Detecting Danger in Roads: An Immune-Inspired Technique to Identify Heavy Goods Vehicles Incident Hot Spots

Grazziela P. Figueredo[1,2], Isaac Triguero[1], Mohammad Mesgarpour[3], Alexandre M.Guerra[1], Jonathan M. Garibaldi[1,2], Robert I. John[1]
1. The School of Computer Science, The University of Nottingham, NG8 1BB, UK
2. The Advanced Data Analysis Centre, The University of Nottingham, NG8 1BB, UK
3. Microlise, Farrington Way, Eastwood, Nottingham NG16 3AG, UK

***Abstract*—We report on the adaptation of an immune-inspired instance selection technique to solve a real-world big data problem of determining vehicle incident hot spots. The technique, which is inspired by the Immune System self-regulation mechanism, was originally conceptualised to eliminate very similar instances in data classification tasks. We adapt the method to detect hot spots from a telematics data set containing hundreds of thousands of data points indicating incident locations involving heavy goods vehicles (HGVs) across the United Kingdom. The objective is to provide HGV drivers with information regarding areas of high likelihood of incidents in order to continuously improve road safety and vehicle economy. The problem presents several challenges and constraints. An accurate view of the hot spots produced in a timely manner is necessary. In addition, the solution is required to be adaptable and dynamic, as thousands of new incidents are included in the database daily. Furthermore, the impact on hot spots after informing drivers about their existence has to be considered. Our solution successfully addresses these constraints. It is fast, robust, and applicable to all different incidents investigated. The method is also self-adjustable, which means that if the hot spots configuration changes with time, the algorithm automatically evolves to present the most current topology. Our solution has been implemented by a telematics company to improve HGV safety in the United Kingdom.**

Keywords*—*Hot Spots, Road incidents, Instance selection, Telematics, Big Data, Artificial Immune Systems

I. Introduction

Despite govern, industrial and societal efforts to improve road safety indicators, traffic incidents still reach unacceptable levels across the globe. The Road Safety Foundation reported that the total cost of road accident in 2014 in the UK was estimated to be 14.7 billion, with 67 people being killed or seriously injured on the roads every day [1]. In particular, a high frequency of heavy goods vehicles (HGVs) incidents is observed, with many implications beyond the financial burden [2]. These incidents are aggravated by the fast growth of motorisation, and they occur mostly due to human error, mechatronics faults, as well as bad weather and road conditions. For HGVs, the emergence of complex logistics and transportation networks has required the widespread use of sensors, tracking devices, and mobile communication equipment to improve performance, economy and safety. These devices constantly gather information of vehicles and their journeys, including safety hazards and driving behaviour. As data availability increases, opportunities and challenges to extract useful information that benefit industry and society take place. In this work we present our immune-inspired instance selection solution applied to a real-world big data problem of determining HGV incident hot spots across the United Kingdom (UK) roads. The problem addressed is defined by Microlise [3] — a UK-based company that provides telematics solutions to help fleet operators to reduce their costs and environmental impacts. Currently, Microlise controls over 25% of the UK HGV fleet. Telematics have traditionally been used to track the position of vehicles via their Global Positioning System (GPS). However, with the increasing power of cloud data storage and computing, telecommunication and data analytics, various other services, such as fuel saving, fleet performance management, driving behaviour monitoring, dynamic routing, diagnostics and prognostics are being offered by telematics providers [4]. Nowadays, telematics is perceived as a wireless communication system encompassing a range of different tracking and management features, options and devices that generate data and enable the vehicles' internet of things. Microlise's telematics solutions allow the capture and processing in real-time of a whole range of HGV safety incidents (such as over speed, harsh braking and harsh cornering) with their date, time and location of occurrence. The company is faced with the challenge of transforming millions of data records into actionable knowledge to enhance their business. Part of this enhancement involves providing clients with bespoke software products and analytics that detect and manage risks of danger to vehicles tracked. Current literature does not effectively address the problem for big data and therefore an alternative solution has to be defined. It is determined thereby that the creation of a product to warn drivers about areas of high likelihood of incidents, based on historical records, is necessary and timely. In order accomplish this product, knowledge regarding incidents must be extracted, interpreted and summarised in a visual manner. We are assigned with the task of performing this analysis and determining the roads hot

spots. In addition, the following requirements regarding the solution are specified:

- Accuracy: hot spots should represent incidents that occur in the same road, in the same direction, within a certain mileage limit.
- Generality: all types of incidents considered (harsh cornering, harsh braking, speeding) have to be addressed in a similar manner to facilitate the implementation of the solution.
- Coverage: the hot spots should be defined for the entire United Kingdom (UK) map and therefore all data points collected are considered.
- Robustness: the results produced should be easily adaptable/modifiable to future incident data collected.
- Performance: although millions of instances are considered, a low-complexity, fast-running algorithm is desirable.

We successfully address the above requirements by adapting an immune-inspired instance selection mechanism [5] to our problem. The mechanism is derived from the Immune System self-regulation features, where redundancies as well as extra elements that are no longer necessary in the system are eliminated; therefore, only valuable elements are kept. We employ this principle to remove noise and repeated or similar values in the data. In addition, we extract the relevant points that characterise hazardous areas. Results are verified and validated through several interactions with transport experts; and the solution has been adopted by Microlise. In the next sections we introduce the problem (Section II), the related work (Section III), the solution provided (Section IV), followed by experiments and results (Section V), conclusions and opportunities for extension of the method to other areas (Section VI).

II. Problem Description

Given a large telematics data set of HGV incidents containing incident type, date, time and location, those areas of high likelihood of incidents should be determined. As mentioned previously, the indication of the hot spot location has to be accurate and encompass all types of incidents for all locations. In addition, it is desirable that the method runs fast and adapts to changes in roads and driving behaviour over time. For the development of our method, we investigate a large data set containing 1,000,612 incidents collected for three HGV companies over a three-month period in 2015, across the UK. The data is distributed between incidents, as follows: **(i) 773,323 speeding incidents; (ii) 213,697 harsh braking incidents; and (iii) 13,592 harsh cornering incidents**. Intuitively, the solution should provide the means to somehow "cluster" the incident points to determine dangerous areas in the UK road map. In addition, a distance measure calculation between each incident point should be employed, as the clusters are defined based on proximity. Furthermore, in situations such as those illustrated in Figure 1, several different cluster configurations could be determined; however, not all of them provide satisfactory solutions. Traditional clustering methods, Bayesian and spacial clustering (such as density-based spatial clustering [6] and other techniques [7]) are not effective for our problem, as they require either a pre-definition of the number of clusters (which limits the number of hot spots and compromises accuracy), produce elliptical clusters as those indicated by the green circles in Figure 1, require parameters to be adjusted or require an adaptation for big data problems [8]. Furthermore, as discussed in the next section, current literature on the identification of hot spots presents limitations with regards to our researched scenarios. Observing the smaller clusters in the figure below, the idea of adapting an existing instance selection (IS) method as an alternative to solve the problem comes therefore from the assumption that, given the size and character of the information collected for our study, there is a high likelihood of noise and redundancy in the data set. We hypothesise therefore that once reduction is performed, the hot spots are uncovered – they are represented by the remaining data points. In addition, it is required an IS approach with low complexity for timely results.

