**Deep Analytics and Visualization**

**Task 3: Wifi Locationing/Internet of Things**

**Diana Amador**

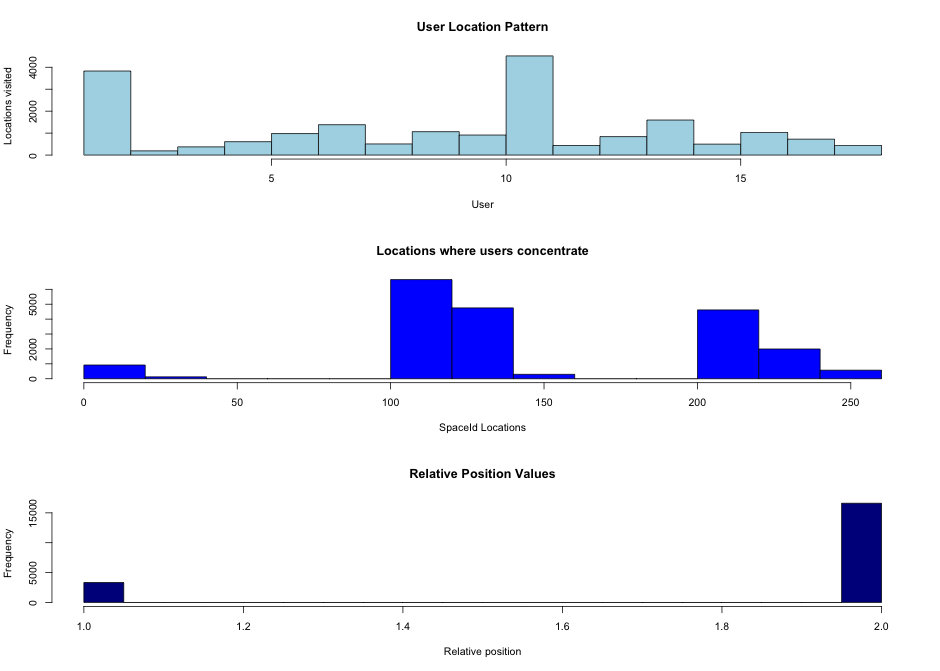
**CASE STUDY DESCRIPTION**

The UJIIndoorLoc database covers three buildings of Universitat Jaume I with 4 or more floors and almost 110.000m2. It can be used for classification, e.g. actual building and floor identification, or regression, e.g. actual longitude and latitude estimation. It was created in 2013 by means of more than 20 different users and 25 Android devices. The database consists of 19937 training/reference records (trainingData.csv file) and 1111 validation/test records (validationData.csv file).   
  
The 529 attributes contain the WiFi fingerprint, the coordinates where it was taken, and other useful information.   
  
Each WiFi fingerprint can be characterized by the detected Wireless Access Points (WAPs) and the corresponding Received Signal Strength Intensity (RSSI). The 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 WAP was not detected. During the database creation, 520 different WAPs were detected. Thus, the WiFi fingerprint is composed by 520 intensity values.   
  
Then the coordinates (latitude, longitude, floor) and Building ID are provided as the attributes to be predicted.

**GOALS:** construct a machine learning predictive model for missing values in the validation data (spaceid, relativeposition and userid).

**DATA VISUALIZATION**

Using exploratory data visualization, we confirmed that there was not evident linear correlation between the target variables. None of them showed a normal distribution.



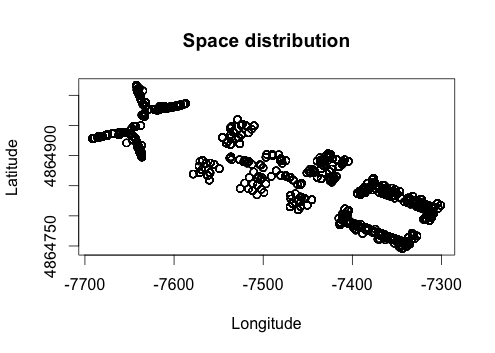
Graph 1. From the first graph “User Location Pattern” we can observe that most of the 18 users spend most of their time in a small set of locations, whereas only two users visit most of the available areas.

