Scalable Data Analytics,
Scalable Algorithms, Software Frameworks
and Visualization ICT-2013 4.2.a

Project **FP6-619435/SPEEDD**
Deliverable **D8.5**
Distribution **Public**

http://speedd-project.eu

Intermediate Evaluation Report of SPEEDD Prototype for Traffic management

Chris Baber, Sandra Starke, Natan Morar, and Andrew Howes (University of Birmingham)
Alain Kibangou (CNRS)
Marius Schmitt, Chithrupa Ramesh, John Lygeros (ETHZ)
Fabiana Fournier (IBM)
Alexander Artikis (NCSR)

Status: Revised version

August 2016

Project

Project Ref. no	FP7-619435
Project acronym	SPEEDD
Project full title	Scalable ProactivE Event-Driven Decision Making
Project site	http://speedd-project.eu/
Project start	February 2014
Project duration	3 years
EC Project Officer	Stefano Bertolo

Deliverable

Deliverable type	Report
Distribution level	Public
Deliverable Number	D8.5
Deliverable Title	Intermediate Evaluation Report for traffic management
Contractual date of delivery	M26 (March 2016)
Actual date of delivery	April 2016
Actual date of delivery of the revised verion	August 2016
Relevant Task(s)	WP8/Tasks 8.5
Partner Responsible	CNRS
Other contributors	UoB, NCSR, IBM, ETHZ
Number of pages	31
Author(s)	C. Baber, S. Starke, N. Morar, A. Howes, A. Kibangou
Internal Reviewers	NCSR
Status & version	Revised
Keywords	Evaluation, User Interface Design, Human Factors, Eye Tracking

Contents

Executive Summary

This deliverable reports the evaluation of various components of SPEEDD prototype for road traffic management use case. Version III of the user interface was presented to DIRCE control room operators and their comments invited. On the basis of these comments, the design was revised to version IV which will be used in the second SPEEDD prototype. In addition, discussion with the operators also focused on the relationship between the use cases used in SPEEDD (specifically relating to ramp metering and demand management) and their current work. A problem that we faced is that Grenoble has yet to implement the ramp metering system that it had planned. This means that the operators remain unsure as to how their work will be affected by ramp metering. They assume that the system will be entirely automated and that, as a result, they will have little opportunity for intervention. Given the assumptions that informed the design of the user interface (i.e., that ramp metering rates could be 'trimmed' or adjusted by operators); this made it difficult for them to fully relate to the design. In order to determine how the User Interface (UI) designs developed in the SPEEDD project have evolved, an experiment has been conducted in which participants perform a series of tasks using the initial prototype (described in D8.1) and the most recent prototype (described in D8.5). The aim of this evaluation was to compare the two UI designs in terms of their usability and their impact of user performance. The Event Processing Component have been also evaluated using both real data and simulated data from the Grenoble traffic network simulator. Finally quantitave evaluation of the real-time capability of the Decision making component has been carried out.

1 Introduction

History of the Document

Version	Date	Author	Change Description
0.1	22/02/2016	Chris Baber	First version of the document
0.2	23/02/2016	Alain Kibangou	Review and additional material
1.0	24/02/2016	Chris Baber	Second version of the document
1.1	20/04/2016	Alain Kibangou	Final version of the document
2.0	26/07/2016	Alain Kibangou	Inclusion of quantitative evaluation of all the components of the SPEEDD prototype.

Purpose and Scope of Document

The purpose of this document is to report an interim evaluation of the SPEEDD prototype in the Road Traffic Management use case. In terms of evaluation, the aim is to show how the prototype is evolving. The target audience of this document will be all parties involved in the implementation of the road traffic use case.

Relationship with Other Documents

As noted in the previous section, this document is related to the following deliverables: 8.1 User Requirements, 8.3 Initial Evaluation Report, D5.1 Design of User Interface for SPEEDD Prototype, D5.2 Design of User Interface for SPEEDD Prototype, D4.2. Second version of real-time decision-making technology, D3.2 Second version of event recognition and forecasting technology.

Sources of Information

In this report, the most recent user interface design for the SPEEDD road traffic use case was presented to DIRCE control room operators. They were invited to comment on the design and to explain how they could imagine using this in their work.

In the Description of Work for Work Package 8, it is proposed that "Every version of the integrated prototype will be followed by technical and user-oriented evaluation to obtain the necessary feedback for the functions to be included or altered in the next version". Over the course of the previous 12 months, the user interface for SPEEDD prototype has undergone several changes. In part these changes have been motivated by the need to support the tasks in the use case, i.e., ramp metering and monitoring, but more significantly, the changes have reflected our interpretation of the requirements of the control operators.

