

Forest Fire Prediction

Predicting size fires with k-nn algorithm

- Environmental impact
  - Fauna and Flora
  - Air quality
  - Greenhouse effect
- Economic impact
- Human deaths and injuries

The Economic Effects of Large Wildfires

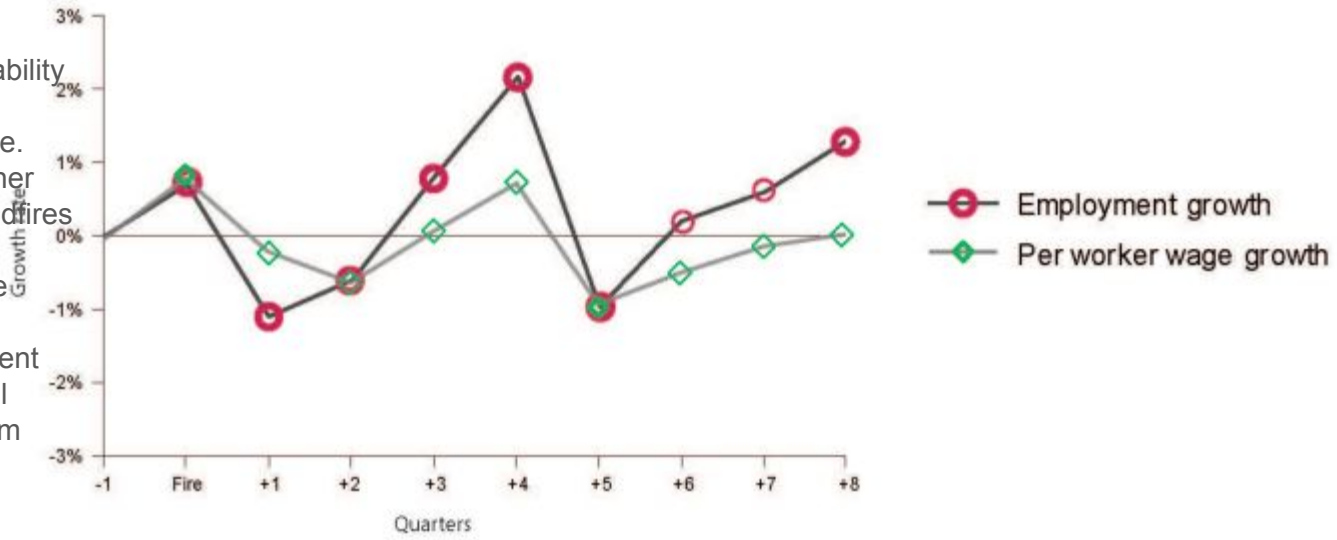
Final Report: JFSP Project Number 09-1-10-3

Large wildfires were defined as those with total suppression costs greater than $1 million

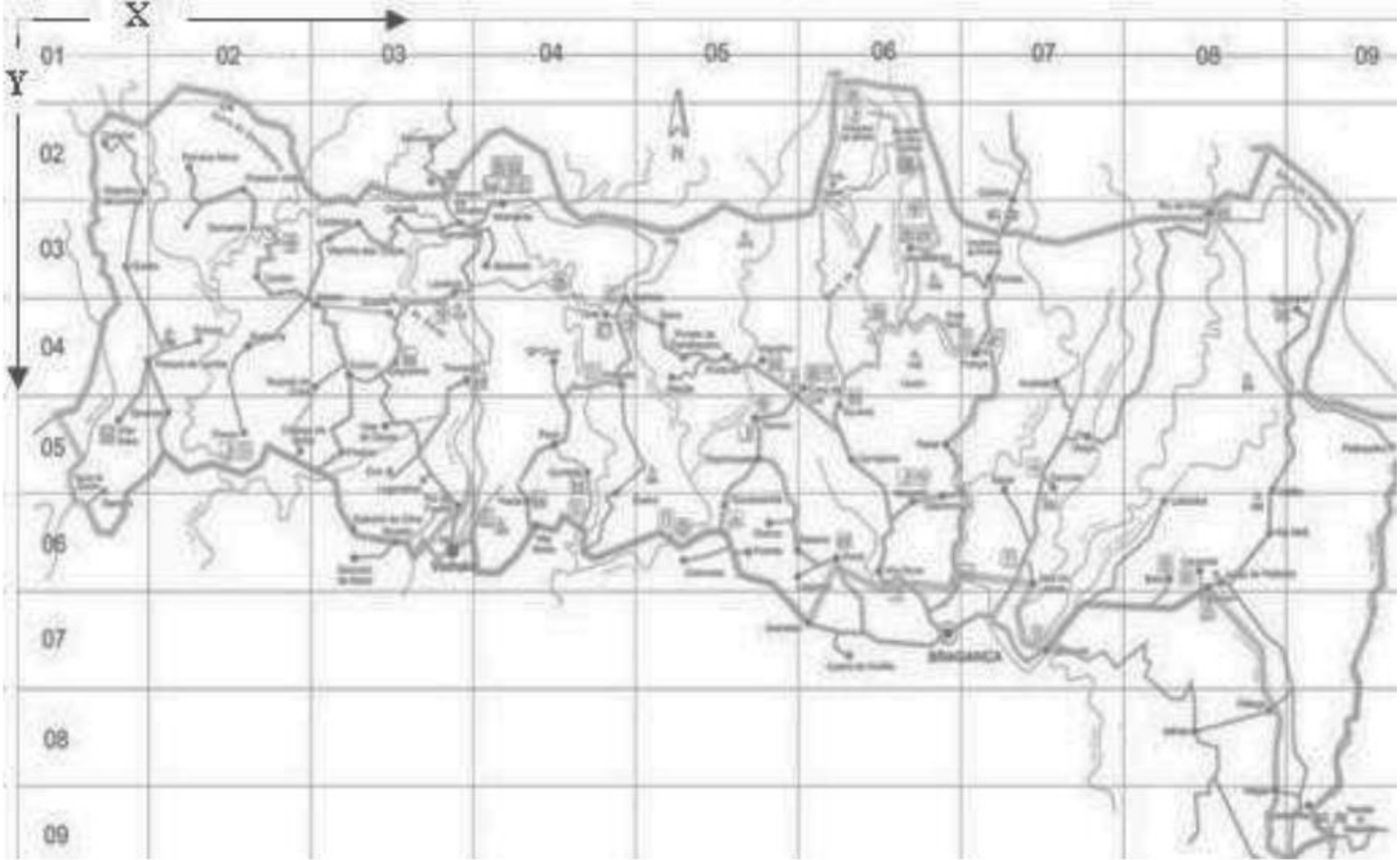
Figure 1. Large wildfires and costs, 2004-2008. The 135 wildfires in the spending breakdown sample are shown in orange.