Graph 2. In graph 2 it can observed that users concentrate in specific sets of spaces, whereas some other spaces are not visited at all.

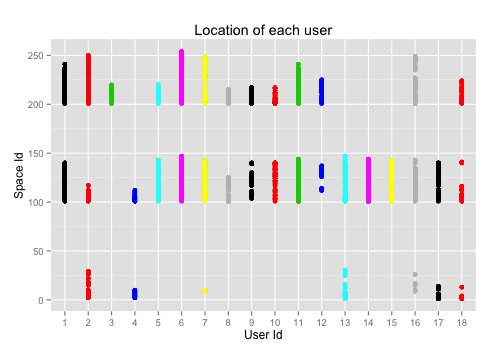
Graph 3. Although relative position has an integer value, it can be considered either a Binary or a Boolean, and this consistent with the attribute definition: the relative position (inside/outside the space) where the capture was taken have been recorded. In this case, the relative position attribute has low significance since we already have the WiFi fingerprint (space coordinates, floor and building) where the space is located.

Since the WiFi fingerprint has been already calculated for all the spaceIds and it is provided in the data set by longitude, latitude, floor and building our main goal is to understand how the 18 users relate to the spaces.

Graph 4. Plot longitud~latitude showed physical location in buildings (building-shaped)



Graph 5 . Plot UserId ~ SpaceId. This graph shows the location patterns for each of the users. It can be observed that there is no linear relation between the space and the user. This kind of patterns suggests that using a classifier instead of a regression model would be the best approach to predict the target attributes: userid, spaceid, and relativeposition.



Given that the target variables did not have a direct correlation with the set of WAP, but can be identified by the WiFi fingerprint provided by the tuple longitude-latitude-floor-building, we also consider the possibility of using a 9-attribute set eliminating the WAP’s attributes to improve performance by reducing sample size.

**DATA SCIENCE FRAMEWORK**

\* All the code in R referred in this document can be found in IndoorLocPrd\_Project.R

**ATTRIBUTION SELECTION**

Creation of simplified datasets for training

WEKA’s CFsSubsetEval was used to find the more related attributes to each target attribute.

## ATTRIBUTE SELECTION USING WEKA

### CFsSubsetEval algorithm was used to select highly predictive attributes for SPACEID, RELATIVEPOSITION

### and USERID

### Evaluates the worth of a subset of attributes by considering

### the individual predictive ability of each feature along with the degree of redundancy between

### them.

### A training subset with better predictors was created for each attribute. Code available in IndoorLocPred\_project.R

tdSpaceId <- tdata[attSpaceId] #Subset of SpaceID Attribute selection (33 attributes)

tdRelPos <- tdata[attRelPos] #Subset of RelativePosition Attribute selection (115 attributes)

tdUserId <- tdata[attUserId] #Subset of UserId Attribute selection (83 attributes)

write.csv(tdSpaceId, file = "tdSpaceId.csv")

write.csv(tdRelPos, file = "tdRelPos.csv")

write.csv(tdUserId, file = "tdUserId.csv")

\*Subsets were to be used in model testing and will be referred as simplified datasets.

Creation of a 9-attribute dataset for training

Given that the target variables did not have a direct correlation with the set of WAP, but can be identified by the WiFi fingerprint provided by the tuple longitude-latitude-floor-building, we also consider the possibility of using a 9-attribute set eliminating the WAP’s attributes to improve performance by reducing sample size.

td <- select(tdata, LONGITUDE:TIMESTAMP)

colnames(td) <-tolower(colnames(td))

write.csv(td, file = "trainingDataSmp.csv")

**REGRESSION MODELS: SMOreg and Linear Regression Testing**

Given that all the data attributes were numeric, we tested two regression methods.

WEKA SMOreg and Linear Regression showed extreme poor performance in predicting target features from original dataset and from simplified datasets(\*).

