

[Home](#)[Table of Contents](#)[Titles & Subject Index](#)[Authors Index](#)

Crime vs. demographic factors revisited: Application of data mining methods

Xingan Li

Law School, Tallinn University, Estonia. E-mail: xingan.li@yahoo.com

Henry Joutsijoki

Computer Science, School of Information Sciences, University of Tampere, Finland.

Jorma Laurikkala

Computer Science, School of Information Sciences, University of Tampere, Finland.

Martti Juhola (Corresponding author)

Computer Science, School of Information Sciences, University of Tampere, Finland.

Tel: +358401901716; fax: +35832191001. E-mail: martti.juhola@sis.uta.fi

Received May 12, 2015; Accepted June 15, 2015

Abstract

The aim of this article is to inquire about correlations between criminal phenomena and demographic factors. This international-level comparative study used a dataset covering 56 countries and 28 attributes. The data were processed with the Self-Organizing Map (SOM), assisted other clustering methods, and several statistical methods for obtaining comparable results. The article is an exploratory application of the SOM in mapping criminal phenomena through processing of multivariate data. We found out that SOM was able to group efficiently the present data and characterize these different groups. Other machine learning methods were applied to ensure groups computed with SOM. The correlations obtained between attributes were chiefly weak.

Keywords

Data mining; Self-organizing map; K-means clustering; Discriminant analysis; K-nearest neighbor classifier; Naïve Bayes classification; Decision trees; Support vector machines (SVMs); Crime; Demographic factors

Introduction

Data mining has in recent decades been an approach to research of approximately every major discipline. Law, in the sense of a scientific field dealing with the topics related to branches of laws, is increasingly in quest of facilitation from data mining as well. Crime, as one of the most attractive research fields, requires processing of data on wide-ranging factors, including demographic, socio-economic, and historical indicators. Data mining, clustering and visualizing techniques, have broadly shown their practical value in a variety of domains, and can be considered to play an essential role in the study of crime. The self-organizing map, which employs an unsupervised learning approach to cluster and visualize data in accordance with patterns identified in a dataset, is a competent instrument meant for such data exploration. The interconnection between artificial intelligence and the study of crime makes an innovative study achievable.

In the past, in identifying abnormality of certain acts, the SOM has found its application in a broad range, for example, in the detection of automobile bodily injury insurance fraud (Brockett, Xia & Derrig, 1998), homicide (Kangas et al., 1999; Memon & Mehboob, 2006), mobile communications fraud (Hollmén, Tresp & Simula, 1999; Hollmén, 2000; Grosser, Britos & García-Martínez, 2005), murder and rape (Kangas, 2001), burglary (Adderley & Musgrave, 2003; Adderley, 2004), network intrusion (Axelsson, 2005; Lampinen, Koivisto & Honkanen, 2005; Leufven, 2006), cybercrime (Fei et al., 2005; Fei et al., 2006), and credit card fraud (Zaslavsky & Strizhak, 2006). This is the chief field where the application of the SOM has formerly been emphasized in the research associated with criminal justice.

In Li and Juhola (2013) and Li (2014), the SOM, assisted with some additional data mining techniques, was applied in the research of crime based on international databases. The research, composed of a series of papers, dealt with the relationship between crime and demographic factors (Li & Juhola, 2014a), economic factors (Li & Juhola, 2015), historical developments (Li & Juhola, 2014b), and that between a particular offence, homicide and its social context (Li et al., 2015).

This article endeavors to make inquisition into correlations between criminal phenomena and demographic factors. This international-level comparative study uses a dataset covering 56 countries and 28 attributes. The data will be processed with the Self-Organizing Map (SOM), assisted with additional clustering methods, and several statistical techniques for obtaining comparable results. The article is an exploratory application of the SOM in mapping criminal phenomena through processing of multivariate data.

Following this section, the next section of the article will briefly introduce the methods used in processing crime-related data. In the third section, information will be given about how the

experiments are designed. The fourth section will present results and discussions. The final section concludes the article with findings from the data mining of crime and its demographic factors.

Methodology

The SOM, developed by Kohonen (1979) to cluster and visualize data was used in this study. The SOM is an unsupervised learning mechanism that clusters objects with multi-dimensional attributes into a lower-dimensional space, in which the distance between every pair of objects captures the multi-attribute similarity between them. Upon processing the data, maps will be generated using software packages. By observing and comparing the clustering map and feature planes, there is the potential to explore into the correlation between crime and demographic indicators. These results, including clustering maps, feature planes as well as correlation tables constitute the fundamental ground for further analysis.

In addition to the SOM, *k*-means clustering, discriminant analysis, *k*-nearest neighbor classifier, Naïve Bayes classification, decision trees and support vector machines (SVMs) will also be used to validate the clusters and analysis by calculating how accurately these methods assemble the same countries into the same clusters as the SOM does.