Large wildfires had persistent effects on local economies. Although the shortterm effects of wildfires on local economies were generally positive, in the medium to longer-term, local economies experienced increased volatility in employment and average wages (see figure2). Wildfires tended to amplify existing seasonal economic patterns, reducing local economic stability in communities for a year or two after the fire. Similar to findings form other natural disasters, the large wildfires created more drastic seasonal patterns in the years following the event. Although increased employment and wages during wildfires will not likely negate these longer-term impacts, they may indicate increased local capacity to contribute to suppression activities and adapt to growing wildfire risk.

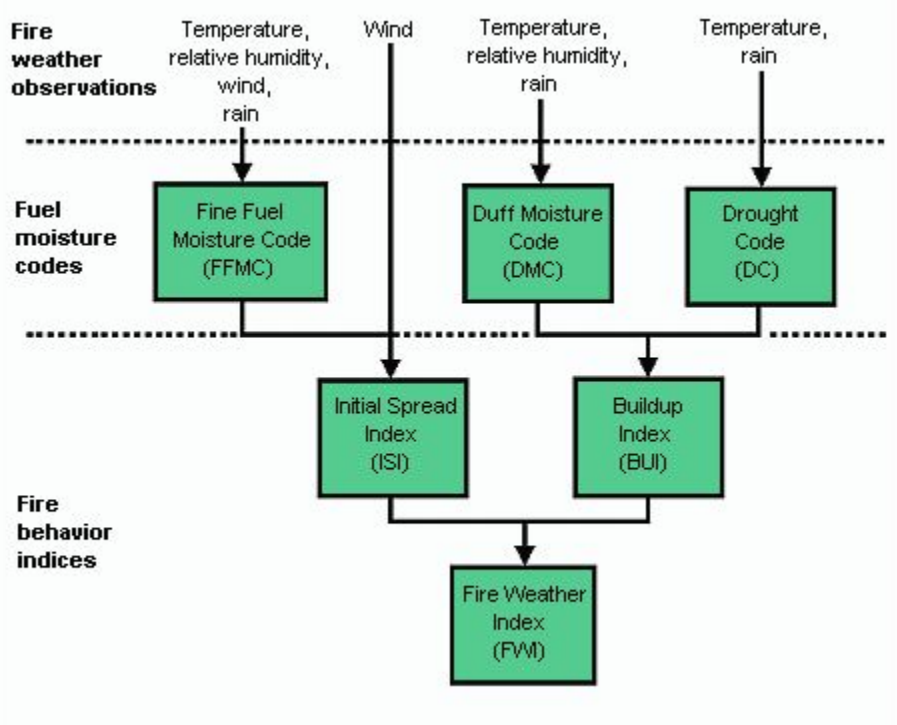
Figure 2. Percent change in average employment and per-worker wage growth rates during and after large wildfires

