Analysis of Criminal Behaviour Using a Logistic Regression Model
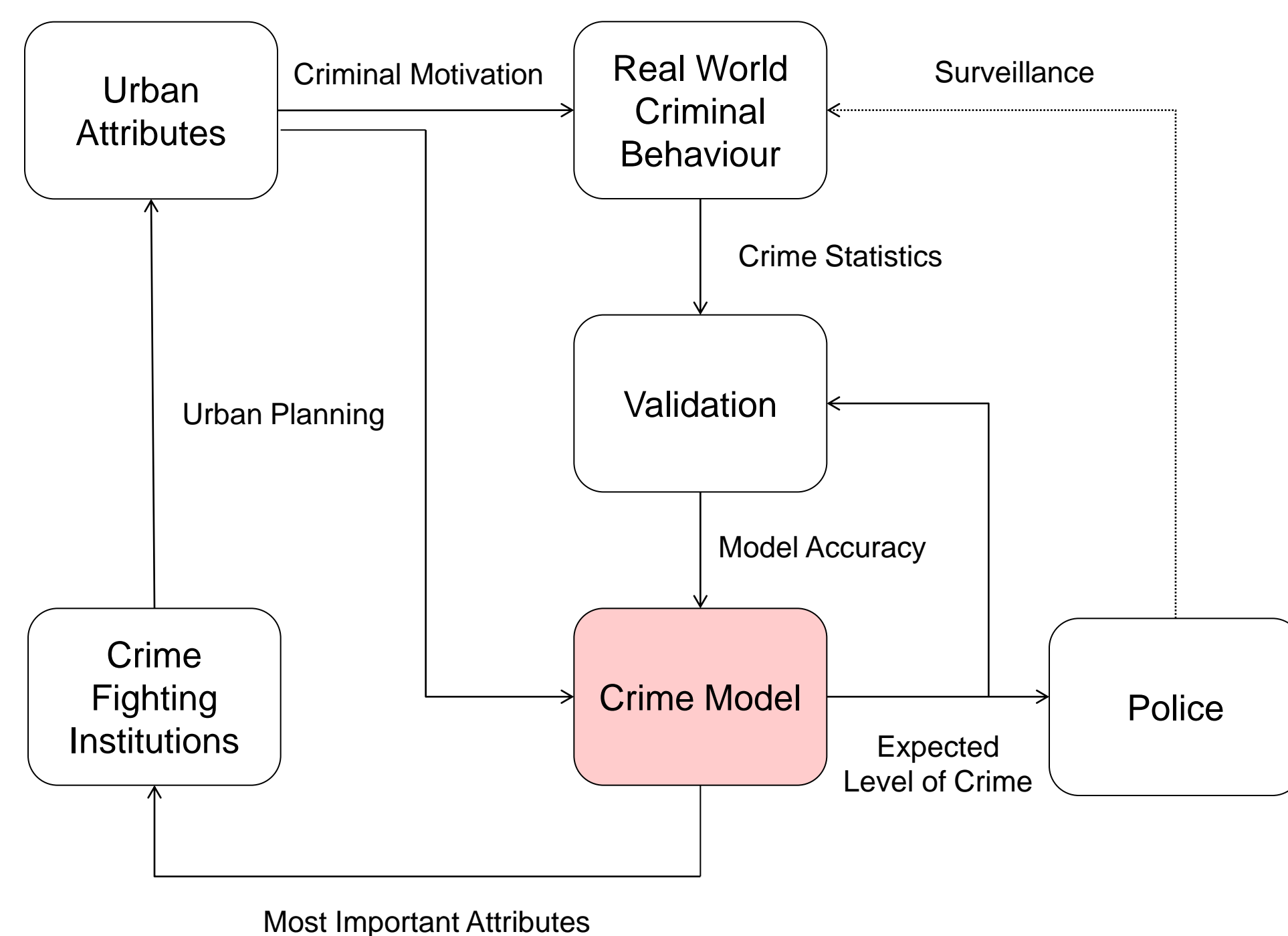
Maria Jofre

Department of Industrial Engineering, University of Chile

mjofre@ing.uchile.cl

Introduction

Understanding the crime phenomenon is fundamental for crime fighting institutions. Therefore, statistical evidence has acquired a critical importance when analysing criminal behaviour.