Table 1. Countries included

Country	Code	Country	Code	Country	Code	Country	Code
Australia	AU	Spain	ES	Korea, Republic of	KR	Romania	RO
Azerbaijan	AZ	Finland	FI	Kazakhstan	KZ	Russian Federation	RU
Bulgaria	BG	France	FR	Lithuania	LT	Saudi Arabia	SA
Belarus	BY	United Kingdom	GB	Latvia	LV	Slovenia	SI
Canada	CA	Georgia	GE	Moldova, Republic of	MD	Slovakia	SK
Switzerland	CH	Greece	GR	Mauritius	MU	Thailand	TH
Chile	CL	Hungary	HU	Mexico	MX	Turkey	TR
Colombia	CO	Indonesia	ID	Netherlands	NL	Ukraine	UA
Costa Rica	CR	Ireland	IE	Norway	NO	United States	US
Czech Republic	CZ	India	IN	New Zealand	NZ	Uruguay	UY
Germany	DE	Iceland	IS	Papua New Guinea	PG	Yemen	YE
Denmark	DK	Italy	IT	Poland	PL	South Africa	ZA
Dominica	DM	Jamaica	JM	Portugal	PT	Zambia	ZM
Estonia	EE	Japan	JP	Qatar	QA	Zimbabwe	ZW

Design of experiments

1. Countries included

The data used in this study covers those of 56 countries, coded in Table 1. These codes will be shown in the maps as “labels”. These countries were selected based on the availability of data on their selected indicators. In general, the ratio of available data on indicators of individual countries was controlled above 80%, and mostly above 90%.

2. Demographic factors

Demographic factors have been studied since the eighteenth century (South & Messner, 2000, p. 83). Demographic factors such as age, sex, and race play an important role in understanding variation in crime rates with regard to temporal and spatial elements. In Li and Juhola (2014a), demographic factors were roughly divided into three categories: population structure, population quality, and population dynamics. Concerning population structure, three rates are selected, including population older than 64 (years old), unemployment rate, and urban population. Concerning population quality, the following factors are taken into account: adult illiteracy, health expenditure per capita, infant mortality rate, life expectancy, population growth rate, population undernourished, and under-five mortality rate. Concerning population dynamics, factors such as birth rate, death rate, fertility rate, net migration, and population density are selected. A synopsis of all variables that were used in this study is given in Table 2. Fifteen of these variables are demographic factors and the rest thirteen are crime-related indicators.

The selection of the contents of these indicators was principally based on availability of data. Another consideration was put on the traditional concept on what might in actual fact cause the occurrence of offences, because in this research pre-determined and presumed correlations were temporarily ignored. Consequently, in this research, some of these factors might traditionally be considered closely related to crime, but some others might be considered quite irrelevant. Both of these categories of factors were revisited in this research with a view to search potential clues for new enlightenment.

Table 2. The country demographic situation measured by 28 different attributes.

Demographic attributes	Name	Codification	Crime-related indicators	Name	Codification
1	Adult illiteracy %	ADUILL	16	Prisoners per Capita	PRIPER
2	Birth rate per 1000	BIR RAT	17	Share of prison capacity filled %	PRIFIL
3	Death rate per 1000	DEARAT	18	Rape per 100,000 people	RAPPER
4	Fertility rate (children born per woman)	FERRAT	19	Robbery per 100,000 people	ROBPER
5	Health expenditure per capita (USD)	HEAEXP	20	Software piracy per 100,000 people	SOFPIR
6	Infant mortality rate per 1000	INF MOR	21	Total crime per 100,000 people	TOTCRI
7	Life expectancy in years	LIFEXP	22	Police per 100,000 people	POLPER
8	Net migration per 1000	NETMIG	23	Murder per 100,000 people	MURPER
9	Population density per km ²	POP DEN	24	Jails per 100,000 people	JAIPER
10	Population growth rate %	POP GRO	25	Fraud per 100,000 people	FRAPER
11	Population older than 64 %	POP OLD	26	Convicted per 100,000 people	CONPER
12	Population undernourished %	POP UND	27	Assault per 100,000 people	ASSPER
13	Under five mortality rate per 1000	UNDFIV	28	Burglary per 100,000 people	BURPER
14	Unemployment rate total %	UNERAT			
15	Urban population %	URBPOP			

There have not been standard abbreviations in use for shortening variables. For this study, data from two different sources were combined. Information about most items was derived from the database of United Nations Development Program (UNDP). Information about some items, which was missing in UNDP database, was derived from the World in Figures of the Statistics Finland (Tilastokeskus). In case information about some items of some countries was unavailable in UNDP database, but it was available in Statistics Finland database, information about such items was supplemented by Statistics Finland data. Such items include: birth rate, death rate, net migration, marriage rate, divorce rate, and population density. Unavailable items still appeared in the last datasheet and were labeled “NaN,” (not a number) as required by SOM program. The sources of data are listed below in Table 3.