2 Evaluation by Subject Matter Experts

2.1 Version III of SPEEDD traffic user interface

As part of the evaluation exercise, operators from DIRCE Grenoble were shown a demonstration of the user interface version III. An annotated version of this is shown in Figures 1 and 2. The ring road (Rocade Sud) is shown on the left-hand panel. This is the coloured line running from right to left. Around the ring road is a set of concentric circles. Each circle is divided into segments, and each segment contains a bar graph. The segments are connected to the locations of ramps in the ring road and the bar graphs indicate the rate of change of the traffic control, the percentage occupancy of the ramp and the speed of vehicles at that location. The intention is for the graphs to change as the data are arriving from the sensors. The inside circle represents the current state of the road. The other two circles represent a 'historical' state, e.g., the state of the road at the same time last week or last month, and a 'predicted' state, e.g., what the sensors might indicate in one or two hours time.

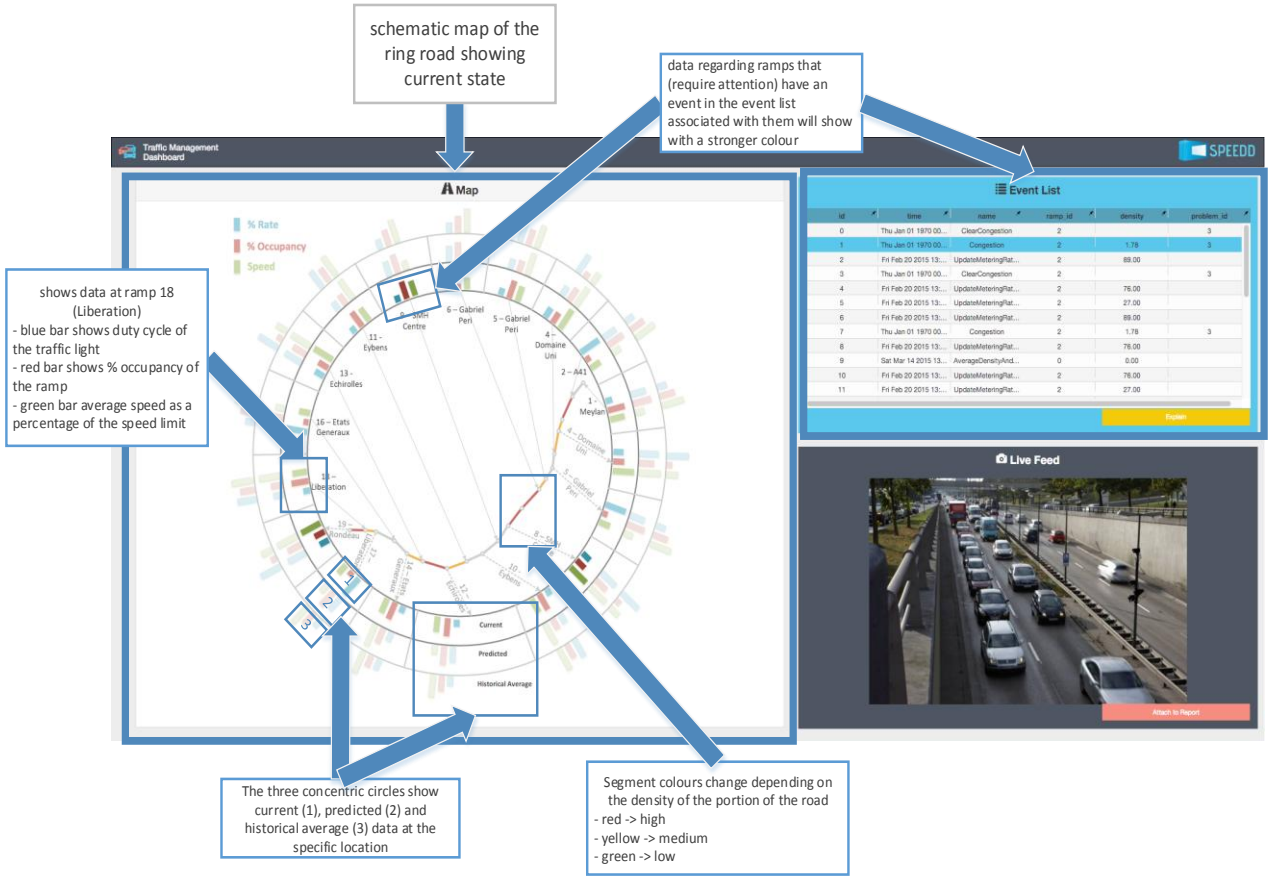

Figure 1: Version III of SPEEDD road traffic user interface

On the right of the display (Figure 2), there is a table showing events that an automated system has flagged for the operator to deal with and, below that, footage from one of the CCTV cameras.

