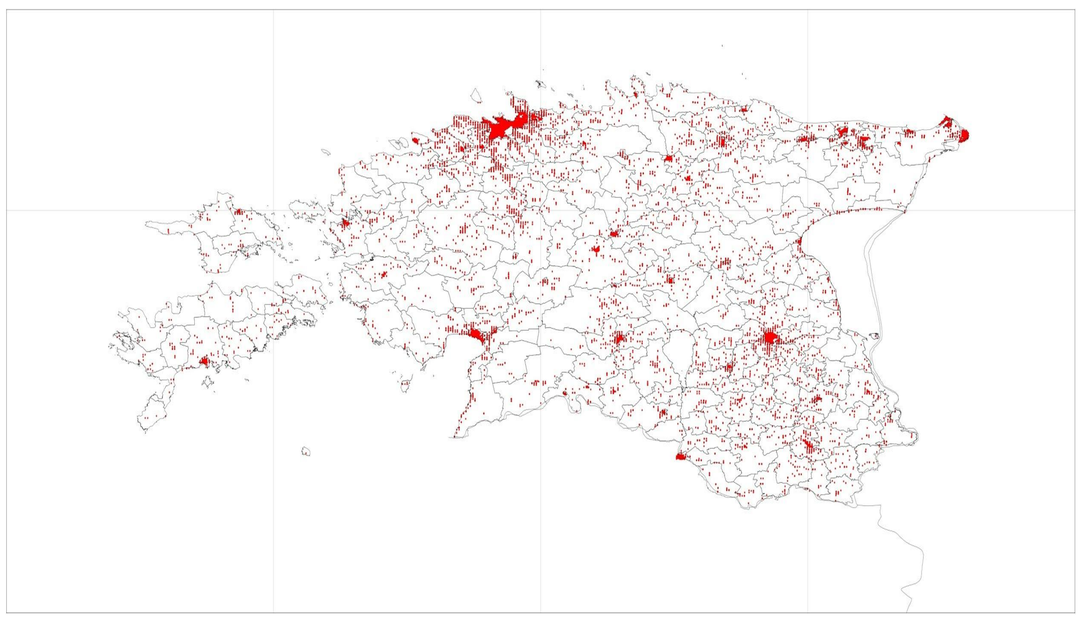
Estonian Property Crimes

Egert Georg Teesaar, Anders Martoja, Ergo Nigola

Introduction

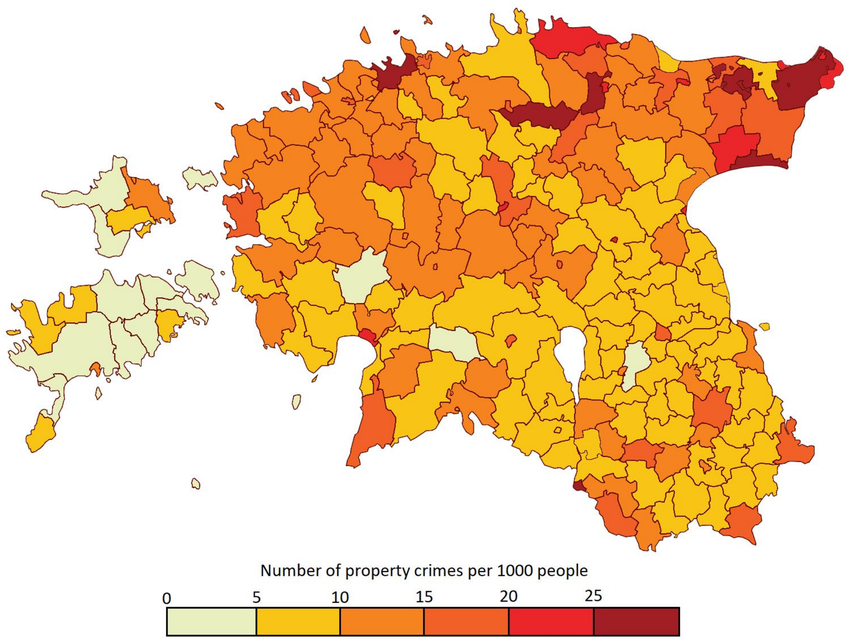
The idea behind our project stems from the possibility to predict the possible cost of a crime against property. Estonian Open Government Data Portal has information about property crimes from 2013 to 2017 freely available, including data about the location, type, cost, time etc. of the crime . The location data was also put to good use and used to polt the locations of the crime scenes all over Estonia. Visualizing the frequency of property crimes relative to the population of every municipality was done to highlight the places where crime is flourishing. Results of our project could be used to locate problematic areas or to predict the loss of a potential or ongoin crime. All of the code for this project is freely available at https://github.com/egertteesaar/Data-Science-CSI-edition.git