Table 3. Sources of data

Name of sources	Websites
United Nations Development Program (UNDP), Statistics of the Human Development Reports, statistical update 2008	http://hdr.undp.org/en/statistics/
The Statistics Finland (Tilastokeskus), World in Figures, updated January 22, 2009	http://www.stat.fi/tup/maanum/index_en.html

The purpose of current study was to map the contemporary crime situation of countries through clustering countries according to their crime and demographic factors, and to verify correlation between crime and demographic background. It required information to be up to date. Information about most items was from UNDP's Human Development Report 2007/2008, with information dated to the year 2005. Information about some items requires a time span. In such cases, the time span ranges from 2 to 10 years. Some items were depicted with information dated 2004, 2006, 2007, or 2008. These items were seen as most relevant data in UNDP database in the sense of time (even though other sources have quite up-to-date information).

The dataset was retrieved from different online sources primarily of 2005, but information of some items was dated 2004, 2006, 2007 or 2008. Twenty-eight variables covering demographic situation were selected on the basis of usual statistical items available on international online platforms. However, figures of sex (gender) and race were excluded because their relationship with crime requires large-scale in-depth study and much has been done by other researchers. Thus, the dataset was composed of 56 rows and 28 columns.

Although the SOM can process a dataset with missing data, in this study, the dataset avoided attributes (in columns) and countries (in rows) with five or more data values unavailable. That is to say, not all attributes with available data are included in this study. All of these countries have no more than five missing values, and most of these attributes have less than ten missing data values. Three of attributes have more missing values, for example, police per 100,000 people. By so doing, missing values have been deliberately controlled to a low rate. The total of missing values was 5.3 % as to all data values when the size of the data matrix applied to all calculations was $56 \times 28 = 1568$ elements. Besides missing values, descriptions presented in Table 4 are mean, standard deviation, minimum and maximum of each attribute.

Table 4. Descriptions of the data used

Attribute	Mean	Std. Deviation	Minimum	Maximum	Missing Values
1	6.6	10.7	0.2	45.9	0 (0%)
2	14.9	7.2	8.3	40.3	0 (0%)
3	9.3	3.9	1.9	21.4	0 (0%)
4	2.0	1.0	1.2	6	1 (1.8%)
5	1381	1313	63	6096	0 (0%)
6	18.6	23.3	2	102	0 (0%)
7	72.8	8.9	40.5	82.3	0 (0%)
8	3.1	25.4	-47	157.9	0 (0%)
9	108.8	123.8	2.8	614	0 (0%)
10	1.0	1.1	-0.4	5.1	2 (3.6%)
11	11.5	5.2	1.3	19.7	1 (1.8%)
12	6.8	9.9	2.5	47	2 (3.6%)
13	24.3	35.0	3	182	0 (0%)
14	10.0	12.5	0.6	80	0 (0%)
15	65.3	18.0	13.4	95.4	0 (0%)
16	116.7	34.8	62.8	245.9	8 (14.3%)
17	194.8	152.1	29	715	11 (19.6%)
18	0.14	0.22	0	1.2	0 (0%)
19	1.1	2.0	0	12.3	0 (0%)
20	52.2	21.0	20	92	6 (10.7%)
21	35.2	32.4	1.2	113.8	4 (7.14%)
22	2.7	1.4	0.4	7.28	14 (25%)
23	0.06	0.11	0	0.62	1 (1.8%)
24	0.08	0.38	0	2.08	8 (14.3%)
25	1.2	1.8	0	10.9	2 (3.6%)
26	6.7	6.7	0.2	33.2	11 (19.6%)
27	2.3	2.8	0.03	12.1	3 (5.4%)
28	5.6	5.9	0	21.8	9 (16.1%)

3. Construction of the map

In this study, the software package used is Viscovery SOMine 6. Compared with some other software packages of the SOM, Viscovery SOMine has almost the same requirements on the format of the dataset. At the same time, requiring less programming, it enables an easier and more operable data processing and visualization.

Missing values were marked with “NaN”. The SOMine software automatically generated maps from the dataset of 56 countries and 28 attributes. The clustering map (Fig. 1) as well as some other detailed statistics, such as correlations as discussed below, can be used in further analysis.