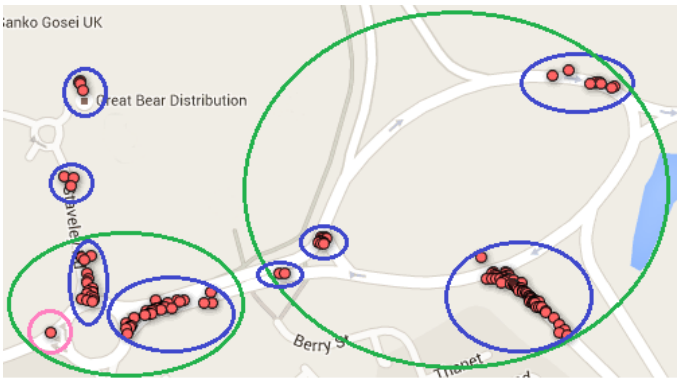

Fig. 1. Examples of data clusters. Clusters indicated in blue represent good candidate solutions. Clusters inside pink (with one instance) and green lines (bigger ellipsis) represent invalid solutions.

III. Related Work

According to Cheng and Washington [9], the objective of hot spot identification (HSID) is to detect transportation system locations with underlying correctable safety problems. These areas are characterised by elevated incident (or accident) frequencies relative to similar sites. Detecting hot spots is the first step of road safety management processes. Effective solutions therefore assist in optimal resource deployment. The literature on accident HSID is vast and the problem is mostly addressed by statistical methods assisted by historical data, as discussed next. In certain cases, government and public participation [10] also contribute to the knowledge regarding hot spots, although the information is inaccurate at times [11]. For accidents, hot spots are defined as *sites at which local risk factors are related to road design and/or traffic control* [12]. We are interested in HGV driving incident hot spots, which also include human error as an important factor. And from the HGV industry point of view, it is preferable to identify all current areas of potential danger/driving errors than to work

on historical assumptions or estimates. The literature regarding incident hot spots for big data sets and safety policies for HGV drivers, however, is still scarce. In this section we therefore present a review of the well-known methods for HSID, mostly involving accidents, and their main conjectures and constraints. It is not our objective to provide a complete review, but rather to point out the main methods and their limitations regarding big data that led to the development of our algorithm.

Montella [12] reviews and compares seven HSID methods using quantitative evaluation: crash frequency, equivalent property damage only crash frequency (EPDO), crash rate, proportion method, Empirical Bayes (EB) [13], EB estimate of severe-crash frequency (EBS) and the potential for improvement (PFI) method. In crash rate, locations are sorted in descending order of accident frequency (simple ranking); to compare areas, the total number of accidents is divided by the length of the road segments. EDPO categorises and ranks accidents according to their severity in terms of damage, costs and injuries. The crash rate normalises crash frequency according to traffic volume. The proportion method, less frequently employed to HSID, prioritises sites depending on their crash probabilities being higher than the threshold proportion, which is calculated from a comparison group. In the empirical Bayesian method, the estimation of the long term safety of a spot is obtained using the history of crashes of the entity and the expected number of crashes from safety performance functions for similar sites. In the EBS, the expected frequency of severe accidents is employed instead. PFI is calculated as the difference between the EB expected accident frequency and a crash prediction model, which is developed to predict accident frequency at locations similar to that being analysed. Montella's case study for comparison employs geometric, traffic and crash records from 2001-2005 for a motorway in Italy. 646 homogeneous segments (343 for each direction) are considered, with a mean length of 395 meters. In the analysis period, 2245 crashes occurred. The author's comparison tests led to the conclusion that the EB method is more suitable to detect priority investigation locations.

Similarly, Cheng and Washington [9] employ simulation data to evaluate three HSID methods based on peer comparison: simple ranking, confidence interval and EB. These methods identify hot spots by establishing a measure of comparison with similar sites. For confidence intervals, a location l is classified as unsafe if the observed crash count of l exceeds the average count of comparable locations. The authors justify the choice for simulated data rather than employing empirical values as it enables prior knowledge regarding safe and dangerous areas. The criteria for evaluation was the number of false positives, false negatives, false identifications and diminishing returns of crash history duration. The authors also conclude that EB in general performs better. However, in low crash count heterogeneity situations, EB is not significantly better. Their analysis also suggests that optimum crash history comprises of three years worth of data, which drastically improves bayesian results when compared to a 1-year period employing simple ranking and confidence interval. The methods investigated however are not directly applicable to our problem, as there is the need of a prior identification of comparable locations and historical data. As our objective is to define all incident hot spots for most roads in the UK, the determination of comparable sites regarding characteristics such as infrastructure, demand, traffic flow and weather conditions would incur in a significant amount of work prior to the HSID.

Anderson [14] introduces a kernel density estimation (KDE) for HSID coupled with a clustering technique determining classes of hot spots and their casual indicators. The author assumes that road accidents are influenced by the density of their occurrences in a specific area. The KDE is therefore employed to establish those areas of high risk of incidents and their spread, which are further classified through cluster analysis. Traffic accident data from London during 1999-2003 is employed. The author provides further knowledge regarding the nature and patterns found within the hot spots by dividing them into categories. The main limitation of the KDE, as pointed out by the author is that it *treats discrete events as a continuous surface.* From a HGV incident perspective, this generalisation might incur in inaccuracies, as not all adjacent events necessarily belong to the same hot spots. In addition, the authors state that the inability to properly determine the statistical significance of the resulting clusters (i.e., whether they are relevant) is still a drawback of the methodology.

Bíl *et al.* [15] improves the KDE cluster detection to overcome the lack of confidence in the results accuracy. Data consists of 7121 traffic incidents collected via GPS by the police in Czech Republic. The analysis is performed on primary roads, excluding highways and urban areas. Roads are separated in 713 sections of around 200m, without intersections. The improved version of cluster significance testing incorporates Monte Carlo to create variations in the incident locations. This allows for better statistical testing and to assess cluster significance. In addition, the authors indicate that to determine the cluster strength, the number of accidents should be contrasted with the length of the cluster and the length of a section. The work is limited, however, as areas in the map are excluded and the roads investigated have been previously segmented. This technique would therefore require further investigation in order to assess its applicability to big data.

Effati *et al.* [16] introduces a geospatial neuro-fuzzy approach for HSIS zones on regional transportation corridors. Historical crash data along with roadway information is used to calibrate and validate the model. Their methodology employs roadway geometry and environmental factors, which are processed through an adaptive neuro-fuzzy inference system. Their case study considers layers of data regarding a highway in the North of Tehran. The correlation between calculated hazardous zones and hot spots obtained using statistical approaches is verified; however, additional hazardous zones are spotted. The method also determines the most important hazardous factors in which crash prevention strategies should be employed. As a final contribution, the authors demonstrate how variations in one or more input factors affect the danger level of the road zones. Although successfully applied to the case study, to the best of our knowledge the method has not yet been exploited for larger data sets. In addition, it requires several layers of information regarding topography, elevation, geometry, weather, accidents and excising hazardous zones.

This data is provided by different stakeholders, which makes the generalisation of the method more difficult.

El-Basyouny and Sayed [17] proposes a depth-based multivariate method to identify and rank hot spots using a full Bayes approach. The authors use 236 signalized intersections from Vancouver, with collision and traffic volume data from 2001-2008. This data is split into two periods: 2001-2005 for ranking and 2006-2008 for evaluation. Markov Chains and Monte Carlo are employed to obtain a set of full Bayes posterior estimators on each multivariate Poisson log-normal model (linear trend, time varying intercept and time varying coefficients). The proposed method identifies dangerous intersections after applying a depth threshold, which is based on the amount of funding available for safety improvement. The performance of their model is compared to analogous methods that are based on depths of accident frequency (AF). Sensitivity, specificity and sum of norms of Poisson means show that the proposed method with full Bayes estimators has better results when compared to the depth-based AF method. This work was limited however to a single data set and this technique requires further research in order to assess its applicability to HSID.

