

Sparse Big Data Problem. A Case Study of Czech Graffiti Crimes

Jiří Horák, Igor Ivan, Tomáš Inspektor and Jan Tesla

Abstract Sparse data sets may be considered as a one of the issues of big data generating extremely uneven frequency distribution. To deal with this issue, special methods must be applied. The study is focused on the Czech graffiti crimes and selected factors (property offences, buildings, flats, garages, educational facilities, and gambling clubs) which may influence the graffiti crimes occurrence. For regression analysis decision trees with the exhaustive CHAID growing method were applied. Grid models with 100, 500 and 1000 m cells were tested. The model of 1 km grid was evaluated as the best. The most influencing factors are the occurrence of secondary schools and gambling devices enhanced for several territorial units. The results of the decision tree for 1 km grid are validated using alternative models of data aggregation—aggregation around the randomly selected building and randomly distributed points.

Keywords Big data · Graffiti · Sparse data · Data mining · Crime · Multidimensional modelling · Decision tree

1 Introduction

The big data faces various problems that are usually not as evident as the high volume of available data. One of the specific issues which causes severe problems in data processing and analysis occurs in a situation of uneven data distribution,

J. Horák (✉) · I. Ivan · T. Inspektor · J. Tesla
Faculty of Mining and Geology, Institute of Geoinformatics,
VŠB-Technical University of Ostrava, 17. Listopadu 15/2172,
708 33 Ostrava-Poruba, Czech Republic
e-mail: jiri.horak@vsb.cz

I. Ivan
e-mail: igor.ivan@vsb.cz

T. Inspektor
e-mail: tomas.inspektor@vsb.cz

J. Tesla
e-mail: jan.tesla@vsb.cz

especially in cases of extremely left skewed distribution (it means a large portion of very small or even zero data).

Sparse data matrices can be effectively stored and managed in multidimensional databases. The multidimensional database is usually understood as a part of Business Intelligence (Horák et al. 2015). Business intelligence (BI) mainly refers to computer-based techniques for identifying, extracting and analysing business data. They are abstracted using Extract, Transform, Load (ETL) tools into a data warehouse (Badard et al. 2012). The ETL processes include selection, harmonisation and aggregation processes. In case of crime events several important harmonisation procedures have to be provided—smart procedures to calculate referential time of event from time interval based on the type of event, length, beginning and ending of the interval; classification of facilities, buildings, age classification etc. (Horák et al. 2016; Loshin 2012).

Aggregated data is stored in fact tables. Data is organised and aggregated according to dimensions. Dimensions usually contain hierarchical structure which can be stored as an explicit (set of normalised tables) or implicit (one table with de-normalised solution) hierarchy (Horák and Horáková 2007). Each element of the hierarchy can be used for grouping facts.

A data warehouse integrates data from multiple data sources. Data warehouse and multidimensional modelling may be applied for different purposes not only for business strategies. The public sector usually repeatedly integrates data from regularly provided data sources and needs effective tools for repeated spatiotemporal analysis of large volume of incremental data. One of such application can be the database for crime prevention, developed for the Ministry of Interior of the Czech Republic (Horák et al. 2016).

In our database, two dimensions are shared among all fact tables—geographical and temporal dimensions. They enable to link fact tables together. The location is considered as a cornerstone of the database and is expressed in each fact table by two geographical dimensions—the administrative dimension and geometrical dimension (square grid). The administrative dimension identifies the administrative units. The basic unit is the part of a municipality which seems to be an appropriate compromise between data availability, privacy protection and the required highest detail. The elementary unit for the geometrical dimension is 100×100 m cell, arbitrarily aggregated. It is consistent with the 4th level of the scale system for public authority where 10 km window and 100 m grids for communes and urban districts are recommended (Bacler 2014). The basic structure of the multidimensional database and structure of the fact table about crimes is described in detail in Horák et al. (2015).

To explore issues and possible solutions of analysing high volume sparse dataset we selected evidence of graffiti crimes in the Czech Republic and explored influences of several factors to evaluate the possibility of improved predictions. Graffiti is a common phenomenon. Often it is represented by simple patterns, signatures (tags) from young people. These people call themselves graffiti artists and consider it as a lifestyle. The graffiti can be seen from several perspectives—as an art (Lachmann 1988), as a reflection of societal customs and attitudes (Stocker et al.

1972), a method to attain notice or fame (Lachmann 1988), as a form of political statement (Ferrell 1995), or as territorial markers (Ley and Cybriwsky 1974).

Unlike sprayers themselves, most of the society considers graffiti as an expression of vandalism as well as the Czech law. The act of vandalism expressed as graffiti is included in §228 par.2 Act No. 40/2009 as an independent fact of a criminal damage offense. In literature and (negative) consciousness of people, the concepts of vandalism and graffiti are often confused. Nevertheless, there is no direct evidence that those who create graffiti, carried out vandalism simultaneously (Thompson et al. 2012). Also Wilson and Healy (1987) have not found statistically significant relationship between graffiti and vandalism in trains. However, what and under which conditions the difference between graffiti and vandalism forms, remains unclear (Halsey and Young 2002). Cohen (1969) distinguished five main motivations for graffiti and vandalism—gaining or sustaining membership within a ‘deviant’ group through anti-social acts that reinforce group membership, self-esteem, self-expression, to disrupt the order of authority, enjoyment and the rush associated with the illegality of the behaviour. Similarly, Bandaranaike (2001) identified following main reasons for graffitiist’s behaviour from questionnaires in Australia—they are asserting identity, in defiance of societal norms, reacting to a heartfelt feeling (need for love or company; anguish of abuse, discrimination), larrikinism and seeking adventure, the adrenaline rush and pleasures in risk taking and association with fad and fashion (imitate peers in other urban environments).

Graffiti is associated with a young age and often linked to school also. Most often mentioned reason is “boring” at school and lack of opportunities for self-expression. This leads to the breaking the rules and disrupting the system by youth (Thompson et al. 2012; Iveson 2007; McCormick 2003). “The Broken Windows” theory (Wilson and Keeling 1982) says that the presence of vandalism, no matter how small, means that unorganised social environment normalises and even supports other disorders. Similarly “contagion theory” principle (Armitage 2002) can be applied. It means the distribution of graffiti should be clustered.