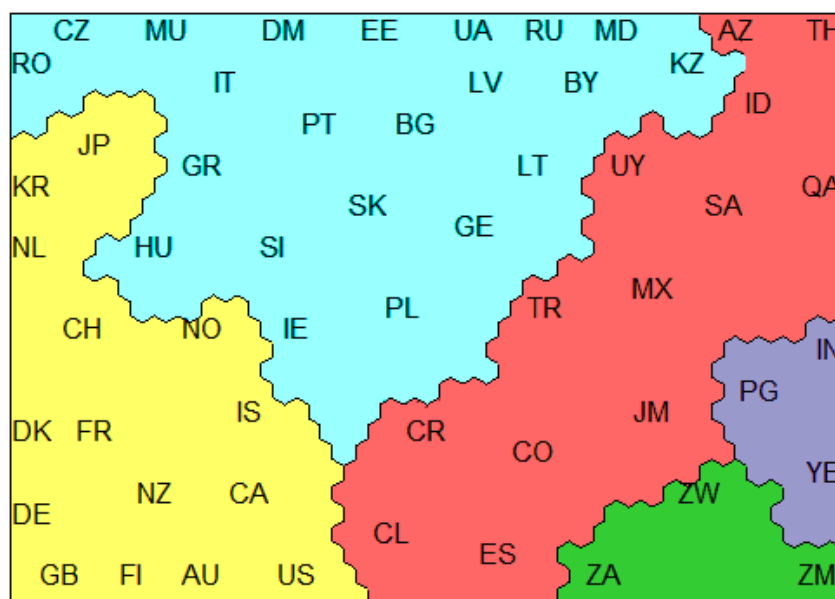


Figure 1. Clustering map with five clusters and labels of countries

4. Multi-class support vector machines and parameter estimation

Support Vector Machine (SVM) (Vapnik, 2000; Burges, 1998; Cortes & Vapnik, 1995) is a well-known supervised classification method originally developed for binary classification problems. The main idea in SVM is to construct a classes separating hyperplane with maximum margin. If classes are non-separable, we need to use kernel functions and so called soft margin SVM (Vapnik, 2000; Burges, 1998; Cortes & Vapnik, 1995). Kernel functions map the data from input space to a higher dimensional feature space (possibly infinite dimensional) by using a nonlinear transformation. In the feature space classes separating hyperplane can again be found. Technical details related to the derivation of SVM classifier in linearly separable, nonlinearly separable and soft margin cases can be found, for example, from (Vapnik, 2000; Burges, 1998; Cortes & Vapnik, 1995). There are several ways to solve the hyperplane optimization problem related to SVM. Firstly, the original approach is to use Quadratic Programming (Vapnik, 2000) for finding

the optimal hyperplane. Secondly, a Sequential Minimal Optimization (SMO) algorithm for solving QP problems was introduced in Platt (1998). Thirdly, there is Least-Squares SVM (Suykens & Vandewalle, 1999) which is a reformulation of Vapnik's SVM.

Since SVMs were mentioned only for binary classification problem, various multi-class extensions (cases where the number of classes is greater than 2) have been presented. We applied four of them in this paper. These were one-vs.-one (OVO), one-vs.-all (OVA), binary complete (BIN) and ordinal (ORD) multi-class extension. Firstly, in OVO method (Galar et al., 2011) a classifier for each class pair is constructed and, hence, the total number of classifiers is $M(M-1)/2$ where $M > 2$ is the number of classes. Secondly, in OVA (Galar et al., 2011) one classifier is trained to separate one class from the rest and, hence, the total number of classifiers to be trained is M . Thirdly, in BIN a classifier for all binary combinations without ignoring any class are constructed and by this means the total number of classifiers is $2^{M-1}-1$ (Mathworks, 2015). Fourthly, in ORD $M-1$ binary classifiers are constructed such that the first binary classifier separates the first class from the other whereas the second one separates the first two classes from the rest etc. (Mathworks, 2015). All the aforementioned multi-class extensions can be presented using error-correcting output codes (ECOC) (Allwein et al., 2000; Dietterich and Bakiri, 1995; Escalera et al., 2010). We can model multiclass classification problems by using ECOC which is a general classification framework and is not specifically designed for SVM (Escalera et al., 2010). The use of ECOC returns to designing a coding matrix (CO) which can be presented in a binary form or ternary form depending on the multi-class extension (Escalera et al., 2010). In coding matrix rows represent codewords for classes and columns represent outputs for binary classifiers. In a binary form elements of coding matrix are -1 or 1 whereas in ternary form coding matrix elements are -1, 0 or 1 where 0 means that the given binary classifier do not consider specific class (Escalera et al., 2010). For instance, OVO method is coded using ternary approach. When test example is classified, the outputs of the binary classifiers are combined so that we obtain a codeword for the test example. Then, test example will be classified to a class having the closest codeword (Escalera et al., 2010; Mathworks, 2015). More details about ECOC can be found, for example, from (Allwein et al., 2000; Dietterich & Bakiri, 1995; Escalera et al., 2010; Mathworks, 2015). In this paper we used ECOC for modeling OVA, OVO, BIN and ORD methods.

The use of SVM requires estimation of parameters. In this research we tested three kernels: linear, second and third degree polynomial kernels (see details Hsu et al., 2013). As a common parameter for SVM is C (also known as boxconstraint), and the applied kernels do not require any other parameters to be estimated than C . Parameter estimation was performed using nested leave-one-out method (NLOO). In NLOO dataset is first divided using leave-one-out method into training and test sets. Then separately for each training set we search optimal parameter value by applying leave-one-out method to the training set. When the optimal parameter value is found for

the training set, SVM classifiers were trained again using full training set and test example was classified. The consequence of using NLOO is that one optimal parameter value might not be found for the whole dataset but optimal parameter value might vary between training sets. For the simplicity, however, we decided that each one of the binary SVM classifiers in multi-class extension was trained with the same parameter value. Altogether, each kernel was tested with 31 parameter values ($C \in \{2^{-15}, 2^{-14}, \dots, 2^{15}\}$). Moreover, we used SMO algorithm in hyperplane optimization. All the tests were made using Matlab 2014b together with Statistics Toolbox and Parallel Computing Toolbox.

