This is the accepted manuscript of the article, which has been published in 2017 IEEE Symposium Series on Computational Intelligence (SSCI). https://doi.org/10.1109/SSCI.2017.8285260

Development of crime in England and Wales 1898-2001: Data mining using self-organising map

Xingan Li
School of Governance, Law and Society
Tallinn University
Narva MNT 29
10120 Tallinn, Estonia
Email: xingan.li@yahoo.com

Henry Joutsijoki, Jorma Laurikkala, Martti Juhola
Faculty of Natural Sciences
University of Tampere
Kanslerinrinne 1, FI-33014,
Tampere, Finland
Email: {henry.joutsijoki,jorma.laurikkala,martti.juhola}@uta.fi

***Abstract*—The aim of this article is to inquire about historical development of criminal phenomena in England and Wales, and relationship between different crime rates, based on a set of English and Welsh historical data. This national-level study uses a dataset covering 103 years (1898-2001, with data of 1939 missing and not counted) and 50 attributes. The collected data are clustered with Self-Organizing Map (SOM) and the features are assessed using Scatter algorithm. Several machine learning methods are applied to verify the clustering result obtained by the SOM. Accuracy of 96.2% gained by one-vs-one least-squares support vector machines shows that the clusters obtained by the SOM are valid. The article is an exploratory application of the SOM in research of criminal phenomena through processing of multivariate data. The research showed that SOM was able to cluster efficiently the present data and to characterize these different clusters.**

I. Introduction

In recent decades data mining has been an approach to research many major disciplines. 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 [1] and [2], the Self-Organising Map (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 [3], [4], economic factors [5], historical developments [6], and that between a particular offence, homicide and its social context [7]. The suitability and the evidence of the performance of SOM in aforementioned studies convinced us to choose SOM as a main machine learning technique for this paper.

The time frame studied in this paper roughly covers the 20th century, during which some considerable stages could be marked in the world, particularly in the United Kingdom (UK): the First World War of 1914-18, the Great Depression of 1929-32, the end of the Second World War in 1945, Austerity of 1945-1950, beginning of the modernization of the UK in the 1950s and 1960s, transformation from welfare state to affluent society in the end of the 1940s to 1960s, as well as, however, loss of its position as a superpower, and final decolonization by the 1970s [8], [9], [10], [11], [12]. The global economy witnessed another decline in the 1970s, when the UK also suffered. The exploitation of coal, which triggered the industrial revolution two centuries ago, now gave way to gas and oil that was exploited from the North Sea, which created the financial basis for the new economic boom. A severe recession took place between 1990 and 1992, but the latter part of the 1990s witnessed a starting of a phase of continuous economic growth that lasted over one and half decade [13].

When we talk about the UK, the European Union (EU) must be an inevitable topic potentially affecting every aspect of social lives, including crime. The British application to join the Common Market in 1961 and application to join the European Community in 1967 were both vetoed by French President Charles de Gaulle, whose absence led to the installation of the British membership of the Community in 1973, but with great division of public opinion [8], [14], [15] that triggered a referendum on 5 June 1975. The proposition to continue membership was passed with a substantial majority [15]. The Single European Act (SEA), the first major revision of the 1957 Treaty of Rome, was enacted into UK law in 1987. In 1992, the UK ratified the Maastricht Treaty, which transformed the European Community into the European Union.

The crime rate in Britain was in a process of lowering when the 20th century started. The worsening of economic situation during the Great Depression of the 1930s brought only a slight increase in crime [16]. During the first two decades of the 20th century, recorded crime in England and Wales (the numbers of Scotland and North Ireland are not included in the statistics)

was on an average of 90,000 indictable offences each year, which increased to over 500,000 during the 1950s, with crime rate quadrupled from 250 crimes per 100,000 people in 1901 to 1,000 by 1950 [17].

Since the late 1950s, both economy and crime of England and Wales were dominated by a sharp rise, a tendency even accelerated in the 1960s when crime doubled. Crime continued to rise for much of the latter half of the 20th century, with an average of over one million crimes recorded each year in the 1960s, two million during the 1970s, and 3.5 million in the 1980s [17]. The number of recorded crimes continued to increase, until it reached the 2003 peak of 6 million [17]. Similar to that in the United States, a steady decrease started after the crime climax of the early 2000s [6].

While the total number of recorded crimes as well as crime rates were changing the same way, the different types of crimes could increase or decrease differently. One thing to mention is that, over the period, there were significant changes to the types of offences recorded as crime, and how they are counted [17]. Another thing to mention is that, while most kinds of recorded crime, with particularly steep falls in some offences such as burglary, new types of crime at different stages, such as those involving cars, motorcycles, and computers were increasing [17].

This article endeavors to inquire about historical development of criminal phenomena in the England and Wales, and relationship between different crime rates, based on a set of English and Welsh historical data. An exploratory multivariate analysis of national-level crime rate data covering the years 1898-1938 and 1940-2001 was performed with SOM, refined by Scatter algorithm [18]. The year 1939 was excluded due to unavailability of data. The clustering result was assessed using several machine learning methods from baseline algorithms to state-of-the-art methods. In other words, our aim is to show the feasibility of SOM in the context of clustering historical data of England and Wales and to verify the clustering result using machine learning methods. The results show that SOM is able to form suitable clusters which are well separable in terms of classification and reflect well the historical facts.