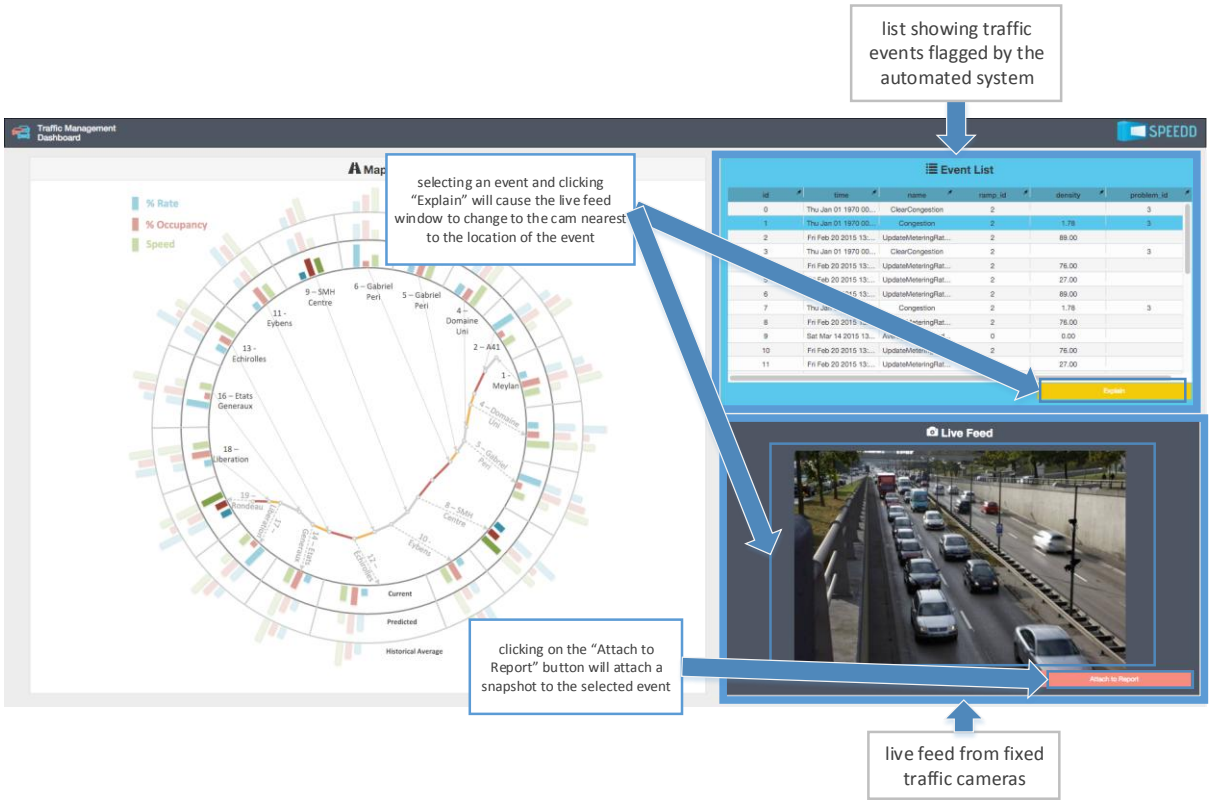

Figure 2: Version III of SPEEDD user interface

2.2 Comments from Operators

The operators found the circular design confusing at first. This is not something that they had seen before. We felt that this was positive, since the aim of the user interface design for year two was to produce a novel configuration. Of course, if the novelty leads to reduced understanding then this would be problematic. From discussion with the operators, we highlight issues that require resolving.

2.2.1 Geographical Information

Figures 1 and 2 only show the ring road. The lack of geographical context in this interface was felt to be a problem for two reasons. First, the operators need to be aware of the other motorways running north to south, with a junction as the western end of the ring road, and running west to east, with a junction at the northern end of the ring road. While they have no control over these motorways (which are managed by operators of a private company in other control rooms), it could help to see when congestion from north or east might spread. However, they have no means of seeing the congestion on these roads because they do not have access to the CCTV placed on them and they don't have access to

sensors in these roads, so it is not easy for them to directly view such congestion. If geographic information is not used, it might be more appropriate to replace the road with a schematic and the circles with a linear display over the top of the road. In this way, the relationship between sensor data and ramp could be explicit. However, this would lose the geographical context of the display and thus the link to the other motorways could be lost. The linking of the graphs to locations was not immediately obvious. We pointed out the arrows linking graph to ramp, but the operators felt that this would involve effort to interpret as it was not clear whether the arrow was pointing to a ramp or to a section of the road.