Results

Upon processing of data, five clusters have been generated, each representing groups of countries sharing similar characteristics. As a default practice in self-organizing maps, values are expressed in colors: warm colors denote high values, while cold colors denote low values.

1. Clusters

Clusters were given in Fig. 1. Due to the feature of the software package, countries were not completely shown in the map. In order to give a full picture of these clusters, the following lists all the countries in each cluster:

Cluster C1 consists of 22 countries: CZ, MU, DM, EE, UA, RU, MD, RO, IT, LV, BY, KZ, PT, BG, GR, LT, SK, GE, HU, SI, PL, IE

Cluster C2 consists of 13 countries: AZ, TH, ID, UY, SA, QA, MX, TR, JM, CR, CO, CL, ES

Cluster C3 consists of 15 countries: JP, KR, NL, CH, NO, DK, FR, IS, NZ, CA, DE, GB, FI, AU, US

Cluster C4 consists of 3 countries: ZW, ZA, ZM

Cluster C5 consists of 3 countries: PG, IN, YE

Although the Viscovery SOMine software package provides the possibility for adjusting the number of clusters, usually automatically generated clusters represented the results that might occur the most naturally. In other experiments the same number of clusters could be set deliberately, countries in these clusters were still re-grouped slightly one-way or the other. In this experiment, a more significant change of a cluster number was still tolerated, because this was expected to leave a new space where the similar issue could be speculated.

2. Validation of clusters

Total 56 countries times 28 attributes with original 5.3% missing values imputed with clusterwise medians, with new clusters (classes) given by the SOM clustering method. After imputation, the results by the SOM were compared to those given by several methods, including k -means clustering, discriminant analysis, k -nearest neighbor classifier, Naïve Bayes classification, decision trees and support vector machines (SVMs). For these other, mostly methods of supervised learning, cluster labels found by the SOM were used as class labels in training and finally in tests to check whether the SOM and classification results of the others agreed or not. Tests were executed according to the leave-one-out technique. In addition, correlations were calculated. The classifiers were implemented and the statistical tests executed with Matlab.

Table 5. Accuracy rates [%] when the imputed attribute values were original, scaled to interval [0,1] or standardized (for decision trees, parameter minparent means the minimum size of node possibly to be divided into child nodes)

	Not scaled	Scaled	Standardized
Unsupervised, 30 iterations			
<i>k</i> -means			
<i>k</i> =5	55.6±0.6	85.2±5.9	86.3±6.0
<i>k</i> =6	56.5±1.2	84.6±4.4	86.4±7.7
<i>k</i> =7	57.4±1.3	85.8±5.2	89.6±4.3
<i>k</i> =8	59.5±2.3	85.1±4.5	90.7±2.5
<i>k</i> =9	60.6±2.1	86.6±5.6	89.9±4.7
<i>k</i> =10	61.6±2.1	86.6±5.6	89.9±4.7
Supervised, 30 iterations			
<i>k</i> -means			
<i>k</i> =5	49.2±2.1	74.0±2.7	76.0±3.4
<i>k</i> =6	47.0±2.7	73.9±2.5	81.5±2.8
<i>k</i> =7	44.3±4.0	74.6±4.6	82.9±3.5
<i>k</i> =8	44.9±2.6	75.5±3.7	84.3±2.7
<i>k</i> =9	44.9±4.3	75.3±4.0	84.3±3.0
<i>k</i> =10	46.8±3.6	77.6±3.2	84.6±3.2
Discriminant analysis			
Linear	83.9	83.9	83.9
Logistic	73.2	73.2	73.2
<i>k</i> -nearest neighbour searching			
<i>k</i> =1	58.9	73.2	73.2
<i>k</i> =2	50	89.3	89.3
<i>k</i> =3	51.8	89.3	89.3
Naïve Bayes with 'kernel' for 'dist'	80.4	80.4	80.4
Decision trees			
minparent=6	57.1	57.1	58.9
minparent=3	66.1	66.1	66.1
minparent=2	67.9	67.9	67.9

ECOC-OVA-SVM			
Linear	69.6	87.5	94.6
Polynomial degree 2	26.8	87.5	87.5
Polynomial degree 3	30.4	85.7	76.8
ECOC-OVO-SVM			
Linear	71.4	83.9	85.7
Polynomial degree 2	28.6	80.4	78.6
Polynomial degree 3	58.9	80.4	80.4
ECOC-ORD-SVM			
Linear	76.8	80.4	78.6
Polynomial degree 2	26.8	89.3	89.3
Polynomial degree 3	25.0	91.1	76.8
ECOC-BIN-SVM			
Linear	76.8	89.3	92.9
Polynomial degree 2	25.0	85.7	89.3
Polynomial degree 3	25.0	85.7	76.8

Unsupervised k -means clustering gave accuracy rates between 55.6%-61.6% when data were not scaled. Scaling of data made the results significantly better, between 84.6%-86.6%. When data were standardized (attribute by attribute, by subtracting with the mean and dividing with the standard deviation of each attribute), overall results still bettered off, between 86.3%-90.7%. Supervised k -means behaved worse than unsupervised, accuracy rates are generally 10-15% lower, though in a few cases the differences are smaller.

