Techniques for Wifi Locationing
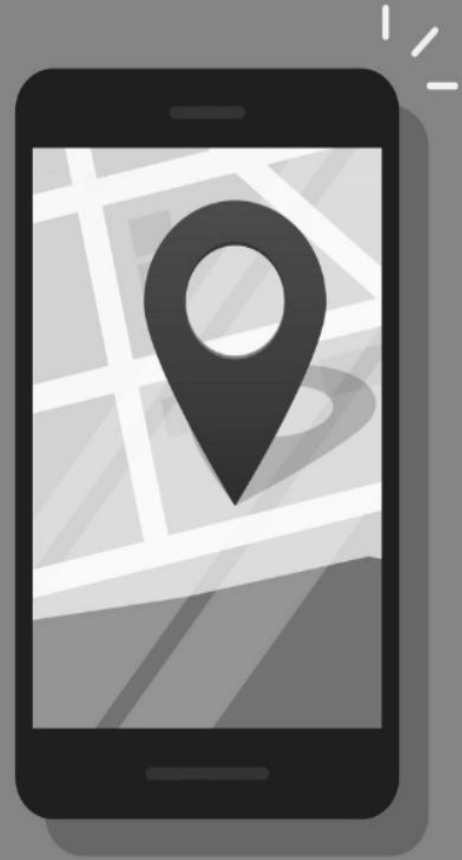
Rhys Hewer

Outline

Project Goals
Data Science Process
Data Quality
Data Preprocessing
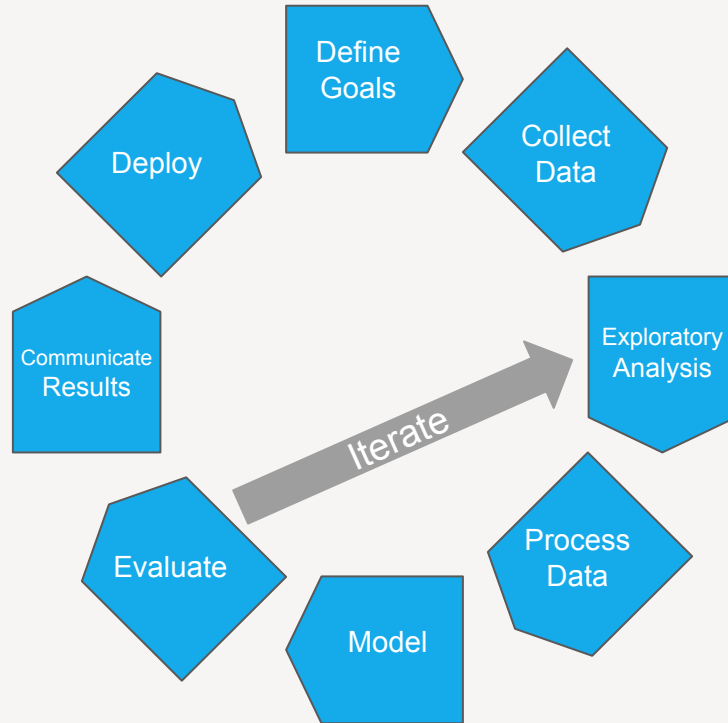Modelling
Recommendations

Project Goals

Project Goals

Use "wifi fingerprinting" to determine a person's location in indoor spaces by using signals from multiple wifi hotspots.

Wifi Fingerprinting = using signal strengths of surrounding wifi access points, as received by device, as a 'fingerprint' of location.

Data Science Process

Data Science Process

Data Collection

Dataset Overview

Used [UJIIndoorLoc database](#) which covers three buildings of Universitat Jaume I (Valencia, Spain) with 4 or more floors and an area of almost 110.000m^2.

529 features: 520 related to wifi signal strength per location,
9 identify location, time and characteristics of collection.

The wifi intensity values are represented as negative integer values ranging -104dBm (extremely poor signal) to 0dbM. The positive value 100 is used to denote when a wifi signal was not detected.

During the database creation, 520 different wifi wireless access points (WAPs) were detected. Thus, the WiFi fingerprint is composed by 520 intensity values.

Split into Training (19937 obs.) & Validation (1111 obs.) - collected in different ways.