According to Tygart (1988), vandalism peaks in seventh grade of primary school and progressively decreases with further successfully completed year. From a socio-economic perspective, the typical school vandal comes from middle economy class with low income (Howard 1978) or was fired from the school (Yankelovich 1975 in Goldstein 1997). Schools are not only a place of concentration of young people, but the campus is also a suitable environment for graffiti. Offenders are attracted to such public places where their exhibition may have a wide audience, i.e. a form of exhibitionism (Buck et al. 2003). Also other factors should be taken into account. According to Gibbons (2004), graffiti comes out as one actor which is consistently correlated with measures of fear of crime, neighbourhood decline, property crimes and escalating crime rates.

Spatial analyses are usually made at the level of individual cities. Popular techniques include creating a square grid and comparison of multiple factors using regression models. E.g. Megler et al. (2014) reported 59,000 records of graffiti within 2 years from San Francisco. No temporal pattern in the data set was recognised. Thus, all data was aggregated to square grid 381 m. Finally, they

applied a GWR model which explained more than 2/3 of the variation of graffiti records.

Grid models generated from multidimensional databases may provide an effective solution for sparse big data. The aim of the study was to verify results of grid models in regression analysis, compare them with alternative models and evaluate the stability of significant outputs.

2 Data Sources

The evidence of graffiti comes from the crime register of the Police of the Czech Republic (PCR). Data was exported from centralised information system for event recording (ETR, Evidence of Criminal Proceedings) which contains detail information about each crime commission. All events in the CR with classification as a criminal act “sprejství” (§228 par.2 Act No. 40/2009) in the period 1.1.2014–10.12.2014 were selected. Thus, there are only events registered by PCR as serious events (suspicion of crime act) and we have no evidence of other graffiti manifestation. It is obvious that the number of all graffiti occurrences is much higher but it is almost impossible to evaluate the missing part. Marešová (2011) estimates the number of unregistered (latent) crimes of vandalism in the range of hundreds to thousands per year. Less serious damages are more frequent. Nevertheless, according to German outputs (Marešová 2011), it is valid for the urban environment that main typical features of latent crime are well overlapped with features of registered crime. Thus, the relationships between registered graffiti crime occurrence and influencing factors should be similar as for the latent or total crime. The relationships are also influenced by other factors like different level of tolerance to graffiti (issued on a different level of reporting), differences in mapping (different frequency of visiting places due to the different number of people or officers), the occurrence of police stations (i.e. Megler et al. 2014), etc.

Three types of locations are distinguished in ETR—a place of commitment, place of effect and place of reporting. Only first two types were selected (and the second represents only 4 %). The same time interval and type of locations were applied for the selection of records about property offences.

Other data sources contain referential and potentially explanatory information:

- Register of schools (Min. of Education, Youth and Sports) offers contact information and classification of education level and type (<http://stisko.uiv.cz/registr/vybskolm.asp>). Data was exported on 6/2015 and represents valid records about registered educational facilities in the school year 2014/15.
- Register of gambling devices contains registration ID, addresses and description of placement for all technical lottery devices approved by the Min. of Finance to 10/2014.
- Register of Statistical Units and Buildings (RSO, Czech Statistical Office) contains information about address points and buildings with an indication of

the type of usage, construction classification and number of flats. Used data comes from the version valid to 1.7.2014. Finally number of buildings, flats and garages were processed.

Data from all registers except RSO was harmonised. In the case of the crime register several types of control procedures were applied (integrity constraints, validity check of the time range, geographical range, the validity of codes in the database) (Horák et al. 2015). Address descriptions in all registers were harmonised (e.g. name of the municipality, municipal part, street, numbers).

Further, geocoding of registers of schools and gambling devices were performed using a specially developed SW Geocoder (Fojtík et al. 2016). The quality of geocoding has to be assured on the satisfactory level. The minimal limit of the success rate is 85 % which should not cause a systematic spatial error (Ratcliffe 2004 in Andresen and Malleson 2013). The success rates for our data sets reached almost 100 % but the quality (positional errors) is variable. The quality of geocoding of Geocoder SW was compared with results of geocoding API of web search engines (Fojtík et al. 2016) with a positive evaluation. In total, 1511 graffiti crimes, 26,046 schools and 6342 gambling clubs were selected in the Czech Republic.

3 Grid Evaluation

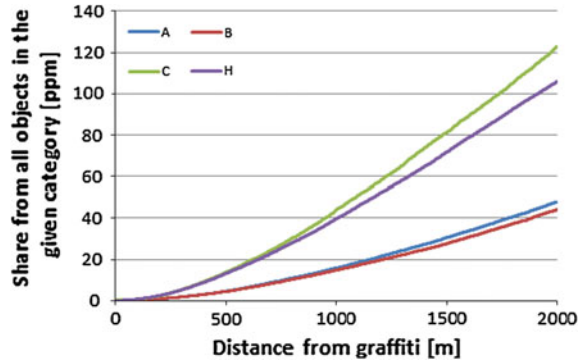
To explore problems of high volume sparse data, we analyse relationships among the graffiti crime acts and the occurrence of property offences or locations of selected objects like buildings, flats, garages, schools, gambling clubs.

To verify factors with higher influence (they are more concentrated around graffiti crimes), the distribution of explored type of objects was analysed using an aggregation of selected objects around graffiti crimes within increasing distance (Fig. 1). It is clear the secondary schools and gambling clubs are much more concentrated around graffiti crimes than lower levels of education facilities.

The problem is such calculation is highly time-consuming because all distances between crime placement and the objects have to be calculated, select the closest one and a share from the total enumerated. The main issues are that such way we can model and explore the influence of only one factor at once (no combination of factors), no non-numerical factors (predictors) (including regional differences) can be included, and no statistical significance of the results is provided.

The main idea was to utilise square grids for analysis of relationships (dependencies) and compare results with alternative methods of data aggregation. Usage of grid analysis (incl. pre-processing) is fast and flexible. Point data is aggregated directly to cells according to its coordinates. All events and objects are aggregated independently and no enumeration of spatial relationships between them is required (the advantage of the implicit topology of the raster model). These features facilitate

Fig. 1 The relative share of kindergartens (A), primary schools (B), secondary schools (C), gambling clubs (H) within an increasing distance from graffiti crimes



spatial analysis and provide great benefits especially for processing of high data volumes.