Different methods of discriminant analysis were tested. Linear analysis got the same rate of 83.9% regardless of the data being scaled, unscaled, or standardized that is natural for discriminant analysis. Quadratic and Mahalanobis analysis got no positive definite covariance matrix, while logistic analysis got 73.2% when data were standardized. Furthermore, k -nearest neighbor searching classifier gave results between 50%-58.9% when data were not scaled. When data were either scaled or standardized, results are always equivalent: 73.2% ($k=1$), 89.3% (in cases when both $k=2$ and $k=3$). Larger k values were not reasonable to run since there were two small clusters of three countries only. Naïve Bayes with parameter value 'normal' did not function for the current data. Naïve Bayes with 'kernel' for 'dist' got results of 80.4%, decision trees got results of 57.1%, 66.1% and 67.9% when minparent= 6, 3, and 2 separately, regardless of data being scaled, not scaled or standardized typical to this method using probabilities.

SVMs obtained the lowest accuracies when data were not scaled or standardized, but the highest rate when data was standardized, and medium rates when data were scaled (as detailed in Table 5). A noticeable detail in accuracies was that the highest accuracy was gained by ECOC-OVA-SVM with the linear kernel with standardized data. This might imply that clusters are well separable in the original input space. When the best agreement of SOM with the other methods in Tables 5 exceeded 90% subject to accuracy values this denoted high consistency.

3. Correlations

Viscovery SOMine could generate a detailed list of correlations, based on which Table 6 was created. These correlations were computed from the original, not yet imputed dataset. Although even strong correlation between two attributes does not necessarily indicate causation, this will bring about materials for further analysis and reference. There are many opportunities that these results can be used to compare with previous studies on crime using other methods. Traditionally, single research on crime did not include so many attributes (or named correlation factors or causes). Even in textbooks, only a dozen or two were introduced. So it shall be highly expected to have such data mining methods to be able to process several dozens of attributes and to provide immediate reference for further analysis. We obtained 10 out of 195 ($p < 0.05$) to be significant after the p values were corrected with the Holm's method. Six of them involved Attribute 20, Software piracy per 100,000 people.

Table 6. Correlations between (not imputed) demographic attributes (A1-A15) and crimes (A16-A28) where those marked with symbol * show the statistically significant

	16	17	18	19	20	21	22	23	24	25	26	27	28
1	0.39	-0.12	0.07	-0.10	0.42	-0.30	-0.17	0.16	0.34	-0.28	-0.28	-0.04	-0.35
2	0.51*	0.01	0.10	-0.06	0.44	-0.23	-0.29	0.19	0.00	-0.26	-0.21	0.04	-0.23
3	0.25	0.40	0.19	0.02	0.16	-0.03	0.07	0.21	-0.08	-0.02	-0.07	0.18	0.02
4	0.47	-0.09	0.10	-0.09	0.37	-0.22	-0.42	0.14	0.16	-0.26	-0.19	0.04	-0.28
5	-0.28	-0.10	0.18	0.02	-0.80*	0.57*	-0.10	-0.29	-0.08	0.44	0.38	0.40	0.37
6	0.30	0.15	0.09	-0.12	0.61*	-0.39	-0.26	0.19	0.07	-0.31	-0.34	-0.03	-0.32
7	-0.39	-0.47	-0.18	0.09	-0.62*	0.36	0.16	-0.35	-0.11	0.30	0.32	-0.06	0.27
8	-0.13	0.00	0.08	0.02	-0.11	0.05	-0.14	-0.15	-0.01	0.11	0.07	0.08	0.09
9	0.09	-0.26	-0.17	-0.11	-0.14	0.01	0.25	-0.15	-0.05	0.17	-0.08	0.03	-0.18
10	0.40	-0.12	0.13	-0.01	0.32	-0.26	-0.43	0.11	0.04	-0.28	-0.19	0.03	-0.34
11	-0.34	-0.10	-0.15	0.06	-0.48	0.35	0.28	-0.30	-0.11	0.37	0.27	-0.05	0.39
12	0.55*	-0.03	-0.01	-0.14	0.56*	-0.32	-0.21	0.09	-0.07	-0.25	-0.24	-0.02	-0.29
13	0.38	0.14	0.09	-0.12	0.57*	-0.35	-0.27	0.15	0.07	-0.28	-0.31	0.00	-0.29
14	0.35	0.07	0.21	0.01	0.38	-0.05	-0.05	0.15	-0.08	-0.14	-0.08	0.20	-0.06
15	-0.32	0.02	0.10	0.19	-0.58*	0.51*	0.01	-0.10	-0.24	0.32	0.35	0.19	0.37

From Table 6, a few correlation values were interesting, while others were very weak. Certainly, it still needs extensive exploration to conclude how demographic factors interconnect with criminal phenomena, affecting their occurrence, or their increase or decrease.