Data collection (simplified)
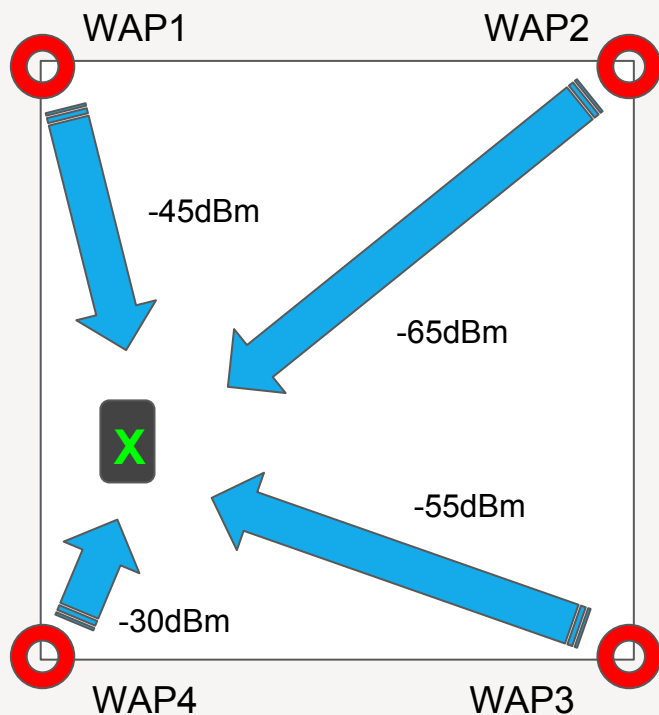
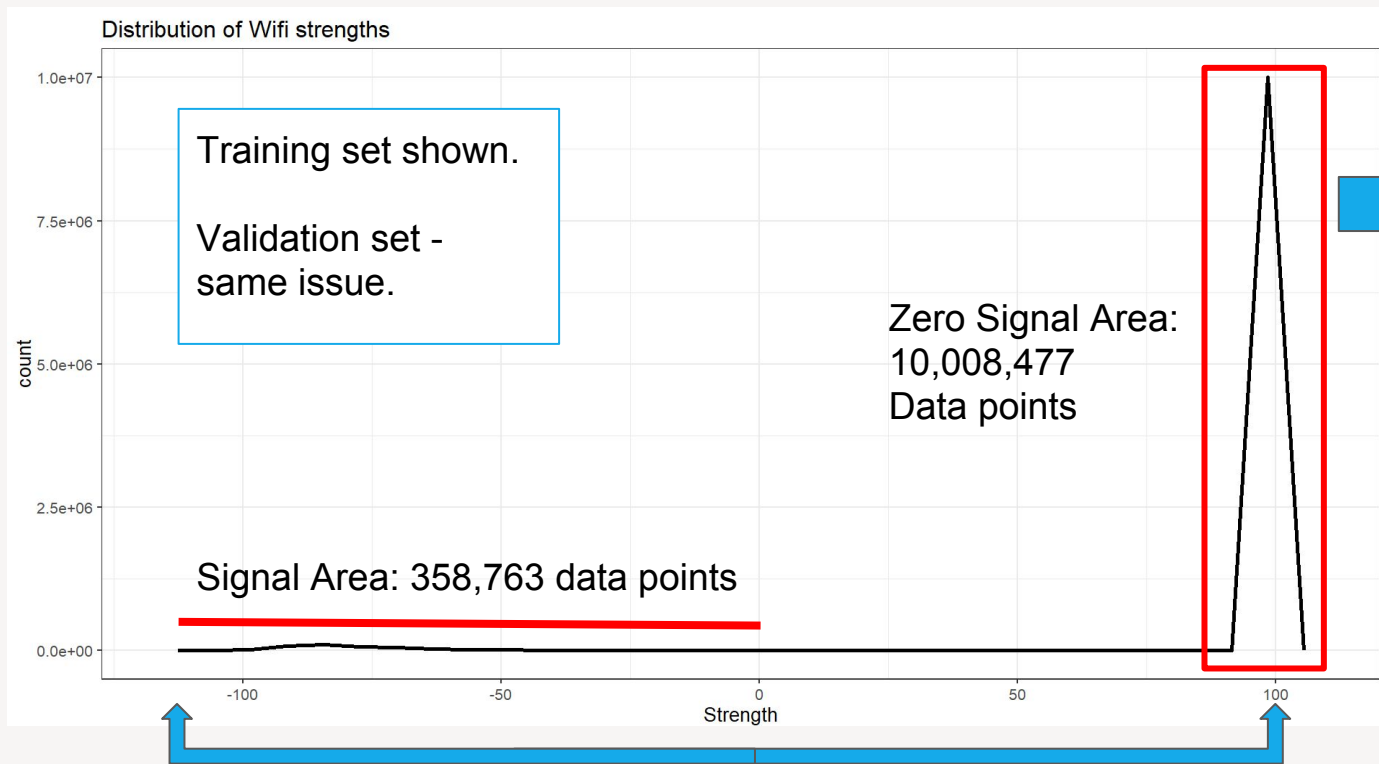| WAP1	WAP2	WAP3	WAP4	Location
-45	-65	-55	-30	X