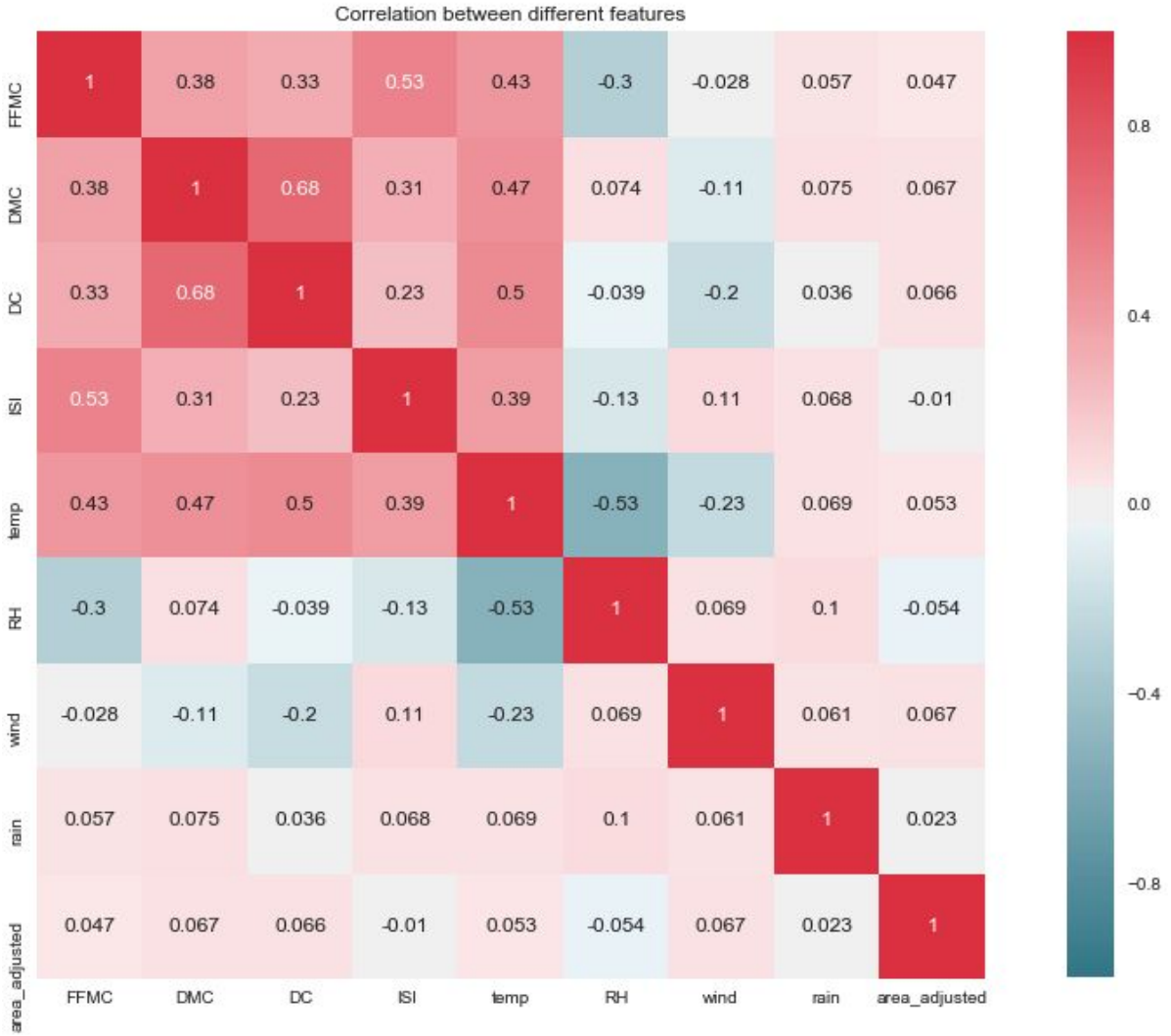


- **Location**
  - **X** - X-axis spatial coordinate within the Montesinho park map: 1 to 9
  - **Y** - Y-axis spatial coordinate within the Montesinho park map: 2 to 9
- **Date**
  - **month** - Month of the year: "jan" to "dec"
  - **day** - Day of the week: "mon" to "sun"
- **Fire Wheater Index - FWI**
  - **FFMC** - Fine Fuel Moisture Code The Fine Fuel Moisture Code (FFMC) is a numeric rating of the moisture content of litter and other cured fine fuels. This code is an indicator of the relative ease of ignition and the flammability of fine fuel: 18.7 to 96.20
  - **DMC** - The Duff Moisture Code (DMC) is a numeric rating of the average moisture content of loosely compacted organic layers of moderate depth. This code gives an indication of fuel consumption in moderate duff layers and medium-size woody material.: 1.1 to 291.3
  - **DC** - The Drought Code (DC) is a numeric rating of the average moisture content of deep, compact organic layers. This code is a useful indicator of seasonal drought effects on forest fuels and the amount of smoldering in deep duff layers and large logs.: 7.9 to 860.6
  - **ISI** - The Initial Spread Index (ISI) is a numeric rating of the expected rate of fire spread. It combines the effects of wind and the FFMC on rate of spread without the influence of variable quantities of fuel.: 0.0 to 56.10
  - **temp** - Temperature °C: 2.2 to 33.30
  - **RH** - Relative Humidity in %: 15.0 to 100
  - **wind** - Wind speed inkm/h: 0.40 to 9.40
  - **rain** - Outside rain in mm/m2 : 0.0 to 6.4
- **area** - The burned area of the forest (in ha): 0.00 to 1090.84


