CENTRE FOR
FINANCIAL RISK
Faculty of Business and Economics

MACQUARIE University

Statistical Machine Learning Analysis of Cyber Risk Data: Event Case Studies

WORKING PAPER 18-02

Gareth W. Peters, Pavel V. Shevchenko, Ruben D. Cohen, Diane Maurice

Statistical Machine Learning Analysis of Cyber Risk Data: Event Case Studies

Gareth W. Peters[a,*], Pavel V. Shevchenko[b], Ruben D. Cohen[c], Diane R. Maurice[d]

[a] *Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, UK*
[b] *Department of Applied Finance and Actuarial Studies, Macquarie University, Sydney, Australia*
[c] *MP Capital, London, UK.*
[d] *Central Bank of Tunisia, US Treasury.*

Abstract

This work explores the common attributes of different types of cyber risk with a view to better understanding the key attributes that contribute to each type of cyber risk category. In doing so we explore event studies on a range of different market sectors, different countries, different demographics over time and categories of cyber risk event type. To perform this study we explore a modern machine learning clustering method to investigate the attributes of cyber risk and how they can be categorised via a statistical method. We then explore the properties of this statistical classification and interpret its implications for the current taxonomies being developed for cyber risk in areas of risk management. In the process we will interpret and analyse the implications our analysis has on both operational risk modelling of cyber risk data, as well as the implications the findings have for cyber risk insurance products. On a broader level, this analysis informs risk behaviour of both traditional and emerging financial institutions such as financial technology (fintech).

Keywords: cyber risk, cyber crime, operational risk, cyber insurance, machine learning, k-means clustering method

*Corresponding author: garethpeters78@gmail.com

1. Cyber Risk Context

The aim of this work is to study the properties of cyber risk from the perspective of event-based studies. The focus will be to understand the attributes that are indicators of enhanced cyber crime frequency and to understand which are the vulnerable areas currently targeted by current cyber crime activity. Such a study can be useful for both mitigation and prevention enhancements through risk management as well as rate setting in cyber insurance contexts.

The aim of this study could be achieved in a number of ways: for instance, one could look at studying the data from the perspective of the different approaches adopted by industry, regulators and academic studies to classify or create a taxonomy for cyber risk and cyber insurance; we refer the reader to Peters et al. (2017). Such approaches are well suited to loss-database studies. In this chapter we will focus on event studies that are more suitable to types of unsupervised cluster analysis methods we adopt.

Managing cyber risk vulnerability and mitigating cyber risk loss events is critical in modern business practices in order to prevent excessive losses due to fraud, business disruption, loss of data and loss of reputation. Information security risk is similar to other business risks in the sense that one can aim to first attempt to remove or eliminate aspects of such risk through modern IT security layering systems. Regulatory guidance that sets out recommendations for such approaches in the banking and insurance sectors is being developed; see, for instance, the cyber security regulatory guidance for banking and insurance sectors from the Federal Financial Institutions Examination Council (FFIEC), which has been developing guidelines for authentication in an Internet banking environment[1] and the Bank of England Prudential Regulatory Authority (PRA) released in mid-2017 a supervisory statement directed at the insurance and reinsurance industries pertaining directly to cyber risk and in particular cyber insurance underwriting risks[2].

In modern banking and insurance practice, it is impossible to treat cyber risk losses purely based on a strategy of elimination; one must also aim to mitigate, control and transfer risks. The traditional approaches to security architecture and design have in the past been relatively successful as a strategy of prevention; however in modern banking environments, the veracity of advancement and sustained persistent attacks from outside sophisticated attackers have rendered such strategies non-viable; in

[1]Guidance Document "Authentication in an Internet Banking Environment" at `https://ffiec.bankinfosecurity.com/new-ffiec-guidelines-full-text-a-3802`

[2]Bank of England, PRA, Supervisory Statement — SS4/17Cyber insurance underwriting risk July 2017 `http://www.bankofengland.co.uk/pra/Documents/publications/ss/2017/ss417.pdf`

other words, the complete prevention of system compromise through technical and procedural means is no longer feasible. As noted in Böhme and Kataria (2006), the elimination of security risks in modern business environments is not possible and consequently risk managers and IT security groups tend to deploy protection technologies such as firewall, antivirus and encryption, and instate appropriate security policies such as passwords, access control, port blocking and so forth to mitigate the probability of a break-in or failure. If the residual risk is manageable it is absorbed; otherwise, it is transferred by either outsourcing security or buying insurance; see discussions on these aspects in Peters et al. (2017).

As observed in Siegel et al. (2002) a modern risk-management approach starts with a complete understanding of the risk factors facing an organisation. It is the intention of this work to perform such assessments utilising modern tools from machine learning and statistics to undertake an unsupervised analysis of core attributes that drive cyber risk events and losses. Such analysis should then aid in developing further risk assessments for security teams to help them to develop and design appropriate control systems based on the most significant risk factors. It will also benefit insurance companies in understanding the underlying key drivers of such losses to help them to enhance the approaches they have developed to price policies for the remediation of harm.

2. Event Case Studies

In this chapter we aim to undertake an unsupervised analysis of empirical data studies based on several event-loss datasets that have been constructed specifically for this analysis from different regulatory reports.

1. DATA BREACH STUDY: Cyber Security Breaches of Health and Human Services Sector in US. Source: *Office for Civil Rights of the US Department of Health and Human Services*.
2. DATA BREACH and IMPACT STUDY: Cyber Security Events reported by the Internet Crime Complaint Center (IC3) working group of the FBI. Source: `www.IC3.gov`.

We note that there is a limited number of papers that study cyber risk based on empirical analysis of data breach information. We first comment on why such studies can be of benefit, followed by the machine learning methods we adopt to study this data, explaining these methods for practitioners to use. Then we follow up with detailed analysis of the results and conclusions.

2.1. ***Why is it useful to study this type of data and perform clustering?***

The reason we study these datasets and undertake the machine learning analysis we perform is that we seek to obtain information that will aid in answering questions such as:

- Are particular regions of space more likely to present increased incidence of cyber loss events or breaches?

  - *which would have implications for rate setting in different regions of the US.*
- Are particular organization types more targeted for cyber attacks or susceptible to cyber incidents?

  - *which could affect the types of lines of business covered or the premiums charged for these sub-businesses.*
- Are there different patterns to the cyber breaches over time?

  - *which could tell us if past history is meaningful to utilise as a measure of historical record to evaluate premiums.*
- Are particular cyber events/breach types more likely to cause excessive cyber breach events compared with others?

  - *which could tell us which event types should be covered and which event types may be excluded.*

These types of questions/ideas can be useful in property and casualty insurance in rate making/premium setting via the approach of class rating or manual rating. This rating means that exposures with similar characteristics are placed in the same underwriting class, and each is charged the same rate. The advantage of class rating lies with its easy application and ability to quickly be obtained (especially with sparse data).

3. Unsupervised Clustering for Cyber Risk: Non-linear Kernel k-means

In this section we outline briefly an important class of nonlinear clustering methods that can be adopted to perform analysis of the data sets discussed previously. These are based on kernel k-means methods which come under the generic title of "Partitioning Methods". These are primarily statistical methods developed in order *to organise the data objects into a required number of preselected clusters amounts (k-clusters) that optimises a certain similarity measure*. It may often be more narrowly defined as a method that is implemented by iterative algorithm with the similarity measure based on the distance between data objects.

3.1. ***Understanding non-linear kernel k-means***

In general in kernel methods we seek to exploit the notion that a nonlinear data transformation into a higher dimensional feature space increases the probability of the linear separability of the transformed data. Achieving this via kernel methods basically exploits properties of dot products

in feature space in terms of kernel functions in input space; for a detailed review of unsupervised learning methods of this type, see Peters (2017).

There are two core approaches to kernel k-means. The first is based on assuming the data considered is equiweighted, ie, appears as uniform independent realisations. The second assumes the data to be clustered is weighted with a typically non-uniform weighting, emphasising the significance of some regions of the feature, or attributes space as more significant for cluster analysis and unsupervised learning compared with other regions. This is both practically meaningful for purposes of interpretation, and also statistically important to consider in order to help remove what is known as Breiman's bias in clustering; see further discussion in Dhillon et al. (2004), Jain (2010) and the tutorial coverage in Peters (2017).

3.1.1. ***Equiweighted kernel k-means***

For kernel k-means we follow the succinct exposition provided in Dhillon et al. (2004), Jain (2010) and Marin et al. (2017). Instead of clustering data points $\{\boldsymbol{x}_p | p \in \Omega\} \subset \mathbb{R}^d$ in their original space, kernel k-means uses mapping $\varphi(\cdot) : \mathbb{R}^d \mapsto \mathcal{H}$ embedding input data $\boldsymbol{x}_p \in \mathbb{R}^d$ as points $\varphi_p \equiv \varphi(\boldsymbol{x}_p)$.