Training Set:
- Pre-selected points used to collect data
- 10 readings per location taken (approx)

Validation Set:
- User defined locations
- 1 reading per location taken (approx)

Data Preprocessing

Data Preprocessing - no signal coding

Distribution of Wifi strengths

Training set shown.

Validation set - same issue.

Zero Signal Area: 10,008,477 Data points

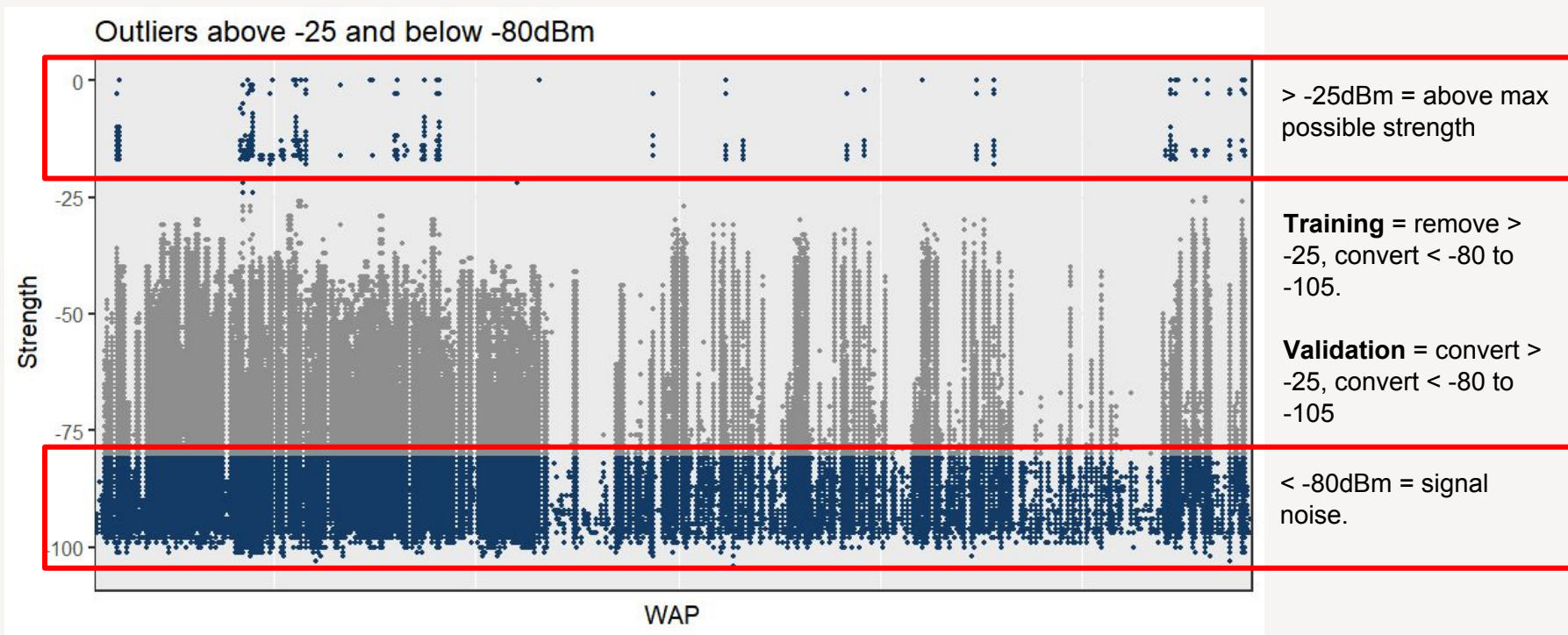Signal Area: 358,763 data points

Converted +100dBm values to -105dBm

Done on both Training & Validation sets.