The correlations in Table 7 were calculated from the imputed dataset. For this dataset, we obtained 11 correlations ($p < 0.05$) to be significant after the p values were corrected with the

Holm's method. Compared to Table 6 there are only few differences: two correlations for Attribute 16 are no longer significant, but for Attributes 17, 20 and 24 one correlation became significant.

Table 7. Correlations between (imputed) demographic attributes (A1-A15) and crimes (A16-A28) where those with symbol * show the statistically significant

	16	17	18	19	20	21	22	23	24	25	26	27	28
1	0.34	-0.13	0.07	-0.10	0.44	-0.33	-0.24	0.15	0.63*	-0.29	-0.34	-0.04	-0.33
2	0.42	0.07	0.10	-0.06	0.44	-0.26	-0.34	0.19	0.28	-0.27	-0.28	0.05	-0.24
3	0.25	0.48*	0.19	0.02	0.12	-0.05	0.13	0.21	-0.10	-0.03	-0.06	0.16	0.07
4	0.36	-0.02	0.09	-0.09	0.40	-0.26	-0.46	0.14	0.42	-0.25	-0.25	0.06	-0.29
5	-0.27	-0.11	0.18	0.02	-0.79*	0.60*	-0.10	-0.29	-0.19	0.45	0.37	0.41	0.40
6	0.29	0.18	0.09	-0.12	0.60*	-0.41	-0.29	0.18	0.32	-0.32	-0.37	-0.03	-0.33
7	-0.39	-0.41	-0.18	0.09	-0.63*	0.39	0.15	-0.35	-0.24	0.31	0.34	-0.04	0.27
8	-0.13	-0.01	0.08	0.02	-0.12	0.11	-0.11	-0.15	-0.03	0.12	0.13	0.09	0.08
9	0.04	-0.27	-0.17	-0.11	-0.15	-0.02	0.25	-0.15	0.05	0.15	-0.08	0.05	-0.15
10	0.29	-0.07	0.12	0.00	0.28	-0.27	-0.41	0.10	0.24	-0.26	-0.24	0.06	-0.37
11	-0.31	-0.08	-0.14	0.06	-0.49*	0.37	0.33	-0.29	-0.28	0.37	0.32	-0.06	0.43
12	0.43	0.09	0.01	-0.15	0.58*	-0.34	-0.30	0.10	0.39	-0.27	-0.28	-0.04	-0.28
13	0.36	0.20	0.09	-0.12	0.57*	-0.37	-0.31	0.15	0.28	-0.29	-0.34	0.00	-0.30
14	0.33	0.23	0.21	0.01	0.35	-0.07	-0.09	0.15	0.05	-0.14	-0.13	0.19	-0.03
15	-0.29	0.00	0.10	0.19	-0.59*	0.53*	0.03	-0.10	-0.43	0.33	0.38	0.19	0.38

Conclusions

This paper dealt with macroscopic data for international comparison. Conventionally, analysis in the study of crime, either on general issues or on particular issues, did not handle large-scale of multidimensional data. Specifically, when international comparison was carried out, discussion was much abstract and theoretical, lack of systematic data processing. With the self-organizing map, multidimensional comparison was realized. The research objects, countries, could be grouped into different clusters with more convergent features.

By using *k*-means clustering, discriminant analysis, *k*-nearest neighbor classifier, Naïve Bayes classification, decision trees and support vector machines (SVMs) to verify the SOM results, findings of the study gave additional proof that the self-organizing map was an interesting tool for assisting research on individual types of crime. The clustering results were easily visualized and convenient to interpret, facilitating practical comparison between countries with diversified socio-economic and criminal features. The article provided broad potential for applying data analysis and visualization methods in the field of the study of crime, where in turn would find significant methodological value of this application.