The harmonised data was aggregated to square grids and stored in a multidimensional database. The multidimensional database effectively stores such data due to the specific structure containing only non-zero counts. However, when we expand data set to cover the whole country we obtain massive data sets with rare occurrences of investigated events. Such data can be classified as a sparse big data.

The main question is how to deal with such data.

Due to the small number of events we did not explore the temporal pattern. Finally, three spatial aggregation levels were applied—grid 100 m (basic units), 500 m and 1000 m. The grid 100 m in the Czech Republic contains 7,423,937 cells (the graffiti crimes occur only in 1189 cells, total sum is 1511), the grid 500 m includes 297,134 cells (the graffiti crimes occur only in 803 cells), and finally the number of cells in 1 km grid is 80,131 (only 578 “positive” cells). The data volume (above 7 mil. records) is too big for processing various statistics methods in software like SPSS (e.g. insufficient memory, too big data volume); in these cases, the random selection of 10 % of data was applied.

The frequency distribution of gridded graffiti crimes is extremely curved and skewed (with a very high portion of NULL cells), obviously more with higher spatial resolution (Fig. 2). The power distribution is highly skewed with a long tail (Brown and Liebovitch 2010) and follows a straight line in log-log graphs. Shapes of histograms indicate that their middle part can follow a power law distribution, more with more coarse spatial resolution, but both tails of the distribution are more extreme than the power law distribution. Such extreme type of distribution makes impossible to use many of typically used statistical and spatial methods (e.g. geographically weighted regression) requiring the normal distribution (Table 1).

The other variables aggregated to cells (selection from the multidimensional database) are number of property offences, buildings (all), flats, garages, kindergartens, primary schools, secondary schools, gambling clubs and gambling devices.

Due to such extreme imbalance between cells with and without graffiti crimes, the proximity measures between variables are explored only for 1 km grid.

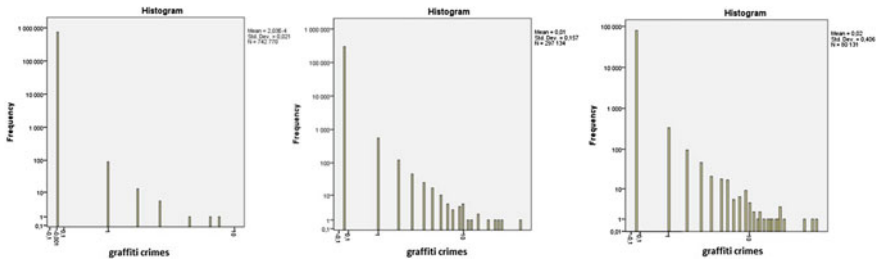


Fig. 2 Histograms of gridded graffiti data sets (100, 500, 1000 m; log-log scales)

Table 1 Statistical characteristics of gridded graffiti data sets

Grid (m)	N	Share from all data (%)	Maximum	Mean	Variance	Skewness	Kurtosis
100	742,770	10	8	0.0002	0.000	193	55,027
500	297,134	100	34	0.01	0.025	88	12,590
1000	80,131	100	45	0.02	0.165	55	4470

The partial correlation cannot be evaluated because it requires a normal distribution. The calculation of Spearman’s coefficient of correlation was impossible for all data because the volume of the data matrix is out of the SPSS capabilities and thus, 10 % random sample was used. The results (Table 2) show the most important dependencies between graffiti and secondary schools (0.268), gambling clubs or gambling devices (0.251 and 0.250 resp.), followed by primary schools (0.207), garages (0.192) and kindergartens (0.181). The correlation with property crimes is surprisingly low.

Due to the fact of small numbers of graffiti and several other factors, also a proximity matrix for categorised (binary) variables was calculated using Phi 4-point correlation for the full data set (Table 3). This index is a binary analogue of the Pearson correlation coefficient and it has a range of -1 to 1 (SPSS 2007). The differences are better expressed than Spearman’s rho in the previous table. The occurrence of graffiti crimes shows the highest proximity to secondary schools (0.307) and gambling clubs (0.306), followed by primary schools (0.206), garages (0.195) and kindergartens (0.189). It well corresponds to the distance analysis (Fig. 1).

For evaluation of factors influencing graffiti crimes, decision trees were selected. The decision tree procedure creates a tree-based classification model. It classifies cases into groups or predicts values of a dependent (target) variable based on values of independent (predictor) variables. The decision tree in SPSS uses various growing methods (SPSS 2007):

- CHAID (Chi-squared Automatic Interaction Detection) chooses the independent variable that has the strongest interaction with the dependent variable. Categories of each predictor are merged if they are not significantly different on the dependent variable.

Table 2 Spearman's coefficients of correlation between explored variables

Correlations										
Type	spearman's rho, Statistics = correlation coefficient									
	Graffiti oc.	Property of.	Gambling	Gamblmach	Kindergarten	Primary	Secondary	Building occ.	Flat occur.	Garage
Graffiti oc.	1.000	.170**	.251**	.250**	.181**	.207**	.268**	.126**	.130**	.192**
Property of.	.170**	1.000	.315**	.315**	.352**	.339**	.225**	.563**	.554**	.338**
Gambling oc.	.251**	.315**	1.000	1.000**	.340**	.352**	.375**	.240**	.239**	.274**
Gamblmach	.250**	.315**	1.000**	1.000	.339**	.352**	.375**	.240**	.239**	.273**
Kindergarten	.181**	.352**	.340**	.339**	1.000	.713**	.269**	.354**	.360**	.255**
Primary sch.	.207**	.339**	.352**	.352**	.713**	1.000	.325**	.318**	.325**	.284**
Secondary s.	.268**	.225**	.375**	.375**	.269**	.325**	1.000	.167**	.171**	.264**
Building occ	.126**	.563**	.240**	.240**	.354**	.318**	.167**	1.000	.946**	.301**
Flat occur.	.130**	.554**	.239**	.239**	.360**	.325**	.171**	.946**	1.000	.295**
Garage occ.	.192**	.338**	.274**	.273**	.255**	.284**	.264**	.301**	.295**	1.000