This is more coherent with scale of lower = weaker.

Data Preprocessing - Outlier Values

Outliers above -25 and below -80dBm

> -25dBm = above max possible strength

Training = remove > -25, convert < -80 to -105.

Validation = convert > -25, convert < -80 to -105

< -80dBm = signal noise.

source

Data Preprocessing - Zero Variance Features

194 zero-variance features removed from Training set

Data removed:

- Outliers: 475 observations x 520 features = 247,000 values
- Zero-variance: 194 features x 19,462 observations = 3,775,628 values

38% total Training Set data removed

Unsuccessful Preprocessing

Signal boosting (signal^2)

Removing >95% correlated WAPS

Boruta feature selection (random forest variable importance)

Model tuning KNN

Aggregation

Normalisation of signal strength

Modelling

Algorithm choice - Classification

Dataset: sample of all data (original data, n = 2000)

Gradient Boosting Machine

Building
Accuracy = 0.9978

Building.floor
Accuracy = 0.9508

K-Nearest Neighbour

Building
Accuracy = 0.9967

Building.floor
Accuracy = 0.849

Decision Tree

Building
Accuracy = 0.8856

Building.floor
Accuracy = not run

Test set: sample of training data

Algorithm choice - Regression

Dataset: subset (building 0 - floor 1, n = 1356)