The idea of k-means and kernel k-means is then to proceed with an objective function, for instance one may choose to minimise the squared errors in the embedded space corresponding to objective function, denoted by $O(\mathcal{S}, m)$ to be minimised, for K total possible clusters, given by

$$O(\mathcal{S}, \boldsymbol{\mu}) = \sum_{k}^{K} \sum_{p \in S^k} ||\varphi_p - \boldsymbol{\mu}_k||^2_{\mathcal{H}} \tag{3.1}$$

where $\mathcal{S} = \left(S^1, S^2, \ldots, S^K\right)$ is a partitioning (clustering) of the data space Ω into K clusters and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K)$ is a set of parameters for the clusters with $|| \cdot ||_{\mathcal{H}}$ the Hilbertian norm.

If one first optimizes $O(\mathcal{S}, \boldsymbol{\mu})$ with respect to parameters $\boldsymbol{\mu}$ it will produce for each cluster class S_k a closed form solution corresponding to cluster means in the embedded space:

$$\boldsymbol{\mu}^*_k = \frac{\sum_{q \in S^k} \varphi_q}{|S^k|} \tag{3.2}$$

where $|\cdot|$ is the cardinality of a cluster.

If we now evaluate the objective function at this point, we obtain a new objective function given

by $O(\mathcal{S}, \boldsymbol{\mu}^*) = O(\mathcal{S})$ which is only a function of the data and cluster groups $\mathcal{S}$, given by

$$\begin{aligned} O(\mathcal{S}) &= \sum_k \sum_{p \in S^k} \left|\left| \varphi_p - \frac{\sum_{q \in S^k} \varphi_q}{|S^k|} \right|\right|_{\mathcal{H}}^2 \\ &\stackrel{c}{=} -\sum_k \frac{\sum_{p,q \in S^k} \langle \varphi(\boldsymbol{f}_p), \varphi(\boldsymbol{f}_q) \rangle_{\mathcal{H}}}{|S^k|} \\ &\stackrel{c}{=} -\sum_k \frac{\sum_{p,q \in S^k} k\left(\boldsymbol{f}_p, \boldsymbol{f}_q\right)}{|S^k|} \end{aligned} \tag{3.3}$$

where $\stackrel{c}{=}$ denotes equality up to an additive constant and the function $k\left(\boldsymbol{f}_p, \boldsymbol{f}_q\right)$ is known as a kernel, often selected to be from the Mercer family of kernel functions, see detailed examples, references and discussion in Peters (2017) and Marin et al. (2017).

Remark 3.1. *One can see from this result that optimization of $O(\mathcal{S})$ enables optimization in high-dimensional Reproducing Kernel Hilbert Space $\mathcal{H}$ that only uses kernel computation and does not require computing or even known the explicit form of the embedding feature map $\varphi(\boldsymbol{x})$.*

3.1.2. ***Weighted kernel k-means***

A weighted version of kernel k-means can be developed in which each data point or feature vector to be clustered is given a non-uniform weight to denote its likelihood of occurrence. Using a weighted scheme which is non-uniform can be beneficial in adding stability and non-trivial diversity in cluster assignments, avoiding a feature of equiweighted set-ups known in clustering as Breiman's bias in the cluster assignment, see further discussion in Marin et al. (2017).

Then the objective function to be minimised is

$$O(\mathcal{S}, \boldsymbol{\mu}) = \sum_{k}^{K} \sum_{p \in S^k} w_p ||\varphi_p - \boldsymbol{\mu}_k||_{\mathcal{H}}^2 \tag{3.4}$$

where $\mathcal{S} = \left(S^1, S^2, \ldots, S^K\right)$ is a partitioning (clustering) of Ω into K clusters and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K)$ is a set of parameters for the clusters with $||\cdot||_{\mathcal{H}}$ the Hilbertian norm and w_p are positive weights for each data point.

As before, one can optimize $O(\mathcal{S}, \boldsymbol{\mu})$ with respect to parameters $\boldsymbol{\mu}$ for each cluster which produces a closed form solution corresponding to cluster means in the embedded space:

$$\boldsymbol{\mu}_k^* = \frac{\sum_{q \in S^k} w_q \varphi_q}{\sum_{q \in S^k} w_q} \tag{3.5}$$

Now the Euclidean distance from φ_p to the cluster center $\boldsymbol{\mu}_k$ is given by

$$\begin{aligned} &\varphi_p \cdot \varphi_p - \frac{2\sum_{q\in S^k} w_q \varphi_p \cdot \varphi_q}{\sum_{q\in S^k} w_q} + \frac{\sum_{q,l\in S^k} w_q w_l \varphi_q \cdot \varphi_l}{\left(\sum_{q\in S^k} w_q\right)^2} \\ &= K_{pp} - \frac{2\sum_{q\in S^k} w_q K_{pq}}{\sum_{q\in S^k} w_q} + \frac{\sum_{q,l\in S^k} w_q w_l K_{ql}}{\left(\sum_{q\in S^k} w_q\right)^2} \end{aligned} \tag{3.6}$$

where the dot products $\varphi_q \cdot \varphi_l$ are computed using kernel functions.

3.1.3. ***Kernel k-means algorithm***

The kernel k-means then follows the algorithmic iterative solution steps given by:

- **Input:** kernel matrix (Gram matrix) K, number of clusters k and initial cluster centers $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$.
- **Output:** Final cluster centers $\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_k^*$ and clustering error (objective function).
- Iterative Steps of the Algorithm:

  1. For each point $\boldsymbol{x}_p$ and every cluster S^k compute $||\varphi_p - \boldsymbol{\mu}_i||^2$ using Equation 3.6.
  2. Find for each point $\boldsymbol{x}_p$ the optimal cluster label

  $$S^*(\boldsymbol{x}_p) = \arg\min_i \left(||\varphi_p - \boldsymbol{\mu}_i||^2\right)$$

  3. Update clusters as $S^k = \{\boldsymbol{x}_p | S^*(\boldsymbol{x}_p) = k\}$
  4. Repeat until convergence (no more clusters change allotments).

There are global solutions to such an algorithm that also search over the number of clusters, see discussions in Marin et al. (2017).

A key property of the kernel k-means and the reason it is called a nonlinear classifier is the fact that after kernalisation, the kernalised features are separated or clustered according to, for instance, a linear boundary. However, when this linear boundary for the classification is mapped back to the feature space or data state space the resulting boundary of the clustering algorithm becomes a nonlinear function.

Furthermore, it is clear that if one increases k without bound and does not impose any parsimony penalty then this will in principle always reduce the amount of error in the resulting clustering. In the most extreme case one would then just achieve zero error by allocating each data point to an individual cluster, clearly an inadequate solution.

Consequently, one must therefore strike a balance between maximum compression of the data using a single cluster and maximum accuracy by assigning each data point to its own cluster. There

are several criteria one may adopt to select the optimal choice of number of clusters; see discussions in Jain (2010).

Now we will commence to demonstrate how these machine learning methods can be utilised to aid in analysis of the kernel k-means clustering method when applied to cyber security data analysis.

4. Case Study 1: Kernel K-Means clustering of Cyber Security Breaches in US Health and Human Services

In this section we consider cyber security breaches of the health and human services sector in US. Since October 2009 organizations in the US that store data on human health are required to report any incident that compromises the confidentiality of 500 or more patients/human subjects (45 C.F.R. 164.408). These reports are publicly available.

The data source: "HHSCyberSecurityBreaches" was downloaded from the Office for Civil Rights of the US Department of Health and Human Services, 2015-02-26 [3].

Data Overview

A data frame containing 1151 observations of 9 variables:

- **Name.of.Covered.Entity**: A character vector identifying the organization involved in the breach.
- **State**: A factor giving the two-letter abbreviation of the US state or territory where the breach occurred. This has 52 levels for the 50 states plus the District of Columbia (DC) and Puerto Rico (PR).
- **Covered.Entity.Type**: A factor giving the organization type of the covered entity with four levels:
  - "Business Associate",
  - "Health Plan",
  - "Healthcare Clearing House", and
  - "Healthcare Provider".
- **Individuals.Affected**: An integer giving the number of humans whose records were compromised in the breach. This is 500 or greater; US law requires reports of breaches involving 500 or more records but not of breaches involving fewer.

