ANOBADA
Anomaly Detection On Vehicle Operational Data

Public report

Project within: FFI – Big Data Analytics (BADA)
Author: Peter Lindskoug Scania CV AB
Date: 2017-08-23

Content

1. Summary

The project object is to apply and adopt methodology from the field of statistical anomaly detection on the accumulated operational data that is continuously collected in vehicle ECU's (electronic control unit).

The number of connected vehicles are increasing and the technique for high frequency connectivity is developing. This will lead to the possibility to, more or less, continuously (or very frequent) fetch operational data from the vehicles in the field. By applying methods for anomaly detection on the frequently received data from connected vehicles, changes in vehicle behaviour or anomalies can be detected within a short time span. A changed vehicle behaviour or anomaly could indicate that the vehicle service interval needs to be adjusted or that there are a risk that the vehicle will have a future problem.

The project investigated a number of statistical methods to see how anomaly detection could be applied on accumulated ECU data in the form of a matrix. The data matrix has inherent dependencies between adjacent elements that needs to be taken care of prior to the actual anomaly detection. The project has investigated different approaches for handling the dependencies in addition to dimensional reduction, clustering and anomaly detection. Methods that the project has investigated includes:

- Kullback-Leibler divergence
- Principal Component Analysis (PCA)
- Gaussian Mixture Models (GMM)
- Bi-variate Gaussian Mixture Models (GMMA)
- Incremental Stream Clustering for anomaly detection and classification (ISC)
- Markov Fields
- Density-based Spatial Clustering of Application with Noise (DBSCAN)

The project decided to drop Kullback-Leibler divergence early on because the adoption of the method to frequent data was deemed to be too time consuming and limit the advances for the other methods within the timespan of the project.

All investigated methods show promising results and are able to detect data readouts that are different from the majority of the data readouts that were used in the project.

PCA and clustering using GMM is a straightforward method giving results that can easily be visualized.

The GMMA method is still in development but adaptions that was done to the method during the project show that the stability of the method was increased.

Markov random fields were added to the ISC framework to handle dependencies within a data matrix.

DBSCAN was used to cluster the PCA result. The method use two main parameters as input for the algorithm. It is important to set these parameters correct to sort out noise and detect outliers. Setting these parameters proved to be a challenge and the project could not get a reasonable output from the algorithm.

2. Sammanfattning på svenska

Fordonsdriftdata sparas kumulativt i fordonens styrenheter i form av skalärer, vektorer och matriser. Data läses ut ur fordonets styrenheter vid verkstadsbesök eller kan fjärrutläsas över det mobila telefonnätet. Fjärrutläsningar av fordonsdriftdata kan göras mer frekvent och därmed kan monitorering av fordon göras. Monitorering möjliggör att förändrade fordonsbeteenden eller begynnande fel kan upptäckas tidigt.

Projektet syftar till att utveckla en statistisk metodik för klustring och anomalidetektion applicerat på fordonsdriftdata. Vektorer och matriser, från fordonsdriftdata, har en inneboende beroendestruktur mellan elementen. För att kunna göra avancerad analys på denna typ av data måste beroendestrukturen hanteras på lämpligt sätt för att få rättvisande resultat. Vi har i projektet undersökt olika sätt att hantera beroendestrukturen i, den för projektet utvalda, driftdatamatrisen. Projektet har följt projektplanen och levererat de specificerade leverablerna. Projektet har arbetat med en bred ansats där fler olika statistiska metoder utvärderas för att se vilken eller vilka metoder som lämpar sig bäst för klustring och anomalidetektion och för att hantera beroendestrukturer i fordonsdriftdata. Vissa av metoderna är välkända men även metoder i forskningsstadiet har varit i fokus.

I ansökan för ANOBADA angavs mål relaterade till FFI-området Effektiva och uppkopplade transportsystem, och den särskilda satsningen BADA. Nedan beskrivs dessa mål samt deras måluppfyllelse:

ANOBADA ska öka forskning- och innovationskapactiteten i Sverige genom att bidra till forskningsfronten i området: Nya algoritmer för anpassning av Gaussiska mixturer har utvecklats i projektet. Detta förväntas leda till en publikation.

ANOBADA ska främja samverkan mellan industri och institut: Samverkan mellan parterna har fördjupats. Industriparten Scania och institutsparten SICS har under projektet haft en nära samverkan och tätt samarbete.