From the literature it is possible to identify the following gaps in current research: (i) the number of instances (accidents) investigated is very limited regarding its size and locations across the studies; (ii) the experiments are mostly conducted within a small number of routes and journeys and/or considers simulated data; (iii) there is disregard for road bearing (direction); and (iv) to the best of our knowledge, there is very little literature regarding HGV incidents for the UK. Our work therefore aims at contributing in filling these research gaps by employing a immune-inspired IS, as further discussed next.

IV. The Instance Selection Mechanism

In data mining, IS [18] aims at determining optimal subsets of data with two fundamental properties: (i) The new obtained set is smaller than the original data set; and (ii) a set containing the most significant instances must be selected in order to build accurate machine learning (classifier, regression, clustering, etc.) models. Furthermore, IS plays an important role in knowledge discovery tasks, as it is supposed to determine the most significant samples in the data set; it accelerates the process of training machine learning methods; it reduces costs associated with data processing; and it is capable of removing noise and redundancy from the original data. In addition, IS is employed in big data sampling, as they choose instances in a more effective manner when compared to random selection [19]. It is not our objective in this work to provide an extensive review on the IS field. Instead, we focus on the technique chosen for our problem. Further information regarding IS approaches is given in [20], [18] and [21]. In Wilson and Martinez [20], a large study with the issues that may be encountered when tackling IS is introduced. In addition, the authors suggest a framework for the analysis and discussion of existing algorithms. Cano *et al.* [18] conduct another important review in the area. The authors list some of the main IS algorithms and categorise them into four sets: (1) techniques based on Nearest Neighbour rules, (2) methods based on ordered removal, (3) methods based on random sampling and (4) evolutionary-based approaches. An updated review in the area with further modern techniques is found in Lopez *et al.* [21].

For our work with road incident hot spots we adapt an existing immune-inspired IS technique [5], [19], namely, SeleSup. This technique has been successfully applied to select instances in data classification tasks [5], [22] (tested in data sets with up to 45,222 instances), data sampling [19] and text classification [23] (datasets with up to 18,300 instances). The SeleSup algorithm is chosen as its features seem to better match the requirements of the problem, when compared to other knowledge discovery techniques. Initially, solutions such as alternative IS methods, common techniques for HSID, spatial clustering and the distance calculation between all data set points were considered (sections II and III). These techniques were discarded, however, as they are mostly not suitable for the large amounts of data considered and require a considerable amount of processing time. Furthermore, they generally loose accuracy as the data increases. SeleSup, on the other hand, performs accurately in large data sets with low complexity (see Section IV-D). Experiments show that overall it requires less computational resources than other data reduction methods [5], [24]. In addition, SeleSup reduces the instances based on a measurement of proximity. This characteristic makes this approach particularly suitable for our problem, as the determination of clusters of hot spots is based mainly on spacial proximity of incidents. Further details regarding the adapted version of SeleSup are given next.

A. The Biological Inspiration

The biological process of mounting an immune response and restoring the homeostasis of the body, namely self-regulation mechanism, has served as inspiration to the development of the SeleSup algorithm. This process is rather complex and involves the interplay between different types of immune cells and molecules [25, Chapter 1], which details are out of the scope of this paper. For simplification, rather than an in-depth description of the biological phenomena, we present an overall view of the concepts that have led to the creation of the method. Whenever there is danger present in the human body, the Immune System recruits numerous cells to assess the risks to the organism, followed by the activation and execution of an appropriate response [26, Chapters 1,2 and 7],[27, Sections I and II]. The most successful set of cells in mounting the immune response receive stimulus to survive (and proliferate) to accelerate danger elimination. The least effective cells, on the other hand, are removed from the organism. Within this mechanism, the majority of successful immune cells that are no longer needed in the body after the immune response receive signals to die [25, Parts I and II]. Only a small proportion of those types of cells is kept to form a immune memory, so that fronts of defense are prepared whenever there is recurrence of similar danger. Existing suppressive signals also balance the numbers of different types of danger-specific cells, based on the risks inside the organism. This favours the proliferation of those immune cells mostly needed for defense at a certain

point. Once the danger is controlled, the excessive specific cells are eliminated. Groups of cells that are slightly different but perform the same function are overtaken by clones of those most fitted cells, in some sort of natural selection process [28]. Our method is inspired by the suppression aspect of this phenomena, where what is no longer useful is eliminated.

B. The Adapted Method

The SeleSup algorithm was originally conceptualised to eliminate very similar instances in data classification tasks. The objectives were to select a smaller set of significant samples and reduce the number of examples processed to build a classifier model, with low detriment to the classification accuracy [5]. For hot spots, however, the same process is modified to establish groupings of similar instances and select those to be identifiable as the cluster centres. Once the centres are defined, the other remaining data points are removed. From an immune perspective, a cluster centre can be interpreted as a cell that receives surviving signals and stays in the organism to fight a specific danger. This cell is supposed to encompass the immune capacities of similar clones eliminated. In our problem, the cluster centres (or surviving cells) represent a summary of where dangerous areas in the road for HGV drivers are located. A hot spot is therefore the location of one instance, which is the centre of a group of data points with similar characteristics (similar/same values for latitude, longitude, direction, address, etc.), within a certain distance in a road. A schematic representation of a hot spot is shown in Figure 2.

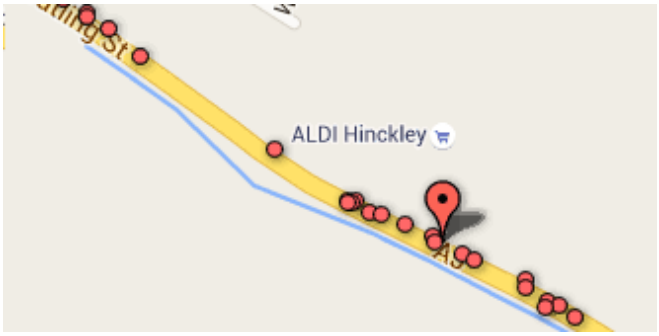

Fig. 2. Hot spot example (represented in the figure by the red balloon). For the example displayed in the map, the hot spot indicates an area of incidents within half a mile range. The small red circles are instances of incident points that occurred within that hot spot area. These points become redundant information to the system once the hot spot is defined; therefore, they should be eliminated.

C. The Algorithm

The SeleSup algorithm starts with the idea that the system's model must identify the best group of "surviving cells". For IS problems, the set of instances selected has to be the most representative (i.e., the most effective for building classification models). The search for these instances therefore has to be informed. For our problem, however, we are only interested in removing similarities. This means that all data records have the same potential of being centres of hot spots (cluster centres) and therefore they can be determined randomly. To establish the cluster centres, the mechanism divides the original data in two subsets, suppressor set and set to be reduced. The first group represents those instances of the data set (or immune cells) that are meant to be kept in the system. The second subset contains the elements to be suppressed. The suppressor group is considerably smaller than the set to be reduced. Initially, the instances for the first group are chosen randomly. As the algorithm progresses, to ensure that all clusters (hot spots) are contained in the smaller subset, instances from the bigger set being reduced can be transferred to the suppressor group. The cells (instances) are represented by an array of attributes; those to be eliminated are associated to the closest cell from the suppressive group. This proximity is determined by a measure of distance. In our case, we consider attributes such as latitude, longitude, vehicle course or bearing (angle of the heading direction of the road), address, day of the week and time of the day. The distance calculation is shown in Algorithm 1.