**Correlation is significant at the 0.01 level (2-tailed)

Table 3 Phi 4-point correlation between explored variables

	Proximity matrix									
	Fourfold point correlation									
	Graffiti oc.	Property off	Gambling oc.	Kindergar oc.	Primary sch.	Second sch.	Building oc.	Flat occur.	Garage oc.	
Graffiti oc.	1.000	.145	.306	.189	.206	.307	.065	.072	.195	
Property off	.145	1.000	.274	.330	.317	.192	.354	.366	.269	
Gambling oc.	.306	.274	1.000	.343	.377	.396	.121	.133	.262	
Kindergar oc.	.189	.330	.343	1.000	.725	.269	.187	.211	.252	
Primary sch.	.206	.317	.377	.725	1.000	.325	.167	.188	.254	
Second sch.	.307	.192	.396	.269	.325	1.000	.082	.092	.221	
Building oc.	.065	.354	.121	.187	.167	.082	1.000	.882	.155	
Flat occur.	.072	.366	.133	.211	.188	.092	.882	1.000	.171	
Garage oc.	.195	.269	.262	.252	.254	.221	.155	.171	1.000	

This is a similarity matrix

- Exhaustive CHAID is a modification of CHAID that examines all possible splits for each predictor.
- Classification and Regression Trees (CRT) splits the data into segments that are as homogeneous as possible on the dependent variable. A terminal node in which all cases have the same value for the dependent variable is homogeneous, “pure” node.
- Quick, Unbiased, Efficient Statistical Tree (QUEST) is a method that is fast and avoids other methods’ bias in favour of predictors with many categories. QUEST can be specified only if the dependent variable is nominal.

CHAID and exhaustive CHAID enables multiway node splitting of trees while CRT and QUEST only binary. The main advantages of decision trees are—no assumptions about the type of distribution, the data population can be heterogeneous, the effect of predictors can be nonlinear (Hendl 2006), data can be measured on different scales (Alkhasawneh et al. 2014) and it is also possible to predict values of the dependent variable.

It is recommended to designating one or more categories as target categories. The selected categories are treated as the categories of primary interest in the analysis and enable to use some classification rule options and gains-related output. It has no effect on the tree model, risk estimate, or misclassification results (SPSS 2007). In our case, the category “yes” for graffiti crimes was selected of primary interest.

In all models, a category (occurrence) of graffiti crimes was the dependent variable and independent variables were property offences, gambling clubs, kindergarten, primary schools, secondary schools, all buildings, all flats, garages and name of district.

The testing showed the usage of scale variables as useless, due to prevalent dividing trees in nodes according to the values 0 and 1. Therefore, we reclassified all numerical scale variables into binary values (0, 1) for the occurrence of given phenomena in a cell. The best results (approved by validation) are obtained using exhaustive CHAID method of tree growing. Significance levels for splitting nodes and merging categories are both 0.05. The maximal depth of trees for CHAID and exhaustive CHAID is three. The results for three different spatial resolutions (100, 500 and 1000 m) indicated best characteristics for a model with 1000 m grid (Tables 4 and 5). Larger grid cells were not explored due to the limited capacity of full comparison with alternative aggregation models and decreasing of model performance found for 2000 m grid (88 %).

It seems that model of 100 m is less probable according to 0 correct predictions, low maximal response but also the provided selection of predictors where secondary schools are absent and the maximal influence is given to property offences.

The best results are reached by 1000 m grid model. The most influencing variable was the occurrence of secondary schools, followed by the districts and the occurrence of gambling. Other variables may have a local effect (garages, property offences and primary schools).

Table 4 Parameters of classification trees for prediction of graffiti crime occurrence in various grids

Grid (m)	List of significant independent variables	Nodes	Percent correct prediction "yes" (%)	Max. response (prediction % for yes)	1st variable	2nd variable	3rd variable
100	Property offences, district, all flats, garages, gambling clubs, kindergartens, all buildings, primary schools	158	0	18	Property offences = 1	District = Brno-město	Gambling clubs = 1
500	Gambling clubs, district, property offences, garages, all flats, primary schools, kindergartens, secondary schools	160	6.6	58.2	Gambling clubs = 1	District = Brno-město	
1000	Secondary schools, district, property offences, primary schools, garages, kindergartens, gambling clubs, all flats	171	26.6	71.1	Secondary schools = 1	District = Písek, Brno-město, Jeseník, Prostějov, Ostrava-město, Beroun	

Table 5 Classification results for three grid models

Model (m)	Observed no – predicted no	Observed no – predicted yes	Observed yes – predicted no	Observed yes – predicted yes	% Correct for yes	Overall % correct
100	7,422,764	0	1,173	0	0	100
500	296,294	38	749	53	6.6	99.7
1000	79,454	99	424	154	26.6	99.3

Table 6 Classification results for three models

Model	Observed no – predicted no	Observed no – predicted yes	Observed yes – predicted no	Observed yes – predicted yes	% Correct for yes	Overall % correct
Grid1km2	79,454	99	424	154	26.6	99.3
RB1km2	78,992	1642	3876	3310	46.1	93.7
RP1km2	79,508	41	528	54	9.3	99.3

The whole tree cannot be visualised due to a large number of nodes (171). The detail is in the Fig. 3. Each statistically significant splitting of the tree is documented by the name of variable (predictor) with adjusted *p*-value, the values of Chi-square test and degrees of freedom. Each tree branch is headed by the value of this variable valid for the whole branch (i.e. district is one of the following names: Zlín, Přerov, Hodonín, Česká Lípa, České Budějovice, Praha-západ; the occurrence of graffiti club is 1). The box describes the node as a data group defined by this tree branch. It contains the identifier of the node, frequency of No and Yes answers in absolute and relative values and the total share of this group to the whole dataset again in absolute and relative values.

The highest positive response in graffiti crimes prediction was reached for the following combination:

- 71 %: (secondary schools occur.) AND (districts: Brno-město BM, Ostrava-město OT, Prostějov PV, Písek PI, Jeseník JE, Beroun BE)
- 56 %: (secondary schools occur.) AND (gambling clubs occur.) AND (districts: Praha AB, Olomouc OL, Šumperk SU, Karviná KA, Kroměříž KM, Znojmo ZN, Blansko BK, Třebíč TR, Jihlava JI, Havlíčkův Brod HB, Litoměřice LT, Uherské Hradiště UH, Žďár nad Sázavou ZR).
- 47 %: (secondary schools missing) AND (district = Brno-město BM) AND (garages occur.)
- 32 %: (secondary schools missing) AND (district = Praha AB) AND (gambling devices occur.)