The rest of the paper is organized as follows. Section II outlines briefly the SOM algorithm. In Section III we present an overview of the dataset and features. Moreover, we describe preprocessing of the data and SOM clustering, classification procedure and the parameter settings related to the classification methods tested. Section IV is for the classification results and Section V focuses on the discussion about crime in the English and Welsh history as seen in the SOM. Section VI concludes the paper and presents directions for future work.

II. Self-Organising Map

Self-Organising Map [19], [20], [21] is an unsupervised artificial neural network algorithm introduced by Teuvo Kohonen. SOM differs from many other commonly used artificial neural networks by its structure. SOM does not have hidden layers such as multi-layer perceptrons [21] or radial basis function networks [21] have. Instead the main idea is to map continuous input space usually into 1D or 2D lattice in a topologically ordered way [21]. However, lattice can be higher dimensional than 1D or 2D. Lattice is constructed from a set of nodes and each one of them includes a weight vector. In SOM there is only the input layer and the whole lattice can be considered as an output layer. SOM algorithm has four main steps [21]:

1) Initialization of weight vectors.
2) Finding the winning neuron for the input vector.
3) Updating the weight vectors.
4) Repeating steps 2-3 until the number of iterations is met or map do not change.

Assume that we have weight vectors $\mathbf{w}_j(t)$ where the index $j = 1, 2, \ldots, M$ points out to a node in a lattice, M is the total number of nodes and t means the iteration round. Moreover, in the initialization step we have $t = 0$. Initialization of weight vectors is an important step since it effects on the convergence of map to a stabilized stage. Weight vectors can be initialized with random values [21]. However, it must be noticed that choosing "purely" random values for weight vectors may lead to a situation where the learning process is non-convergent or very slow. Hence, random values would be good to choose so that they represent quite closely the distribution of the input space.

The second step is actually the first step of the iterative learning process of SOM. We match the input vector $\mathbf{x} \in \mathbb{R}^n$ with all weight vectors and seek a node where the distance between weight vector and input vector is the smallest. In other words, we seek a winning neuron $r(\mathbf{x})$ where

$$r(\mathbf{x}) = \arg\min_j \|\mathbf{x}(t) - \mathbf{w}_j(t)\|,\ j = 1, 2, \ldots, M$$

at the iteration round t [21].

After finding the winning neuron $r(\mathbf{x})$, we update the weight vectors for all neurons in a lattice as follows [21]:

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \eta(t) h_{j,r(\mathbf{x})}(t)[\mathbf{x}(t) - \mathbf{w}_j(t)]$$

where $0 < \eta(t) < 1$ is a monotonically decreasing learning rate parameter and $h_{j,r(\mathbf{x})}(t)$ is a dynamically changing neighborhood function around the winning neuron $r(\mathbf{x})$ [21]. Neighborhood function $h_{j,r(\mathbf{x})}(t)$ is again defined with the following way [21]:

$$h_{j,r(\mathbf{x})}(t) = \exp\left(-\frac{d_{j,r}^2}{2\sigma^2(t)}\right), t = 1, 2, \ldots,$$

where $d_{j,r}^2 = \|\mathbf{s}_j - \mathbf{s}_r\|^2$ is the Euclidean distance between the jth neuron in a lattice and the winning neuron r. In addition, for the $\sigma(t)$ we have

$$\sigma(t) = \sigma_0 \exp\left(-\frac{t}{\tau_1}\right)$$

where σ_0 user-defined parameter value and τ_1 is a time constant [21]. Learning rate parameter $\eta(t)$ is again defined as follows [21]:

$$\eta(t) = \eta_0 \exp\left(-\frac{t}{\tau_2}\right)$$

where τ_2 is another time constant.

III. Design of experiments

A. Dataset and features

The purpose of the current study was to inquire about the historical development of criminal phenomena in the United Kingdom, and the relationship between different crime rates, based on a set of English and Welsh historical data. The data used in this study covers a period of 103 years from year 1898 till 2001. However, the data of 1939 is missing. The years were selected based on the availability of data on their selected indicators.