The distribution of thefts varies over space, suggesting that urban characteristics are key variables when a crime occurs. This study provides analytical evidence of the impact of urban attributes in the expected level of crime.

Aim

The objective of this study is to determine the most important features that contribute to a high level of criminal activity.

In particular:

- Develop a logistic regression model to predict the level of crime in a region based on its urban characteristics.
- Identify the variables that define a higher probability of theft occurrence.
- Provide recommendations related to police resource organisation and urban planning for crime fighting institutions.

Method

Data Collection

Two databases are used in this study: (i) police crime reports of thefts, and (ii) urban attributes related to commerce, tourism, culture, transport, government and religion, among others.
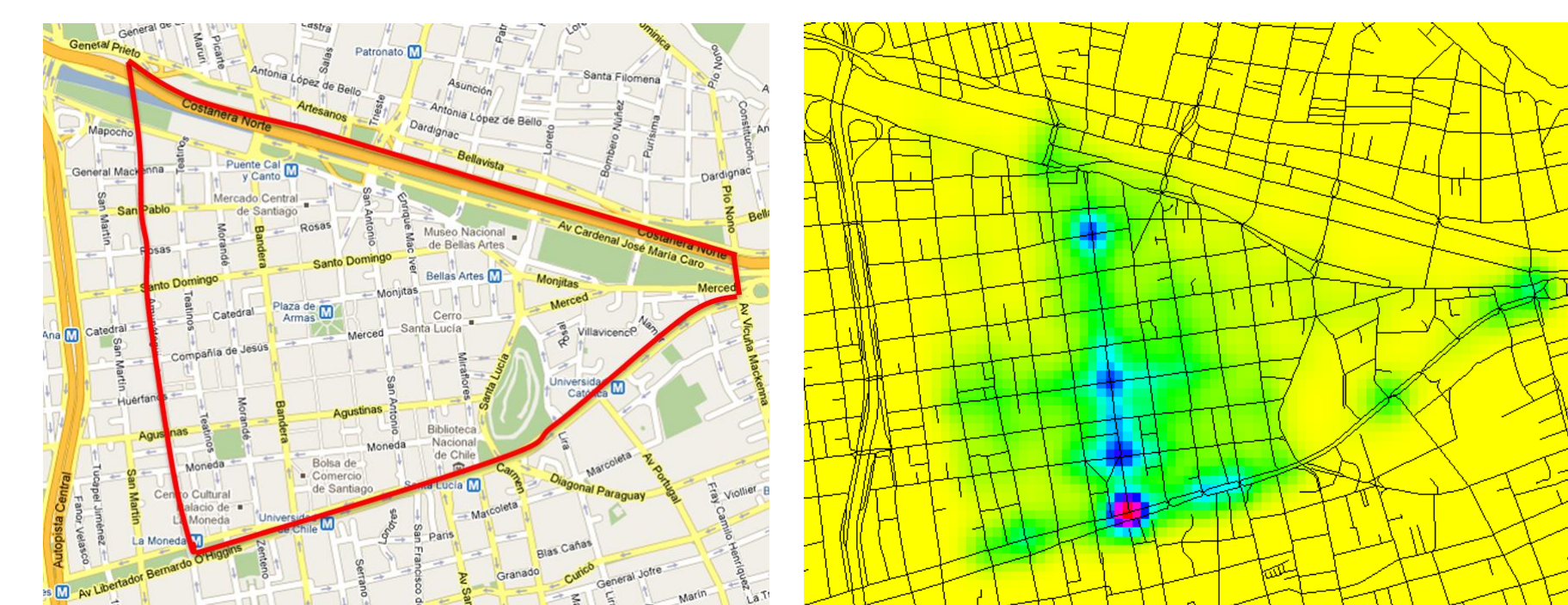
(a) Area of interest   *(b) Theft distribution in the studied area*