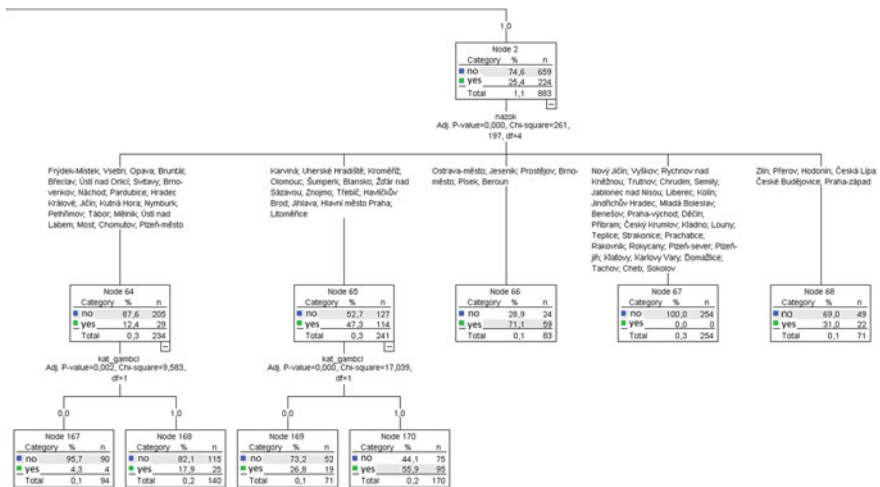


Fig. 3 Detail of decision tree for grid 1 km around the maximal probability of positive graffiti crime results (node 66)

4 Comparison of Three Different Models of Aggregation

To independently verify if the results of the selected grid model are relevant, we performed alternative ways of aggregations and build other data sets for decision trees.

4.1 Aggregation Around Randomly Selected Buildings

The random sample of 85,589 buildings from 2,765,726 was created using mod operation from the meaningless artificial identifier of buildings in RSO. The sample size is the approximation of the number of cells for 1 km grid. The uniformity of distribution in the CR was checked and evaluated as satisfactory.

The aggregation was performed for several distances; finally for this study we use aggregation within the distance 564.19 m which provides a circle with the same area as the square 1×1 km. Furthermore, the model is labelled as RB1km2. The aggregation of buildings around each selected building was quite slow and without optimisation, the calculation takes 2 weeks (PC AMD 8320 8 cores, 8 GB RAM).

4.2 Aggregation Around Randomly Generated Points

The set of randomly distributed points was generated in ArcGIS using a function “create random points”. The number of points (80,131) is the same as cells in 1 km grid, the distance for aggregation was 564.19 m to assure the same aggregation area as other two methods. Furthermore, the model is labelled as RP1km2.

4.3 Decision Trees

The variables and the settings for the decision tree method are the same as in the above analysis. The example displaying a part of the decision tree for the RB1km2 model is given in Fig. 4. The maximal probability of graffiti crime incidence is 94.8 % (91 occurrences) for data groups with a positive occurrence of secondary schools in the Prostějov (PV) district.

The resulted trees look differently showing different lists of independent variables, in a different order, different selected districts (Table 7). The RP model seems to be weaker according to the low probability for yes target category (Table 6). The random distribution of points in the country leads to slightly higher distances to the location of crimes or descriptors, but it does seem to fully explain the worse behaviour of this model.

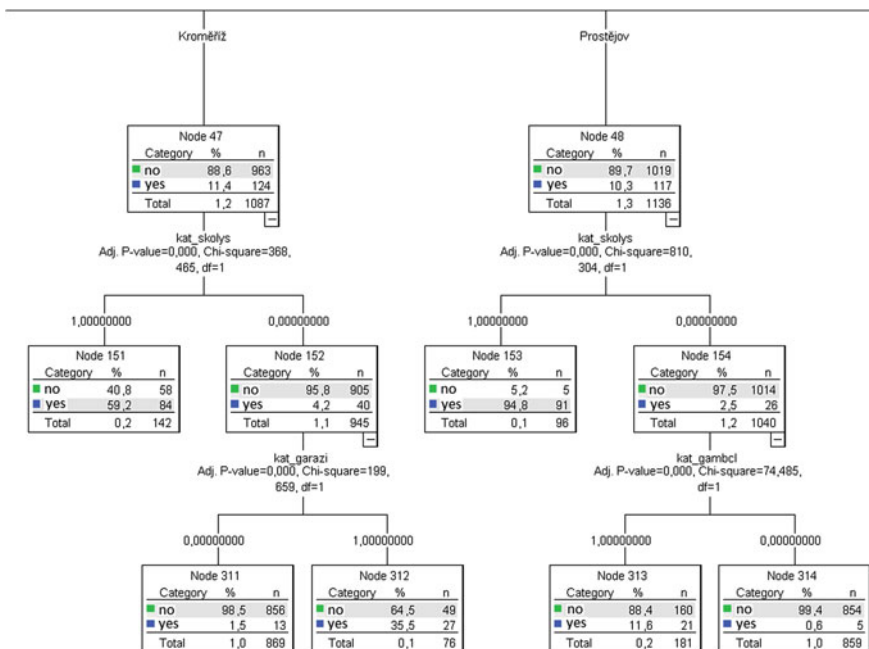


Fig. 4 Detail of decision tree for RB1km2 around the maximal probability of positive graffiti crime results (node 153)

Further, the differences between models were evaluated in more details. SQL WHERE conditions were written for all nodes in the classification trees with probability p for yes higher than 1 %. In the next step, districts were selected for those models where the positive occurrence of secondary schools (secondary = 1) is the part of SQL WHERE condition. The similar selection was made for positive occurrences of gambling clubs. The outputs follow.