A synopsis of all attributes that were used in this study is given in Table I where each one of the attributes is provided with an unique code "Xnumber". There are no commonly accepted abbreviations for the attributes so they are presented in complete form. The word "total" is used in the definition of an attribute when there are also sub-categories of the same attribute. This form of presentation is ordinarily used in statistics in England and Wales. All the 50 attributes concern crime rates with various perspectives. The selection of the contents of these indicators was principally based on availability of data. Overall, the dataset was composed of 105 rows and 50 columns. The difference why dataset has 105 rows and only 102 years are included can be explained by the fact that how the statistical methods have changed during the years [1]. Information about the most of the attributes was derived from the database of UK Home Office.[2]

B. Preprocessing and generating clusters

The first phase was to generate clusters from the collected data using SOM. For this purpose we used a software called Viscovery SOMine 6 [22] which applies SOM-Ward algorithm[3] [23], [24] for generating clusters. Overall, 50 nodes were used in SOM and training parameter "Tension" was set to 0.5. SOM-Ward algorithm is not the only method which hierarchically clusters the SOM map and for more detailed discussion on this topic a reader can look at the article by Vesanto and Alhoniemi [25]. Missing values in a dataset were marked as "NaN". The SOMine software can automatically handle the missing values and generate a map from the given data. SOM was selected for this study due to its good performance in publications [1], [2], [3], [4], [5], [6], [7].

The default value was to divide the map into seven clusters. However, we also tested the numbers of clusters from five to ten which are explained in detail as follows.

- Case 1 (seven clusters): With seven clusters, there formed a big group of years including the period from late 1890s to 1930s, with the exception of some separate years spreading from 1900s to 1920s. Other clusters usually cover consecutive years.
- Case 2 (eight clusters): If the number of clusters was adjusted to 8 clusters, years from the end of the 1920s and 1930s were separated from late 1890s and 1900s, 1910s and 1920s. However, the exceptional years of 1909, 1911, 1915, 1918, 1919, and 1924 were still in an independent cluster. All the other clusters compared to seven clusters situation did not change.
- Case 3 (nine clusters): In the case of nine clusters, the original clusters covering the 1950s and 1960s would have been divided into almost two clusters composed of the years of the 1950s and 1960s, with 1949 in the cluster of mainly the 1950s, while 1959 in the cluster of mainly the 1960s. Significantly, 1969 is in another cluster of mainly the 1970s.
- Case 4 (ten clusters): For ten clusters case, the original cluster composed of the 1990s and the early 2000s would be divided into a cluster of the early 1990s and another one of the latter part of the 1990s and the early 2000s.
- Case 5 (six clusters): If six clusters were selected, the cluster of 1970s and that of the 1980s would be combined, with 1969 and 1990, which were already in these two original clusters.
- Case 6 (five clusters): If the number of clusters had been decreased to five, further combination would have been occurred: the cluster of the 1940s would be combined into the big cluster originally composed of the late 1890s, the 1900s, 1910s, 1920s and 1930s, with exception of 1909, 1911, 1915, 1918, 1919 and 1924 originally already in an independent cluster.

Cluster evaluation and the selection of the number of the clusters can be performed with various ways (for example, using intrinsic clustering measures such as silhouette or elbow methods). However, our approach is research interest based and established on historical and societal facts. Hence, in our study, it is preferable to have a separate cluster for the 1930s due to the fact that there was a significant event of crisis in economy and, thus, in the society in the Western world, including England and Wales. Thus, eight clusters in which the years of the 1930s were separated from those years before the 1930s, can better reflect our concern. Figure 1 presents the result where the instances have been grouped into eight clusters. Clusters are also presented in Table II.

After finding the clusters, the second step in preprocessing was to process the possible missing values. The database where the data were collected did not provide all values for all attributes. Thus, missing values had to be imputed some way before machine learning methods could be applied to verify the obtained clusters. We used attribute means to replace the missing values. Overall, the missing values were occurred in attributes X6, X7, X10, X15, X19, X36, X42, X43, X44 X45 and X47 when following the notation of Table I. The respective number of missing values together with the proportion in percentage for the aforementioned

[1]https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/116649/rec-crime-1898-2002.xls

[2]A summary of recorded crime data from 1898 to 2001/02, at https://www.gov.uk/government/statistics/historical-crime-data)

[3]https://www.viscovery.net/download/public/The-SOM-Ward-cluster-algorithm.pdf

Table I
CRIMINAL PHENOMENA INDICATED BY 50 DIFFERENT ATTRIBUTES.

Name	Name	Name	Name	Name	Name
X1: Homicide (includes murder, manslaughter and infanticide)	X2: Attempted murder	X3: Threat or conspiracy to murder	X4: More serious wounding or other acts endangering life	X5: Endangering railway passangers	X6: Other wounding etc.
X7: Assaults	X8: Abandoning a child under age of two years	X9: Child abduction	X10: Procuring illegal abortion	X11: Total violence against a person	X12: Buggery
X13: Indecent assault on a male	X14: Gross indecency between males	X15: Rape	X16: Indecent assault on a female	X17: Unlawful sexual intercourse with a girl under 13	X18: Unlawful sexual intercourse with a girl under 16
X19: Incest	X20:Procuration	X21: Abduction	X22: Total sexual offences	X23: Total robbery	X24: Total violent crime
X25: Burglary in a dwelling	X26: Burglary in a building other than dwelling	X27: Total Burglary	X28: Theft from a person	X29: Theft by an employee	X30: Theft or unauthorized taking from mail
X31: Other theft and unauthorized taking	X32: Handling stolen goods	X33: Total theft and handling stolen goods	X34: Frauds by company directors	X35: False accounting	X36: Other frauds
X37: Total fraud and forgery	X38: Arson	X39: Going equipped for stealing etc.	X40: Concealment of birth	X41: Bigamy	X42: Blackmail
X43: Riot	X44: Violent disorder	X45: Other offences against the State or public order	X46: Perjury	X47: Libel	X48: Other indictable or triable either way offences
X49: Total other offences	X50: Total recorded crime				

Table II
CLUSTERS GIVEN IN FIG. 1 PRESENTED IN A LIST FORM.