Algorithm 1: Distance Calculation

inputs: Latitudes, Longitudes and angles in the road of suppressor instance s and candidate instance to be reduced r; Address, Weekday, Time of s and r if considered;and $MileageRange$;

output: Determine if r is within range of s

```
if Address_s = Address_r then
    Delta ← AngleDifference(Angle_s, Angle_r);
    if Delta ≤ 60 then
        H ← HaversineDistance(Latitude_s, Longitude_s,
        Latitude_r, Longitude_r);
        if H ≤ MileageRange then
            if Weekday_r = Weekday_s OR
            WeekdayNotApplicable then
                if Time_r is within Time_s interval OR
                TimeNotApplicable then
                    r is within s range;
```

In Algorithm 1, for latitude and longitude we employ the Haversine distance [29], [30]. Angles of similar points need to be within a sixty degree range. Nominal variables, such as full address (or just partial address) and weekday must have the same value. Time of the day is separated according to peak and off peak hours for HGV traffic, which are defined by Microlise. No data normalisation is necessary. The pseudo-code for our new version of SeleSup, i.e, SeleSup HSID can be seen in Algorithm 2.

The method applied to the hot spot problem (SeleSup HSID) is shown in figures 3 to 8. Figure 3 shows the first step, where the centres of the clusters are chosen randomly. Figure 4 exemplifies how similar values within these centres are removed, for increased performance. Figure 5 displays the process of identifying points belonging to a cluster by their proximity to a centre. Figure 6 shows the elimination of points belonging to a centre, as they are not necessary to identify the

Algorithm 2: The Adapted SeleSup for HSID

inputs: The entire data set $\mathbb{D}$; a fraction f of *Suppressor Cells* (default $f = 0.1$); and the maximum distance range of a cluster;

`// If constraints such as address, weekday, time are to be considered, they should also be inputs.`

output: A *reduced* set $\mathbb{D}'$

1 Assign $\lceil f \cdot |\mathbb{D}| \rceil$ randomly selected samples as *SupressorCells* (*suppressor set*);
2 The remaining samples are *CellsToBeEliminated* (*set of redundancies, to be reduced*);
3 **forall the** *SuppressorCells* **do** *fitness* $= 0$;
`// Suppressor set redundancy removal`
4 **foreach** *Suppressor cell* s_i *from SuppressorCells* **do**
5 *SetOfRedundantSupressors* $\leftarrow$ suppressor cells within the similarity range (Algorithm 1) of s_i;
6 *SuppressorCells* $\leftarrow$ *SuppressorCells* - *SetOfRedundantSuppressors*;
7 s_i's fitness $\leftarrow$ size(*SetOfRedundantSuppressors*);
`// Finding closest suppressor cell to cells to be removed (redundant)`
8 **foreach** r_j *from* $CellsToBeEliminated$ **do**
9 *Nearest suppressor* $s_k \leftarrow$ Find the *SuppressorCell* within the similarity range of r_j;
10 **if** *NearestSuppressor* $\neq \varnothing$ **then**
`// Cell is redundant`
11 $CellsToBeEliminated \leftarrow CellsToBeEliminated$ - r_j;
12 increase s_K fitness;
`// Adding instances not yet represented in` $SuppressorSet$
13 **forall the** *CellsToBeEliminated* **do** *fitness* $= 0$;
14 **if** $CellsToBeEliminated \neq \varnothing$ **then**
15 **foreach** r_l *from* $CellsToBeEliminated$ **do**
16 *RedundantSet* $\leftarrow$ other cells from $CellsToBeEliminated$ within the mileage (and constraints) of r_l);
17 *CellsToBeEliminated* $\leftarrow$ *CellsToBeEliminated* - *RedundantSet*;
18 r_l's fitness $\leftarrow$ size(*RedundantSet*);
19 $SuppressorCells \leftarrow SuppressorCells$ + $CellsToBeEliminated$;
`// Output phase`
20 Eliminate those *SuppressorCells* with *fitness* $= 0$;
21 Output the set of *surviving SuppressorCells*s as the reduced set $\mathbb{T}'$ containing the hot spots locations and their fitness.

hot spots. Figure 7 introduces how incident points that were not in the range of a cluster centre are converted into new centres. Figure 8 presents the final outcome of the method, where the hot spots (cluster centres) are established. At each elimination stage, the remaining hot spots are associated with a score (fitness value). This value indicates the number of incidents that occurred in a certain hot spot cluster. This allows for hot spot ranking and comparison and supports decisions. In addition, the the hot spot stores the date in which the last incident took place. This is because once policies are adopted to avoid incidents, it is expected that hot spots disappear with time, and the date log assists in determining those areas that should no longer be regarded as hazardous, because, for instance, no incident occurred in the past 6 months.

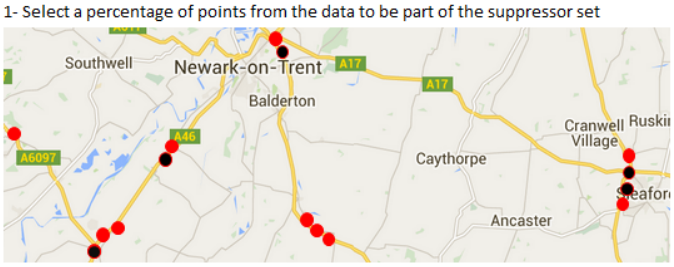

Fig. 3. Step 1. A percentage of random points is selected to form the suppressor set (line 1 in Algorithm 2).

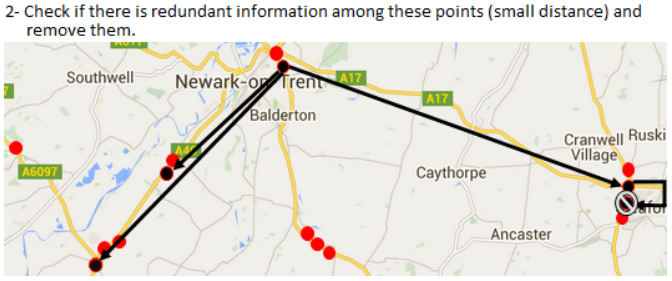

Fig. 4. Step 2. In this step, the distance between points from the suppressor set is calculated in order to remove any redundancies (points representing the same cluster). If two points are close, the deletion of one of them occurs randomly. This corresponds to the first *for-each* loop (lines 4 to 7) in Algorithm 2.

Fig. 5. Step 3. The distance between the set of points to be reduced and the centres of the clusters is calculated. This corresponds to line 9 of the second *for-each* loop in Algorithm 2.