4.4 Influence of Secondary Schools Occurrence

The overlay of results for grid 1km and RB1km2 for $p > 1\%$ reaches 47.8 % (22 from 46 districts). All non-overlayed districts (occurrence only in one model) are provided by the grid model. Thus, the grid model provides more indications. The overlay of results for grid1 km and RP1km2 for $p > 1\%$ is 28.6 % (22 from 77 districts). All three models coincide in 22 following districts: Praha (AB), Beroun (BE), České Budějovice (CB), Písek (PI), Litoměřice (LT), Havlíčkův Brod (HB), Jihlava (JI), Třebíč (TR), Žďár nad Sázavou (ZR), Blansko (BK), Brno-venkov

Table 7 Parameters of decision trees for prediction of graffiti crime occurrence in three models

Aggreg	List of independent variables	Percent correct prediction "yes" (%)	Max. response (prediction % for yes)	1st variable	2nd variable	3rd variable
Grid1km2	Secondary schools, district, property offences, primary schools, garages, kindergartens, gambling clubs, all flats	26.6	71.1	Secondary schools = 1	District = Pisek, Brno-město, Jeseník, Prostějov, Ostrava-město, Beroun	
RB1km2	District, gambling clubs, secondary schools, garages, primary schools, kindergartens, property offences	46.1	94.8	District = Prostějov	Secondary schools = 1	
RP1km2	Gambling clubs, secondary schools, property offences, district	9.3	56.8	Gambling clubs = 1	Secondary schools = 1	District = Přerov, Tachov, Děčín, Ústí nad Labem, Jablonec nad Nisou, Liberec, Jičín, Plzeň-jih, Brno-město, Plzeň-město, Opava, Benešov, Kladno, Mělník, Nymburk, Praha-východ, Praha-západ, Pardubice

(BO), Břeclav (BV), Hodonín (HO), Znojmo (ZN), Olomouc (OL), Prostějov (PV), Přerov (PR), Šumperk (SU), Kroměříž (KM), Uherské Hradiště (UH), Zlín (ZL) and Karviná (KA).

The overlay of results for grid 1 km and RB1km2 for $p \geq 20\%$ reaches 74 % (20 from 27 districts). The majority of partly indicated districts (occurrence only in one model) are provided by the grid model except Břeclav (BV). All three models approve 11 following districts: Beroun (BE), Písek (PI), Litoměřice (LT), Jihlava (JI), Třebíč (TR), Znojmo (ZN), Olomouc (OL), Přerov (PR), Šumperk (SU), Zlín (ZL) and Karviná (KA). Increasing p limit to 30 % does not show any change in district assignments.

The overlay of results for grid1 km and RB1km2 for $p \geq 40\%$ is slightly decreased to 70 % (16 from 23 districts). Districts České Budějovice (CB), Hodonín (HO), Přerov (PR) and Zlín (ZL) are declared only for RB model, while Brno-město (BM), Jeseník (JE) and Ostrava-město (OT) are only in the grid model. The p limit of 40 % seems to be too high for grid models—four missing districts are indicated in the grid model but with lower p . The model RP1km2 provides totally different results, the coincidence with grid models is in one district (Brno-město BM) and with the RB1km2 model also in one district (Přerov PR).

4.5 Influence of the Gambling Devices Occurrence

The overlay of results for grid1 km and RB1km2 for $p > 1\%$ is 41.9 % (18 from 43 districts). Majority of partly indicated districts (occurrence only in one model) is provided by the grid model, and only following districts are indicated only by RB1km2 model: České Budějovice (CB), Strakonice (ST), Plzeň-jih (PJ), Brno-město (BO), Hodonín (HO), Jeseník (JE), Přerov (PR), Zlín (ZL), Ostrava-město (OT). All three models coincide in 18 districts (23.4 % from all indicated districts)—Hlavní město Praha (AB), Tábor (TA), Litoměřice (LT), Most (MO), Havlíčkův Brod (HB), Jihlava (JI), Pelhřimov (PE), Třebíč (TR), Žďár nad Sázavou (ZR), Blansko (BK), Brno-venkov (BO), Břeclav (BV), Znojmo (ZN), Olomouc (OL), Šumperk (SU), Kroměříž (KM), Uherské Hradiště (UH) and Karviná (KA).

Increasing p limit to 20 % provides better coincidence with the RB model (48.1 %, 13 from 27 districts). All partly indicated districts [České Budějovice (CB), Brno-město (BM), Hodonín (HO), Přerov (PR), Zlín (ZL)] are generated by the RB model; no surplus districts are indicated by the grid model. Coincidence of three models is found only in 9 districts [14.5 %; Litoměřice (LT), Jihlava (JI), Třebíč (TR), Žďár nad Sázavou (ZR), Blansko (BK), Znojmo (ZN), Olomouc (OL), Šumperk (SU) and Karviná (KA)].

4.6 Simultaneous Combination of Both Secondary Schools and Gambling Devices Occurrences

The overlay of results for grid1 km and RB1km2 for $p > 1\%$ is 39.4% (15 from 38 districts). The majority of partly indicated districts (occurrence only in one model) is provided by the grid model, and only following districts are indicated only by RB1km2 model: České Budějovice (CB), Hodonín (HO), Přerov (PR), Zlín (ZL). All three models coincide in 15 districts (19.7% from all indicated districts)—Hlavní město Praha (AB), Litoměřice (LT), Havlíčkův Brod (HB), Jihlava (JI), Třebíč (TR), Žďár nad Sázavou (ZR), Blansko (BK), Brno-venkov (BO), Břeclav (BV), Znojmo (ZN), Olomouc (OL), Šumperk (SU), Kroměříž (KM), Uherské Hradiště (UH) and Karviná (KA).

Increasing p limit to 20% improves the overlay. The coincidence reaches 68.4% (13 from 19 districts). All partly indicated districts (occurrence only in one model) is provided by the RB model and the grid model was unable to discover them. All three models agree only seven districts (only 12.1%)—Litoměřice (LT), Jihlava (JI), Třebíč (TR), Znojmo (ZN), Olomouc (OL), Šumperk (SU) and Karviná (KA). The level of coincidence of two main factors for p above 30% is portrayed in following figures (Figs. 5, 6).