X
Y
01
02
03
04
05
06
07
08
09
N
BRAGANÇA

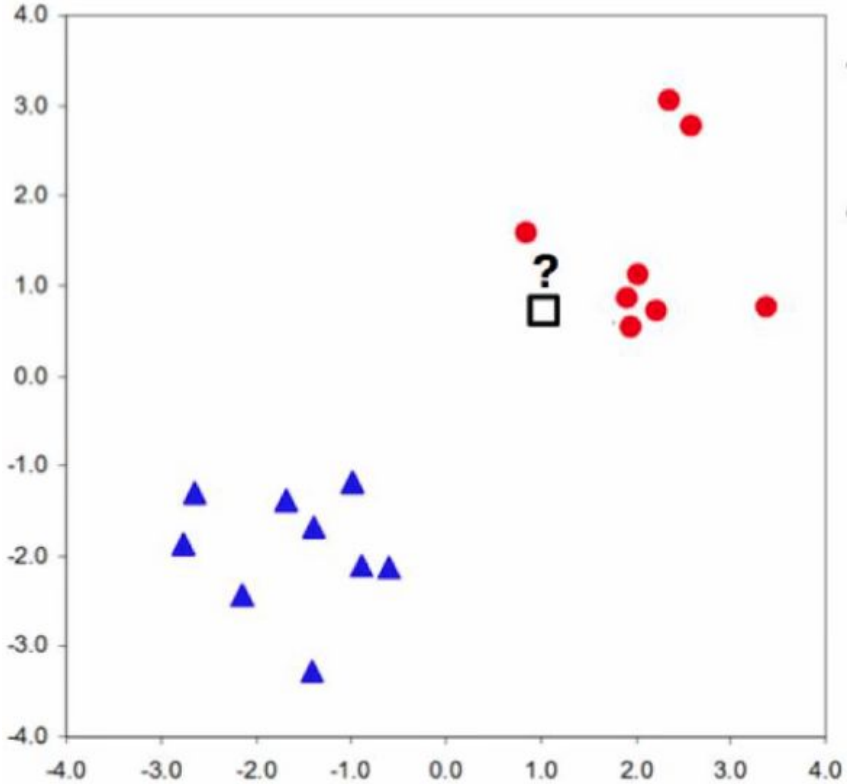
Fire Weather Index - FWI




Correlation between different features
FFMC
DMC
DC
ISI
temp
RH
wind
rain
area_adjusted
1 0.38 0.33 0.53 0.43 -0.3 -0.028 0.057 0.047
0.38 1 0.68 0.31 0.47 0.074 -0.11 0.075 0.067
0.33 0.68 1 0.23 0.5 -0.039 -0.2 0.036 0.066
0.53 0.31 0.23 1 0.39 -0.13 0.11 0.068 -0.01
0.43 0.47 0.5 0.39 1 -0.53 -0.23 0.069 0.053
-0.3 0.074 -0.039 -0.13 -0.53 1 0.069 0.1 -0.054
-0.028 -0.11 -0.2 0.11 -0.23 0.069 1 0.061 0.067
0.057 0.075 0.036 0.068 0.069 0.1 0.061 1 0.023
0.047 0.067 0.066 -0.01 0.053 -0.054 0.067 0.023 1
0.8
0.4
0.0
−0.4
−0.8

K-NN

- Method based on distance
- Simple algorithm
- How did you do?
  - Calculated a P(red|blue) or P(blue|red) - Naive Bayes?
  - Decision Tree?
  - Fit in a hyperplane?
- Natural intuition
  - Nearby points are related in the same concept