2.2.2 Managing Ramp Metering

The operators struggled with the management of ramps on and off the ring road. Their immediate concern related to the role that the operators might play in ramp metering. For example, if the system was fully automated, then the only intervention that they could make would be to turn off the automated system. This raises the question of how best to integrate operators in the decision loop. They also felt that the ramp metering could be focused on isolated ramps but that they would need to understand the consolidated impact across ramps, i.e., in terms of appreciating how activity on one ramp could affect activity before or after this ramp on the ring road. As congestion could extend from the ramps back into the city, there might be problems with managing the efficiency of the system in order to maintain balance across ring road and city. It is likely that the ramp metering system would be separate from the traffic control in the city since two different authorities are in charge of these two networks (peri-urban and urban).

2.2.3 Using bar graphs to indicate sensor data

The use of bar graphs to indicate multiple dimensions was potentially confusing for several reasons. First, the parameters in the graphs are not typically used by the operators. Second, presenting all of the values implies that all are of equal value. We pointed out that the fading of some of these graphs was an attempt to reduce their relative importance, so that key information would stand out for the operators. If the graphs are showing different values between the current, historical and predicted states (as is intended in the design) then the operators are not sure which represents the 'correct' version and how they are meant to interpret these differences. One suggestion was simply to use a single indicator (either a colour or a length of bar) to show change in state. The operators were concerned with the concept of 'historical' or 'predicted' states. They could accept that there could be nominal congestion at certain times of day but noted that these would vary according to other events or to weather conditions. They were sceptical that it would be possible to produce a 'normal' pattern to serve as the reference here. In terms of 'predicted', the operators were not sure how a model could be developed that could accurately predict the state of the road.

2.2.4 Overall impression

The operators felt that the user interface presented several interesting ideas. Their concerns lay in the accuracy of data presentation (particularly in terms of historical and predicted states) and in the effort that they felt would be required to interpret the information displayed to them. They noted that much of the state information is available through CCTV and through the automated alert system that they

now use. This is discussed in the next section. They felt that having some indication of occupancy of the ramps could be useful, e.g., in terms of knowing whether there was any spare capacity in the system.

3 Development of User Interface Design

Following the discussion with the operators, the user interface design was revised. Initially, the revision involved overlaying the circles over a map of the city. This was intended to provide geographic context and to highlight the position of the other motorways. In Figure 3, the ring road is shown in blue. This is as it appears on the map of the region. Below this blue line, a red-orange-and-grey line is depicted. This line was the original schematic used in version III. Figure 3 shows several problems with this arrangement. First, the two roads are not aligned and so there is a need to change the schematic. Second, and more importantly the arrows linking ramps to graphs become confused. This is particularly apparent in the regions indicated by the red squares where the lines pointing to on or off ramps overlap.