Linear Regression

Latitude
RMSE = 7.5939909

Longitude
RMSE = 7.5706319

K-Nearest Neighbour

Latitude
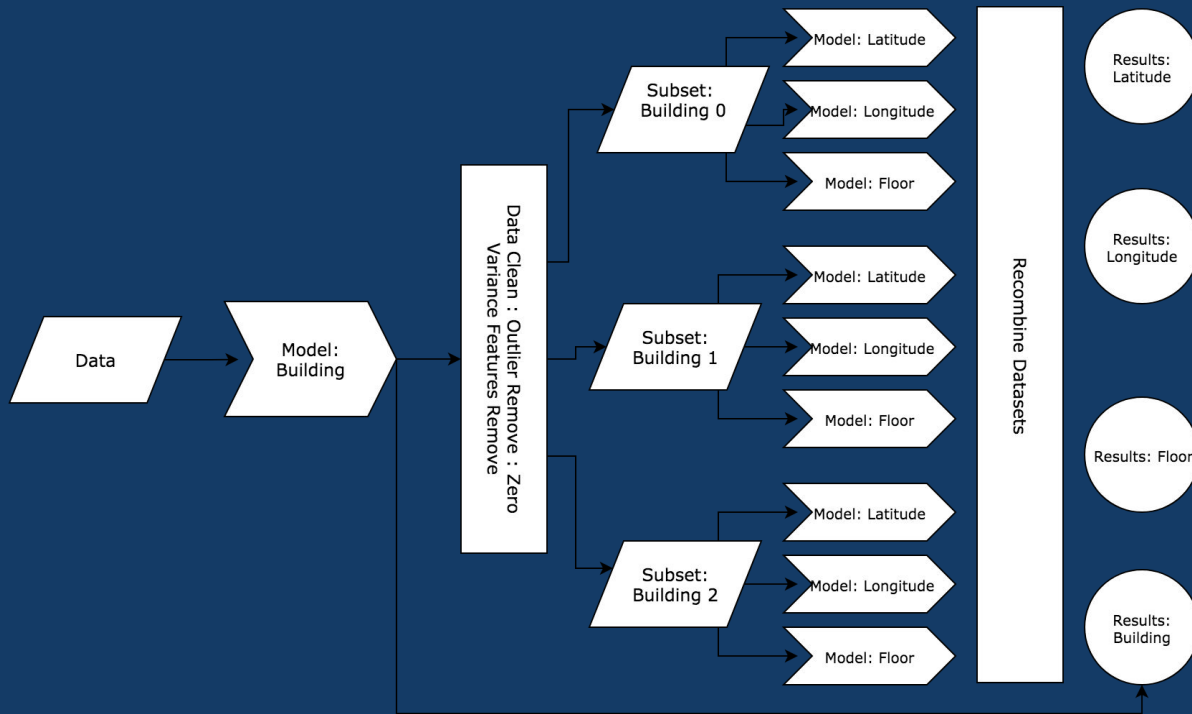RMSE = 4.5743112

Longitude
RMSE = 4.805804

Support Vector Machine