=== SMOreg Performance on full training set ===

Correlation coefficient 0.4853

Mean absolute error 3.5739

Root mean squared error 4.541

Relative absolute error 85.198 %

Root relative squared error 91.0271 %

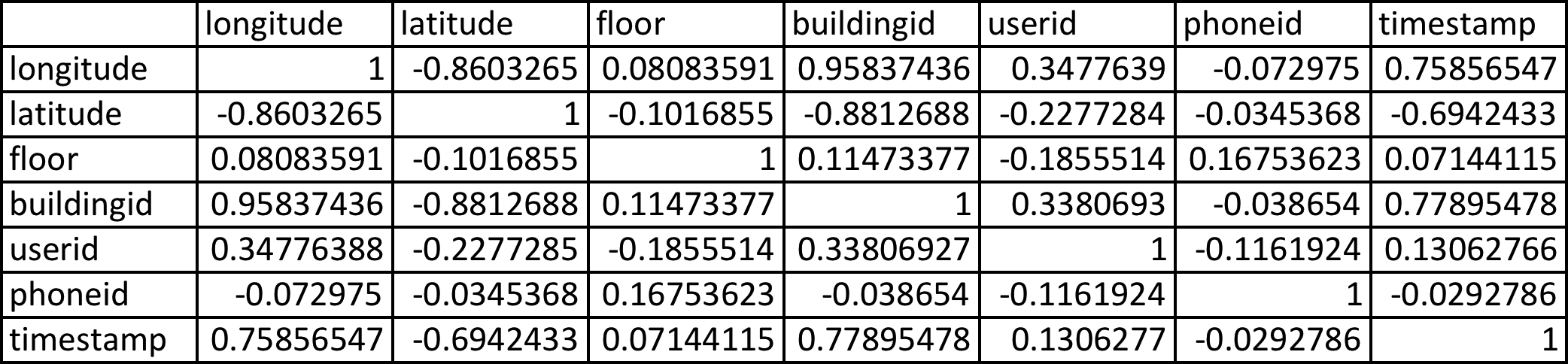
Total Number of Instances 19937

We created a simplified data set with 9 variables by eliminating WAP attributes and use it to generate a linear model in R and test prediction for target attributes. The file name is “trainingDataSmp.csv” and it was subsequently used to run classifier tests as well.

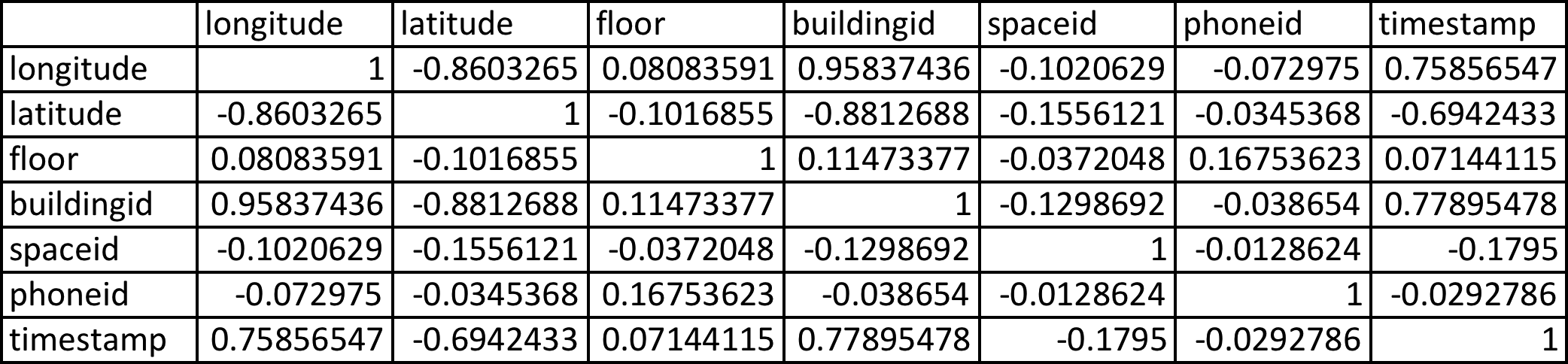
The correlation tables generated by the linear model showed no significant correlation between any of the attributes. Therefore we abandoned regression models in favor of classifiers.

Correlation table using R

* cor(td[c("longitude", "latitude", "floor", "buildingid", **"userid"**, "phoneid", "timestamp")])



* cor(td[c("longitude", "latitude", "floor", "buildingid", **"spaceid",** "phoneid", "timestamp")])



By checking the correlation tables generated in R, we could confirm that linear correlation between target variables and the rest of the variables was weak, as we expected from the exploratory analysis, and regression modes were not optimal to predict the target variables.