Steps 2 (Figure 4) and 6 (Figure 8) overcome a limitation of the original SeleSup, which is having a fixed parameter for the data reduction percentage (which also defines the size of the reduction set). For our case study, the cluster centre represents all other points within its range; therefore, we can remove all redundancies rather than just a fraction of them. Moreover, if there is data left with no centre, the suppression set increases to encompass this data. The reduction percentage therefore

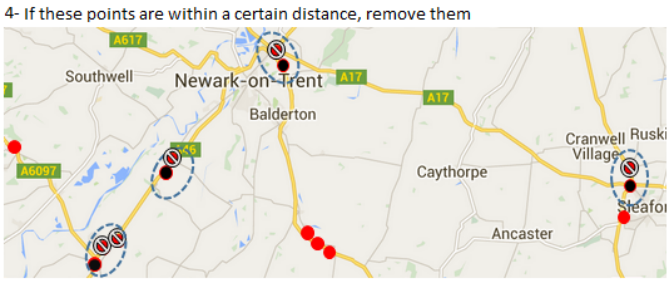

Fig. 6. Step 4. If the distance between a point from the set to be reduced and a centre is smaller than a certain limit, then this point is suppressed (lines 10 to 12 in Algorithm 2). At this stage, distance as well as other constraints can be adopted. For instance, apart from belonging within a distance range of a cluster centre, the point must also have the same information about time of the day, day of the week, direction of the road, road address, etc.

Fig. 7. Step 5. After step 4 (Figure 6), there is the possibility that some areas of where incidents occur, as displayed in the map, are not covered by the centres. To overcome this issue, the points from the set to be reduced with no cluster assigned need to be turned into cluster centres. If they are within a close range, they can be merged into one hot spot, similarly to step 2 (Figure 4). This step corresponds to lines 13 to 19 of Algorithm 2.

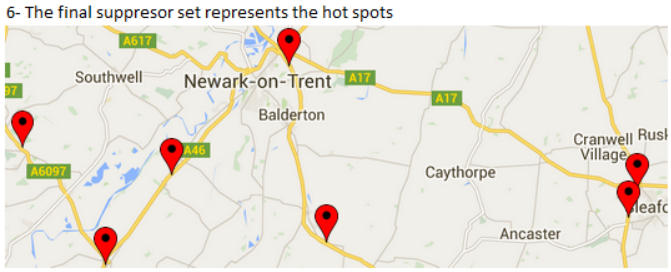

Fig. 8. Step 6. The hot spots are determined as the centres of the clusters of similar points. In addition, a score (fitness) measure is associated with each hot spot. The fitness is equal to the number of incidents that occurred around the hot spot (lines 20 and 21 of Algorithm 2).

changes with time, which makes this variable less relevant to this problem. The suppression approach also provides an effective way of determining clusters of incidents. When compared to most existing clustering techniques such as K-means [31], Partition Around Medoids (PAM) [32], etc., it has the advantage of not requiring the number of clusters as input. The final number of cluster centres (hot spots) is an emerging effect of the elimination process. Furthermore, the method covers all areas in the map and the minimum number of hot spots necessary to cover the entire search space is obtained.

D. Complexity Analysis

Let n be the size of the the entire data set $\mathbb{D}$ of incidents ($n = |\mathbb{D}|$), s the number of suppressor instances ($s = \lceil f \cdot |\mathbb{D}| \rceil$), and r the instances to be reduced, $r = \mathbb{D} - (\lceil f \cdot |\mathbb{D}| \rceil)$ (see Algorithm 2). Since the time consuming operation in the algorithm is the computation of distance between elements of these sets, the time complexity analysis is focused on that operation. The time complexity, therefore, is expressed in terms of *number of distance calculations* for a given input size. From Algorithm 2 it is possible to notice that the time consuming code is performed in the suppression phase on the *for-each* loops. Reducing the suppressor (first *for-each* loop), finding the nearest suppressor for an instance from the set to be reduced (second *for-each* loop) and removing redundancies from the set to be reduced (third *for-each* loop) mean that, in the worst case scenario, when there is no possible reduction, the whole sets need to be searched. Consequently, there are $\frac{(1+s)\cdot s}{2}$ distance calculations per suppressor cell; $s \cdot r$ calculations for the set to be reduced; and $\frac{(1+(r))\cdot(r)}{2}$. Hence, the maximum total number of distance operations O carried out on those loops is: $O = \frac{(1+s)\cdot s}{2} + (s \cdot r) + \frac{(1+r)\cdot r}{2}$; but $s = f \cdot n; r = n - f \cdot n$. Therefore, $O = \frac{f \cdot n + f^2 \cdot n^2}{2} + (f \cdot n^2 - f^2 \cdot n^2) + \frac{+n + f \cdot n + n^2 - 2 \cdot f \cdot n^2 + f^2 \cdot n^2}{2}$. As the parameter f belongs to the interval $[0, 1]$, the expression $2 \cdot f \cdot n$ achieves its maximum at $f = 1$. With $f = 1$ the O expression is: $O = \frac{3 \cdot n^2 + n}{2}$. Finally, considering the asymptotic case where n, the number of input instances, goes to infinity, the complexity of Algorithm 2 is given by $O(n^2)$. As for our problem, we know that there is redundancy in the data and the processing time of the algorithm is always smaller than the worst case scenario. The benefit of SeleSup HSID to our problem lies therefore on the fact that it considerably reduces complexity by splitting the data, constantly removing instances and thereby decreasing the size of the data to be processed. Rather than comparing all instances against each other, they are compared against a subgroup, which has its size reduced in every iteration. Different types of incidents as well as mileage and other constraints, however, affect the method's performance. In those cases where less constraints are considered, more data reduction is expected. Next section presents a study on the impact of constraints and parameters, such as the initial suppressor set size and mileage range on SeleSup HSID performance.

V. Experiments and Results

In order to assess the correctness and performance of the proposed method for HSID, we employ our method to four real-world data sets of speeding, harsh cornering, harsh braking and contextual speeding incidents. The data refers to three months of incidents collected by Microlise telematics. All data sets contain the same attributes (latitude, longitude, course, address). The mileage limit ranges for the clusters definition are set to 0.5, 2 and 5 miles for speeding and contextual speeding incidents; and 0.1, 0.2 and 0.5 mileage limit for harsh braking and harsh cornering incidents. We consider the following constraints: mileage limit, course and address, under two different scenarios: (a) considering mileage and course as

TABLE I. HOT SPOTS IDENTIFIED WITH SELESUP HSID CONSIDERING $f = 0.1$ FOR INITIAL SUPPRESSOR CELLS

Dataset	Incidents	Constraint	Mileage	Raw Hot Spots		Fitness > 0		Fitness > 10		Runtime (s)	
				Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.
speeding	3139	mileage and course (a)	0.5	1403.31	19.11	612.58	13.76	20.84	2.75	3.05	0.05
			2	785.28	18.52	514.33	26.94	35.30	5.67	2.79	0.02
			5	491.06	17.60	365.13	17.23	62.59	7.54	2.70	0.01
		mileage, course and address (b)	0.5	1413.19	25.43	610.38	13.52	21.29	2.37	2.79	0.01
			2	811.34	22.14	521.82	25.71	34.32	5.12	2.67	0.00
			5	547.51	13.57	391.07	16.57	61.86	5.68	2.61	0.00
harsh cornering	13568	(a)	0.1	3906.76	47.58	1863.86	44.78	195.70	8.99	4.80	0.01
			0.2	3446.74	29.49	1827.00	33.82	214.68	9.34	4.42	0.01
			0.5	3102.20	37.56	1733.89	35.02	220.20	9.24	4.20	0.01
		(b)	0.1	4208.92	57.35	1944.51	26.11	186.18	7.75	3.71	0.01
			0.2	3815.58	24.44	1940.43	19.01	202.89	6.52	3.59	0.01
			0.5	3585.39	35.22	1903.78	32.11	208.38	5.44	3.52	0.01
harsh braking	213697	(a)	0.1	62455.73	1013.56	29163.82	759.91	3030.10	204.81	430.01	17.35
			0.2	50219.04	913.52	27393.08	896.55	3382.61	174.12	304.20	50.93
			0.5	34769.19	749.77	22859.59	1083.46	3830.10	305.27	174.40	11.32
		(b)	0.1	65350.23	749.88	30505.43	812.17	2891.27	182.88	151.67	4.22
			0.2	54152.23	616.18	29455.06	619.90	3237.08	174.93	105.06	1.34
			0.5	40684.22	776.83	26504.14	640.32	3669.06	228.96	61.08	1.42
contextual speeding	770184	(a)	0.5	58026.77	1801.40	45690.93	1992.43	14813.30	1525.27	874.88	281.88
			2	21623.59	1289.80	19316.79	1461.09	9510.91	873.70	279.63	55.53
			5	10234.10	834.35	9430.69	708.59	5915.69	638.91	127.16	7.45
		(b)	0.5	59933.59	2884.44	46932.68	2609.04	14836.40	1672.42	182.10	16.60
			2	26066.29	1299.13	23236.41	1162.06	9991.40	603.90	68.07	4.35
			5	18251.47	693.25	16714.45	713.61	7531.22	342.15	49.58	4.13