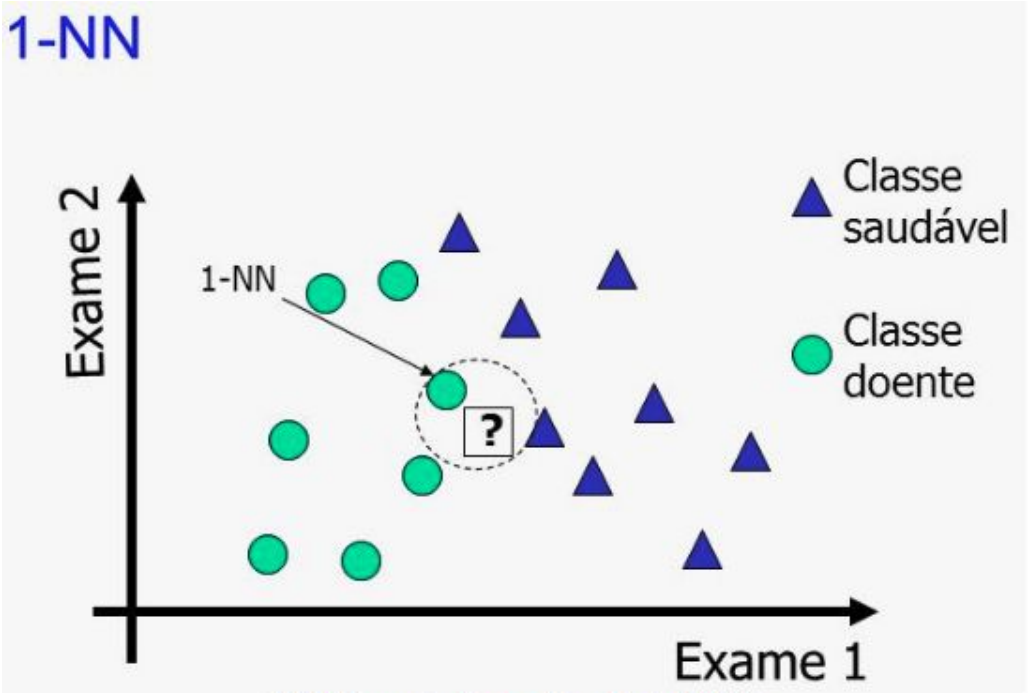
1-nearest neighborhood

Input:
A training set: D={(Xi,Yi), i =1,...n}
A test object to be classified: t={Xt,Yt = ?}
The distance function between objects: d(Xa,Xb)
Output:
Yt: Given class related to t

```
Dmin <- +infinity

For each i in 1,...,n do
    If d(Xi,Xt) < Dmin so
        Dmin <- d(Xi,Xt)
    End
End
Yt = Yidx
```