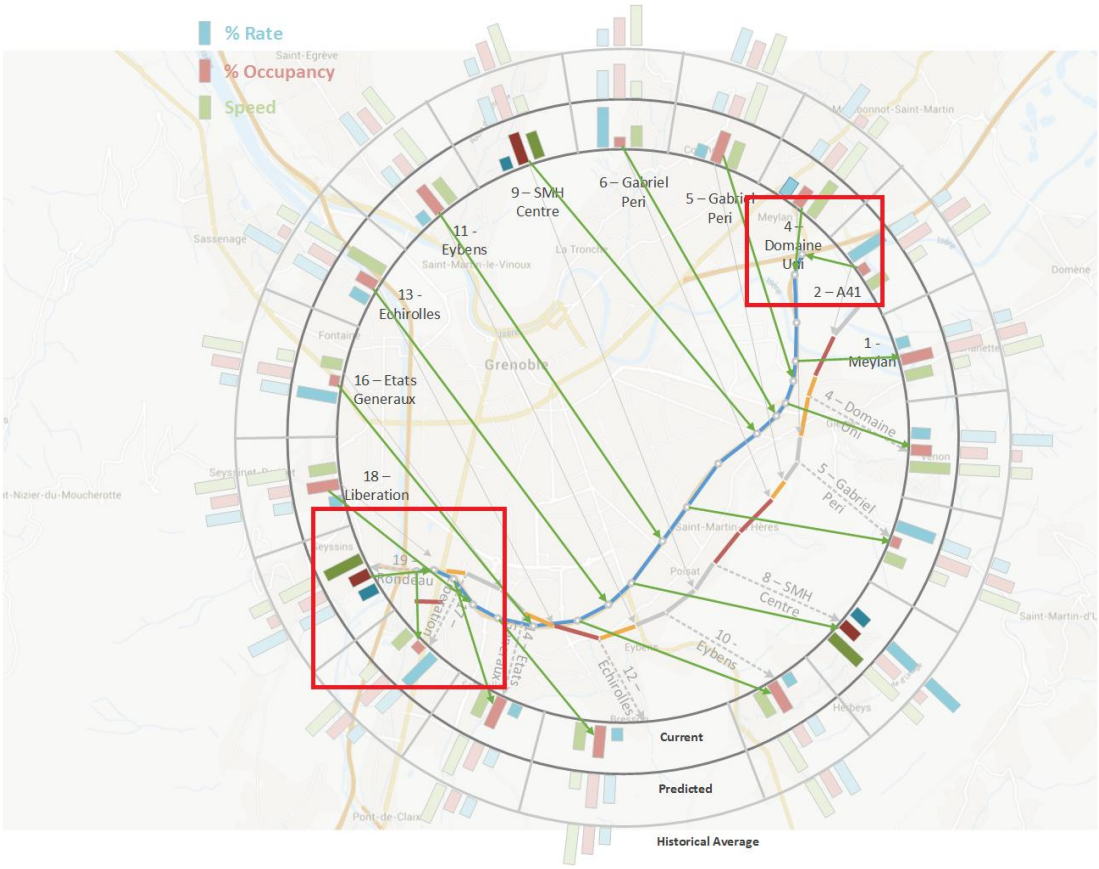

Figure 3: Overlaying the user interface on to a map of the region

In order to address the problem of overlapping arrows, two alternative arrangements were proposed. These are shown in Figure 4. On the left, the circle is divided in to two crescents that are positioned

either side of the road. We felt that while the bottom crescent followed the road, the top one seemed less well matched. By mirroring the crescents, as on the right, the relationship between ring road and graphs retained the graphical integrity that we had originally defined while also removing the overlapping arrows problem (by splitting the two halves of the display).

Figure 4: Alternative solutions to the overlapping arrows problem

In order to relate the graphs to the sensor data, we renamed the labels in the legend and in order to relate the locations to CCTV locations we resorted to a camera icon as indicator and highlighed the segment that this camera was displaying. This is shown in Figure 5.

Figure 5: Relating CCTV locations to the graphical display

The final version of the user interface for the second prototype is shown in Figure 6. This retains the layout of the earlier design but have two important changes. The first relates to the presentation of the sensor data relative to the road and its positioning on the map. The second relates to the set of CCTV images that correspond to the specific location (indicated by the camera icon). The intention is for the CCTV images to allow operators to maintain their current practice of developing situation awareness of congestion by referring to collections of CCTV images.