The majority of selected districts are the same in both figures. Overall, the model of aggregation around randomly selected buildings (RB1km2) provides stronger

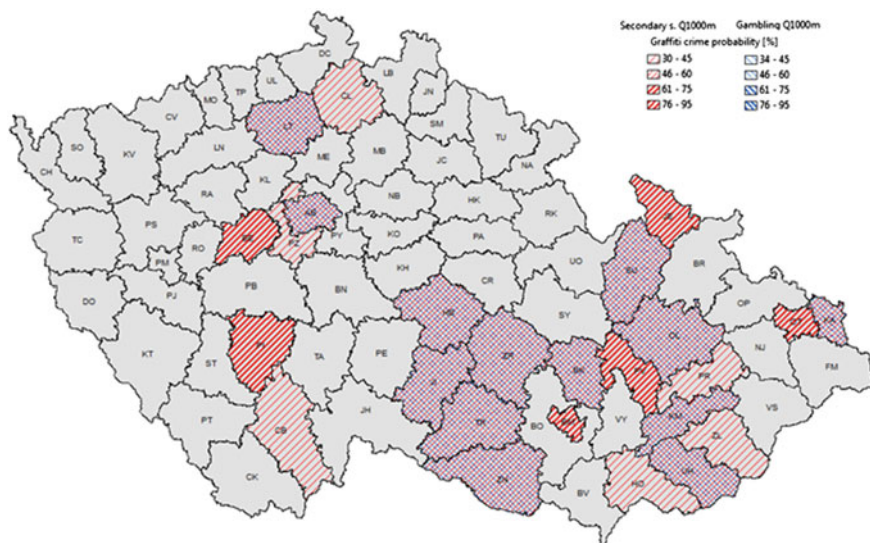


Fig. 5 Probability of the influence of secondary school occurrence and gambling device occurrence on graffiti crime occurrence by grid1km2 model

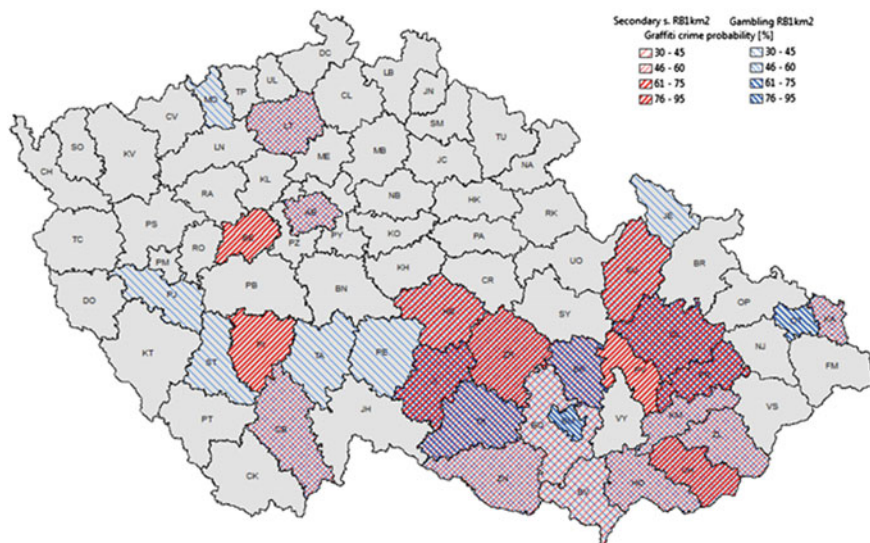


Fig. 6 Probability of the influence of secondary school occurrence and gambling device occurrence on graffiti crime occurrence by random building1km2 model

relationships than grid1km2. It is the reason why several districts (MO, PJ, ST, TA, BO, BV) are added to the selection in this model (while only PZ and CL are missing). More interesting is the fact that three districts (Jeseník JE, Ostrava-město OT, Brno-město BM) show opposite dominance of explored factors in both models.

In these districts, according to the grid1km2 model, the influence of the occurrence of secondary school is more significant than the occurrence of gambling clubs while the opposite is true according to the RB1km2 model (but the probability is higher for grid1km2). Additionally, several districts show the influence of both factors with $p > 30\%$ in the RB1km2 model while the only influence of secondary schools is significant for the grid1km2 model. The differences are related to the different method of data aggregation where the RB1km2 model tends to strengthen occurrences in an urban environment due to a higher frequency of buildings.

The selection of districts with significant influence of explored factors is determined mainly by the count of graffiti crimes in the district; obviously less frequent crimes cannot be evident as a significant relationship. The higher occurrence is reported in eastern part of our country which is reflected in the figures.

Despite the evident differences between obtained classification trees, the coincidence in indicating places with significant influence of secondary schools and/or gambling clubs occurrence is on the satisfactory level and it strengthens our trust to these models.

5 Conclusion

The grid models are easy to create using the multidimensional modelling and mainly they facilitate to combine various factors together. Other models grouping objects around point or building are extremely time-consuming and this fact is disqualifying them for operational analytical usage. The study highlighted the differences between these models and indicated the suitability of grid based regression models.

The study focused on the graffiti crimes and selected factors (property offences, buildings, flats, garages, educational facilities, and gambling clubs) which may influence the graffiti crimes occurrence. For regression analysis decision trees with the exhaustive CHAID growing method were selected.

Three different spatial resolutions of the grid were tested (100 m, 500 m, 1 km). The model of 1 km grid was evaluated as the best. The most influencing factors are occurrence of secondary schools and gambling devices, especially in following districts: Brno-město (BM), Ostrava-město (OT), Prostějov (PV), Písek (PI), Jeseník (JE), Beroun (BE), Praha (AB), Olomouc (OL), Šumperk (SU), Karviná (KA), Kroměříž (KM), Znojmo (ZN), Blansko (BK), Třebíč (TR), Jihlava (JI), Havlíčkův Brod (HB), Litoměřice (LT), Uherské Hradiště (UH), and Žďár nad Sázavou (ZR).

To verify the results of the decision tree for grid1km2, alternative aggregation models were built and analysed in a similar way. The comparison approves the satisfactory level of a coincidence especially for probability above the limit of 20 %.

The utilisation of decision trees for analysis of sparse big data should be further studied and verified using other data sets. The sensitive issues are mainly the ability to detect statistically significant relationships of investigated phenomena reported in sparse data and the stability of results using different aggregation approaches.

Acknowledgments Data is provided by the courtesy of the Czech Statistical Office, Police of the Czech Republic, Czech Ministry of Finance. The research is supported by the research of the Czech Ministry of Interior, project VF20142015034 “Geoinformatics as a tool to support integrated activities of safety and emergency units”.