ANOBADA ska utveckla skalbara algoritmer för dataanalys som kan tillämpas på Big Data i fordons- och trafikdomänerna på ett robust och kommersiellt gångbart sätt: Alla metoder som har undersökts och utvecklats inom projektet är skalbara för Big Data. Under projektet har verklig data från fordon under normala transportuppdrag använts. Därmed säkerställs att de utvecklade metoderna fungerar på ett robust och kommersiellt gångbart sätt.

ANOBADA ska utvärderas genom demonstration av projektets resultat på utläst fordonsdriftdata i en realistisk miljö: Metodiken har testats på Scania i deras miljö och med standardprogramvara som används inom Scania. Den föreslagna metoden för anomalidetektion har utvärderats med data från andra bilar än de som den har utvecklats på, och ett antal avvikelser har framgångsrikt identifieras.

Metoder som projektet valde att undersöka inkluderar:

- Kullback-Leibler divergence
- Principal Component Analysis (PCA)
- Gaussian Mixture Models (GMM)
- Bi-variate Gaussian Mixture Models (GMMA)
- Incremental Stream Clustering for anomaly detection and classification (ISC)
- Markov Fields
- Density-based Spatial Clustering of Application with Noise (DBSCAN)

Projektet beslöt att inte gå vidare med Kullback-Leibler metodiken efter initiala försök. Anledningen till detta var att vi såg att de nödvändiga anpassningar som krävdes för att hantera stora och strömmande datamängder skulle kräva mycket projekttid och projektet ville fokusera på att undersöka fler möjliga metodiker inom projektets löptid.

Av de inkluderade metoderna är kombinationen av PCA för dimensionsreducering och GMM för klustring den metodik som är enklast att utföra. Metodiken detekterar intressanta anomalier och producerar kluster som verkar intuitivt förklarande för olika fordonsanvändning.

Till ISC har metod med Markov fält implementerats och metodiken ger intressanta resultat och är lovande. Resultaten i form av anomalier är inte exakt desamma som för PCA/GMM metodiken vilket gör resultatet intressant. Det förefaller som att de olika metodikerna upptäcker olika tendenser i data och därmed känsliga för olika saker i data. Mer arbete krävs för att förstå dessa skillnader i resultaten.

Metodiken med bi-variat Gaussian Mixture Models för dimensionsreducering är också lovande. Resultaten liknar de från PCA/GMM med avseende på klustringsresultaten.

Vid försöken med klustringsmetoden DBSCAN visade det sig svår att ställa in de parametrar som efterfrågas och resultaten från metoden gav antingen en överväldigande

stor andel outliers, eller väldigt stora kluster. Vår slutsats är att metoden inte lämpar sig i vårt fall där datapunktertern i PCA-rymden formerar kluster som ligger nära varandra.

Metoderna som har använt inom projektet är alla skalbara till Big Data. En utmaning är dock att anpassa klustermetoderna till inkrementell beräkning. Detta problem har inte adresserats i projektet utan lämnas till framtida arbete.

Ytterligare framtida arbete skulle vara att använda sannolikhetsberäkningar vid tilldelningen av klustertillhörighet för driftdatautläsningarna.

3. Background

Vehicle operational data is stored in the vehicle ECU's (electronic control unit) as scalars, vectors or matrices and are accumulated to its nature. The construction of the vectors and matrices can be compared to histograms with buckets consisting of defined intervals describing a particular state for the signal of interest.

The operational data are extracted from the ECU's either when the vehicle is visiting a workshop or over the air by a remote connection to the vehicle. Remote connections to vehicles could lead to more frequent operational data readouts and hence vehicles may be monitored in a close fashion. This will lead to a setting where monitoring of the vehicle can be done and deviations from a normal utilization could be detected. Deviations from normal utilization is of interest since it could have influence on vehicle service intervals, increased understanding of repairs and quality issues. Deviations could also indicate problems with the vehicle that could lead to future quality issues. Detecting these deviations is the first step in understanding the vehicle usage and to prevent problems.

The project aims to develop a statistical methodology to be able to detect deviations between readouts of operational data. The project take start in the field of statistical anomaly detection and investigate a number of different approaches to detect deviations and anomalies.