Cluster 1={1898, 1899, 1900, 1901, 1902, 1903, 1904, 1905, 1906, 1907, 1908, 1910, 1912, 1913, 1914, 1916, 1917, 1920, 1921, 1922, 1923}
Cluster 2={1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968}
Cluster 3={1991, 1992, 1993, 1994, 1995, 1996, 1997, 1997/8, 1998/9(or), 1998/9(nr), 1999/2000, 2000/01, 2001/02}
Cluster 4={1925, 1926, 1927, 1928, 1929, 1930, 1931, 1932, 1933, 1934, 1935, 1936, 1937, 1938, 1940}
Cluster 5={1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990}
Cluster 6={1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977}
Cluster 7={1909, 1911, 1915, 1918, 1919, 1924}
Cluster 8={1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948}

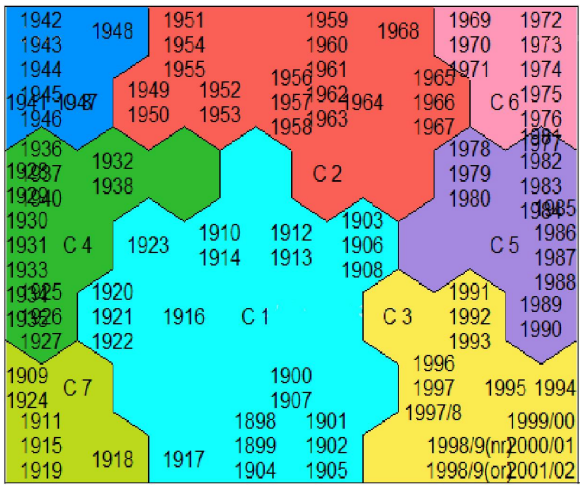

Figure 1. Eight clusters given by the SOM-Ward algorithm.

features were: 4(3.81%), 15(14.29%), 2(1.90%), 1(0.95%), 11(10.48%), 4(3.81%), 2(1.90%), 18(17.14%), 28(26.67%), 4(3.81%) and 3(2.86).

The third step in preprocessing was to investigate the relevance of features. In other words, we searched for the possible redundant features which could be excluded from the dataset. There are several algorithms for feature importance evaluation available in the literature and we selected to use Scatter algorithm in this study. Scatter algorithm [4] is presented in detail in [18] so we do not review the algorithm here. One of the outputs of Scatter algorithm is separation power which measures the separability ("goodness") of an attribute compared to the rest of the attributes.

Since Scatter algorithm is a supervised algorithm, we needed the imputed and labeled version of the dataset. A commonly encountered process in data mining and machine

[4]Free-to-use implementation of this algorithm called "ScatterCounter" can be found from http://www.uta.fi/sis/cis/researcg_groups/darg/publications.html

learning when the dataset is not labeled, is to first cluster the dataset and then label the clusters obtained. We followed this process in our study. Hence, class label for an instance is determined based on the cluster to which it belongs. According to the separation power values given by the Scatter algorithm no attribute should be removed and the further processing of the data will be performed with all 50 attributes.

C. Classification and parameter settings

Our study is dichotomous from the methodological perspective. Firstly, with the help of SOM we examine the structure of collected data in order to find the natural clusters. Secondly, we verify, by means of machine learning methods, the results of SOM clustering. For the latter objective we selected a large collection of classification methods to be used in this study. More specifically, we tested the following methods: Linear Discriminant Analysis (LDA) [26], Classification and Regression Tree (CART) [27], Naïve Bayes (NB) [28] with and without kernel density estimation (KDE) [29], Multinomial Logistic Regression (MNLR) [30], [31], k Nearest Neighbor method (KNN) [27], [32], Random Forests [33] and Least-Squares Support Vector Machines (LS-SVMs) [34], [35].

Some of the methods tested require parameter value testing. For CART there are two main parameters called "minparent" and "minleaf". Minleaf expresses the minimum number of leaf node instances whereas minparent describes the minimum number of branch node instances. For the minleaf parameter we used value of 1 and for minparent the values of 1 and 5 were tested. Both minparent values yielded the same results. Moreover, the best split attribute at each branch node was performed using interaction test [36].

Naïve Bayes classifier was tested with and without KDE. When KDE was applied, we selected triangle kernel [29] to be used in this study. In the case of KNN we performed wide experiments. Firstly, we tested six distance measures (Chebychev, Manhattan, correlation, cosine, Euclidean and Spearman). Secondly, each one of the distance measures was further tested with the odd k values of $k \in \{1, 3, \ldots, 15\}$. For Random Forests classifier the essential issue is the number of trees in a forest since Random Forests classification is an ensemble learning method. We decided to test the number of trees from 1 to 30.