**CLASSIFICATION MODELS: IBK**

Based on

[Application of an Improved K Nearest Neighbor Algorithm in WiFi Indoor Positioning](http://link.springer.com/chapter/10.1007/978-3-662-46632-2_45)

[Enhanced weighted K-nearest neighbor algorithm for indoor Wi-Fi positioning systems](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6268565)

[Redpin](http://redpin.org/index.html)

We created a group of K-nearest neighbor trials using different datasets and K values.

**KNN model in R**

A KNN model was created in R to determine performance on a 9-attribute dataset from its crosstable, statistical summary and by calculating TP (30%) and TN (70%). The original dataset was divided into training and testing subsets to confirm performance. To create this model target attributes were labeled (converted to factors). However, we didn’t applied any normalization procedure and believe that might have impacted negatively the model performance. For practical reasons, we decided the next trials in WEKA. However, the significant improvement in prediction accuracy that we observed in this trial against regression models, the fact that no linear correlation was observed among attributes, and the empirical data from research literature on the topic, made us decide to move towards a classification algorithm.

Since we didn’t run a normalization pre-processing, we proceed to use WEKA algorithm that normalize data as a standard procedure. We decided to do that in order to obtain a benchmark for classifier comparison.

**IBK Training Trails**

Trial models and output can be found on xxx folder.

**Trial 1**: IBK running with K=141 on simplified dataset (33 attributes) for target attribute SPACEID. (We used the recommended value for K as the square root of the number of observations, but did not record a significant improvement in the algorithm performance.)

Output File: IBK-K-141-train-33A-spaceid

WEKA did not provide a confusion matrix for this dataset size

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient 0.9587

Mean absolute error 0.584

Root mean squared error 1.4239

Relative absolute error 13.9226 %

Root relative squared error 28.5428 %

Total Number of Instances 19937

**Trial 2**: IBK running with K=10 on simplified dataset (33 attributes) for target attribute SPACEID

Output File: IBK-K-10-train-33A-spaceid

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient 0.9609

Mean absolute error 8.444

Root mean squared error 16.1521

Relative absolute error 16.9951 %

Root relative squared error 27.6858 %

Total Number of Instances 19937

**Trial 3**: IBK running with K=1 on simplified dataset (33 attributes) for target attribute SPACEID. This trial had the best performance for all IBK trails on 33 attributes + target.

Output File: IBK-K-1-train-33A-spaceid

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient 1

Mean absolute error 0

Root mean squared error 0

Relative absolute error 0 %

Root relative squared error 0 %

Total Number of Instances 19937

**Trial 4**: IBK running with K=1 on 9-attribute (8 attribute + target) dataset for target attribute SPACEID. Same performance as Trial 3.

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient 1

Mean absolute error 0

Root mean squared error 0

Relative absolute error 0 %

Root relative squared error 0 %

Total Number of Instances 19937

**Trial 5**: IBK running with K=1 on simplified dataset (81 attributes + target) for target attribute USERID

Output File: IBK-K-1-train-82A-userid

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient 1

Mean absolute error 0

Root mean squared error 0

Relative absolute error 0 %

Root relative squared error 0 %

Total Number of Instances 19937

**Trial 6**: IBK running with K=1 on 9-attribute (8 attributes + target) dataset for target attribute USERID

Output File: IBK-K-1-train-9A-userid

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient 1

Mean absolute error 0

Root mean squared error 0

Relative absolute error 0 %

Root relative squared error 0 %

Total Number of Instances 19937

**Trial 7**: IBK running with K=1 on 9-attribute (8 attribute + target) dataset for target attribute RELATIVEPOSITION.

Output File: IBK-K-1-train-9A-relposition

=== Evaluation on training set ===

=== Summary ===

Correlation coefficient 1

Mean absolute error 0

Root mean squared error 0

Relative absolute error 0 %

Root relative squared error 0 %

Total Number of Instances 19937

**Results:** IBK trials for k=1 and 9-attribute dataset showed consistent 100% accuracy in prediction for every target variable. Attributes were normalized by WEKA as part of the IBK algorithm so we didn’t have to create any special preprocessing for attributes that were not nominal class.