Latitude
RMSE = 11.5698076

Longitude
RMSE = 9.2211412

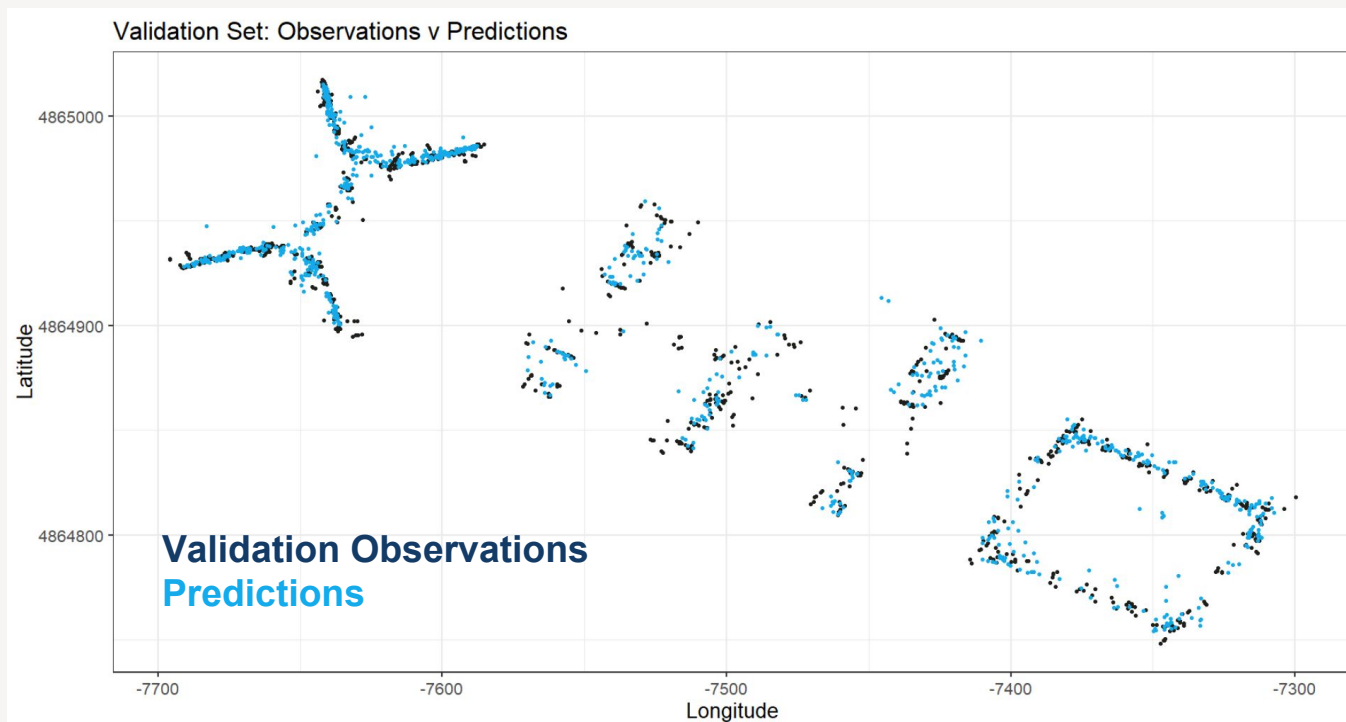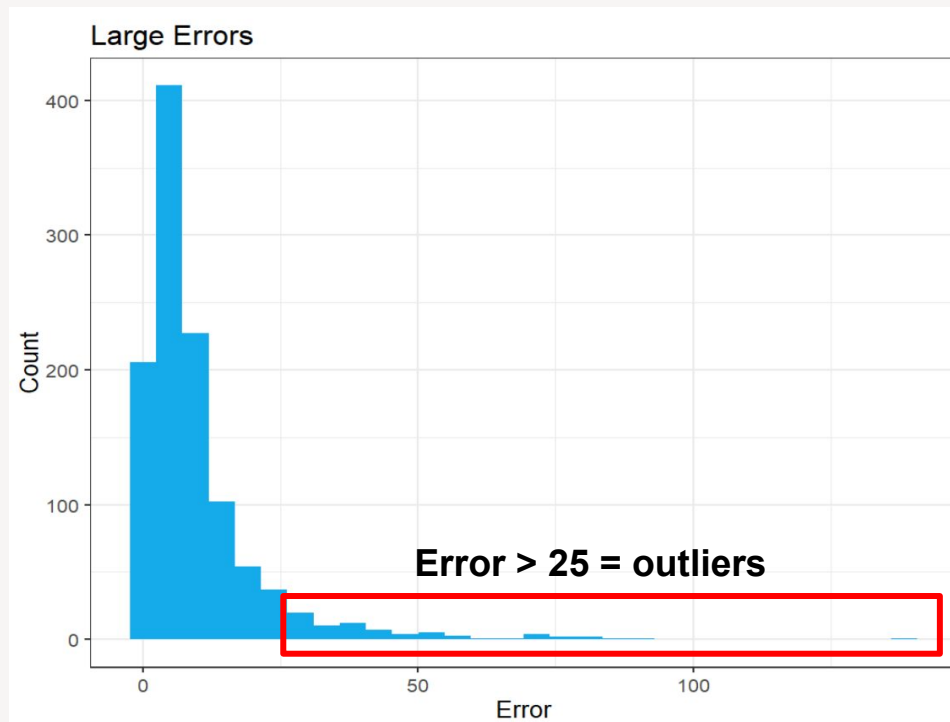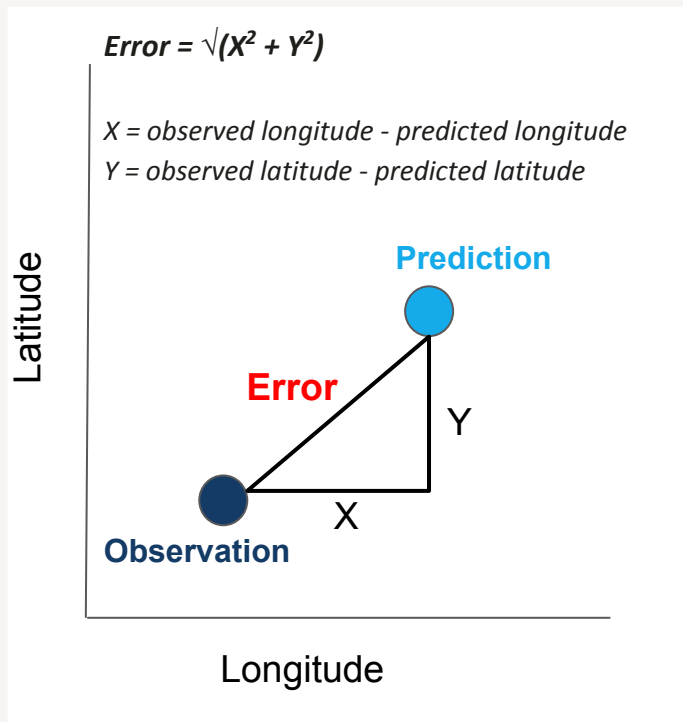**Test set:** subset of training data