[3]Source of data "HHSCyberSecurityBreaches" from CRAN package "Ecdat": `https://cran.r-project.org/web/packages/Ecdat/Ecdat.pdf`

- **Breach.Submission.Date**: Date when the breach was reported.
- **Type.of.Breach**: A factor giving one of 29 different combinations of 7 different breach types:
  - "Hacking/IT Incident",
  - "Improper Disposal",
  - "Loss",
  - "Other",
  - "Theft",
  - "Unauthorized Access/Disclosure", and
  - "Unknown".
- **Location.of.Breached.Information**: A factor giving one of 47 different combinations of 8 different location categories:
  - "Desktop Computer",
  - "Electronic Medical Record",
  - "Email",
  - "Laptop",
  - "Network Server",
  - "Other",
  - "Other Portable Electronic Device", and
  - "Paper/Films".
- **Business.Associate.Present**: Logical = (Covered.Entity.Type == "Business Associate")
- **Web.Description**: A character vector giving a narrative description of the incident.

Split of the data records by type of incident, state and location is presented in Tables 1, 2 and 3 respectively.

Next we plot in Figure 1 the cluster outcome of kernel k-means obtained from a radial basis function kernel with estimated hyper parameters. In particular the first study we undertake shows the effect of optimising the kernel parameters over the entire dataset for different choices of $K \in \{2, 3, 4, 5\}$, based on a global estimation of the kernel parameters. The results of the clustering on all the features described above are then presented for the case of the radial basis function kernel, first for two attributes: number of days since first event and number of people affected.

We repeat this analysis for the different number of clusters, this time using a different approach to selection of the kernel parameters. In this case we use a local kernel parameter selection approach, in which a local scaling is used. This provides a more advanced approach by setting a width parameter for every point in the dataset. We show in Figure 2 that there is quite a substantial effect of the

Table 1: Split of the data records by type of incident.

Hacking/IT Incident	Improper Disposal
94	51
Loss	Other
111	111
Theft	Unauthorized Access/Disclosure
633	240
	Unknown
	16

Table 2: Split of the data records by state of incident.

AK	AL	AR	AZ	CA	CO	CT	DC	DE	FL	GA	HI	IA	ID	IL	IN
5	17	7	27	128	20	18	6	1	69	41	1	7	3	57	37
KS	KY	LA	MA	MD	ME	MI	MN	MO	MS	MT	NC	ND	NE	NH	NJ
7	26	9	35	17	1	25	27	24	6	6	34	3	6	4	17
NM	NV	NY	OH	OK	OR	PA	PR	RI	SC	SD	TN	TX	UT	VA	VT
11	9	72	34	8	16	45	28	7	13	2	33	100	11	22	1
WA	WI	WV	WY												
28	11	5	4												

Table 3: Split of the data records by location of breached records.

Desktop Computer	Electronic Medical Record
157	49
Email	Laptop
88	273
Network Server	Other
158	197
Other Portable Electronic Device	Paper/Films
142	278

selection of the kernel parameters. We prefer the second approach of kernel parameter estimation, which is more advanced; see detailed explanation of this choice in Ng et al. (2002).

Based on the outcomes from this analysis in Figure 2 we will focus on the case of $K = 3$ and we will continue to further explore the effects of the kernel choice. We demonstrate in Figure 3 the results of using different kernels in the clustering framework for the local hyper parameter estimation setting. The kernels considered were linear (standard k-means), and nonlinear k-means given by radial basis function (square exponential); the tanh kernel and the Bessel function kernel. We see that these differences in the Gram (proximity) matrix presented in Figure 3 have a clear effect on the cluster outcomes achieved.

We see there is a significant difference between the linear and the nonlinear k-means; however, in this case the choice of kernel does not present a significant difference in cluster solution. We proceed with this analysis therefore with the RBF (square exponential) kernel.

Next we will consider the following questions.

- Are any particular types of breach grouped in this clustering?
- Are any states more risky than others?

We present results for the global kernel parameter estimation analysis for $K = 3$ compared to the same analysis with a local kernel parameter estimation approach, these are demonstrated side by side in the following output analysis. In Figure 4 we present the results for pairwise analysis of the number of records stolen vs the state of the incident. In Figure 5 we present the results when looking at the states vs the number of days since the start of the records and in Figure 6 we present the results split in a different way, we show the location (state) where the event occurred vs the type of cyber attack. Clearly, we see that the choice of kernel parameters can have a pronounced effect on the clustering outcomes. We believe that the analysis utilising the local scale kernel parameter estimation is more desirable. Furthermore, we see in Figure 7 the results of the analysis displayed geographically in the US by state. We show the state versus average number of attacks per cluster $K = 3$ and global kernel parameter estimation. It shows that certain states have high concentrations or there is a greater prevalence for attack in these states. It may be due to increased opportunity or it may be due to lax controls relative to other states with regard to preventative measure or risk management. This is an aspect to be explored further with regard to the types of attacks in each state.

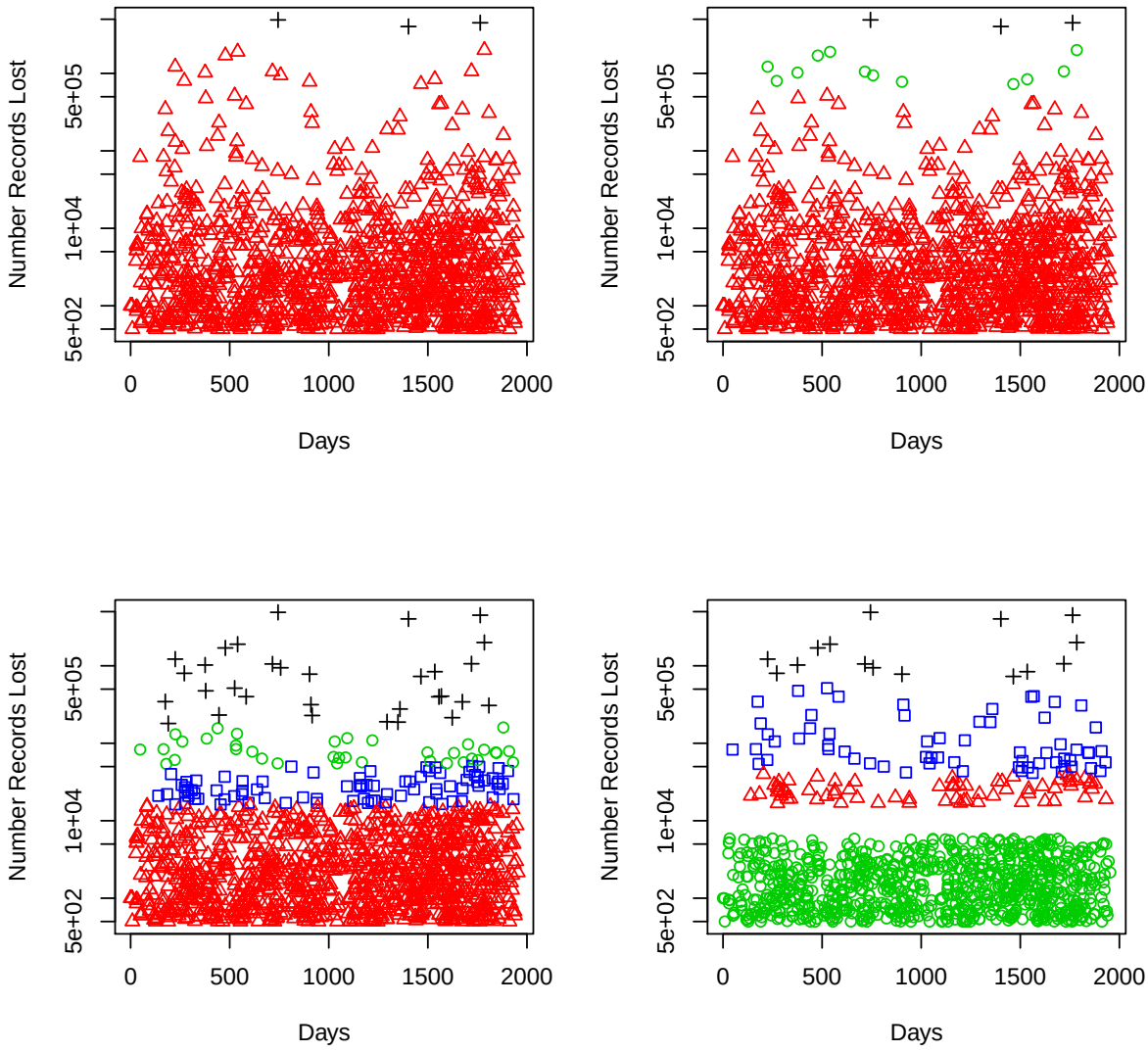

Figure 1: Number of days since 2009-10-21 vs log of number of records stolen. Global kernel hyper-parameter estimation results. Panels – Top Left: K=2 clusters; Top Right: K=3 clusters; Bottom Left: K=4 clusters; and Bottom Right: K=5 clusters