**Validation Trials**

Validation trials ran on 9-attribute validation set.

**Trial 1**: IBK model (Trial 4) k=1 running on validation dataset for SPACEID.

Output File with predicted values: IBK-K-1-test-9A-spaceid.

WEKA didn’t provide a confusion matrix option for this model. We assume this is due to size of the sample. However because the model had a 100% TP, the rest of the metrics were not required. See statistical performance metrics for trial 4.

**Trial 2**: IBK model (Trial 6) k=1 running on validation dataset for USERID.

Output File with predicted values: IBK-K-1-test-9A-userid.

Relevant trial metrics are the same as trial 6.

**Trial 3**: IBK model (Trial 7) k=1 running on validation dataset for RELATIVEPOSITION.

Output File with predicted values: IBK-K-1-test-9A-relposition.

Relevant performance metrics are the same as trial 7.

=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances 19937 100 %

Incorrectly Classified Instances 0 0 %

Kappa statistic 1

Mean absolute error 0

Root mean squared error 0

Relative absolute error 0.0129 %

Root relative squared error 0.0106 %

Total Number of Instances 19937

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

1 0 1 1 1 1 1

1 0 1 1 1 1 2

Weighted Avg. 1 0 1 1 1 1

=== Confusion Matrix ===

a b <-- classified as

3329 0 | a = 1

0 16608 | b = 2

**Results**: IBK model was used to predict values for the target attributes successfully. See WEKA output buffers for each target attribute.

**CONCLUSION**

Users in the buildings will be located at certain locations at certain times, that fact shows us that the data already has a pattern, and WAP information will not necessarily provide a WiFi fingerprint already calculated. The WiFi fingerprint created by longitude, latitude and floor attributes provided enough information to determine SpaceId. The same situation was found for UserId and RelativePosition. That allowed us to extremely simplify the dataset attributes, and create a more efficient model based on weighted labeling following RedPin and IEEE models.

As KNN provided the best prediction, although other classifiers can also be considered to create predictive models using labeling. An alternative might be using an ensemble with tree algorithms to classify WiFi fingerprints into spaceid labels, and a linear model to associate spaceid labels or to users.

**Required Resources**

**Indoor Locationing Data Set (Training and Validation Sets):**

<http://archive.ics.uci.edu/ml/datasets/UJIIndoorLoc>

**Supplemental Resources:**

Textbook(s):

*Predictive Analytics for Dummies* (ISBN-13: 978-1118728963)

Optional Textbook(s):

*Machine Learning with R* (ISBN-13: 978-1782162148)

*Machine Learning with R* – Chapters 6-12

Wifi Locationing Papers

[Advanced support vector machines for 802.11 indoor location](http://www.researchgate.net/profile/Antonio_Caamano/publication/256994119_Advanced_support_vector_machines_for_802.11_indoor_location/links/02e7e5267d593585e3000000.pdf)

[Application of an Improved K Nearest Neighbor Algorithm in WiFi Indoor Positioning](http://link.springer.com/chapter/10.1007/978-3-662-46632-2_45)

[Enhanced weighted K-nearest neighbor algorithm for indoor Wi-Fi positioning systems](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6268565)

Resource for Confusion Matrix and the above five measures: <http://en.wikipedia.org/wiki/Confusion_matrix>

Ensemble Learning

<https://en.wikipedia.org/wiki/Ensemble_learning>

[An Intro to Ensemble Learning in R](http://www.r-bloggers.com/an-intro-to-ensemble-learning-in-r/)

[Predictive Model Selection and Assessment using R](http://stat.fsu.edu/~fchen/model-selection.pdf)

[Resources for learning how to implement ensemble methods](http://stats.stackexchange.com/questions/32703/resources-for-learning-how-to-implement-ensemble-methods)

R References

[R Reference card](http://cran.r-project.org/doc/contrib/Short-refcard.pdf)

[R Manuals](http://cran.r-project.org/manuals.html)

[R Project](http://www.r-project.org/)