Model Splitting

Results

Model Metrics

Latitude
RMSE: 11.372944
R^2: 0.973840
MAE: 6.429971

Longitude
RMSE: 10.7104703
R^2: 0.9921076
MAE: 6.3366795

Building
Accuracy: 0.9982
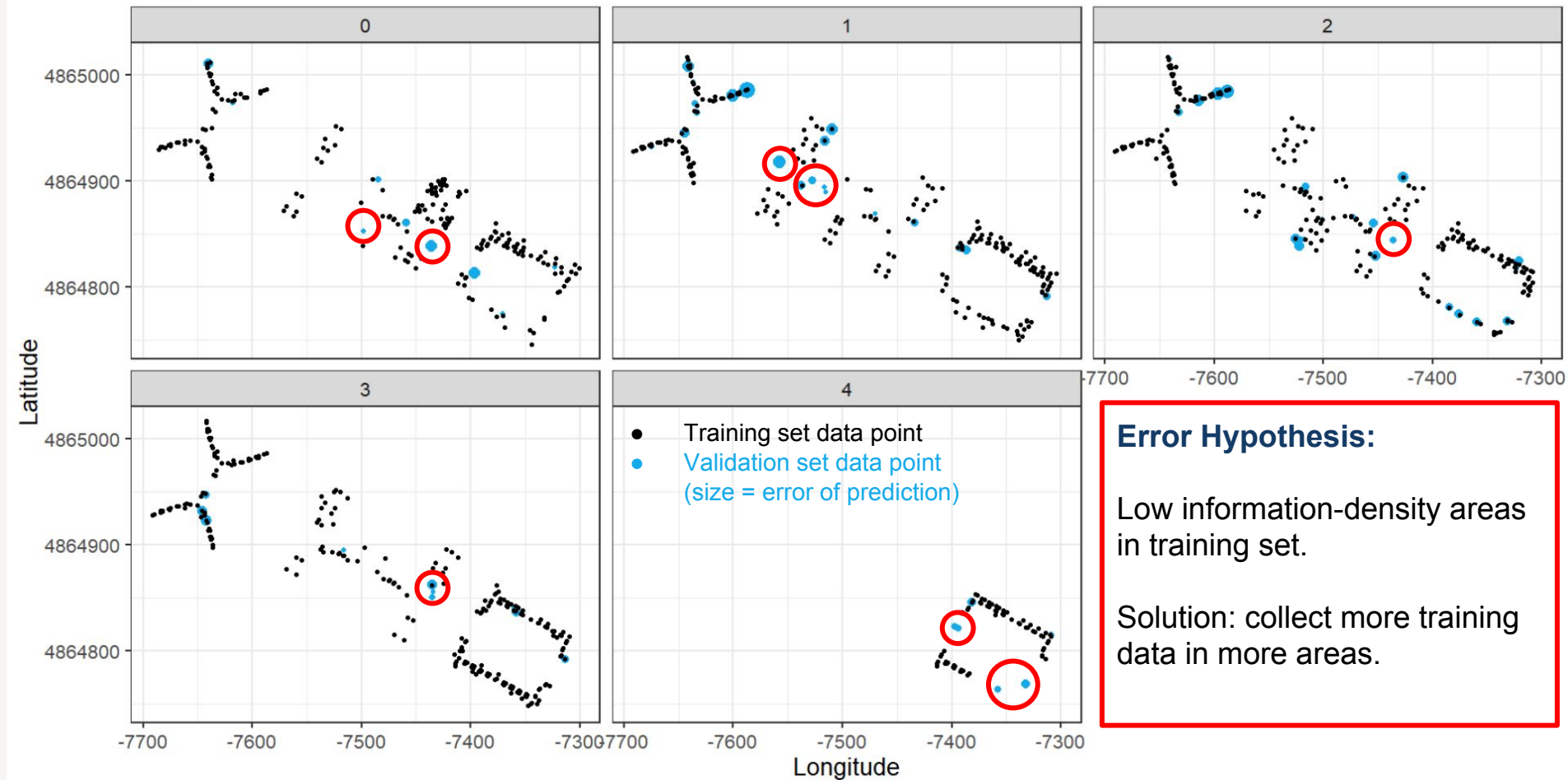Kappa: 0.9972

Building & Floor
Accuracy: 0.9055
Kappa: 0.8943

Outcome Visualisation - Latitude & Longitude

Validation Set: Observations v Predictions

Error Metric - Latitude & Longitude

$Error = \sqrt{(X^2 + Y^2)}$

X = observed longitude - predicted longitude
Y = observed latitude - predicted latitude

Prediction

Error

Y

X

Observation

Latitude

Longitude

Large Errors

Error > 25 = outliers

Count

Error

Large Error Locations

Error Hypothesis:

Low information-density areas in training set.

Solution: collect more training data in more areas.