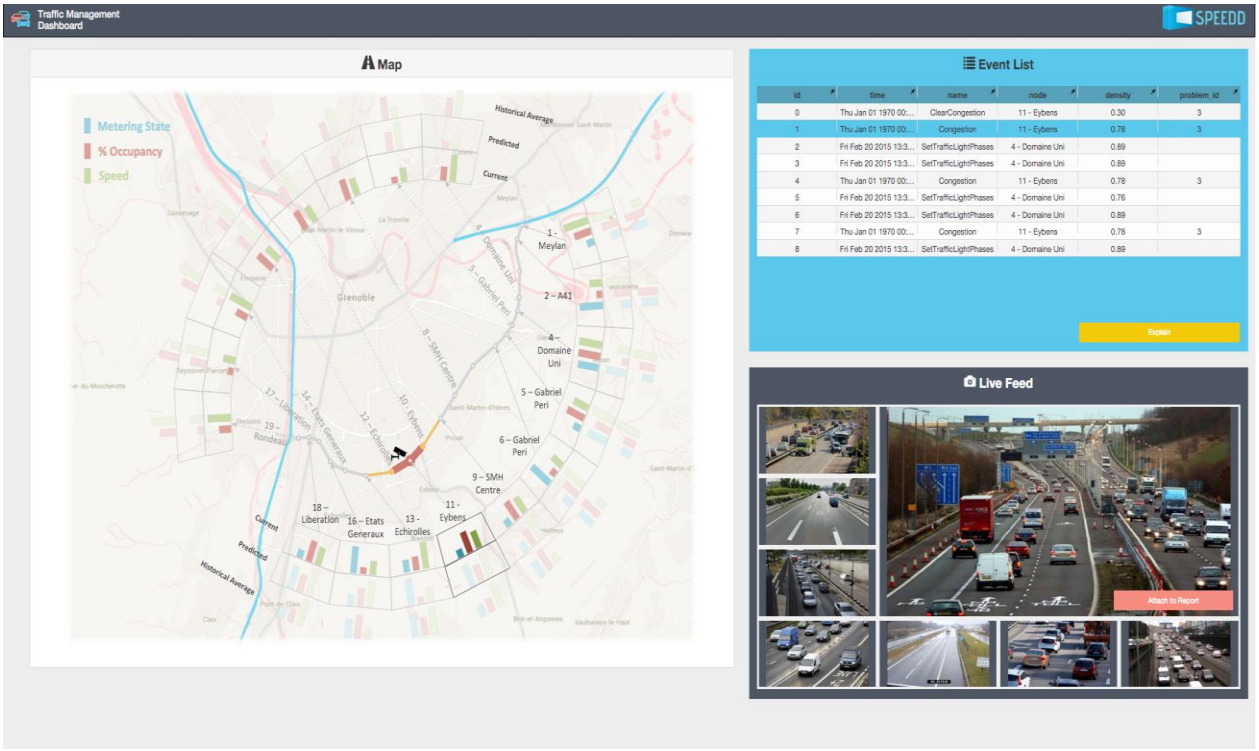

Figure 6: User interface for second prototype

4 Evaluation of User Interface Designs

4.1 Introduction

In order to determine how the User Interface (UI) designs developed in the SPEEDD project have evolved, an experiment has been conducted in which participants perform a series of tasks using the initial prototype (described in D8.1 and depicted in Fig. 7) and the most recent prototype (described in D8.5 and depicted in Fig. 6). The aim of this evaluation was to compare the two UI designs in terms of their usability and their impact of user performance. Usability was measured using the Software Usability Scale (SUS), which was described in D8.1 and from a short questionnaire completed at the end of the experiment. Performance was evaluated in terms of time to make a decision, correctness of the decision and user workload. The objective performance measures (time and correctness) are still be analysed and will be presented in a later report.

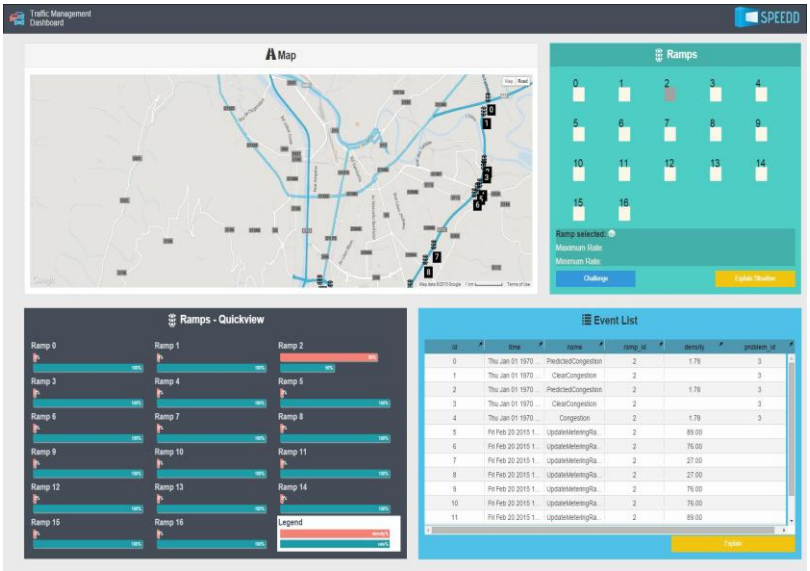

Figure 7: User interface for SPEEDD Prototype 1 (2015)

4.2 Description of Experiment

Following the previous evaluations of the Road Traffic UI, the experiment was designed to present users with data concerning ramp metering and the requirement to decide whether to accept the computer recommendation or make a different decision. This paradigm allows us to use the functionality of the SPEEDD architecture. The tasks involved in ramp monitoring are (at present) not performed by the expert operators at DIRCE. Consequently, it was felt that presenting the task to participants who had received training in this task (to criteria) would be a reasonable substitute. We accept that the final arbiters of the UI design would be the experts at DIRCE and plan to perform a version of this experiment with them in the autumn. However, we could only expect to recruit 4 or 5 operators during the visit and this would be insufficient to perform statistical analysis on the data. So, we have to make a trade-off between statistical reliability, on the one hand, and expertise opinion, on the other.

The experiment involved 21 participants recruited using opportunity sampling. The age of participants ranged from 22 to 45 years. None of the participants had any involvement in the SPEEDD project or prior

experience of road traffic management. All participants described themselves as computer literate and had normal (or corrected to normal vision). The experimental design conforms to the University of Birmingham ethics statement for the project.

On arrival, participants received a brief explanation of the experiment and instructions on how to manage the ramp metering task. They were provided with an aide memoire which they could consult at any time, and which described the conditions under which ramp rates should be increased or decreased. These instructions have been previously described in D8.2. Once participants confirmed that they understood the rules, they were given a 5 minute practice session in which they responded to up to 20 ramp metering problem using one of the UI designs. If they were able to perform the task correctly 5 times in a row, then they were deemed to have met criteria and the experiment using one UI began. Once the experiment with one UI was completed, this process (of practice and then experiment) was repeated with the other UI. The order of presentation of UIs was counter balanced across participants on order of appearance.