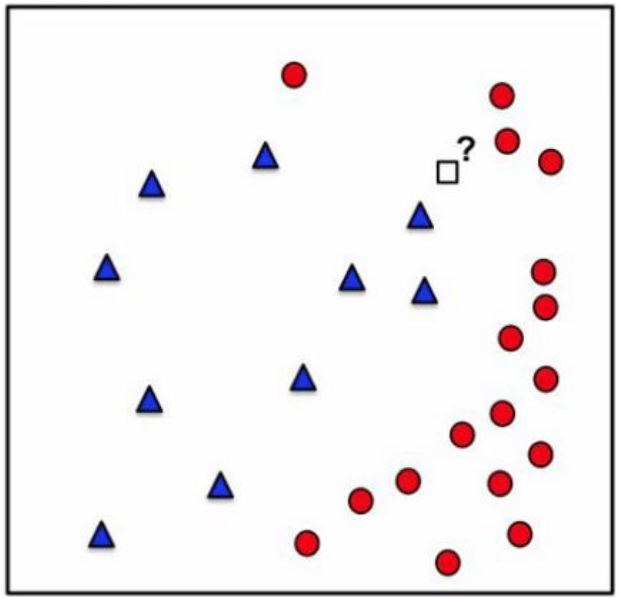
Return: Yt

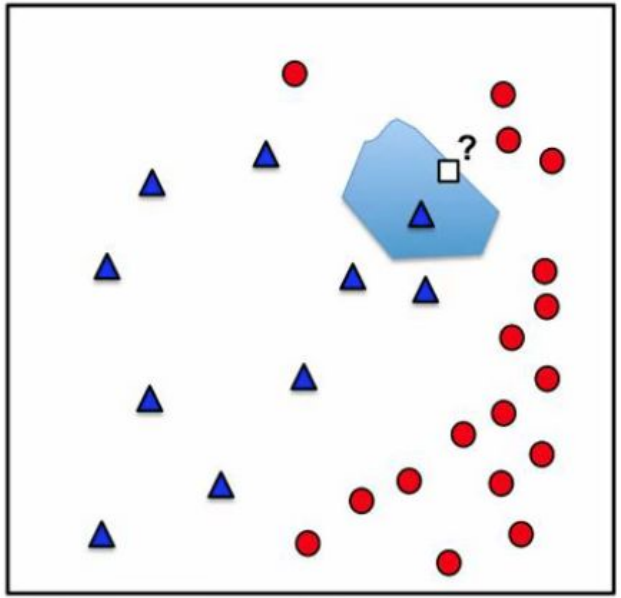


1-NN Example. Source: Facelli et al, 2011

Decision Boundaries

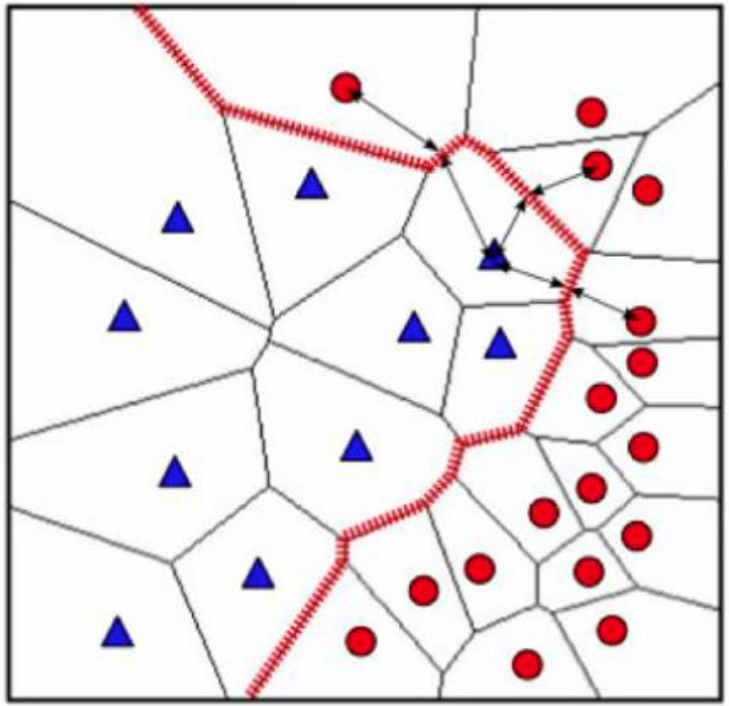
And now, Is it red or is it blue?





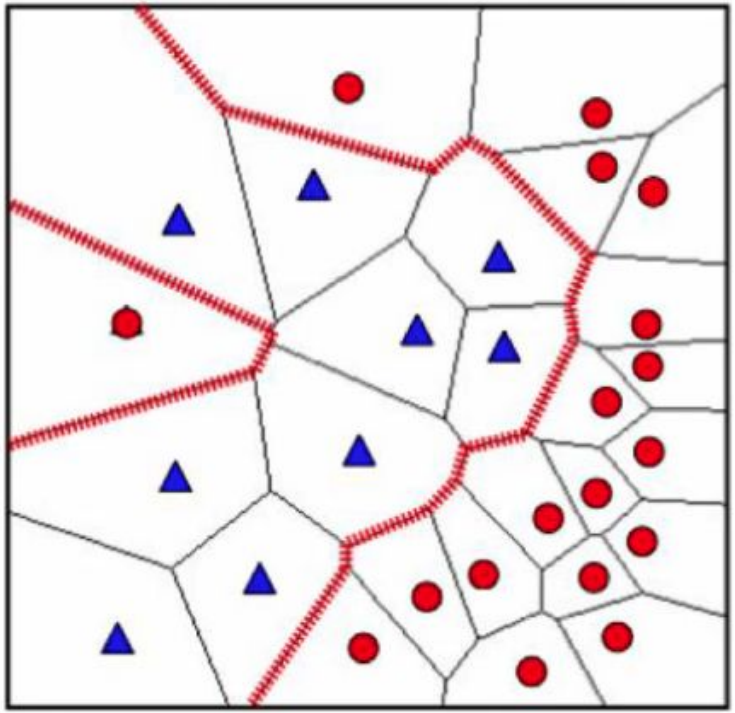
Voronoi Diagram

- Euclidian distance
-



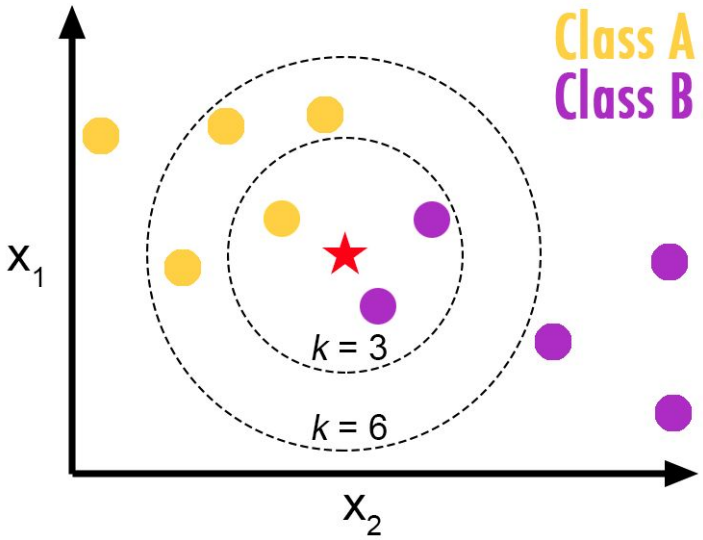
Outliers

- Algorithm is sensitive to outliers
- No confidence P(Y|X)
- To improve robustness
  - Use more than one nearest neighbor to make classification.
  - Generally, larger values of *k* reduce the effect of noise on the classification



K-NN

- A extension to 1-NN is to consider k nearest neighbors
- Classification problems:
  - The most voted class
- Regression problems:
  - Cost function is quadratic error
  - Cost function is standard deviation



Feature Selection

For high-dimensional data (e.g., with number of dimensions more than 10) dimension reduction is usually performed prior to applying the k-NN algorithm in order to avoid the effects of the curse of dimensionality.

