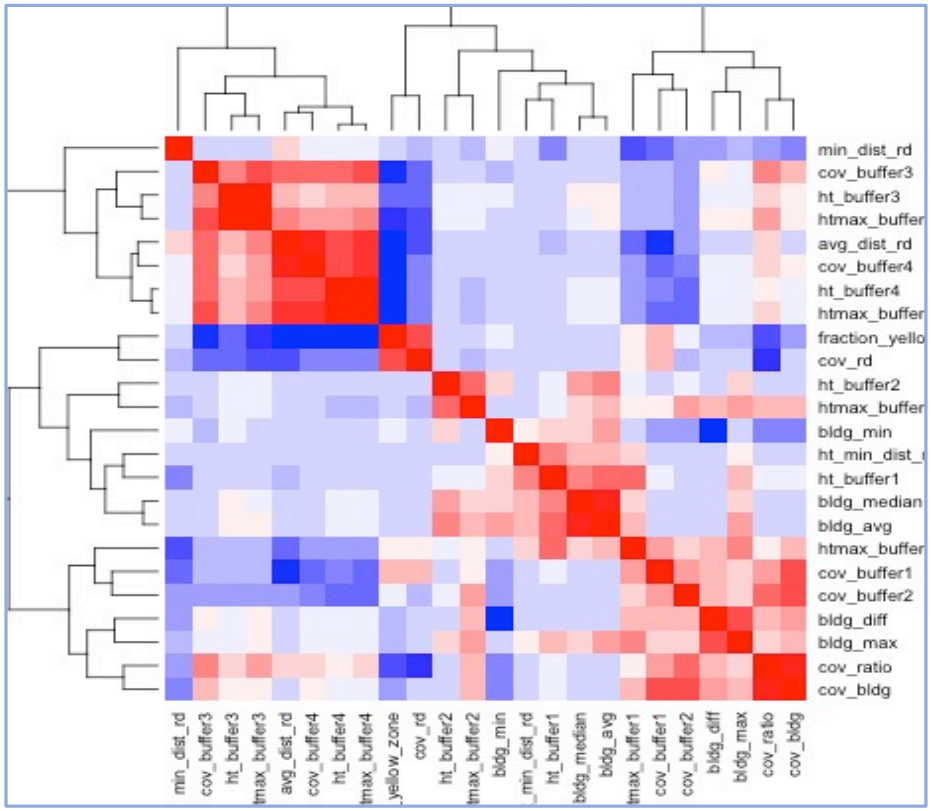


Feature Engineering - correlation

- Features that correlated with each other were optimally selected by using feature importance generated by the Random Forest model

- Final feature selection:

cov_rd	cov_bldg
cov_ratio	htmax_buffer1
htmax_buffer3	htmax_buffer4
cov_buffer3	cov_buffer1
cov_buffer4	avg_dist_rd



Model Analysis and Performance

Data splitting

- Non-specific data and specific data were combined to give 4750 observations
- Split of 80% training data and 20% testing data made based on the following reasons:
    - specific dataset only about 5% and so not enough to use for testing.
    - specific dataset were observations from the test site and thus combining it with the non-specific data would be more efficient in training the models

Model Analysis and Performance

Technical Skills:
Data Wrangling, Data Visualization, Statistical Analysis, Regression Analysis, Machine Learning

Technical Tools:
R & RStudio with ggplot, dplyr, caret package
GitHub version Control

Model Analysis and Performance

Output measure/Analysis

As the goal of the project is to minimize noise in an urban area, Root Mean Square Error (RMSE) was selected as the error measurement so that large errors would be given more weight.

$$\text{Root Mean Square Error} = \sqrt{\frac{1}{N}\Sigma(actual - prediction)^2}$$

The Baseline model used for comparison represented the average of the entire fraction yellow zone in the training data with this:

Minimum Value	Maximum Value	Overall Average
0.2526	0.9983	0.6572

All the models were initially run and tested before performing hyper parameter tuning for better results.