4.2.1 Experimental Task

The experiment was adapted from the 'automation bias' experiment (described in D5.2 and in the paper submitted to the Big Data Research special issue). Participants are presented with information concerning the state of traffic on the main road and the switching rate of the ramps leading onto the roads. If the main road is congested then increasing ramp rate would increase congestion, and so it would be sensible to either leave the ramp rate as it is or to decrease it. If the ramp is congested then the ramp rate would need to be increased in order to alleviate this. The task is supported by a computer recommendation which is intended to simulate the advice offered by the SPEEDD system. In these experiments, we do not directly employ the SPEEDD architecture because we need to manipulate the reliability of the advice provided. In other words, the system can be correct to high (80%), medium (50%) and low (20%) level of reliability. Participants will be to evaluate the information presented to them and the advice of the system (they are not told the reliability of the system that they are using) in order to make a recommendation. Manipulating the level of reliability allows us to perform Signal Detection assessment of performance, through comparison on agreement or disagreement with the system when it is right or wrong.

4.3 Usability Evaluation

As in D7.2, D8.3 and D8.5 a usability evaluation of the UI design was performed using the Software Usability Scale questionnaire (Brookes, 1988). The SUS scale consists of 10 simple questions concerning the potential usefulness and benefit that users feel that the User Interface might provide them. Each statement is rated on a scale of 0 to 4. The scoring of responses then involves subtracting 1 from odd-numbered questions and subtracting scores of even-numbered questions from 5. This is because the questions alternate between positive and negative connotations. Scores are then summed and multiplied by 2.5, to give a final score out of 100. As a rule of thumb, scores in excess of 65 are deemed 'acceptable'. Figure 8 compares the evaluation of version III with the previous versions.

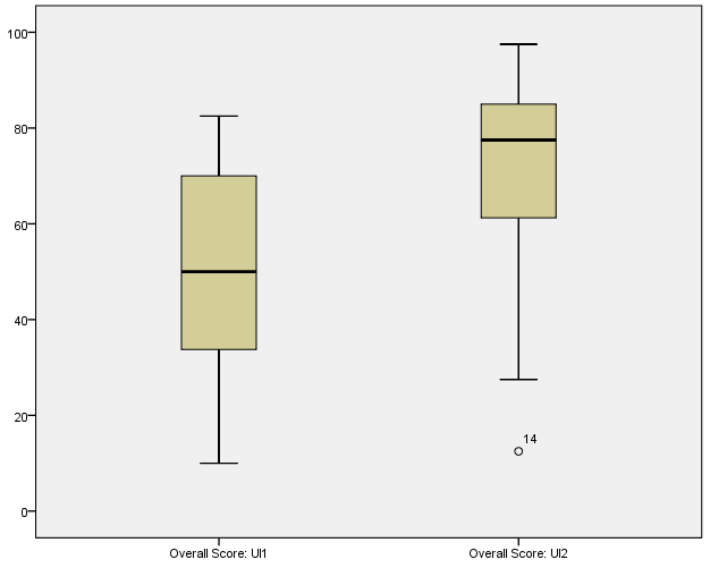

Figure 8: Comparison of median SUS scores for UI1 and UI2

We can see that the median is significantly higher for UI2 meaning that it scored higher on usability. The scores for UI1 were more spread out; scores for UI2 were generally higher. It was decided to remove participant 14 from the data at this point as they showed as an outlier for most questions concerning UI2. The median score for UI2 is 71; according to SUS, anything over 68 is above average. The median for UI1 appears as 48, which according to SUS would be below average for usability. (although removing this participant did not affect subsequent analysis).

The two UI designs were compared for each individual question, using two-tailed Student t-tests. The results are shown in table 1.

Table 1: Comparison of median response to each SUS question for UI1 and UI2 (higher scores shown in bold)

Question	Median SUS score		
	UI1	UI2	Sig. Diff.
I think that I would like to use this system frequently	2.2 (.9)	**3.57 (.9)**	0.0001
I found the system unnecessarily complex	**3.34(1.2)**	2.09 (1)	0.007
I thought the system was easy to use	2.74 (1)	**3.96(1.1)**	0.0001
I think that I would need the support of a technical person to be able to use this system	2(3)	2 (3)	n.s
I found the various functions in the system were well integrated	2.91 (1.2)	**3.78 (.9)**	0.004
I thought there was too much inconsistency in this system	2.39 (.9)	2 (.9)	n.s.
I would imagine that most people would learn to use this system very quickly	3 (1.2)	**3.83(1.03)**	P=0.027
I found the system very cumbersome to use	**3.26(1.3)**	1.91(1)	0.007
I felt very confident using the system	2.87 (1)	**3.96 (1.2)**	0.001
I needed to learn a lot of things before I could get going with this system	**2.87 (1.1)**	2.17 (1.2)	0.027

4.4 Workload

While there are many ways to measure the cognitive effort (workload) that people experience in performing mentally demanding tasks, a popular set of measures rely on participants providing subjective estimates of their workload. These measures can be surprising robust, sensitive to changes in demands and correlate well with physiological measures. One commonly used subjective workload measure is the NASA TLX (Task Load Index) (Hart and Staveland, 1988). This is a rating scale with six workload dimensions. It can be administered in either a computer or paper based format. We used the paper and pencil version of the test[1]. The rating scales are presented as questions that the participants scores on a scale of 1 (low) to 20 (high). The questions relate to mental demand, physical demand, temporal demand, effort, performance and frustration (figure 4).