Data Processing

The studied region is divided in smaller areas defined by Voronoi polygons centred on street intersections.
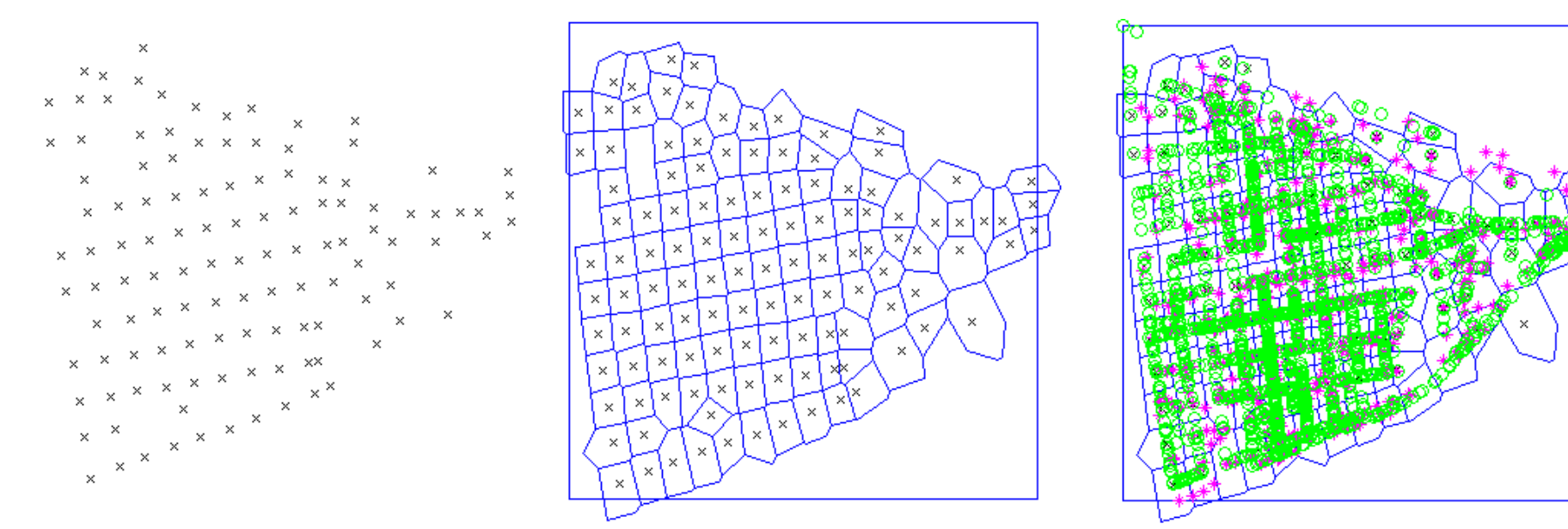
(a) Corners   *(b) Voronoi Polygons*   *(c) Thefts and Attributes*

Logistic Regression Model

To achieve the pursued objective, a multivariate logistic regression model is proposed. The idea is to determinate the relationship between a dependent variable (Y) and one or more independent explanatory variables (X_i). In this case, the dependent variable is the level of crime and the independent variables are the urban attributes distributed along the studied area.

The model determines a coefficient (β_i) for each attribute (i), that defines the contribution to the predicted probability of high level of crime ($P(Y=1)$), which is calculated as follow:

$$P(Y = 1) = \frac{1}{e^{-(\sum \beta_i x_i)} + 1}$$

For this particular data mining technique, a Maximum Likelihood method is used to estimate the coefficients of the model. The idea of this method is to find the values of β_i that maximise the probability of the real data being generated by the model.
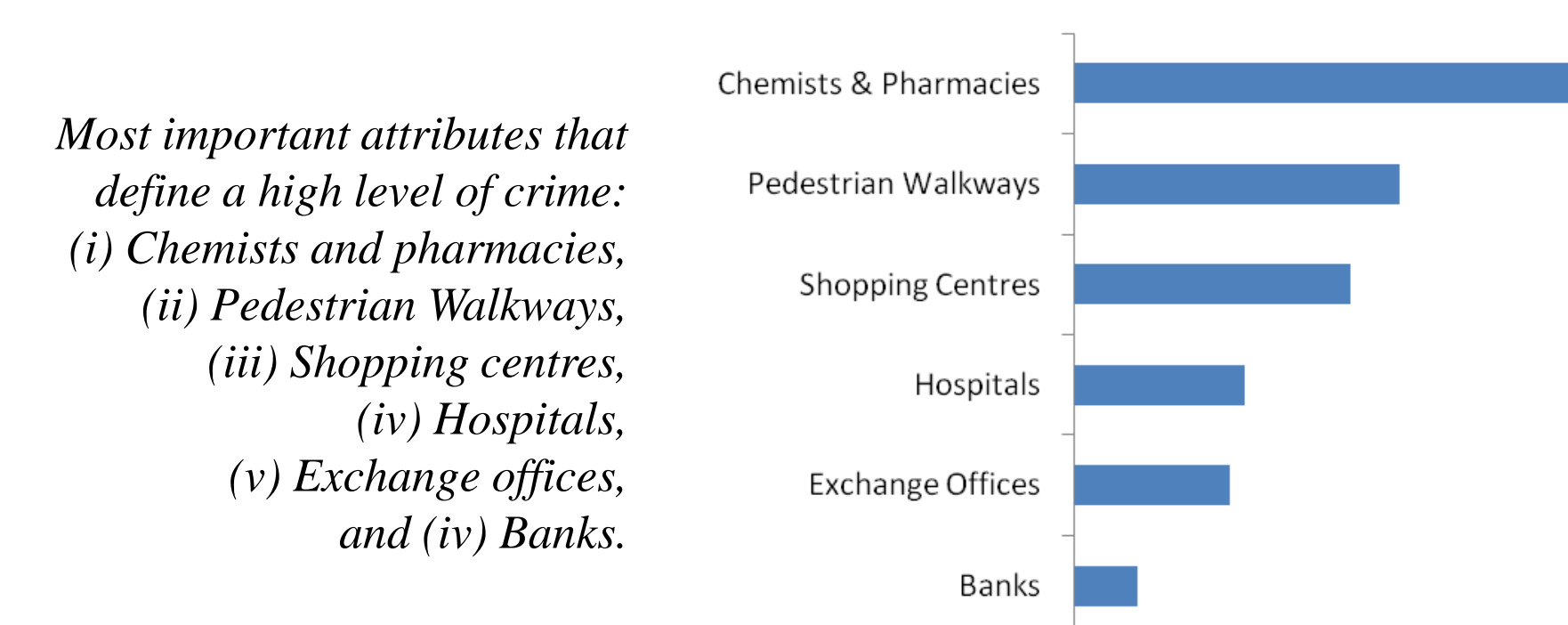
Subsequently, the estimated coefficients are used to calculate the probability of theft occurrence in each Voronoi polygon.

Once the probability of occurrence is defined, then each polygon is classified in deciles depending on how likely is to present a high level of crime.

Results

The application of the proposed model determines the relationship between 18 considered urban attributes and the expected level of crime.

The results show that 6 urban characteristics are the most important to explain the phenomenon. All of these attributes are directly proportional to the level of crime. Therefore, if these variables are present in a particular area, then a higher amount of thefts is expected in that region.

Most important attributes that define a high level of crime: (i) Chemists and pharmacies, (ii) Pedestrian Walkways, (iii) Shopping centres, (iv) Hospitals, (v) Exchange offices, and (iv) Banks.

Once we obtain the scores for all the attributes, we can proceed to calculate the probability of theft occurrence in each Voronoi polygon and the decile to which it belongs.

The polygons belonging to the first decile are the zones that are more likely to suffer thefts, i.e. most dangerous corners.