The construction of histograms for collecting data leads to interesting questions regarding how to effectively analyse the data to be able to draw conclusions on vehicle usage or vehicle health.

4. Purpose, research questions and method

The overarching question that the project aims to answer is if it is possible to apply anomaly detection on a current readout of operational data from a particular vehicle. If an anomaly is detected, it would signal that either the vehicle is used in a new way or in an anomalous way. An anomaly may indicate that something might be wrong with the vehicle. To be able to determine what the anomaly is reflecting, other analysis of data may be needed and is not covered within the project.

In general, anomaly detection consist of different steps that needs to be taken; data preparation; clustering and determine an anomaly score.

The purpose for the project is to develop a statistical methodology for anomaly detection applied on vehicle operational data stored as histograms. The histogram that was selected by the projects has inherent dependencies between adjacent buckets. A research question for the project was to choose a methodology that could handle those dependencies. These dependencies could be handled as part of the data preparation. Dependencies could also be taken care of as part of the methodology for anomaly detection. This depends on the chosen approach and method for determining anomaly.

The project investigated different statistical methodologies that handles the dependencies between adjacent histogram buckets as well as different clustering methods and methods for determining anomalies.

The main focus of the data analysis here is to detect anomalies, and, as a step towards this, clustering of the data. Both anomaly detection and clustering is ultimately about distances - how “far away” a sample is from the other samples. The problem with using the raw vectors and matrices can therefore be described in terms of the distance metrics: the strong dependencies between neighbouring element makes a normal Euclidean space unsuitable. We need to find a better metric between the samples that compensates for these dependencies.

One approach of doing this is to select another basis with much less correlations between axes, typically of lower dimensionality than the original space to get rid of redundancy and noise. This approach is taken by two of the used methods: Principal Component Analysis (PCA) and Gaussian Mixture Model Approximation (GMMA). Another approach is to try to model the dependencies between elements explicitly, and is used by a third method: to model the dependencies with a Markov Random Field. This means that we assume that each element is directly correlated primarily with its neighbouring elements, and this correlation can then be compensated for when calculating distances.

Anomaly detection can be used in a few different ways, potentially discovering different types of anomalies. For the two methods that used dimensionality reduction (PCA and GMMA) one approach is to use the residual, that is the distance from original representation to its projection on the lower dimensional subspace. This means that a sample which is far outside of the selected subspace is considered anomalous. Note that this is different from being far away from the majority of samples within the subspace, which may occur even if the residual is zero. Figure 1 illustrates these different kinds of anomalies. The axes x1 and x2 represents the two strongest PCA components, and most sample points are positioned close to this plane in the original space. However, the purple point is positioned far away from this plane, and thus has a high residual. Note that its projection is here in the middle of a cluster of points, so this anomaly could not be detected within the lower dimensional space. The red point on the other hand is right on this plane, so its residual is zero, but it is located far away from the other points in the subspace.

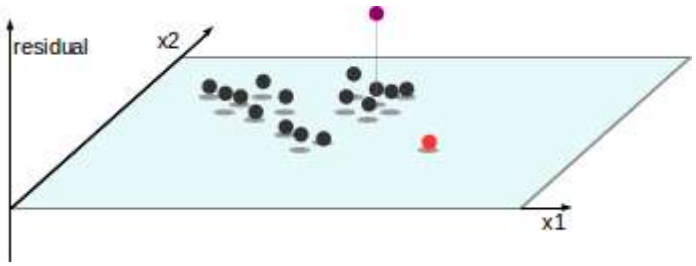

Figure 1. Graphical explanation of residuals.

To measure the distance from other samples within the subspace, it is typically necessary to use *clustering*, since there is not just one way of being "typical", and we wish to detect as anomalies samples that are far away from all of these "groups" of different kinds of typical samples. To find these groups we have tried both clustering by a Gaussian Mixture Model (GMM) using Expectation Maximization (EM) and by Density-based spatial clustering of applications with noise (DBSCAN). For the Markov field method, for which none of these clustering methods are applicable, a *Mixture of Markov Fields* trained with a variant of Expectation Maximization was used instead.

Finally anomaly detection was performed relative to the found clusters. For DBSCAN, the outlier detection in DBSCAN was used, for the Gaussian Mixture Models and the Mixture of Markov Fields, the parametric statistical anomaly detection method *Bayesian Principal Anomaly* was used.