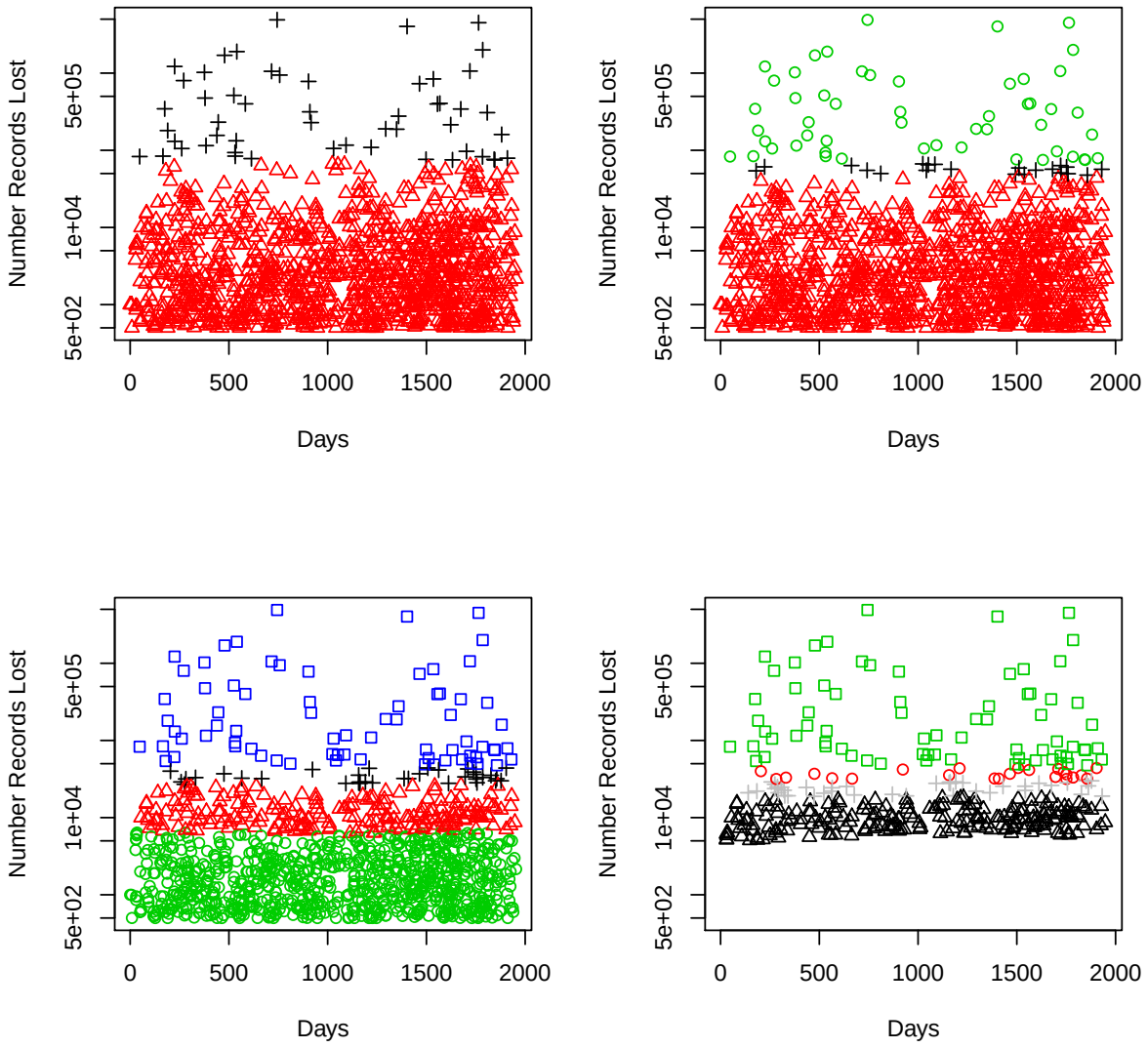

Figure 2: Number of days since 2009-10-21 vs log of number of records stolen. Local kernel hyper-parameter estimation results. Panels – Top Left: K=2 clusters; Top Right: $K = 3$ clusters; Bottom Left: $K = 4$ clusters; and Bottom Right: $K = 5$ clusters

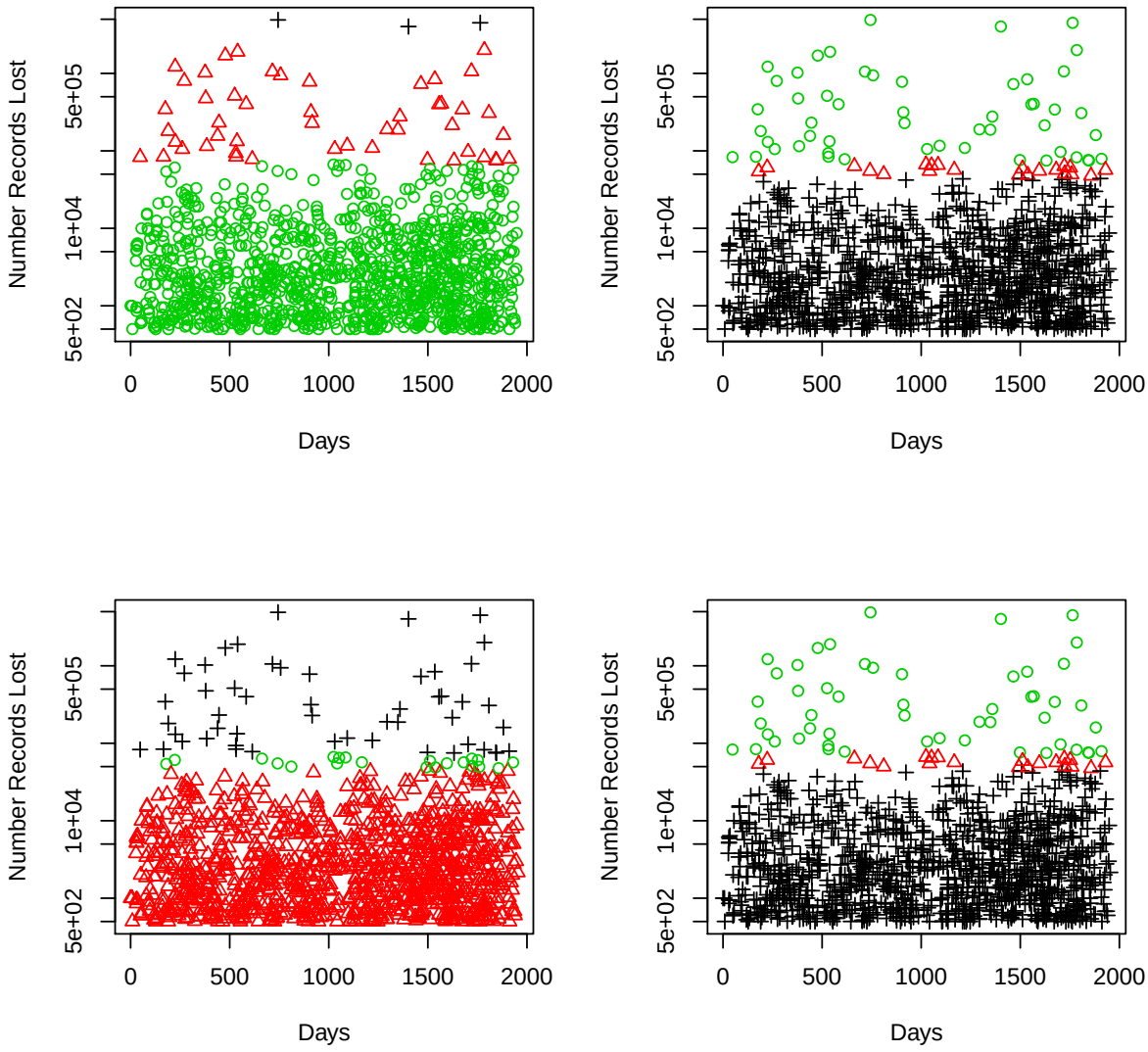

Figure 3: Number of days since 2009-10-21 vs log of number of records stolen. Effect of kernel covariance function on clustering outcome. Panels – Top Left: Linear kernel; Top Right: RBF kernel; Bottom Left: Tanh kernel; and Top Left: Bessel kernel