constraints and (b) considering mileage, course and address. We want to investigate: **(i) the impact on the results across multiple independent runs; (ii) the influence of the initial size of the suppressor set; and (iii) how the mileage range affects the final number of hot spots.** In our experiments, we employ a parallel implementation[1] of SeleSup HSID based on Apache Spark (see more details in [33]) to deal efficiently with the size of the selected data sets. The experiments are carried out in a single node with an Intel(R) Xeon(R) CPU E5-1650 v4 processor (12 cores) at 3.60GHz, and 64 GB of RAM. We use the Cloudera's open-source Apache Hadoop distribution (Hadoop 2.6.0-cdh5.4.2) and Spark 1.6.2, with a total number of 8 concurrent tasks.

A. Results for Multiple Independent Runs: Table I shows the results of SeleSup HSID applied to the data sets using $f = 0.1$. As the initial selection of the suppressor set is random, the number of hot spots and their location vary slightly. Thus, we run the method 100 times, obtaining average (Avg.) and standard deviation (Std.) values for the number of hot spots found. The table reports the number of original incidents for each data set, the number of raw hot spots ($fitness \geqslant 0$), hot spots with $fitness > 0$ and those with $fitness > 10$. In addition, for illustration on the complexity of the method, we also include the average runtime to process the data sets from our parallel implementation. Results for the numbers of hot spots and their small percentage of standard deviation (less than 0.5% in average) show how the method is consistent and robust across multiple runs. Furthermore, there appears to be a general standard deviation decline as fitness increases; this indicates that those hot spots of high incidence are detected consistently in the experiments. Results for runtime also confirm how quickly large amounts of data are processed. The impact of the constraint is observed by the increased number of hot spots and processing time. Figure 9 shows an example of the resulting hot spots (with fitness $\geqslant 10$) determined in two different runs for harsh braking incidents around Cambridgeshire, UK. In the figure it is possible to observe that the hot spots detected in Figure 9(a) and Figure 9(b) represent the same location, with slight variations, as indicated by the areas circled in red in the map. It is also possible to observe that the coverage provided by the outcomes remains unchanged and that is what was of interest for Microlise.

B. Mileage Range Impact: From Table I we can observe that the number of identified hot spots decreases when the mileage radius considered is increased. We conduct further tests to assess the impact of the mileage range in the final number of hot spots for the largest data set (contextual speeding), also considering $fitness \geqslant 0$, $fitness > 0$ and $fitness > 10$. Figure 10 shows the number of hot spots identified as the mileage radius increases. We can observe that the number of raw hot spots tend to decrease drastically as we consider a distance greater than one mile radius. However, the number of hot spots with $fitness > 10$ (i.e. more relevant ones) seems to have far less variation. Our experiments show therefore that those hot spots with high fitness appear to be the same within multiple replicates, variations in mileage and some constraint values (Table I). This coincides with current literature that employs frequency calculations to determine hot spots. However, no prior information is necessary in our case to identify those areas. And although our goal is to determine all areas of high likelihood of incidents, results suggest that we can assess how significant to road safety the hot spots are not only by their fitness values, but also by how stable hot spots are under different experimental set ups. This steady state reached by hot spots numbers allows for the identification of hazardous areas proportional to HGV traffic volume in certain areas, with no need for prior traffic volume information. This is another advantage when compared to current methods that require this

[1]Source code available at: https://github.com/triguero/Immune-HotSpot

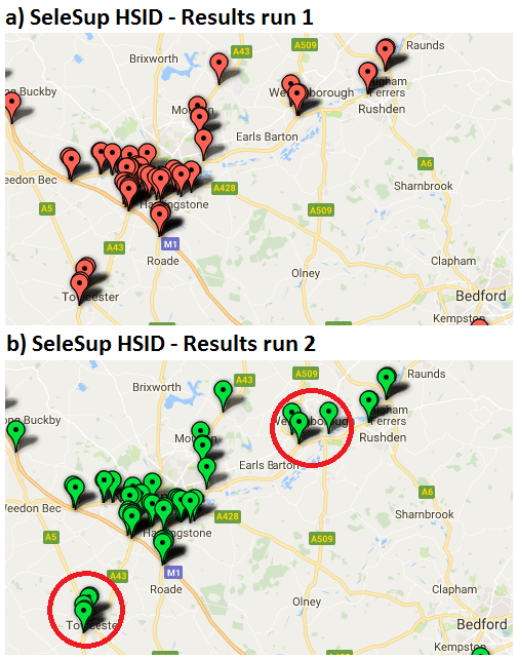

Fig. 9. Results for harsh braking for two runs. The areas surrounded by red circles show variations in the outcomes, with no detriment to coverage. Differences observed in the top circle in (b) compared to (a) are due to the fact that in (b) the three hot spots are more spread apart. However, for both runs, three hot spots are detected.

extra layer of information.

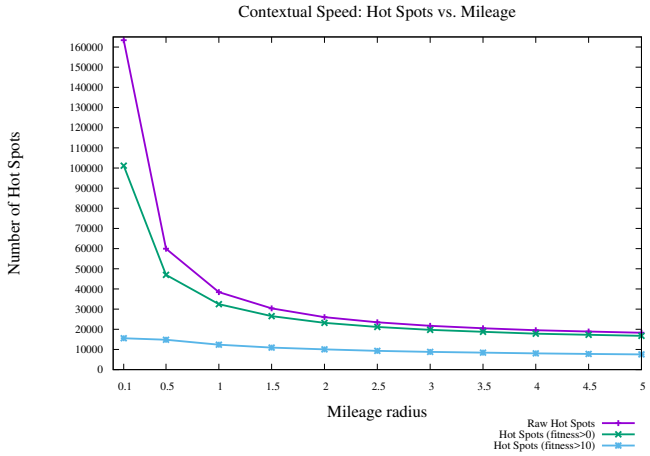

Fig. 10. Number of Hot Spots identified with respect to the mileage radius on Contextual Speeding data set.

C. Impact of the Initial Size of the Suppressor Set: We assess the SeleSup HSID consistency in detecting hot spots regardless of its parameter value. The only parameter associated to SeleSup is the initial number of randomly selected suppressors (*suppressor set size* = $\lceil f \cdot |\mathbb{D}| \rceil$). In Figure 11, we plot the number of identified hot spots (fitness>10) for all data sets studied according to different percentages of initial suppressors ($f \in [0.05, 0.1, 0.1, 0.3, 0.4]$). We can observe that the algorithm is stable regarding the number of resulting identified hot spots, independently of the initial number of selected data points.