5. Objective

As stated in the research application the main objectives for project ANOBADA are to:
Develop scalable algorithms applicable to accumulated vehicle operational data leading to advances in the ability to process and exploit Big Data within the transportation sector.
Develop a closer cooperation between research institute and vehicle industry.
Develop algorithms that are robust and commercially viable for Big Data within the transport sector.
Demonstrate the developed methodology on vehicle operational data within a realistic setting.

6. Results and deliverables

Data collection in vehicles and the ability to retrieve data on a frequent basis opens opportunities to monitor vehicles on a continuous basis. The purpose of developing methodologies for anomaly detection on vehicle data is to capture vehicle behaviour that stand out from normal and to do that in a timely manner. An odd vehicle behaviour could indicate that the vehicle is used in a way which it is not suited for which could lead to unplanned stops due to vehicle failures. Through the project, and the selected methods for anomaly detection, we try to adopt a statistical methodology that take advantage of the current way of collecting data and the future possibility of frequent remote data readouts. Ultimately, the ANOBADA project aims to quickly and effectively transform the accumulated vehicle data into information describing vehicle usage and signal if a vehicle behaviour is out of the normal.

The project group choose to have a broad approach in the selection of methods for anomaly detection. Each selected method are presented and the resulting outcome is discussed below.

The method using PCA/GMM was demonstrated for the ANOBADA steering group and for a Scania internal audience. The demonstration used a random selection of vehicles that was not used for the development of the methodology. The demonstration showed that the methodology can detect operational data that are dissimilar.

6.1 Kullback-Leibler Divergence

The idea with using Kullback-Leibler divergence as a means to measure the distance between matrices is that they can be considered as probability distributions of in which state the vehicle may be at each moment. Thus, rather than using some weighted Euclidean distance in an N*M-dimensional space, it may be more appropriate to compare the matrices with the Kullback-Leibler divergence. This distance measure can then be used for both anomaly detection and clustering.

However, this idea was only preliminary tested and then abandoned. The reason is that it in itself still does not consider the dependence between elements in the matrix. Such dependences must be covered by the statistical model used to describe the probability distribution over a matrix, and the suggestions for how this can be done are the same as discussed in relation to the other methods. The additional value of using the Kullback-Leibler metric on those models were considered small compared to the work required to do it. Therefore focus was instead put on other methods.

6.2 PCA/GMM

A traditional way to handle high dimensional data with strong correlations is to use Principal Component Analysis to replace the original axes with a set of de-correlated axes. The axes with the smallest variation can then be considered noise and removed from further analysis, thus also reducing the dimensionality of the space. In this project, the seven first components were addressed, which covers over 90% of the total variation.

Figure 2 plots the resulting PCA, principal component 1 and principal component 2 from a total of 9043 operational data matrix readouts. It can be seen that the readouts form structures within the reduced PCA-space. Using GMM as a method for clustering the PCA results, and forming four clusters, the PCA results are shown in figure 3.

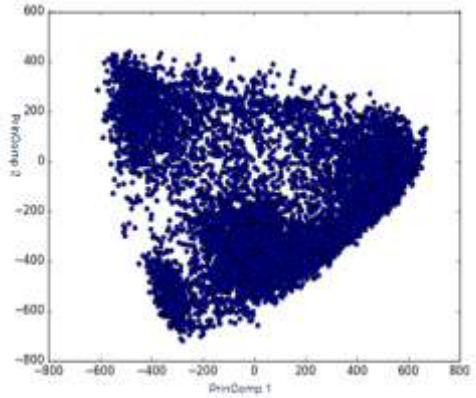

Figure 2. PCA-component plot.

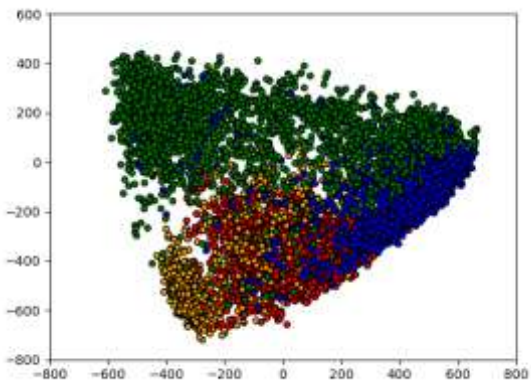

Figure 3. GMM clusters on PCA plot.