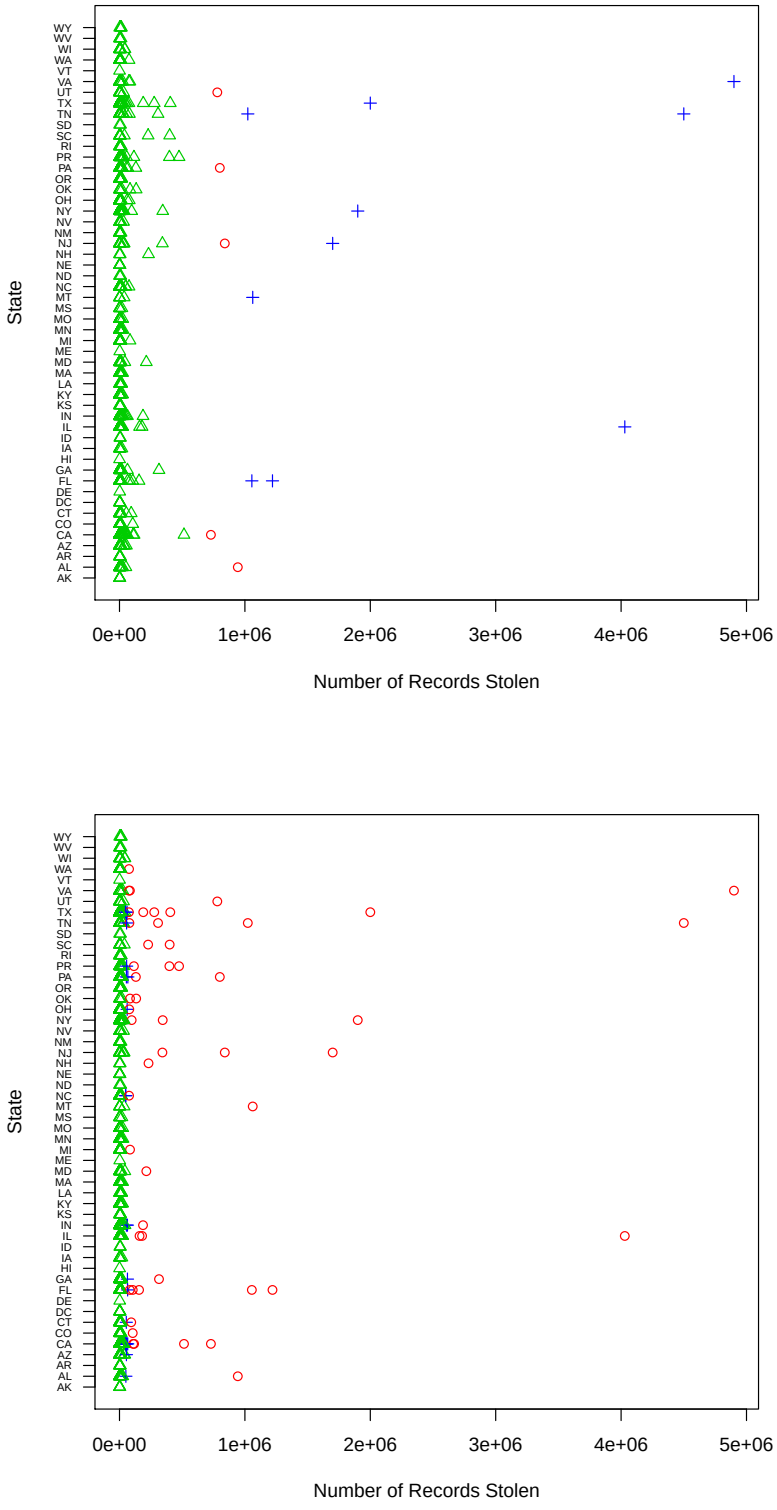

Figure 4: Number of records stolen versus state for $K = 3$ clusters, global hyper parameter (top), and local hyper parameter (bottom) estimation in kernel.

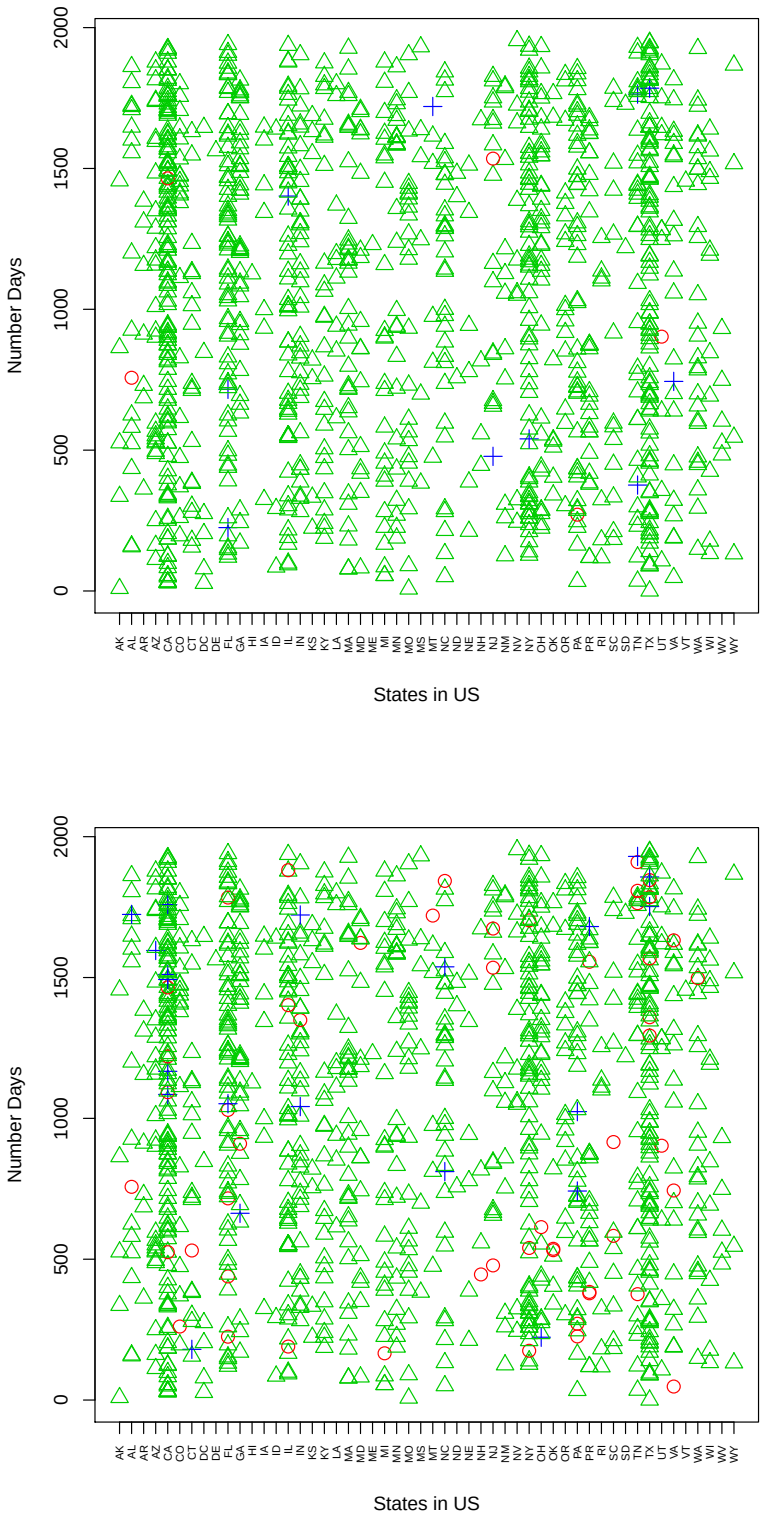

Figure 5: State versus number of days since first event for $K = 3$ clusters, global hyper parameter (top) and local hyper parameter (bottom) estimation in kernel.