Finally, we applied LS-SVMs in classification. Because originally LS-SVM is for binary classification tasks and our classification problem is a multi-class problem, we used one-vs-one (OVO) [37] classification scheme together with majority voting method. In majority voting scheme each individual binary classifier gives a predicted class label for a test instance and the class which obtains the highest number of votes will be assigned for the final class label for a test instance. However, majority voting does not prevent the occurrence of possible ties. In this study we solved the ties with the following procedure:

1) Find out classes which occur in a tie.
2) Extract the corresponding training data from those classes.
3) Divide the interval $[0, 1]$ into smaller non-overlapping segments with respect to the class sizes encountered in a tie.
4) Generate a random number from uniform distribution $U(0, 1)$.
5) Decide the final class label based on the segment to which the random number belongs.

In order to illustrate the procedure we present a simple example of it.

- Let there be a tie between two classes C_1 and C_2.
- Let the class sizes of C_1 and C_2 be 40 and 60 in a training set.
- Now we divide the $[0, 1]$ interval such a way that $[0, 0.4]$ belongs to class C_1 and interval $(0.4, 1]$ is for class C_2.
- If a random number s derived from a uniform distribution $U(0, 1)$ belongs to segment $[0, 0.4]$, final class label for the test instance will be C_1 and, otherwise, class label will be C_2.

The success of LS-SVMs is highly dependent on the parameter values selected. In our research we tested, altogether, five kernels (the linear, the quadratic, the cubic, the RBF and the Sigmoid). For the polynomial kernels we tested $C \in \{2^{-12}, 2^{-11}, \ldots, 2^{17}\}$ where the C is the regularization parameter common for all kernels. For the RBF kernel a hyperparameter, σ, is needed and for it we selected the same parameter value space as for C. In other words, we have $\sigma \in \{2^{-12}, 2^{-11}, \ldots, 2^{17}\}$. Finally, for the Sigmoid kernel we have two hyperparameters, κ and δ for which $\kappa \in \{2^{-12}, 2^{-11}, \ldots, 2^{17}\}$ and $\delta \in \{-2^{17}, -2^{16}, \ldots, -2^{-12}\}$. Overall, polynomial kernels were tested with 30 values, the RBF kernel with 30^2 (C, σ) combinations and the Sigmoid kernel with 30^3 (C, κ, δ) triplets.

Verification of clustering (classification of clusters) was performed using the leave-one-out method with all classification methods tested. Before performing the leave-one-out procedure data used in classification was scaled to $[-1, -1]$ using min-max scaling (also known as min-max normalization) [38]. If a method required parameter value search, we performed leave-one-out procedure with all parameter value(s) (combinations) and the optimal parameter value (combination) was selected based on the highest accuracy. Accuracy is here defined as the sum of diagonal elements of a confusion matrix divided by the sum of all elements of a confusion matrix. Besides accuracy, sensitivities (also known as true positive rate) are reported for all clusters (represent the classes in our study) in result tables.

IV. Classification results

Table III presents the classification results when LDA, CART, NB and MNLR classifiers were applied to the imputed dataset. A clear exception in Table III results is NB when KDE was used. It obtained only 55.2% accuracy and the sensitivities are very low except with cluster 2 where 85.0% sensitivity was achieved. CART algorithm with different parameter settings yielded around 86% accuracy which is a good result. Furthermore, clusterwise sensitivities were mainly above 80.0%

Table III
RESULTS OF THE LDA, CART, NB AND MNLR CLASSIFIERS. SENSITIVITIES OF EACH CLUSTER AND ACCURACIES HAVE BEEN PRESENTED.

	LDA	CART	NB Normal	NB KDE (triangle)	MNLR
Cluster 1	95.2%	76.2%	66.7%	90.5%	85.7%
Cluster 2	100.0%	95.0%	100.0%	85.0%	95.0%
Cluster 3	100.0%	92.3%	100.0%	30.8%	100.0%
Cluster 4	93.3%	86.7%	93.3%	53.3%	100.0%
Cluster 5	92.3%	92.3%	92.3%	53.8%	92.3%
Cluster 6	100.0%	100.0%	100.0%	22.2%	88.9%
Cluster 7	66.7%	50.0%	66.7%	0.0%	50.0%
Cluster 8	100.0%	87.5%	100.0%	12.5%	87.5%
Accuracy	**95.2%**	86.7%	89.5%	55.2%	90.5%

except with clusters 1 and 7. However, cluster 7 had only 6 instances and in classification tasks small classes may easily be lost. NB with Gaussian distribution assumption and MNLR both obtained around 90% accuracy. The difference between the results of these two methods was that NB gained 100.0% sensitivity in four clusters whereas MNLR had the similar results only with two clusters. The best result in Table III gave LDA having 95.2% accuracy which is a very good result. All clusters except the smallest one were well classified with LDA.