References

- Adderley, R. (2004). The use of data mining techniques in operational crime fighting. H. Chen et al. (EDs.), *Intelligence and Security Informatics (ISI) 2004, Lecture Notes in Computer Science*, 3073, pp. 418-425.
- Adderley, R., & Musgrave, P. (2003). Modus operandi modelling of group offending: a data-mining case study. *International Journal of Police Science and Management*, 5(4), 265-276.
- Allwein, E.L., Schapire, R.E., & Singer, Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1, 113-141. Retrieved February 10, 2015, from <http://www.jmlr.org/papers/volume1/allwein00a/allwein00a.pdf>
- Axelsson, S. (2005). *Understanding intrusion detection through visualization*, PhD thesis, Chalmers University of Technology, Göteborg, Sweden.
- Brockett, P. L., Xia, X., & Derrig, R. A. (1998). Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. *Journal of Risk and Insurance*, 65(2), 245-274.
- Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 121-167.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
- Dietterich, T.G., & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2, 263-286.
- Escalera, S., Pujol, O., & Radeva, P. (2010). On the decoding process in ternary error-correcting output codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1), 120-134.
- Fei, B., Eloff, J., Olivier, M., & Venter, H. (2006). The use of self-organizing maps for anomalous behavior detection in a digital investigation. *Forensic Science International*, 162(1-3), 33-37.
- Fei, B., Eloff, J., Venter, H., & Olivier, M. (2005). Exploring data generated by computer forensic tools with self-organising maps. In *Proceedings of the IFIP Working Group 11.9 on Digital Forensics*, 1-15. Retrieved February 10, 2015, from http://www.seralliance.com/enews/vol2no1/pdfs/computer_forensic_tools_jan05.pdf
- Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8), 1761-1776.
- Grosser, H., Britos, P., García-Martínez, R. (2005). Detecting fraud in mobile telephony using neural networks. In: M. Ali, and F. Esposito (Eds.). *Lecture Notes in Artificial Intelligence*, Springer-Verlag, Berlin, Germany, 3533, 613-615.
- Hollmén, J. (2000). *User profiling and classification for fraud detection in mobile communications networks*. PhD thesis, Helsinki University of Technology, Finland.
- Hollmén, J., Tresp, V., & Simula, O. (1999). A self-organizing map for clustering probabilistic models. *Artificial Neural Networks*, 470, 946-951.
- Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2013). *A practical guide to support vector classification, Technical report*. Retrieved May 5, 2015, from <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>

- Kangas, L. J. (2001). *Artificial neural network system for classification of offenders in murder and rape cases*, The National Institute of Justice, Finland.
- Kangas, L. J., Terrones, K. M., Keppel, R. D., & La Moria R. D. (1999). Computer-aided tracking and characterization of homicides and sexual assaults (CATCH). In: *Proceedings SPIE 3722, Applications and Science of Computational Intelligence II*.
- Kohonen, T. (1979). *Self-Organizing Maps*. Springer-Verlag, New York, USA.
- Lampinen, T., Koivisto, H., & Honkanen, T. (2005). Profiling network applications with fuzzy C-means and self-organizing maps. *Classification and Clustering for Knowledge Discovery*, 4, 15-27.
- Leufven, C. (2006). *Detecting SSH identity theft in HPC cluster environments using self-organizing maps*, Master thesis, Linköping University, Sweden.
- Li, X., & Juhola, M. (2013). Crime and its social context: analysis using the self-organizing map. In: *Intelligence and Security Informatics Conference (EISIC)*, 2013 European, 12-14 August 2013, Uppsala, Sweden, IEEE, pp. 121-124.
- Li, X. (2014). *Application of Data Mining Methods in the Study of Crime Based on International Data Sources*, PhD thesis, University of Tampere, Tampere, Finland.
- Li, X., & Juhola, M. (2014a). Country crime analysis using the self-organizing map, with special regard to demographic factors. *Artificial Intelligence and Society*, 29(1), 53 - 68.
- Li, X., & Juhola, M. (2014b). Application of the self-organising map to visualisation of and exploration into historical development of criminal phenomena of the USA, 1960-2007. *International Journal of Society Systems Science*, 6(2), 120 - 142.
- Li, X., & Juhola, M. (2015). Country crime analysis using the self-organising map, with special regard to economic factors. *International Journal of Data Mining, Modelling and Management*, 7(2), 130-153.
- Li, X., Joutsijoki, H., Laurikkala, J., Siermala, M., & Juhola, M., 2015. Homicide and its social context: Analysis using the self-organizing map. *Applied Artificial Intelligence: An International Journal*, 29(4), 382-401.
- Mathworks Documentation Center (2015). *MathWorks*. Retrieved May 11, 2015, from <http://se.mathworks.com/help/>
- Memon, Q. A., & Mehboob, S. (2006). Crime investigation and analysis using neural nets. In: *Proceedings of International Joint Conference on Neural Networks*, Washington, D.C., pp. 346-350.
- Platt, J.C. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines*. Microsoft Research Technical Report MSR-TR-98-14.
- South, S. J., & Messner, S. F. (2000). Crime and demography: multi linkages, reciprocal relations. *Annual Review of Sociology*, 26, 83-106.
- Suykens, J.A.K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3), 293-300.
- Vapnik, V. N. (2000). *The nature of statistical learning theory*. 2nd edition. New York: Springer-Verlag.

Viscovery Software GmbH (2015). *Viscovery SOMine*. Retrieved May 10, 2015, from <http://www.viscovery.net/somine/>.

Zaslavsky, V., & Strizhak, A. (2006). Credit card fraud detection using self-organizing maps. *Information and Security: An International Journal*, 18, 48-63.

Bibliographic information of this paper for citing:

Li, Xingan, Joutsijoki, Henry, Laurikkala, Jorma, & Juhola, Martti (2015). "Crime vs. demographic factors revisited: Application of data mining methods." *Webology*, 12(1), Article 132. Available at: <http://www.webology.org/2015/v12n1/a132.pdf>

Copyright © 2015, Xingan Li, Henry Joutsijoki, Jorma Laurikkala, and Martti Juhola.