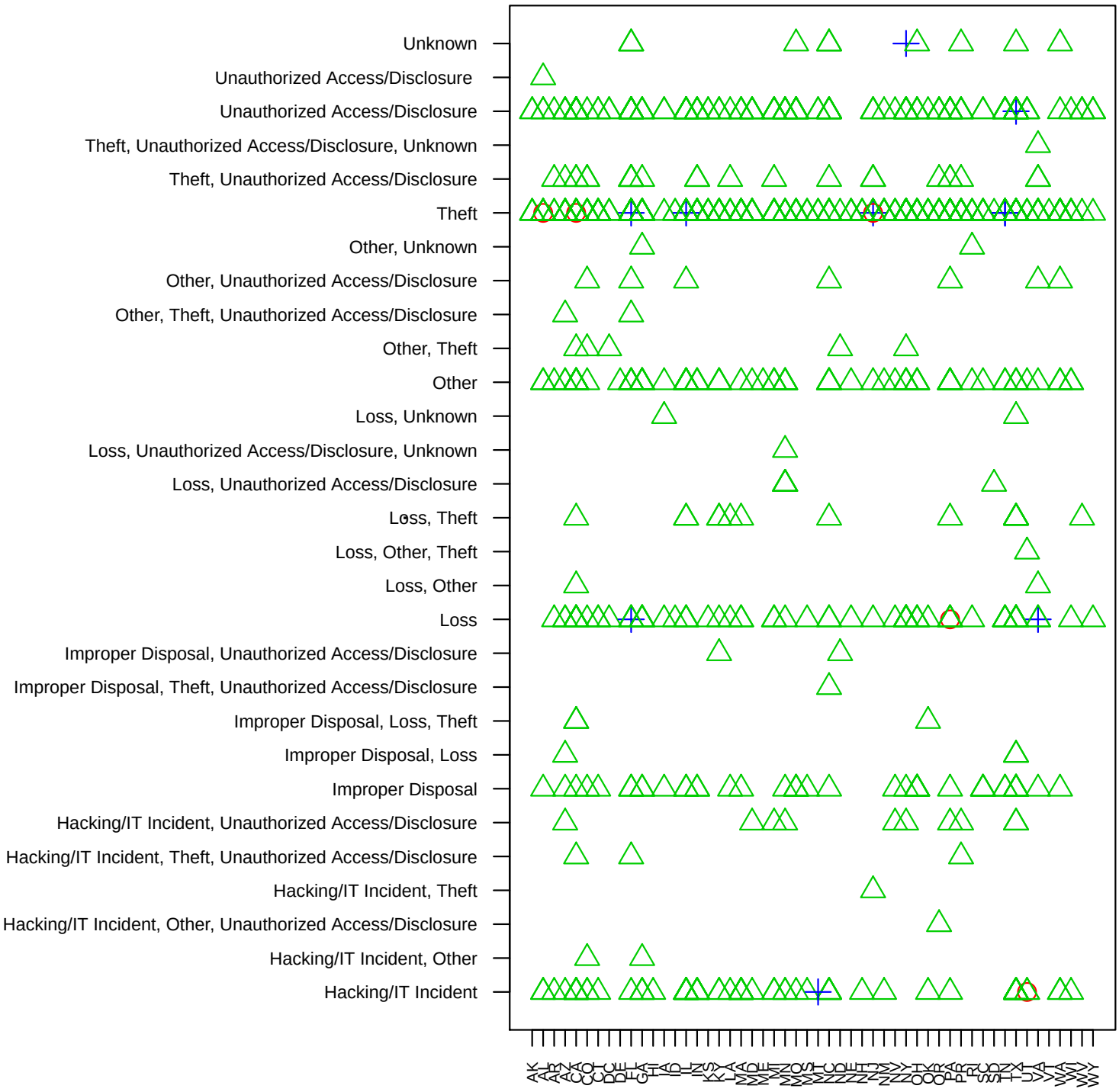

Figure 6: State versus attack type for $K = 3$ clusters and global hyper parameter estimation in kernel.

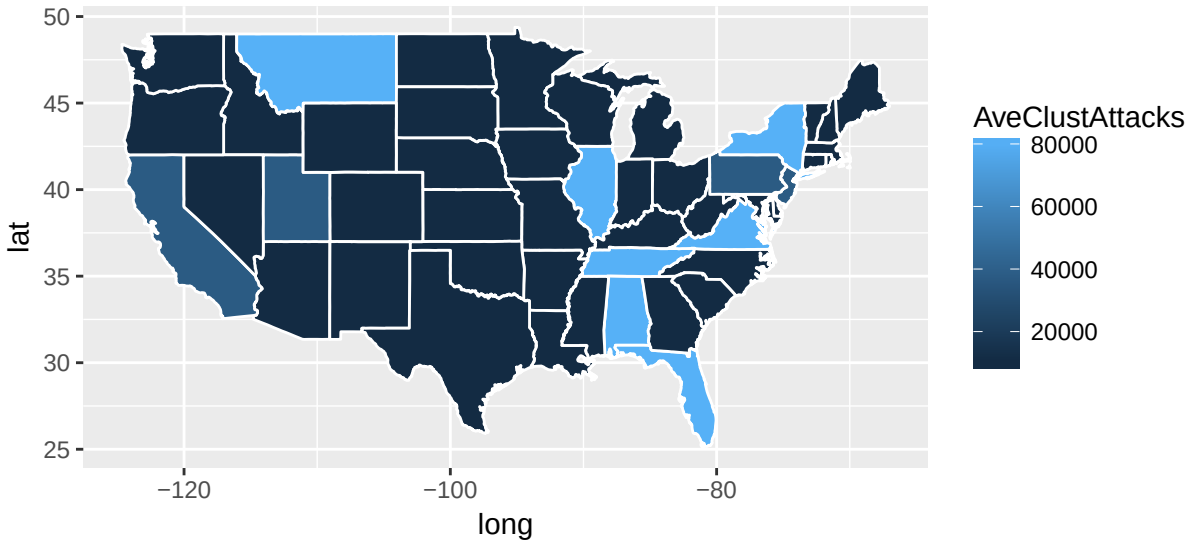

Figure 7: State versus average number of attacks per cluster $K = 3$ and global kernel parameter estimation.

5. Case Study 2: Kernel K-Means clustering of Reported Cyber Crimes in US from FBI IC3

In this section we undertake a second analysis of kernel k-means for understanding clustering of cyber crime and cyber fraud in the US. The data was collected from the annual reports of the IC3 Internet Crime Report of the FBI in the US (see `https://www.ic3.gov/default.aspx`). The years in which the data was taken from the annual reports include 2003 to 2016. We will consider the data corresponding to the following record attributes.

- Total number of cyber complaints filed or reported.
- Average loss per age per event in US.

  `Under 20,20-29,30-39,40-49,50-59,60+`

- Proportion of total complaints by state, where the states with records considered are denoted by abbreviations:

  `AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY`

Throughout this analysis we will consider $K = 2$ clusters based on some preliminary analysis. We will also study the effect of the local vs global kernel hyper parameter selection for the kernels. Furthermore, we will utilize a Matern kernel for the clustering analysis.

We begin with an analysis in Figure 8 showing the clustering of the data for a cross section of two of the attributes, the time of the event (year) vs the number of reported events total. We see the results for the Matern kernel with both global and local kernel parameter estimation. In this case, we see little difference between the cluster results for both types of kernel parameter selection. We will proceed therefore going forward with local kernel parameter selection. Next, in Figure 9, Figure 10 and Figure 11 we show cross sections of the scatter plots for attributes used in the clustering and their cluster assignments. We first see the results for number of cyber crime events versus age stratification. Then we see the results for the time (year) versus the number of complaints and the best and worst states for cyber crime events. Finally, in the third plot set we see the results for attributes of age and state combined. This provides a detailed analysis of the groupings by each of these different attributes. IC3 reports also provide percentage of the number of complaints filed with IC3 from countries around the world that we summarise in Table 4 for 2011-2016 to observe the change over time.

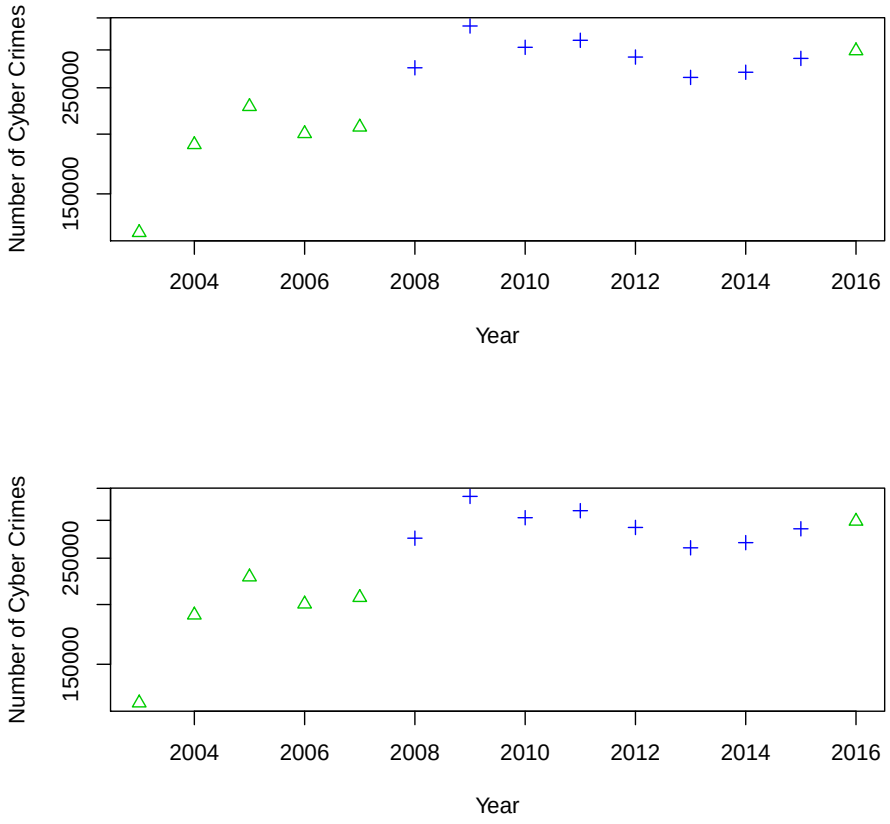

Figure 8: The analysis of clustering with Matern kernel for $K = 2$ with a cross section view of the number of events versus the year of event. Top Panel: global hyper parameter selection in kernel; Bottom Panel: local hyper parameter selection in kernel.