The data material used were weekly readouts during a two year period for a number of vehicles. For a particular vehicle the PCA/GMM method could detect the cluster identity for each week and hence we could plot the cluster identity and cluster change over the time period. Figure 4 shows the cluster identity for a vehicle. It can be seen that the cluster identity for this vehicle changes about midway through the time period. Figure 5 displays the position in the PCA space for each readout. Cluster id number 3 is represented by the green dots and cluster id number 1 represented by blue dots.

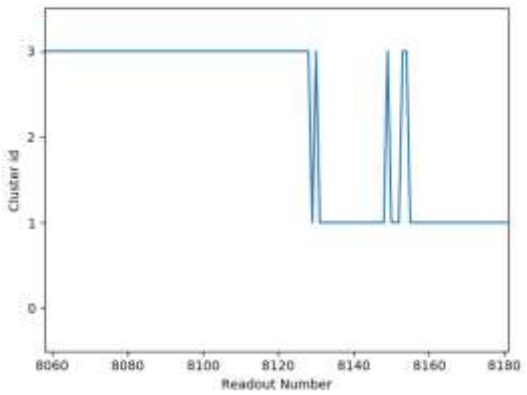

Figure 4. Cluster id change for a vehicle over time

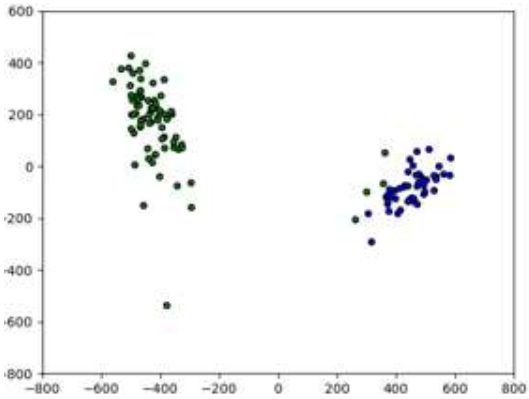

Figure 5. Readout position in PCA-space

Figure 6 is a plot over the residuals from the resulting PCA using seven components. It can be seen that two high residuals are present at observation 993 and 3110.

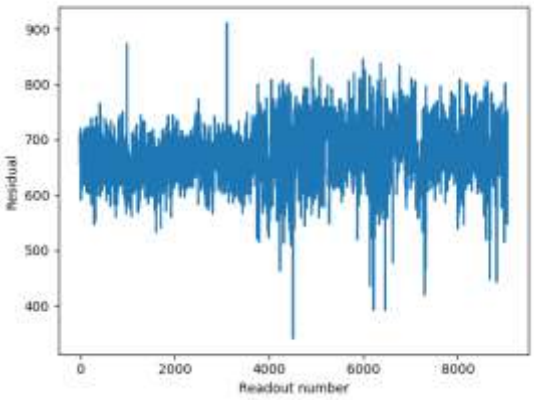

Figure 6. Residual plot from PCA.

Figures 7 below are contour plots of the data matrix for the vehicles with the highest residual from observation 3110 in the residual plot. When comparing the weekly readouts it can be seen that the position of the peaks have shifted and even disappeared in the second readout. The readout from 30NOV2015 which corresponds to the PCA residual (observation 3110) is clearly different from the readout from the same vehicle the week before 24NOV2015.

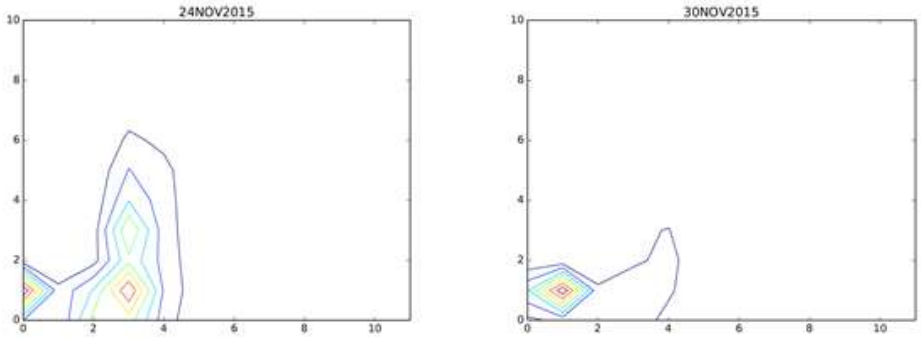

Figure 7. Contour plot of data matrix from two consecutive weeks.