References

- Act No. 40/2009, Zákon trestní zákoník (Penal Code). In: Collection of acts 9. 2. 2009. ISSN 1211-1244
- Alkhasawneh MS, Ngah UK, Tay LT et al (2014) Modeling and testing landslide hazard using decision tree. *J Appl Math*. doi:[10.1155/2014/929768](https://doi.org/10.1155/2014/929768) Article ID 929768
- Andresen MA, Malleson N (2013) Spatial heterogeneity in crime analysis. In: Leitner M (ed) *Crime modeling and mapping using geospatial technologies*. Geotechnologie and Environmnet 8, Springer, Dordrecht. doi: [10.1007/978-94-007-4997-9_1](https://doi.org/10.1007/978-94-007-4997-9_1)
- Armitage R (2002) *Tackling anti-social behaviour: what really works*. NACRO, London
- Bandaranaike S (2001) *Graffiti: a culture of aggression or assertion? The character, impact and prevention of crime in regional Australia*. Australian Institute of Criminology, Townsville

- Baclar LH (2014) EFGS and integration of geography and statistics. In: Proceedings of EFGS Krakow. 22–24 October 2014
- Badard T, Kadiillak M, Percivall G, Ramage S, Reed C, Sanderson M, Singh R, Sharma J, Vaillancourt L (eds) (2012) Geospatial business intelligence (GeoBI). OGC White paper. Ref. Number OGC 09-044r3
- Brown C, Liebovitch L (2010) Fractal analysis. SAGE. 2010, 165 s. ISBN 978-4129-7165-2
- Buck AJ, Hakim S, Swanson Ch, Rattner A (2003) Vandalism of vending machines: factors that attract professionals and amateurs. *J Crim Justice* 31(1):85–95. doi:[10.1016/S0047-2352\(02\)00201-5](https://doi.org/10.1016/S0047-2352(02)00201-5)
- Cohen S (1969) Hooligans, vandals and the community: a study of social reaction to juvenile delinquency. PhD thesis, The London School of Economics and Political Science (LSE)
- Ferrell J (1995) Urban graffiti: crime, control, and resistance. *Youth Soc* 27(1):73–92
- Fojtík D, Horák J, Orliková L, Kocich D, Inspektor T (2016) Smart geocoding of objects. In: Proceedings of ICCG 2016, Tatranská Lomnica
- Gibbons S (2004) The costs of urban property crime. *Econ J* 114(499):F441–F463. doi:[10.1111/j.1468-0297.2004.00254.x](https://doi.org/10.1111/j.1468-0297.2004.00254.x)
- Goldstein AP (1997) Controlling vandalism: The person-environment duet. In: Conoley J, Goldstein A (eds) *School violence intervention: a practical handbook*. Guilford Press, New York, pp 290–321
- Halsey M, Young A (2002) The meanings of graffiti and municipal administration. *Australian*. 35 (2):165–186. doi:[10.1375/acri.35.2.165](https://doi.org/10.1375/acri.35.2.165) ISSN 0004-8658
- Hendl J (2006) Přehled statistických metod zpracování dat. Portál, Praha. ISBN 80-7367-123-9
- Horák J, Horáková B (2007) Datové sklady a využití datové struktury typu hvězda pro prostorová data. In: Proceedings of GIS Ostrava 2007. Ostrava, 28–31 January 2007. ISSN 1213-2454
- Horák J, Ivan I, Drozdová M, Horáková B, Bala P (2016) Multidimensional database for crime prevention. In: Proceedings of ICCG 2016, Tatranská Lomnica
- Horák J, Ivan I, Horáková B (2015) OLAP for heterogeneous socio-economic data—the challenge of integration, analysis and crime prevention: a Czech case study. In: Proceedings of European forum for geography and statistics, Vienna, 10–12 November 2015
- Howard ER (1978) *School discipline desk book*. Parker, West Nyack, NY
- Iveson K (2007) *Publics and the city*, vol xii. Blackwell, Oxford. ISBN 978-140-5127-301
- Lachmann R (1988) Graffiti as career and ideology. *Am J Sociol* 94(2):229–250
- Ley D, Cybriwsky R (1974) Urban graffiti as territorial markers. *Ann Assoc Am Geogr* 64(4):491–505
- Loshin D (2012) *Business intelligence: the savvy managers guide*. MORGAN Kaufmann, Newnes, USA, p 370
- Marešová A (2011) *Resortní statistiky – základní zdroj informací o kriminalitě v České republice. Vybrané metody kriminologického výzkumu*. Institut pro kriminologii a sociální prevenci, Praha. ISBN 978-80-7338-110-3. Available at <http://www.ok.cz/iksp/docs/385.pdf>
- McCormick J (2003) “Drag me to the Asylum”: disguising and asserting identities in an urban school. *Urban Rev* 35(2):111–128
- Megler V, Banis D, Chang H (2014) Spatial analysis of graffiti in San Francisco. *Appl Geogr* 54:63–73. doi:[10.1016/j.apgeog.2014.06.031](https://doi.org/10.1016/j.apgeog.2014.06.031)
- SPSS (2007) PASW statistics 18 command syntax reference. Documentation
- Stocker TL, Dutcher LW, Hargrove SM, Cook EA (1972) Social analysis of graffiti. *J Am Folklore* 85(338):356–366. doi:[10.2307/539324](https://doi.org/10.2307/539324)
- Tygart C (1988) Public school vandalism: toward a synthesis of theories and transition to paradigm analysis. *Adolescence* 23:187–199
- Thompson K, Offler N, Hirsch L, Every D, Thomas MJ, Dawson D (2012) From broken windows to a renovated research agenda: a review of the literature on vandalism and graffiti in the rail industry. *Transp Res Part A Policy Pract* 46(8):1280–1290. doi:[10.1016/j.tra.2012.04.002](https://doi.org/10.1016/j.tra.2012.04.002). ISSN09658564
- Wilson JQ, Keeling GL (1982) Broken windows. *Atl Mon* 249(3):29–38
- Wilson P, Healy P (1987) Research brief: graffiti and vandalism on public transport. Australian Institute of Criminology, Woden, ACT. ISBN 06-421-1868-X