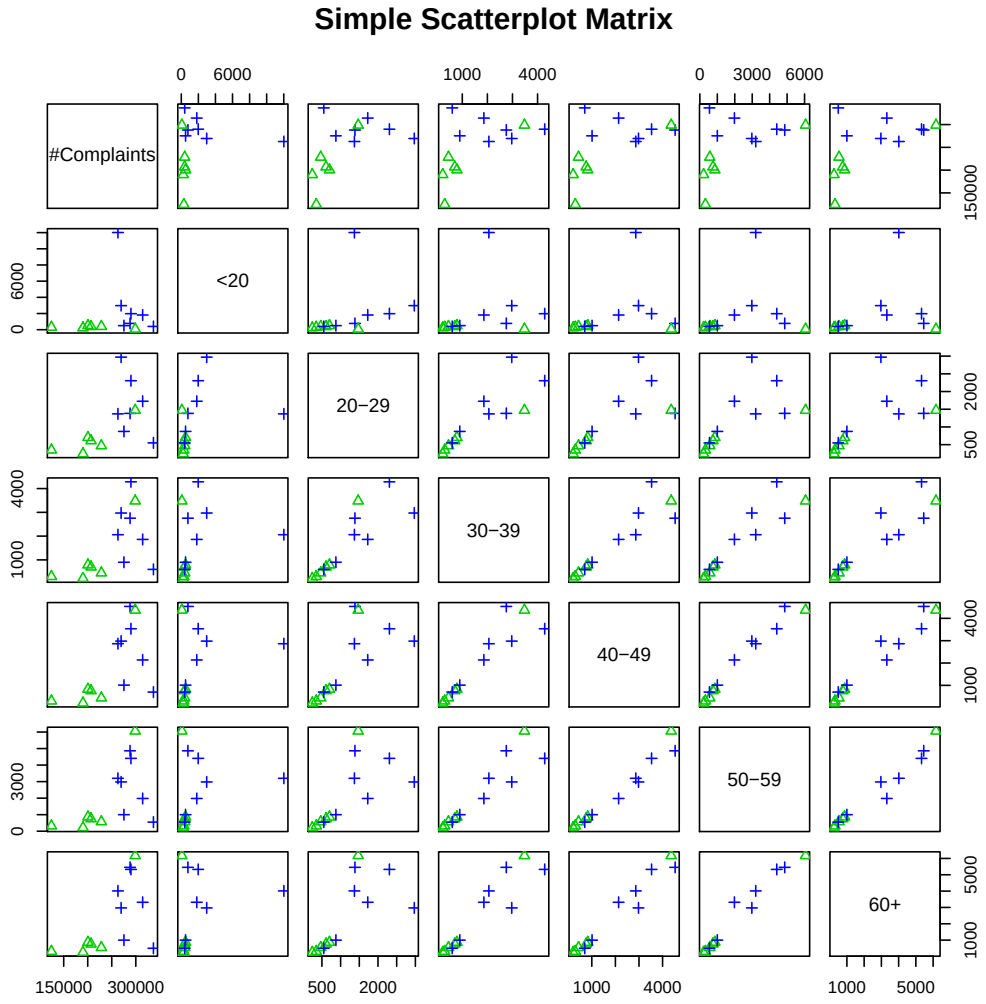

Figure 9: The analysis of clustering with Matern kernel for $K = 2$ with pairwise scatter plots of the cluster results for each of the attributes for number of events versus age stratification (local kernel hyper parameter estimation).

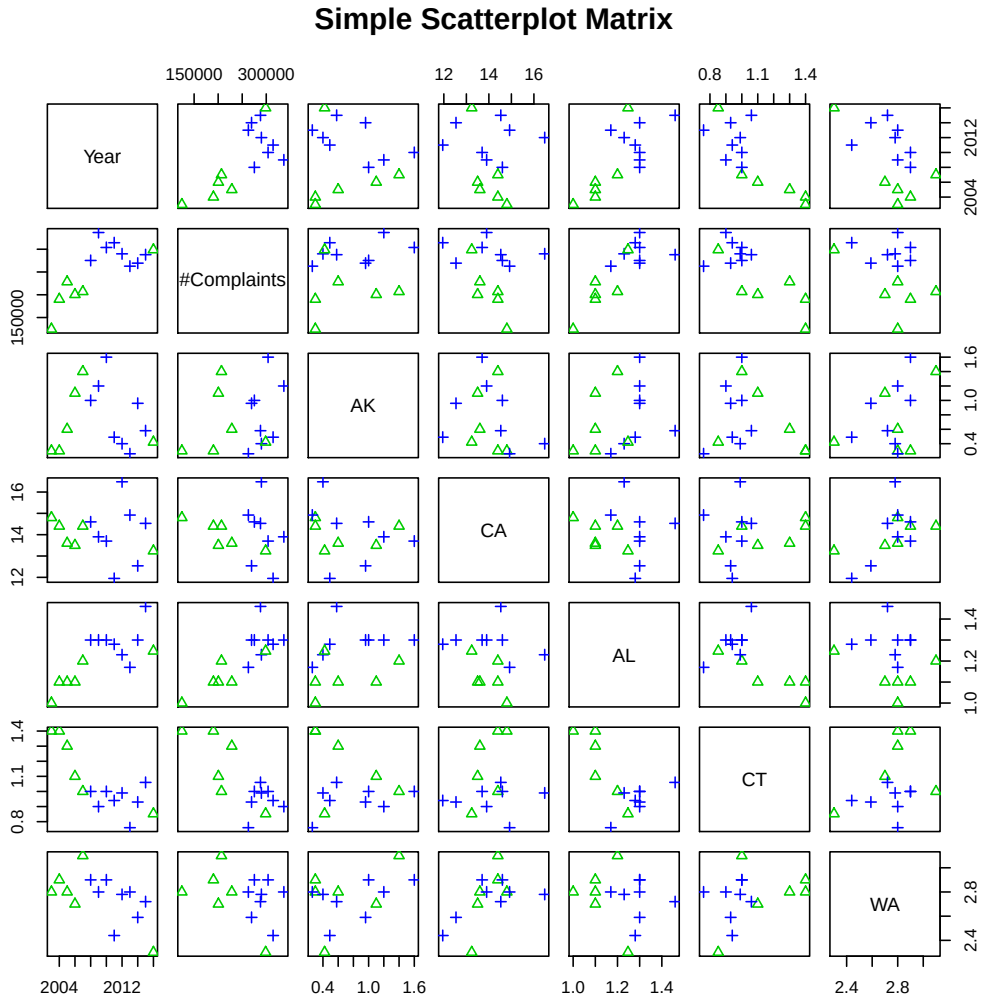

Figure 10: The analysis of clustering with Matern kernel for $K = 2$ with pairwise scatter plots of the cluster results for each of the attributes for number of events versus age stratification (local kernel hyper parameter estimation).