Table IV shows the KNN results with various distance measures. All distance measures achieved good accuracies and the highest ones were obtained by Chebychev, Manhattan, correlation and Euclidean measures. Furthermore, the optimal k values were small in each case. Only values of 3, 5 and 7 occurred within the optimal parameter values. With the aforementioned measures 95.2% accuracy (the same as with LDA) was achieved. Clusters 3 and 5 were perfectly classified with Chebychev, Manhattan, correlation, and Euclidean measures. With cosine measure the results were also good and the difference with respect to accuracy was only around one percentage. A slightly larger decrease in results was gained with Spearman measure since 90.5% accuracy was achieved with it. However, this result is still a good one although it was the worst within Table IV results. A more detailed inspection reveals that the highest dispersion in sensitivities between different distance measures was in the case of cluster 7. The sensitivities varied between 0.0% and 50.0% interval. However, it must be noticed that this cluster is a small one having only six instances.

In Table V the results of Random Forest classifier and OVO-LS-SVMs are given. Overall, a good level in the results continued compared to Table IV results. Random Forest gained the lowest accuracy (89.5%) among the classification methods. Multi-class LS-SVM instead obtained 94.3%, 95.2% or 96.2% accuracies. Accuracy of 96.2% was the best one within the result tables. When considering the sensitivities more closely, it can be noticed that clusters 3, 4, 6 and 8 were the best classified clusters. Overall, it can be said that SOM-Ward algorithm is capable of finding good clusters for the current dataset which are well separable from the classification perspective.

For the results in Tables III-V the results of the Friedman test showed that there were significant ($p < 0.05$) differences in the median sensitivities of the classification methods. Only the best results from the nearest neighbor and least-squares support vector classification were included in the testing, because the results of these methods were very similar. Therefore, the tested methods were LDA, CART, both of the Naïve Bayes methods, MNLR, KNN (Euclidean distance and $k = 5$), Random Forest and LS-SVM RBF. As expected, the post hoc comparisons, with the significance level 0.05 adjusted with the Dunn-Bonferroni method, showed significant differences only between the worst method NB KDE and the other methods ({NB Normal, KNN, LDA, LS-SVM RBF }> NB KDE).

V. DISCUSSION: CRIME IN THE BRITISH HISTORY AS SEEN IN THE SOM

As we can observe from the figure of the SOM, the years between 1898 and early 1920s were clearly grouped in the same cluster (Cluster 1), with the exception of some years of the late 1910s and the early 1920s separated from this cluster and formed Cluster 7: 1909, 1911, 1915, 1918, 1919, and 1924. The late 1920s, 1930s and starting of the 1940s were in Cluster 4. The period in this cluster ranges from the end of the First World War, through the Great Depression of the 1930s, to the start of the Second World War. The most years in the 1940s formed Cluster 8, which covers the period of the Second World War and its aftermath.

Cluster 2 swathes a long span of time, from the end of the 1940s, the 1950s, to the 1960s. This was the longest period of prosperity for the UK after the Second World War. The crime rate of the England and Wales underwent a process of gradual increase, slight decrease and then fast increase. During this period, Britain remained a European leader in both economic power and political influence, but a sudden reduction started [8], [39]. Cluster 6 contains the end of the 1960s and most of the 1970s, among which are the economic crisis of the 1970s. This is also the time when the UK was accepted into the Economic Community. The end of the 1970s and the 1980s fall into Cluster 5, which spread another period of economic growth, but was shortly interrupted by two to three years in the beginning of the 1990s. Cluster 3 covers the years across the 1990s and the early 2000s, when the UK also enjoyed a continuous economic growth. Criminal phenomenon also had a steep increase and then surged to a historical peak.

On the surface, these clusters are well mosaiced into different stages of socio-historical development of England and Wales. Because of the high accuracy value in this study, it can

Table IV
RESULTS OF THE k-NN CLASSIFIER WITH DIFFERENT DISTANCE MEASURES. SENSITIVITIES RELATED TO CLUSTERS AND ACCURACIES ARE PRESENTED. WITHIN THE PARENTHESIS THE OPTIMAL k VALUE IS GIVEN.

	k-NN Chebychev ($k = 7$)	k-NN Manhattan ($k = 3$)	k-NN correlation ($k = 5$)	k-NN cosine ($k = 5$)	k-NN Euclidean ($k = 5$)	k-NN Spearman ($k = 3$)
Cluster 1	95.2%	95.2%	100.0%	95.2%	100.0%	90.5%
Cluster 2	100.0%	95.0%	100.0%	100.0%	100.0%	100.0%
Cluster 3	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Cluster 4	100.0%	100.0%	93.3%	100.0%	100.0%	86.7%
Cluster 5	92.3%	100.0%	92.3%	92.3%	92.3%	100.0%
Cluster 6	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Cluster 7	50.0%	50.0%	50.0%	50.0%	50.0%	0.0%
Cluster 8	100.0%	100.0%	100.0%	87.5%	87.5%	100.0%
Accuracy	**95.2**%	**95.2**%	**95.2%**	94.3%	**95.2%**	90.5%

Table V
RESULTS OF RANDOM FORESTS AND ONE-VS-ONE LEAST-SQUARES SUPPORT VECTOR MACHINES CLASSIFIERS. SENSITIVITIES RELATED TO CLUSTERS AND ACCURACIES ARE PRESENTED. WITHIN THE PARENTHESIS THE OPTIMAL PARAMETER VALUES ARE GIVEN.