Model Analysis and Performance

The following models were run:

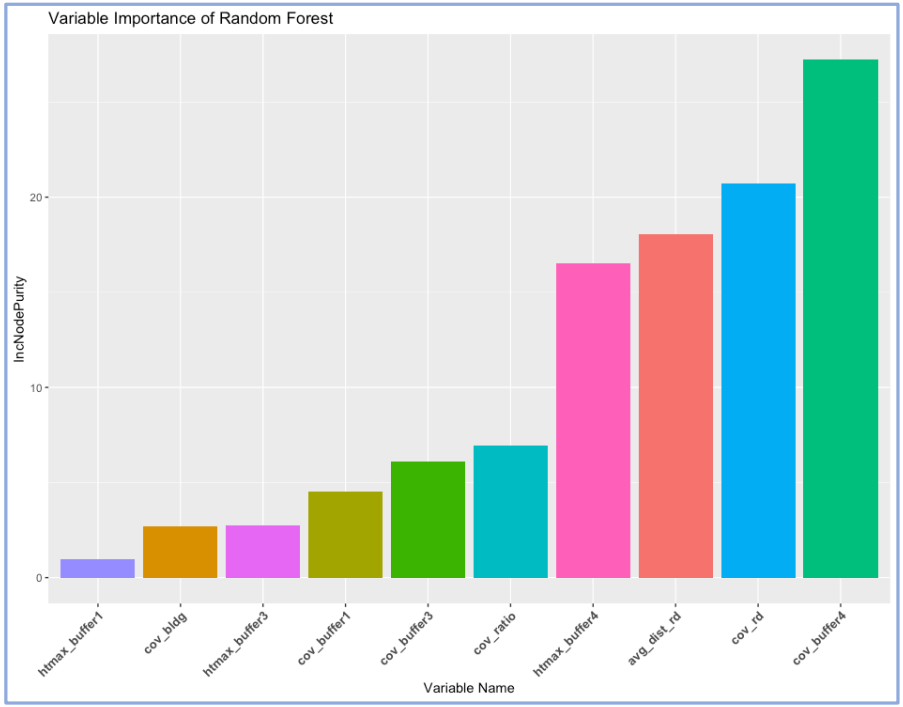
Model name	RMSE	Hyper Parameters
Guess	0.3766663	NA
Average (Baseline)	0.1691386	NA
KNN	0.0857905	K
Linear Regression	0.0750217	NA
Gam Loess	0.0673779	span
Regression Tree	0.0624453	Complex parameter
SVM	0.0582215	tunelength
Random Forest	0.0548306	mtry, ntree

Random Forest showed the best performance with RMSE of 0.0548

Model Analysis and Performance

Random Forest Variable Importance

The noise coverage in Buffer zone 4 was the most important feature in explaining the target variable in the Random Forest model.



Results – Test Site

Random Forest Testing:

The Random Forest model was tested on the test site data and resulted in an RMSE of 0.0516.

Ensemble Model Testing:

To get better results an average of the three models with the lowest RMSE values: Regression Tree, SVM and Random Forest were used as an ensemble model and tested on the test site data resulting in an RMSE of 0.0472.

Conclusion and recommendations

- With a 96% prediction accuracy of the fraction yellow zone, clients would be well informed to be able to predict the fraction yellow zone based on which building and noise configuration their proposal intend to utilize and thereby be able to comply with code and regulations.

- For future work based on our research, neural networks and ensemble models like stacking and blending can be explored. Features could also be trimmed further based on the feature importance and the models re-trained to compare performance.

References

- https://topepo.github.io/caret/model-training-and-tuning.html
- https://stats.stackexchange.com/questions/142873/how-to-determine-the-accuracy-of-regression-which-measure-should-be-used
- https://glassboxmedicine.com/2019/09/15/best-use-of-train-val-test-splits-with-tips-for-medical-data/#:~:text=At%20the%20beginning%20of%20a,10%25%20val%2C%2010%25%20test
- http://r-statistics.co/Loess-Regression-With-R.html
- https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e
- https://library.municode.com/ma/cambridge/codes/code_of_ordinances?nodeId=TIT8HESA_CH8.16NOCO
- https://www.modelop.com/blog/the-importance-of-rapid-iteration/

THANK YOU

Appendix

Distance to Road Calculation