Classification Errors - Building & Floor
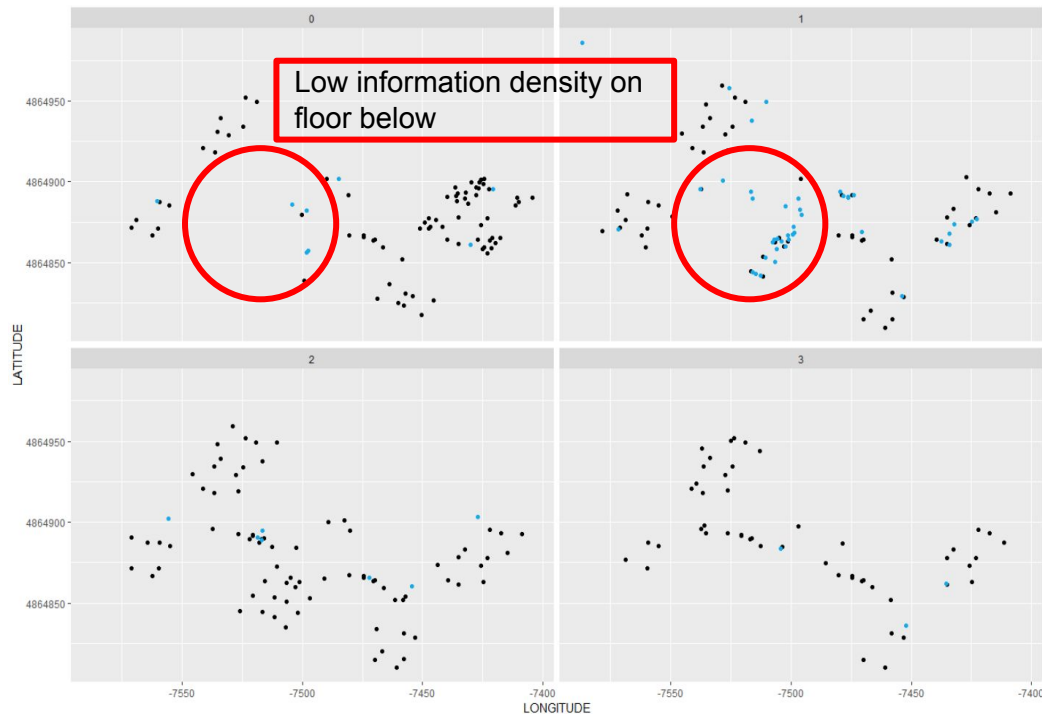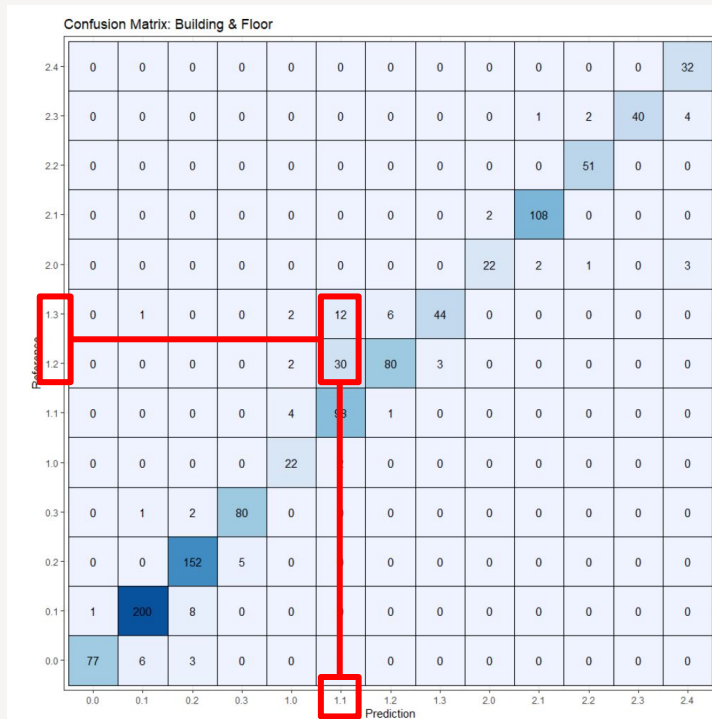
Additional Error Hypotheses

User error introduced in confirming position for validation set.

Differing occupancy levels of buildings within & between training data and validation data collection (people as signal blockers).

Recommendations

Data Collection Recommendations

Collect more training data - ensure that every area within buildings are covered, including open areas.

Validate the validation set - using similar methodology to training set to choose fixed points to collect data (different from the training set).

Analytic Recommendations

Change WAP signal strengths from -105 - 0 to 0 - +105 to allow no signal to be represented by 0.

Consider per-floor modelling in addition to per-building modelling.