The graph below shows that the average expected probability of occurrence for the most dangerous corners is 77% compared to 83% of theft occurrence in the real world.
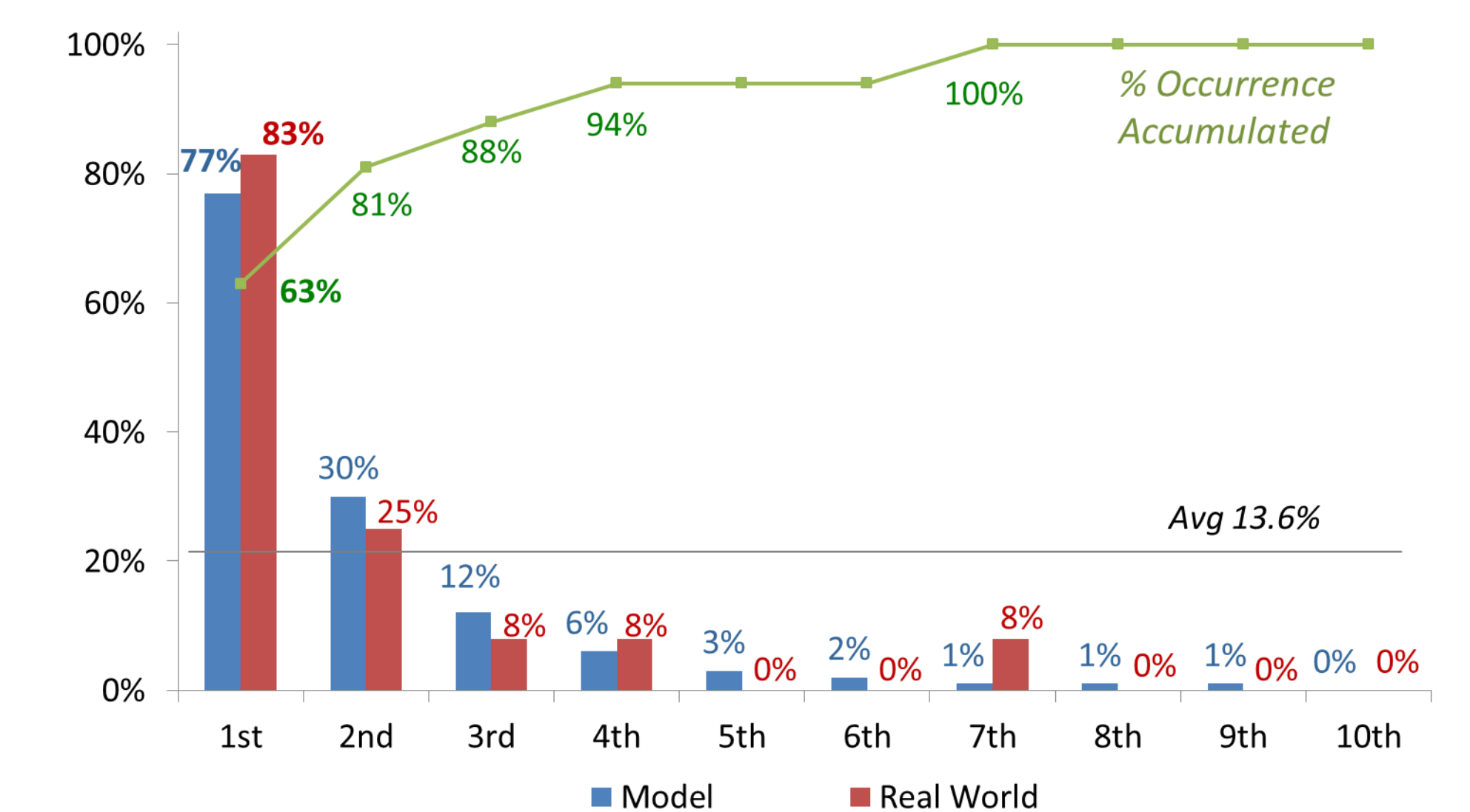
Expected probability of a theft occurrence compared to real world data, and the accumulated percentage of occurrence, by deciles

In addition, we can observe that the first 3 deciles contain 88% of all crimes, whilst the other 7 deciles only contribute with the remaining 12%.

The results of the model are validated using the real world crime data as ground truth, thus obtaining a overall level of accuracy of 93.2%.

The model's precision for high level of crime prediction is better when the analysis is divided in deciles. The level of accuracy of true positives significantly increases in each one of the groups, particularly in the first decile.

Conclusion

The presence of chemists, pedestrian walkways and shopping centres define a greater likelihood of a theft.

Given that the police resources are limited, it is recommended police surveillance in critical areas determine by the model.

Also, we recommend using the results of this study for urban planning, taking in consideration the most dangerous combination of urban characteristics.

As a future work we expect to model crime as an spatial-temporal phenomenon using more complex statistical techniques, such as Gaussian Process. This model can also quantify the uncertainty in crime prediction.