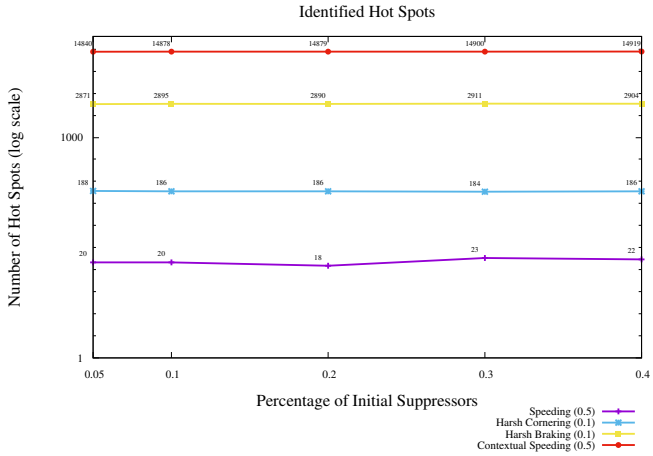

Fig. 11. Number of Hot Spots identified ($Fitness¿10$) with regards to the initial percentage of randomly selected suppressors. For each dataset, mileage limit is indicated between brackets.

VI. Conclusions

Transportation research mostly aims at improving driving performance, economy and safety. Our work contributes to this area by introducing a big data IS method to identify HGV road incident hot spots. We were provided with a large data set containing three months of incidents collected via telematics. Hundreds of thousands of incidents including speeding, harsh braking, harsh cornering and contextual speeding were investigated. Our challenges were to create an accurate, general, adaptable and robust solution, which had to be employed to all kinds of incidents considered. Preferably, due to the large volumes of data and the industrial application, it was requested that the solution was relatively low in complexity and fast to run. The method provided is an adaptation of an immune inspired instance selection mechanism, namely SeleSup. The SeleSup algorithm was chosen as its features better match the requirements of the problem when compared to other knowledge discovery and HSID techniques. Unlike existing traditional HSID approaches, SeleSup is suitable to tackle large data. Furthermore, it does not require road segmentation or public data as input. SeleSup is inspired by the Immune System self-regulation mechanism, where only the fittest immune cells remain in the organism. The method works by establishing a set of data points (suppressor set), which is meant to have the most significant information in the data (in our case it is the set of hot spots). This set size and its data points are initially defined randomly; however, the self-adjustable, self-adapting character of the method allows for the establishment of the optimal number of hot spots, even when new data is logged to the system. The remaining incidents not contained in the suppressor set constitute the set to be reduced, as they

represent redundant information. In our approach, the hot spots are not explicitly ranked, as for our industrial partner it is important to determine all areas of frequent incidents. Instead, a score (fitness) value for each hot spot coupled with the last incident date are determined. This allows for stakeholders in industry to decide which dangerous areas should be tackled (or informed to HGV drivers, depending on the location, HGV traffic, etc.). However, our experiments suggest that hot spot stability over different experiment scenarios might indicate those areas of more relevance for safety measures. In addition, it is possible to determine how long a hot spot should 'survive' the system, based on the last incident logged. Our approach was successfully applied to the problem and further verification and validation was conducted by experts in the HGV industry. Given the data set with incidents and their locations, SeleSup HSID determined all hot spots in the UK map in a timely manner; and our solution has been adopted by Microlise on its HGV fleet. The method developed can be adapted to other problems and presents several opportunities for improvements. For instance, it can be employed to determine areas of high incidences of crime involving HVGs. In addition, hot spots of HGV accidents can also be determined. In unrelated areas, SeleSup HSID can determine hot spots of diseases outbreaks. In addition, it can be applied to mobile phones and fitness trackers data to identify places most frequented for business purposes. As future directions for transport research we intend to aggregate value to hot spots. For instance, the overlaying of hot spots with statistics regarding weather conditions, traffic, types of HGVs (size, age, weight), driver profiles, etc. allows for more accurate reports regarding the causes of incidents in certain locations. In addition, the establishment of how hot spots of accidents correlate to those of incidents is necessary.

REFERENCES

[1] RSF, "Road safety foundation," *URL http://www.roadsafetyfoundation.org/*, last acessed 22-10-2015.

[2] B. Coll, S. Moutari, and A. H. Marshall, "Hotspots identification and ranking for road safety improvement: An alternative approach," *Accident Analysis and Prevention*, vol. 59, pp. 604 – 617, 2013.

[3] Microlise, *www.microlise.com*, Last accessed 8 Jul, 2017.

[4] M. Mesgarpour *et al.*, "Overview of telematics-based prognostics and health management systems for commercial vehicles," *Activities of Transport Telematics*, vol. 395, pp. 123–130, 2013.

[5] G. P. Figueredo, N. F. F. Ebecken, D. A. Augusto, and H. J. C. Barbosa, "An immune-inspired instance selection mechanism for supervised classification," *Memetic Computing*, vol. 4, pp. 135–147, 2012.

[6] M. Ester *et al.*, "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise," ser. KDD. AAAI Press, 1996, pp. 226–231.

[7] J. Han, M. Kamber, and A. K. H. Tung, "Spatial clustering methods in data mining: A survey," in *Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS*, H. J. Miller and J. Han, Eds.

[8] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, *Big Data Clustering: A Review*. Springer Int. Pub., 2014, pp. 707–720.

[9] W. Cheng and S. P. Washington, "Experimental evaluation of hotspot identification methods," *Accident Analysis & Prevention*, vol. 37, no. 5, pp. 870 – 881, 2005.

[10] W. Kowtanapanich, Y. Tanaboriboon, and W. Chadbunchachai, "Applying publich participation approach to black spot identification processs: a case study in thailand," *IATSS R.*, vol. 30, no. 1, pp. 73 – 85, 2006.

[11] B. P. Loo, "Validating crash locations for quantitative spatial analysis: A GIS-based approach," *Accident Analysis & Prevention*, vol. 38, no. 5, pp. 879 – 886, 2006.

[12] A. Montella, "A comparative analysis of hotspot identification methods," *Accident Analysis & Prevention*, vol. 42, no. 2, pp. 571 – 581, 2010.

[13] H. Robbins, "An empirical bayes approach to statistics," in *Third Berkeley Symp on Math Stats and Prob, Vol 1: Contribs to the Theory of Statistics*. Univ of California Press, 1956, pp. 157–163.

[14] T. K. Anderson, "Kernel density estimation and k-means clustering to profile road accident hotspots," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 359 – 364, 2009.

[15] M. Bíl, R. Andrik, and Z. Janoka, "Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation," *Accident Analysis & Prevention*, vol. 55, pp. 265 – 273, 2013.

[16] M. Effati, M. A. Rajabi, F. Samadzadegan, and S. Shabani, "A geospatial based neuro-fuzzy modeling for regional transportation corridors hazardous zones identification," *Int. J. of Civil Engineering*, vol. 12.

[17] K. El-Basyouny and T. Sayed, "Depth-based hotspot identification and multivariate ranking using the full bayes approach," *Accident Analysis & Prevention*, vol. 50, pp. 1082 – 1089, 2013.