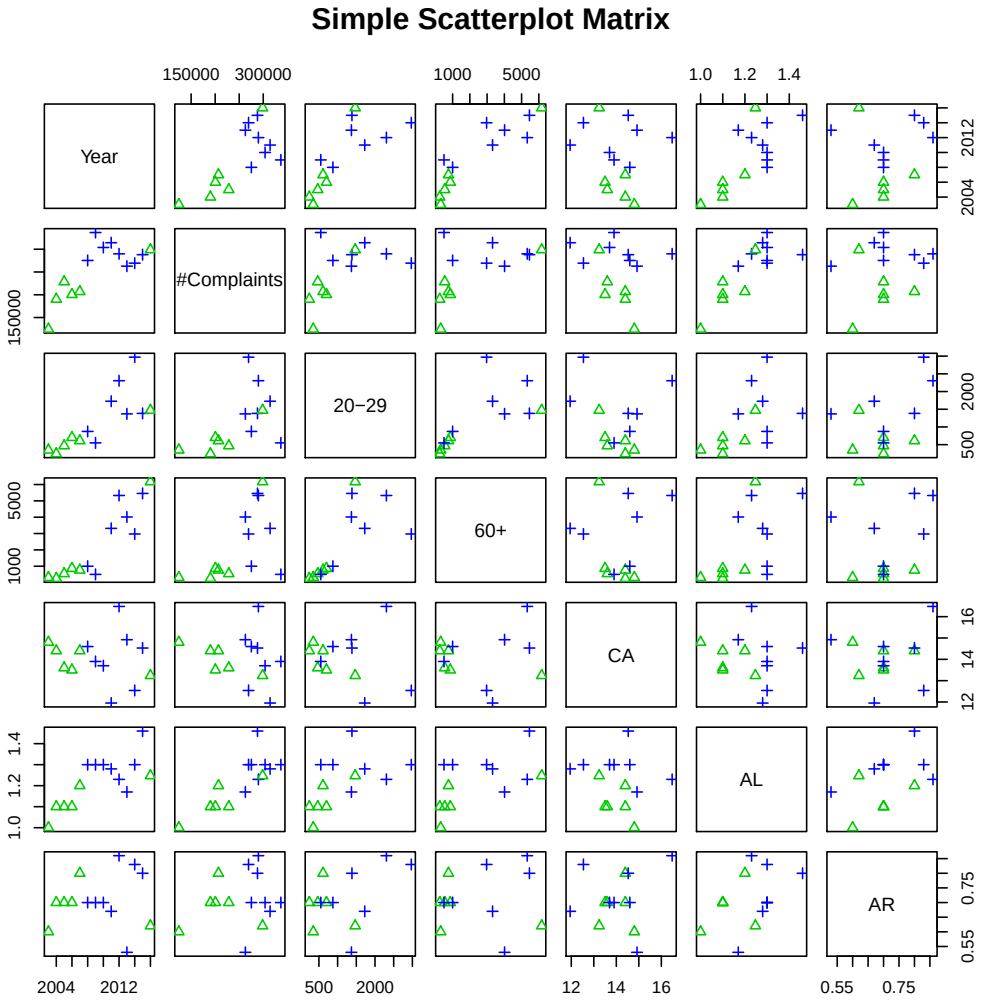

Figure 11: The analysis of clustering with Matern kernel for $K = 2$ with pairwise scatter plots of the cluster results for each of the attributes for number of events versus age stratification (local kernel hyper parameter estimation).

6. Findings and Conclusions

The results of these two case studies are the first detailed study of unsupervised learning methods applied to the empirical study of cyber risk and cyber crime event analysis to our knowledge. The approaches proposed in the study, based around nonlinear kernel k-means methods for partition-based unsupervised clustering, have clearly revealed structure in the cyber event studies, when viewed over the different attributes. Such structure would not have been discerned when standard kernel k-means methods were adopted, as demonstrated when comparing the linear kernel case to the nonlinear kernel settings.

In the event case study relating to health sector crime events, we have observed from the kernel k-means clustering that the states separate into three clusters, where the clusters with more serious numbers of cyber events in cluster 2 and cluster 3 are given for cluster group 2 (CA NJ PA UT) and cluster group 3 (AL FL IL MT NJ NY TN TX VA). The clusters selected are homogeneous in time, meaning that the temporal component is not affecting the cluster groupings. Furthermore, we can see that cyber theft and hacking incidents are what distinguish cluster groups. The plot of the US states' average number of attacks per cluster, shows that the concentration of cyber crime is primarily concentrated on the west coast.

The second case study shows that when one considers general cyber crimes in the US, the temporal component of the clustering has more effect. There is definitely now a change in cluster membership over time for different states as they attempt to tackle the increasing number of attacks using different strategies. When we observe the number of attack complaints by different age ranges, we see that there are definitely systematically two distinct cluster groups in all age brackets.

This work has explored common attributes of different types of cyber risk with a view to better understanding the key attributes that contribute to each type of cyber risk category. The application of unsupervised nonlinear kernel k-means partition clustering techniques has proven successful in developing groupings of the cyber crime event data in each case study that can provide insight to further explore and model such loss processes.

References

Böhme, R. and Kataria, G. (2006). Models and measures for correlation in cyber-insurance. In *WEIS*.

Dhillon, I. S., Guan, Y., and Kulis, B. (2004). Kernel k-means: spectral clustering and normalized

cuts. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, pages 551–556. Association for Computing Machinery.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Marin, D., Tang, M., Ayed, I., and Boykov, Y. (2017). Kernel clustering: density biases and solutions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856.

Peters, G., Shevchenko, P., Cohen, R., and Maurice, D. (November 5, 2017). Understanding cyber risk and cyber insurance. *Available at SSRN: https://ssrn.com/abstract=3065635*.

Peters, G. W. (2017). Statistical machine learning and data analytic methods for risk and insurance. *Available at SSRN: https://ssrn.com/abstract=3050592*.

Siegel, C. A., Sagalow, T. R., and Serritella, P. (2002). Cyber-risk management: technical and insurance controls for enterprise-level security. *Information Systems Security*, 11(4):33–49.

Region (%)	2011	2012	2013	2014	2015	2016
Australia	0.66	0.68	0.19	1.39	NA	0.30
Austria	NA	NA	NA	0.07	NA	NA
Afghanistan	NA	NA	0.22	NA	NA	NA
Argentina	NA	0.07	NA	NA	NA	NA
Brazil	NA	0.19	0.04	0.19	NA	0.20
Benin	NA	NA	0.16	NA	NA	NA
Belgium	NA	0.12	0.05	0.19	NA	NA
Bulgaria	NA	0.03	0.03	NA	NA	NA
Cameroon	NA	NA	0.15	NA	NA	NA
Canada	1.44	1.43	0.68	1.48	NA	1.21
Chile	NA	0.03	NA	0.82	NA	NA
China	NA	0.06	0.99	0.46	1.91	0.15
Cyprus	NA	NA	0.04	NA	NA	NA
Costa Rica	NA	NA	NA	0.06	NA	NA
Colombia	NA	0.06	NA	0.09	NA	NA
Domini. Rep.	NA	NA	NA	0.12	NA	NA
Denmark	NA	NA	NA	0.07	NA	NA
Egypt	NA	0.04	NA	0.07	NA	NA
France	0.19	0.19	0.18	0.10	NA	0.18
Finland	NA	NA	NA	0.07	NA	NA
Germany	0.19	0.15	0.23	0.39	NA	0.11
Ghana	NA	NA	0.30	NA	NA	NA
Greece	NA	0.07	0.03	NA	NA	NA
Hong Kong	NA	0.06	0.13	1.08	NA	0.07
Hungary	NA	0.03	NA	NA	NA	NA
India	0.5	0.59	0.58	0.74	1.46	0.70
Israel	NA	0.1	NA	NA	NA	NA
Italy	NA	0.11	0.10	0.09	NA	NA
Ireland	NA	0.06	0.04	NA	NA	NA
Indonesia	NA	0.07	0.09	0.16	NA	NA
Jamaica	NA	NA	0.10	NA	NA	NA
Japan	NA	0.06	0.09	0.24	NA	0.14
Kazakhstan	NA	NA	NA	0.15	NA	NA
Kuwait	NA	NA	NA	0.07	NA	NA
Macedonia	NA	0.37	0.19	NA	NA	NA
Mexico	NA	0.19	0.14	0.25	NA	0.16
Malaysia	NA	0.08	NA	0.14	NA	0.06
Morocco	NA	NA	0.05	NA	NA	NA

Monaco	NA	NA	NA	0.07	NA	NA
Mongolia	NA	NA	NA	0.25	NA	NA
Netherlands	NA	0.14	0.10	0.21	NA	NA
New Zealand	NA	0.1	NA	0.09	NA	0.06
Nigeria	NA	0.08	1.37	0.37	2.2	0.06
Norway	NA	0.05	NA	0.21	NA	NA
Puerto Rico	0.22	0.21	0.04	0.12	NA	NA
Philippines	NA	0.16	0.27	0.09	NA	0.14
Pakistan	NA	0.1	0.06	0.1	NA	NA
Panama	NA	NA	0.05	NA	NA	NA
Portugal	NA	0.08	0.03	0.11	NA	NA
Poland	NA	0.06	0.04	0.07	NA	NA
Russia Fed.	0.17	0.12	NA	0.13	NA	NA
Romania	NA	0.06	0.05	NA	NA	NA
Rep. Korea	NA	0.04	NA	0.25	NA	NA
South Africa	0.22	0.18	0.2	0.82	NA	0.11
Spain	NA	0.12	0.15	0.58	NA	0.07
Singapore	NA	0.08	0.05	0.18	NA	0.06
Sweden	NA	0.08	0.06	0.20	NA	NA
Senegal	NA	NA	0.03	NA	NA	NA
Saudi Arabia	NA	0.06	NA	0.27	NA	NA
Switzerland	NA	0.05	0.06	0.11	NA	NA
Turkey	NA	0.05	0.06	0.07	NA	0.09
Thailand	NA	0.05	0.05	0.12	NA	NA
Taiwan	NA	NA	NA	0.15	NA	NA
UAE	NA	0.06	0.09	0.36	NA	0.06
UK	0.97	0.88	1.72	1.08	2.47	0.48
US	90.99	91.19	31.89	83.96	80.2	96.22
Ukraine	NA	0.04	0.10	NA	NA	NA
Venezuela	NA	NA	NA	0.13	NA	NA
Vietnam	NA	NA	0.03	NA	NA	NA

Table 4: Percentage of cyber crime complaints filed with IC3 over time around the world.