Using PCA and GMM on the resulting PCA-space we have been able to detect cluster changes for vehicles which indicate that the vehicle performed in different ways on the consecutive weeks. The distance between the clusters may also indicate how different the driving style has been. The residual analysis from the PCA also indicate that high residuals identifies odd vehicle performance.

6.3 Gaussian Mixture Model Approximation (GMMA)

An alternative to use PCA for dimensionality reduction is to use some other set of "basis vectors" for the space, especially selected for the domain at hand. If the basis is appropriately selected, such as for example that the components has some physical meaning, then such a basis may be more useful than the principal components. In the selected matrix (and similarly in the other matrices and vectors on a vehicle that are produced in the same way), the profile plotted as a two-dimensional histogram appears to consist of a number of soft hills (Figure 8. Matrix Histogram).

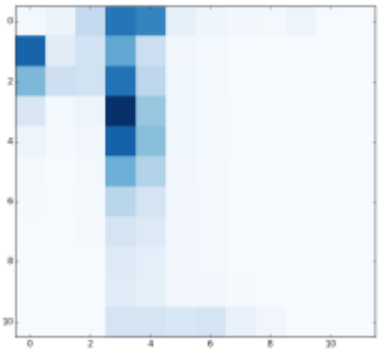

Figure 8. Matrix histogram.

It can be speculated that each of these hills correspond to a specific driving condition, for example one for idling, one for high load driving, one for flat fast driving, etc. Each such hill is spread over several neighbouring elements of the histogram, since there is some random variation in the load and speed in each those driving situation (except possibly idling which is rather concentrated in one element). So the hypothesis is that the driving profile in the matrix is made up of a weighted sum of several such components each approximately Gaussian shaped (if the variation in each driving situation is approximately Gaussian). The weight for each component would correspond to the proportion of time the vehicle has spent in that driving condition, making these weights a good indication of how the vehicle has been used. This leads to the idea of using Gaussian shaped components representing these driving modes, instead of PCA components, to make up the total matrix.

This amounts to finding a two-dimensional Gaussian mixture model that fits the sum of all the matrices (assuming that all vehicles have the same driving modes positioned at the same element in the matrix). This can be done by Expectation Maximization (EM), where the observations are the coordinates of each histogram element, weighted by the value in that element.

The project implemented a GMMA and an EM algorithm in Python 2.7. Upon running the algorithm several times with various run-time condition (such as Log Likelihood difference threshold for convergence and number of components), the algorithm managed to find models that capture the ‘soft hills’ that are visually apparent. The components of such a model can then be used to form a basis for a subspace of the original space, similar to the PCA dimensionality reduction described above. However, a well-documented shortcoming of the EM-algorithm for GMMs is its instability: it is not guaranteed to reach the same solution when run again on the same data. Also, it is inconvenient to have to determine the number of components manually. Thus, in order to use this approach, we needed to adjust the algorithm to get rid of these shortcomings first.

Approaches to improving the stability of EM is a current and active research topic. In the project we made an approach to an evolutionary variant of EM on GMM. The approach is promising and has shown early indications of stability. We describe the approach next.

An initial population of N GMMs are allowed to converge using EM. Of these, the best K GMMs are allowed to survive, i.e., the K GMMs with the highest Log Likelihood with respect to the data. (K=N/3) The next generation consists of the K best ones together with two mutations of each of those: the first mutation removes one random component, and the second mutation splits one random component into two. Thus from generation to generation we allow the number of components to vary depending on the fitness of the GMM models. To avoid the well-known condition that a higher number of components always tend to fit the data better, we also add a small punishment factor, logarithmic in the number of components. This process converges when the K best GMMs are better than their mutations (after running EM to convergence again). The experiments indicate that the evolutionary EM is much more stable than our first variant: it converges to a specific number of components placed roughly in the same places from run to run. Figure 9 below illustrates the three best individuals after convergence of a run of the algorithm. The seven components of the three GMM:s are visible and seem to have captured the significant characteristics of the data.