	Random Forest ($\#trees = 2$)	LS-SVM Linear ($C = 2^{-11}$)	LS-SVM Quadratic ($C = 2^{-8}$)	LS-SVM Cubic ($C = 2^{-11}$)	LS-SVM RBF ($C = 2^3$,$\sigma = 2^0$)	LS-SVM Sigmoid ($C = 2^{-1}$, $\kappa = 2^{-5}$, $\delta = -2^2$)
Cluster 1	95.2%	90.5%	90.5%	81.0%	90.5%	90.5%
Cluster 2	100.0%	95.0%	95.0%	100.0%	100.0%	100.0%
Cluster 3	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Cluster 4	93.3%	100.0%	100.0%	100.0%	100.0%	100.0%
Cluster 5	92.3%	100.0%	92.3%	100.0%	100.0%	100.0%
Cluster 6	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Cluster 7	0.0%	66.7%	66.7%	66.7%	66.7%	66.7%
Cluster 8	75.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Accuracy	89.5%	95.2%	94.3%	94.3%	**96.2%**	**96.2%**

be confirmed from technical point of view that such a result of clustering of stages of development of criminal phenomena copes with socio-economic history as a whole. However, due to the limited scale of this study, indicators of socio-economic development were not included in the data, leaving a great potential for future work.

Some of the iconic historical events could occur out of a sudden, such as the eruption of wars, occurrence of decline, introduction of new inventions and new products, etc. Its prelude could already be reflected in preceding years, while its postlude reflected in the following years as well. Situation of criminal phenomena would not generally change abruptly. However, over a long run, criminal phenomena would follow the tendency of social change. That also is reflected in the clusters generated above, that most clusters across a longer span of time than one year or several years when a historical event simply took place.

In recent years that were not covered by the data in this paper, there have been yet more significant events occurred in the UK, for example, the global economic crisis of 2008 also led to economic contraction of the UK, ending 16 years of continuous economic growth. A referendum on the UK's exiting the EU on 23 June 2016 put an end to the country's membership. All these affect short or long term development of the British society, ultimately on criminal phenomena as well. These subsequent happenings provide an opportunity for future research on the similar topic.

VI. CONCLUSIONS

In traditional way, because of technical limit, criminological research did not handle large-scale multidimensional data. This paper used national statistical data to study historical development of criminal phenomena in England and Wales. The self-organizing map was applied to facilitate multidimensional comparison, with the research objects, years, being classified into different clusters with convergent features.

To verify the results, a large number of methods were used, such as discriminant analysis, k-nearest neighbor classifier, naïve Bayes classification, decision trees, random forests and least-squares support vector machines. Findings of the study proved to be ideal to use the self-organizing map to support research on patterns of historical development of crime. Convenient visualization and easy interpretation facilitated practical division of stages. Using large scale data, the SOM requires for the well-framed data sets. Therefore, it is necessary to exploit high quality statistics, and thus acquisition and preparation for them might take noteworthy efforts. Another limit was that consistent and continuous official historical statistics, over decades, centuries or millennia did not exist. Traditionally, such a situation was remedied by application of qualitative methods. A conflict of ideas between qualitative and quantitative methods could occur when large data sets were dealt with. This posed the necessity for more future research. From the methodological point of view we will examine other clustering methods such as K-means [40], K-means++ [41] or

hierarchical clustering [42] together with different clustering evaluation measures in future for comparison. Hence, we will have even broader insight how other clustering methods will perform from clustering the historical data of England and Wales. Furthermore, we obtain wider perspective about the possible trends within clusters. By this means we will advance the research around the young research area of computational history.

ACKNOWLEDGMENT

The second author is thankful for the Finnish Cultural Foundation Pirkanmaa Regional Fund for the support.

REFERENCES

[1] X. Li and M. Juhola, "Crime and its social context: Analysis using self-organising map," *Proceedings of the European Conference on Intelligence and Security Informatics*, pp. 121-124, 2013.

[2] X. Li, Applications of data mining methods in the study of crime based on international data sources, PhD Thesis, University of Tampere, Finland, 2014.

[3] X. Li and M. Juhola, "Country crime analysis using the self-organizing map, with special regard to demographic factors," *AI & Society*, Vol. 29, No. 1, pp. 53-68, 2014.

[4] X. Li, H. Joutsijoki, J. Laurikkala, M. Siermala, and M. Juhola, "Crime vs. demographic factors revisited: Applications of data mining methods," *Webology*, Vol. 12, No. 1, Article 132, 2015.

[5] X. Li and M. Juhola, "Country crime analysis using the self-organising map, with special regard to economic factors," *International Journal of Data Mining, Modelling and Management*, Vol. 7, No. 2, pp. 130-153, 2015.