[18] J. Cano, F. Herrera, and M. Lozano, "Using evolutionary algorithms as instance selection for data reduction in KDD: An experimental study," *IEEE Trans. on Evolutionary Comp*, vol. 7, no. 6, pp. 561–575, 2003.

[19] G. P. Figueredo, N. F. F. Ebecken, and H. J. C. Barbosa, "An immune-inspired sampling technique for data selection," in *The XXX Iberian Latin American Congress on Computational Methods in Engineering (CILAMCE 2009)*, 2009.

[20] D. R. Wilson and T. R. Martinez, "Reduction techniques for instance-based learning algorithms," *M Learning*, vol. 38, pp. 257–268, 2000.

[21] J. Olvera-López, J. Carrasco-Ochoa, J. Martínez-Trinidad, and J. Kittler, "A review of instance selection methods," *Artificial Intelligence Review*, vol. 34, no. 2, pp. 133–143, 2010.

[22] G. P. Figueredo, N. F. F. Ebecken, and H. J. C. Barbosa, "The supraic algorithm: A suppression immune based mechanism to find a representative training set in data classification tasks," in *ICARIS*, ser. LNCS, vol. 4628. Springer, 2007, pp. 59–70.

[23] M. L. C. Passini, K. B. Estbanez, G. P. Figueredo, and N. F. F. Ebecken, "A strategy for training set selection in text classification problems," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 6, pp. 54–60, 2013.

[24] G. P. Figueredo *et al.*, "A support vector machine-based technique for instance selection," in *XXXIV CILAMCE*, 2013.

[25] C. A. Janeway, P. Travers, M. Walport, and M. Shlomchik, *Immunobiology: The Immune System in Health and Disease*, 5th ed. Garland Science, 2001.

[26] I. Tizard, *Intr. to Veterinary Immunology*, 2nd ed. ROCA, 1985.

[27] A. K. Abbas, A. H. Lichtman, and J. S. Pober, *Cellular and molecular immunology*. Saunders, Philadelphia, 1991.

[28] G. L. Ada and G. J. V. Nossal, "The clonal selection theory," *Scientific American*, vol. 257, no. 2, pp. 50–57, 1987.

[29] G. V. Brummelen, *Heavenly Mathematics: The Forgotten Art of Spherical Trigonometry*, 2012.

[30] F. Cajori, *A History of Mathematical Notations: Vol. II*, 1928.

[31] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100–108, 1979.

[32] L. Kaufman and P. Rousseeuw, *Clustering by Means of Medoids*, ser. Reps. of the Faculty of Mathematics and Informatics, 1987.

[33] I. Triguero *et al.*, "Vehicle incident hot spots identification: An approach for big data," in *Procs of the 11th IEEE International Conference On Big Data Science And Engineering*, ser. IEEE BigDataSE-17. IEEE, 2017.

Dr Grazziela P Figueredo is a Data Scientist at the Advanced Data Analysis Centre (ADAC) within the School of Computer Science at The University of Nottingham. The focus of her research is the development and application of techniques for systems simulation and intelligent data analysis. She has been working with data analysis for a wide range of areas, including academic, medical and industrial partners. As part of ADAC, she specialised in providing consultancy within the University of Nottingham and externally, with the mission to enhance current research and business by providing state-of-the-art tools and the expertise to engineer and interpret data. She has published articles in leading journals, such as Plos ONE, the BMC Transactions in Bioinformatics, and international conferences, such as the IEEE Big Data SE.

Dr Isaac Triguero received the M.Sc. and Ph.D. degree in Computer Science from the University of Granada, Spain, in 2009 and 2014, respectively. He is currently an assistant professor in data science at the University of Nottingham, United Kingdom. He has published more than 25 papers in international journals. His research interests include data mining, data reduction, evolutionary algorithms, semi-supervised learning, bioinformatics and big data learning.

Dr Mohammad Mesgarpour received his PhD degree in Operational Research from the University of Southampton in 2012. He also completed his Masters degrees in Information Systems and Industrial Engineering. He worked for the University of Nottingham as a KTP research associate for two years before joining Microlise in 2014 as a Technical Research Analyst. His main area of research is in the fields of Transport Management, Predictive Modelling, Data Analytics and Combinatorial Optimisation.

Alexandre M Guerra received his B.S. (2016) in electrical engineering from University of Campinas, São Paulo, Brazil. He is currently working on his Master's at the same university in electrical and computer engineering. His master thesis is based on coclustering and association rules. His research interest are data analysis, data mining, machine learning and development of graphical user interfaces.

Professor Jonathan M Garibaldi is Head of School of Computer Science at the University of Nottingham, and Head of the Intelligent Modelling and Analysis (IMA) Research Group. His main research interest is in developing intelligent techniques to model human reasoning in uncertain environments, with a particular emphasis on the medical domain. Prof. Garibaldi has been the PI on EU and EPSRC projects worth over 3m, and CoI on a portfolio of grants worth over 25m. He is Director of the University of Nottingham Advanced Data Analysis Centre, established in 2012 to provide leading-edge data analysis services across the University and for industrial consultancy. His experience of leading large research projects includes his roles as Lead Scientist and Co-ordinator of BIOPTRAIN, a Marie-Curie Early Stage Training network in bioinformatics optimisation worth over 2m, the local co-ordinator of the 6.4m BIOPATTERN FP6 Network of Excellence, lead Computer Scientist on a 700k MRC DPFS (Developmental Pathway Funding Scheme) project to transfer the Nottingham Prognostic Index for breast cancer prognosis into clinical use. Industrial projects include a TSB funded project for data analysis in the transport sector, and a collaborative project with CESG (GCHQ) investigating and modelling variation in human reasoning in subjective risk assessments in the context of cyber-security. He is currently the local PI for Nottingham on the 900k UKCRC Joint Funders Tissue Directory and Coordination Centre, a CoI on the 14m BBSRC/EPSRC Synthetic Biology Research Centre in Sustainable Routes to Platform Chemicals, and was CoI on the 10m BBSRC/EPSRC Centre for Plant Integrative Biology. Prof. Garibaldi has published over 200 articles on fuzzy systems and intelligent data analysis, including over 50 journal papers and over 150 conference articles, three book chapters, and three co-edited books. In January 2017, Prof. Garibaldi was appointed as the Editor-in-Chief of the IEEE Transactions on Fuzzy Systems, the leading international journal in the field of fuzzy methods. He was Publications Chair of FUZZ-IEEE 2007 and General Chair of the 2009 UK Workshop on Computational Intelligence, and has served regularly in the organising committees and programme committees of a range of leading international conferences and workshops, such as FUZZ-IEEE, WCCI, EURO and PPSN.

Professor Robert I John received the B.Sc. (Hons.) degree in mathematics from Leicester Polytechnic, Leicester, U.K., the M.Sc. degree in statistics from UMIST, Manchester, U.K., and the Ph.D. degree in Fuzzy Logic from De Montfort University, Leicester, U.K., in 1979, 1981, and 2000, respectively. He worked in industry for 10 years as a mathematician and knowledge engineer developing knowledge based systems for British Gas and the financial services industry. Bob spent 24 years at De Montfort University. He has over 150 research publications of which about 50 are in international journals with over 6000 citations. Bob joined the University of Nottingham in 2013 where he heads up the research group ASAP in the School of Computer Science. The Automated Scheduling, Optimisation and Planning (ASAP) research group carries out multi-disciplinary research into mathematical models and algorithms for a variety of real world optimisation problems. He is also a member of LUCID.