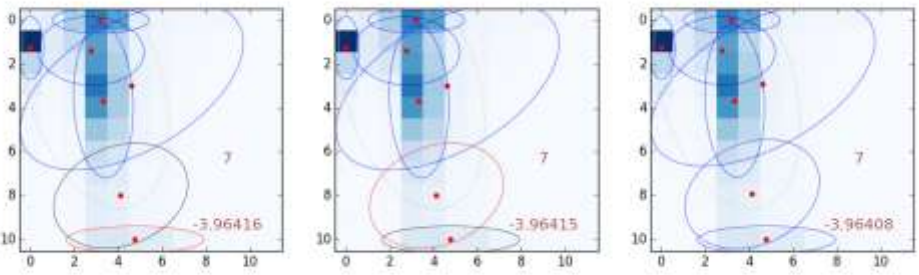

Figure 9. Results of the improved EM.

6.4 Gaussian Clustering with pyISC

The ISC is an anomaly detection and classification framework in C++. The pyISC framework is a Python wrapper and extension to the ISC framework.

ISC is available as open source at http://github.com/sics-dna/isc2 and pyISC is available as open source at http://github.com/stream3/pyisc. Software related to the ISC that is developed in ANOBADA goes into the above two repositories. Documentation for installing pyISC is at https://github.com/STREAM3/pyisc and documentation on how to use it is available as an interactive tutorial (requires jupyter (http://jupyter.org)) or as a webpage at https://github.com/STREAM3/pyISC/blob/master/docs/pyISC_tutorial.ipynb.

The pyISC clustering algorithm for creating *k* number of clusters works basically as follows:

1. Divide the data into *k* distinct clusters
2. Train the model to classify the data according to the cluster membership
3. Use the trained model to reclassify the data and thereby create new cluster memberships
4. Repeat steps 2. and 3. above until no memberships are changed

This is similar to how the k-means algorithm works, which is trained by forming cluster centres from the mean of the instances and then can classify instances with respect to the distance to the closest cluster centre. The above approach fits a statistical model and then classify according to the class with the highest probability.
Two types of clustering methods have been implemented in pyISC: 1) anomaly clustering and 2) outlier clustering. The difference is that in the second type, we can provide a fraction of most anomalous outliers that should be excluded from the model. By removing the outliers, we can get a more robust statistical model.

6.5 Markov Random Field with pyISC

In the used Markov random field, features are assumed to only directly influence the closest neighbours among the features, in this case elements in the matrix. The data are assumed to have a multivariate Gaussian distribution, where the covariance matrix reflects that each matrix element is only influenced by its neighbours.

Mathematically the probability of a matrix *M* using the above assumptions can be written as follows:

$$p(M) = \prod_{i=1}^{n}\prod_{j=1}^{m} p(M_{ij}|\, M_{i+1\,j}M_{ij+1}) \prod_{i=1}^{n+1} p(M_{1,j+1}|M_{i+1,j+1}) \prod_{j=1}^{m} p\big(M_{mj}\big|M_{mj+1}\big)p(M_{n+1,m+1})$$

where $M_{i\,j}$ denotes the slot at row *i* and column *j*.

The Markov random field has been implemented in the pyISC framework, which can easily be created as an anomaly detector

6.6 DBSCAN

Our approach to use DBSCAN were twofold. We wanted to use it to find the amount of clusters for our data set and to group the readouts. Another additional effect would be to consider the remaining outliers as anomaly. Since we wanted to compare the outcome from the different clustering algorithms, the focus of using DBSCAN was to try to have the same amount of clusters as other algorithms, rather than as anomaly detection.

For our use case and in practice, the trucks do not drive distinctly in one type of application, but in many type of applications in different proportions. This makes it difficult to use DBSCAN as a clustering algorithm as the data points blend into each other. DBSCAN might be better suited to only find readouts that stands out from the rest of the "common use", i.e. as anomaly detection, for our use case. Few tests have shown that the two points with high residuals from the PCA are captured as outliers when using DBSCAN in the PCA coordinates. However, further tests need to be conducted.