[6] X. Li and M. Juhola, "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*, Vol. 6, No. 2, pp. 120-142, 2014.

[7] X. Li, H. Joutsijoki, J. Laurikkala, M. Siermala and M. Juhola, "Homicide and its social context: Analysis using self-organising map," *Applied Artificial Intelligence*, Vol. 29, No. 4, pp. 382-401, 2015.

[8] D. McDowall, *An Illustrated History of Britain*, Essex, England, Pearson Education Limited, 2006.

[9] R. Wolfson and J. Laver, *Years of Change: Europe 1890-1945*, London, England, Hodder & Stoughton, 1996.

[10] V.G. Kiernan, *European Empires from Conquest to Collapse, 1815-1960*, Fontana, Douglas, Isle of Man, 1982.

[11] R. Williams, *Culture and Society 1780-1950*, Middlesex, England, Penguin Books, 1961.

[12] D. Thomson, *England in the Twentieth Century*, Middlesex, England, Penguin Books, 1978.

[13] A. Marr, *A History of 20th Century Britain*, Oxford, England, Pan Macmillan, 2011.

[14] J. Steiner and L. Woods, *Textbook on EC Law*, London, England, Blackstone Press, 1996.

[15] J. Pinder and S. Usherwood, *The European Union*, Oxford, UK, Oxford University Press, 2013.

[16] The National Archives, Crime in the 20th Century, 2016. Accessed 15 October 2016, at http://www.nationalarchives.gov.uk/education/candp/crime/g10/default.htm

[17] Home Office, *Historical recorded crime series*, in: G. Thompson, O. Hawkins, A. Dar, M. Taylor (Eds.), Olympic Britain - Social and Economic Change since the 1908 and 1948 London Games, House Commons, pp. 153-154, 2012.

[18] M. Juhola and M. Siermala, "A scatter method for data and variable importance evaluation," *Integrated Computer-Aided Engineering*, Vol 19, No. 2, pp. 137-149, 2012.

[19] T. Kohonen, *Self-Organizing Maps*, New York, USA, Springer-Verlag, 1979.

[20] T. Kohonen, "The self-organising map," *Proceedings of the IEEE*, Vol. 78, No. 9, pp. 1464-1480, 1990.

[21] S. Haykin, *Neural Networks: A Comprehensive Foundation*, second ed., New Jersey, Prentice-Hall, 1999.

[22] Viscovery Software GmbH, Viscovery SOMine (2015), https://www.viscovery.net/somine/.

[23] N. Yorek, I. Ugulu, and H. Aydin, "Using self-organizing neural network map combined with Ward's clustering algorithm for visualization of students' cognitive structural models about aliveness concept," *Computational Intelligence and Neuroscience*, Vol. 2016, No. 2015, Article ID 2476256, 14 pages.

[24] Z. Yao, T. Eklund, and B. Back, "Using SOM-Ward clustering and predictive analytics for conducting customer segmentation," *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, pp. 639-646, 2010.

[25] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, Vol. 11, No. 3, pp. 586-600, 2000.

[26] K.J. Cios, W. Pedrycz, R.W. Swiniarski, and L.A. Kurgan, *Data Mining: A Knowledge Discovery Approach*, New York, Springer-Verlag, 2007.

[27] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, second ed., New York, John Wiley & Sons, 2001.

[28] D.D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," *Proceedings of the European Conference on Machine Learning*, pp. 4-15, 1998.

[29] B.A. Turlach, "Bandwidth selection in kernel density estimation: A review," Working Paper, 1994.

[30] C. Kwak and A. Clayton-Matthews, "Multinomial logistic regression," *Nursing Research*, Vol. 51, No. 6, pp. 404-410, 2002.

[31] Y. Wang, "A multinomial logistic regression modeling approach for anomaly intrusion detection," *Computers & Security*, Vol. 24, No. 8, pp. 662-674, 2005.

[32] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, Vol. 14, No. 1, pp. 1-37, 2008.

[33] L. Breiman, "Random forests," *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001.

[34] J.A.K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, Vol. 9, No. 3, pp. 293-300, 1999.

[35] T. Van Gestel, J.A.K. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. De Moor, and J. Vandewalle, "Benchmarking least squares support vector machine classifiers," *Machine Learning*, Vol. 54, No. 1, pp. 5-32, 2004.

[36] W.Y. Loh, "Regression trees with unbiased variable selection and interaction detection," *Statistica Sinica*, Vol. 12, No. 2, pp. 361-386, 2002.

[37] M. Galar and A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "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*, Vol. 44, No. 8, pp. 1761-1776, 2011.

[38] S. Garcia, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*, Cham, Springer-Verlag, 2015.

[39] M. Sorokina, Great Britain and the European Integration, Master's thesis, Masaryk University, Czech Republic 2014.

[40] A.K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, Vol. 31, No. 8, pp. 651-666, 2010. Pattern Recognition Letters

[41] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," *Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027-1035, 2007.

[42] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, Springer-Verlag, 2013.