For curse of dimensionality understand that as the dimensions of a data set increases, the number of samples required for an estimator to generalize increases exponentially.

Feature Selection

- Select K best
  - Select features according to the k highest scores
  - Chi-Square
    - Chi-square test measures dependence between stochastic variables, so using this function “weeds out” the features that are most likely to be independent of class and therefore irrelevant for classification
- Problems - PCA

Experiments

- Initial approach: “brute force”
  - Small dataset makes this approach feasible
- Even so, large number of dimensions don’t allow to use all possible combinations
- Features list:
  - The ones use it in the reference paper;
  - Each column alone;
  - Combination of FWI indexes and weather conditions in order to use all weather conditions;
  - 20 different columns combinations as features
- K-NN: Numbers of neighborhoods from 1 to 19
- Cross Validation: 9 differents number of folds [3, 5, 7, 9, 10, 11, 15, 19, 23]
- 6840 combinations: ~10 minutes

Experiments

- For each configuration:
  - Used the ‘cross_val_score’ from ‘sklearn.model_selection’
    - RMSE and MAE
  - The results were saved in a .csv file with the configuration description:
    - Average, standard deviation and relative values of the mean
- Configurations with smaller means have large std (and vice versa)
- Created a score column
  - Weights the mean and the std (used the relative values)

Results

```
results.sort_values(by='SCORE_RMSE')[:3][columns]
```

	HYPER_PARAM	FEATURES	K_FOLDS	AVG_RMSE	STD_RMSE	SCORE_RMSE
182	1	FFMC, DMC, DC, ISI	5	55.622226	27.703588	3.242933
5330	16	temp, RH, wind, rain	5	56.972798	28.305571	3.318924
2900	9	X, Y, temp, RH, wind, rain	5	60.524415	24.888559	3.324163

```
results.sort_values(by='SCORE_MAE')[:3][columns]
```

	HYPER_PARAM	FEATURES	K_FOLDS	AVG_MAE	STD_MAE	SCORE_MAE
128	1	rain	5	12.824424	4.130688	0.659871
848	3	rain	5	12.867796	4.094543	0.660152
1568	5	rain	5	12.855549	4.108533	0.660220

Feature Selection

```
lab_enc = preprocessing.LabelEncoder()
encoded = lab_enc.fit_transform(Y)

test = SelectKBest(score_func=chi2, k=4)
fit = test.fit(X, encoded)
print(fit.scores_)
```

```
[   299.036     83.273     40.33    8299.994  24665.44     481.685
    521.351   1262.382    221.665   1972.178]
```

- SelectKBest
  - Chi squared
    - DC: 24.665
    - DMC: 8.299
    - Rain: 1.972
    - RH: 1.262
- The combination of this 4 columns or the 2 best columns weren't in the configuration tested before
  - DC + DMC:
    - **AVG_RMSE: 55,27**; STD_RMSE: 27,92;
    - **AVG_MAE: 16,72**; STD_MAE: 4,27
  - DC + DMC + Rain + RH:
    - **AVG_RMSE: 55,37**; STD_RMSE: 31,58;
    - **AVG_MAE: 19,11**; STD_MAE: 3,92