7. Dissemination and publications

7.1 Dissemination

How are the project results planned to be used and disseminated?	Mark with X	Comment
Increase knowledge in the field	X	The project results has increased the knowledge within Scania about statistical methodologies for anomaly detection. The pyISC program has been updated according to project findings.
Be passed on to other advanced technological development projects	X	The development done on bi-variate Gaussian Mixture Models will be included in ongoing project (BIDAF) which will lead to future publications.
Be passed on to product development projects	X	The project results has invoked great interest from different department within Scania CV AB. This interest will likely be formulated into future development projects between Scania and RISE SICS.
Introduced on the market		
Used in investigations / regulatory / licensing / political decisions		

7.2 Publications

The work in the project on finding ways to stabilize the EM algorithm in high dimensional spaces (as described in Section 6.3 above) is relevant in a broader scope than the Anobada project. The approach is being further developed in a second ongoing project (BIDAF, funded by KKS) and we plan to submit a joint Anobada/Bidaf manuscript for publication of results during the fall of 2017.

8. Conclusions and future research

Of the investigated methods, the combination of using PCA for dimensionality reduction and GMM for clustering, is the most straightforward to use. It manages to detect interesting anomalous usages of vehicles, and it produces clusters that are intuitive and can be connected to the profiles of the vehicles.

The Markov random field method, as implemented in the pyISC package, also provides interesting results, and is therefore promising. The results are not exactly the same as the PCA/GMM method, which is what makes them interesting - there are more than one way in which a vehicle may be anomalous. More work is however required to investigate how this difference can be characterized and the significance of the detected anomalies and clusters.

The bi-variate Gaussian Mixture Models for dimension reduction (GMMA) also looks promising. The results are similar to those of PCA when it comes to clustering, but the residuals are slightly different. Which one is better is a matter of how the found mixture components can be interpreted physically, and thus what the residuals (and detected anomalies) represent. However, this needs further work to establish.

The DBSCAN clustering method is difficult to use since the parameters that needs to be specified is difficult to set appropriately to avoid a large portion of the readouts to be declared as outliers or alternatively everything joined into one cluster. The separation between different clusters in the space is just not large enough for this kind of method to be useful.

The approach of using the Kullback-Leibler distance to measure the distance between matrices was abandoned early on in the project. The approach is viable to use for anomaly detection, however the project decided not to develop the method further because of limited resources, and because it was judged not to add much compared with the other methods.

The three first methods mentioned above are all worth continuing with. The first one, PCA/GMM, is the most "ready" one for real application, whereas both the GMMA

methods and the Markov Field method requires some refinements and more experiments to see whether the detected anomalies and clusters are relevant and intuitive.

One objective in the project was that the selected methods should be applicable in a big-data setting, which here means that: the models should not grow, or grow very slowly, as the amount of data increase; the methods should be tolerant to noise and variability in data, such as in this case highly varying sample times; and the methods should work on streaming data and not require batchwise running.

The two first criteria are met by all three primary methods mentioned above: they are all parametric models so the model size do not grow with the data; and the preprocessing with normalization will make varying sample times possible to handle. Regarding the third criterion, it is only partially fulfilled by the methods in their basic forms. The main complication is the clustering algorithms which all require iteration to convergence. It is possible to formulate the clustering algorithms in an incremental manner, but this has not been tested in this project. This is left for future work.

We have for simplicity in the results above considered hard categorization of each vehicle reading to its most probable cluster. However, the found clusters are not very distinct but overlaps each other, and the finally selected clustering methods are all probabilistic in that they give probabilities of a reading to belong to each cluster. It can be noted from the results that some vehicles jump back and forth between clusters, which may be a result of using hard categorization rather than probabilistic - they may vary quite little in behaviour but be positioned on the border between two clusters. A third route to explore in the future is to use probabilistic categorization and thus instead describe each vehicle in terms of in which proportion it belongs to different clusters. This may give a better characterization of the vehicles, and also opens up for detecting gradual shifts in vehicle usage.

9. Participating parties and contact persons

Scania CV AB

Peter Lindskoug

REIA
Scania CV AB
151 87 Södertälje
Sweden

Tel: +46-8-553 831 51
Email: peter.lindskoug@scania.com

RISE SICS AB

Björn Bjurling

RISE SICS AB
Box 1263
164 29 Kista
Sweden

Tel: +46 70 775 15 89
Email: bjorn.bjurling@ri.se