Visualizing property crime scenes

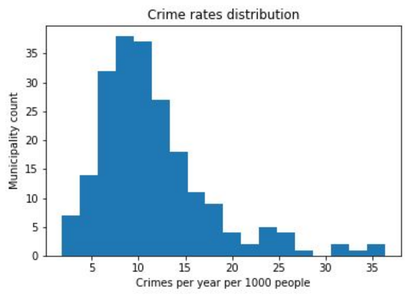


- Data for property crimes was obtained from Estonian Open Government Data Portal
- LAMBERT-EST97 (L-EST97) coordinates from vara_2.csv file were munged into lists for conversion
- pyproj library was utilised to convert L-EST97 coordinates into World Geodetic System coordinates (WGS84)
- Basemap was utilised to draw a map of Estonia in WGS84 projection
- Shapefiles from www.gadm.org were used to draw counties onto the map

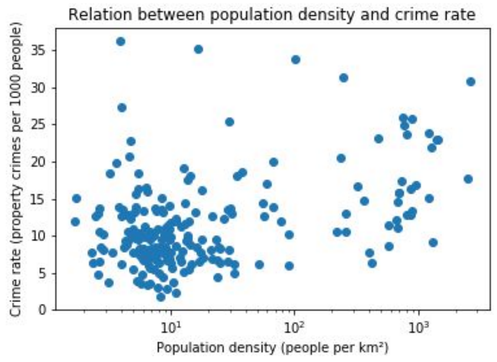
Visualizing property crime rates



The map on the left visualizes property crime rates in different Estonian municipalities. We can see that the lowest crime rates are on the islands, while the highest are in North-Eastern Estonia. In more general, Southern Estonia seems to have lower property crime rates than Northern Estonia. The municipality borders on the map are correct as of the end of 2016 and there are a total of 214 municipalities.



Knowing something about the distribution of population of Estonia, one might notice that areas with higher population, or more precisely higher population density, tend to have also higher crime rates. The relation between crime rates and population density is visualized on the scatter plot to the right. The Pearson correlation coefficient of that data is 0.43.



Municipalities with the highest amount of property crimes per 1000 people:

1. Vaivara vald: 36.31
2. Rakvere vald: 35.11
3. Jõhvi vald: 33.86
4. Narva-Jõesuu linn: 31.25
5. Tallinn: 30.84

Municipalities with the lowest amount of property crimes per 1000 people:

1. Ruhnu vald: 1.81
2. Hiiu vald: 2.35
3. Salme vald: 2.85
4. Muhu vald: 3.37
5. Pöide vald: 3.45

Predicting the loss of the property

Goal: classify crime into specific category

The loss of property (“Kahjusumma”) is divided into 4 classes:

1. 0 - 499
2. 500 - 4999
3. 5000 - 49 999
4. 50 000+

Data is very inbalanced:

```
data['Kahjusumma'].value_counts()
```

Out[143]:

```
0-499          84632
500-4999       21916
5000-49999      2317
50000-x          211
Name: Kahjusumma, dtype:
int64
```

Feature selection:

- Paragraph number
- Section
- Violation
- Cost
- Type of place
- Crime location
- Vechicle type
- Vehicle model
- Vechicle year
- Type of Crime

Data preprocessing:

- One hot encoding of columns every column, because contained categorical values
- Removed NaN columns generated by one-hot-encoder
- Allocating test data (20% of instances were picked from each loss category)

```
training.shape,testing.shape
((87262, 295), (21814, 295))
```

Parameter tuning:

GridSearch exhaustive search from parameters:

- "n_estimators": [5,10,20,30],
- "criterion": ["gini", "entropy"],
- "max_depth": [10,50,100,150,None]
- "min_samples_split":[5,10,2]
- "min_samples_leaf": [1,3,7]
- class_weight: {'0-499': 250, '500-4999': 2500, '5000-49999': 25000, '50000-x': 50000}

Model training:

Sklearn’s RandomForestClassifer

- "n_estimators": 20
- “criterion”: entropy
- class_weight..

(others default)

Model evaluation:

Weighted accuracy:

- training: 0.74
- test: 0.54

Weighted ROC AUC:

- training: 0.92
- test: 0.77