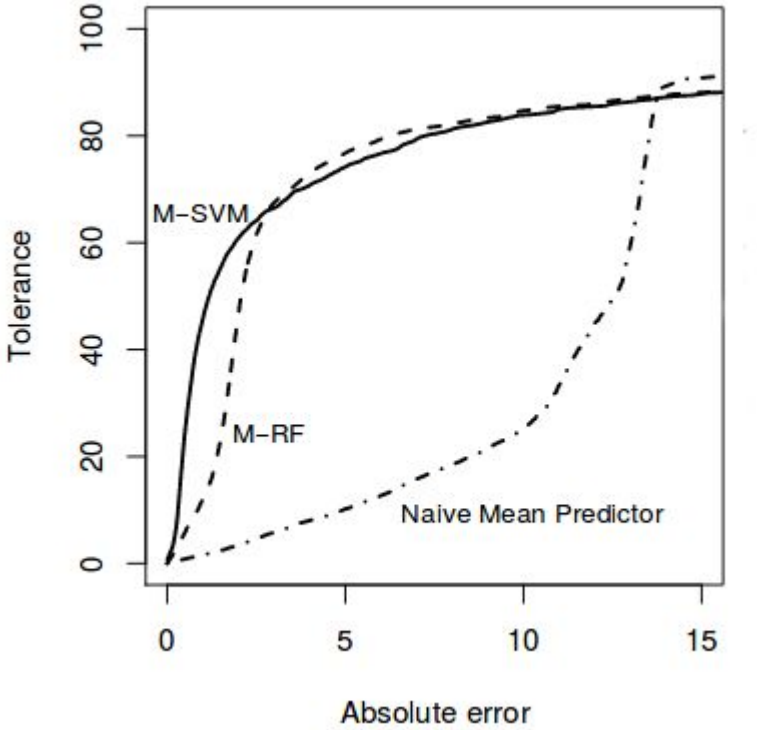
Comparison

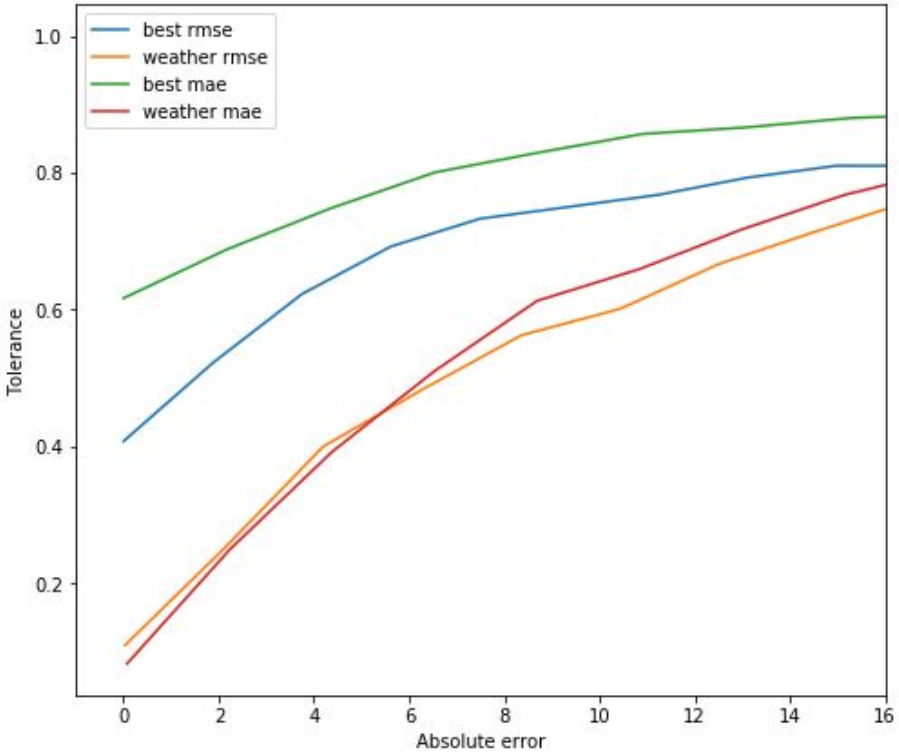
Table 3. The predictive results in terms of the MAD errors ($RMSE$ values in parentheses; underline – best model; **bold** – best within the feature selection)

DM Model	Feature Selection Setup STFWI	STM	FWI	M
Naive	18.61±0.01 (**63.7**±0.0)	18.61±0.01 (**63.7**±0.0)	18.61±0.01 (**63.7**±0.0)	18.61±0.01 (**63.7**±0.0)
MR	**13.07**±0.01 (64.5±0.0)	**13.04**±0.01 (64.4±0.0)	13.00±0.00 (64.5±0.0)	13.01±0.00 (64.5±0.0)
DT	13.46±0.04 (64.4±0.1)	13.43±0.06 (64.6±0.0)	13.24±0.03 (64.4±0.0)	13.18±0.05 (64.5±0.0)
RF	13.31±0.02 (64.3±0.0)	**13.04**±0.01 (64.5±0.0)	13.38±0.05 (64.0±0.1)	12.93±0.01 (64.4±0.0)
NN	13.09±0.04 (64.5±0.0)	13.92±0.60 (68.9±8.5)	13.08±0.05 (64.6±0.1)	13.71±0.69 (66.9±3.4)
SVM	**13.07**±0.04 (64.7±0.0)	13.13±0.02 (64.7±0.0)	**12.86**±0.00 (64.7±0.0)	**12.71**±0.01 (64.7±0.0)

- **Our model:**
  - **Best MAE: 12,82**
  - **Best RMSE: 55,27**

Comparison


Tolerance
0
20
40
60
80
100
M–SVM
M–RF
Naive Mean Predictor
0
5
10
15
Absolute error


best rmse
weather rmse
best mae
weather mae
Tolerance
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
14
16
Absolute error

Conclusions

- K-NN:
  - Simple and efficient for some regression problems
- Feature Selection:
  - Important, but there is no guarantee that will lead to the best model configuration
  - Good understanding of the problem can be necessary
    - Small understanding of the problem could also be good
      - Use a feature selection unlikely to be good
      - Exploration x intensification dilemma