[1] http://humansystems.arc.nasa.gov/groups/tlx/paperpencil.html

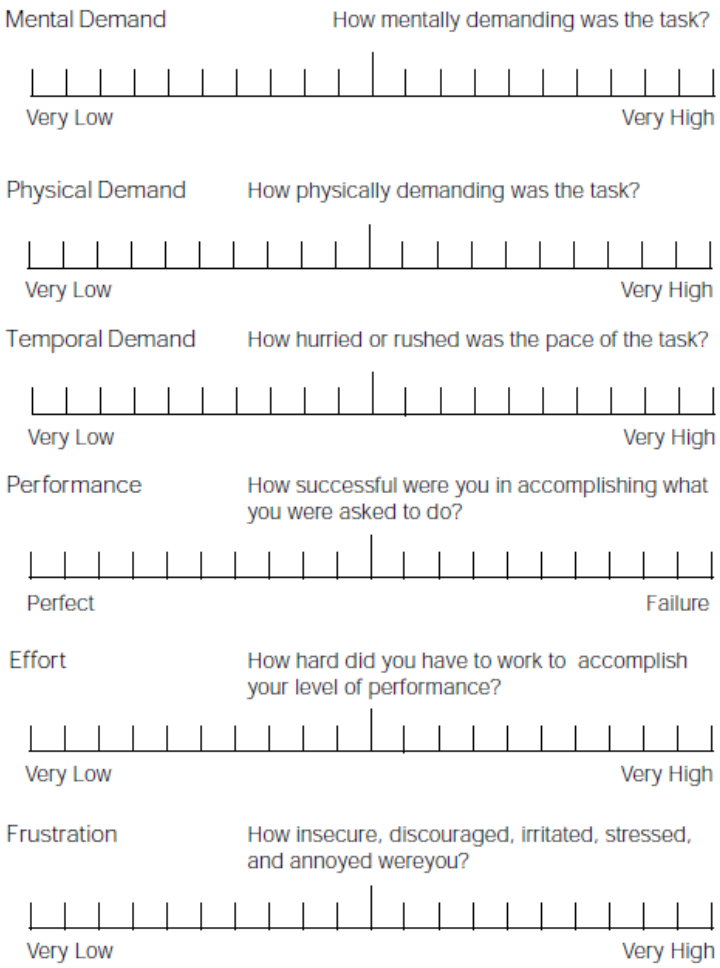

Figure 9: NASA TLX rating form

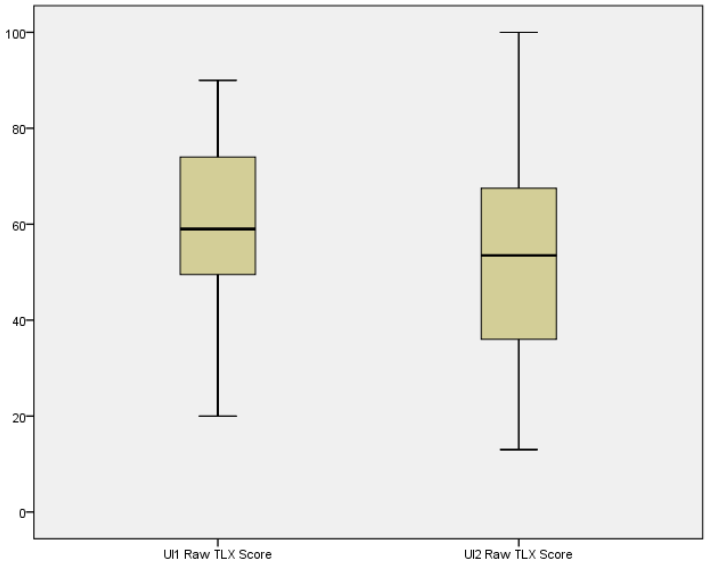

Figure 10: Overall workload for UI1 and UI2

The median is higher for UI1 than for UI2; this shows a higher workload was required. However, scoring is less consistent for UI2, with a larger interquartile range (although this could be attributed to variance in responses to question 2 for the two UIs, as shown in figures 6 and 7).

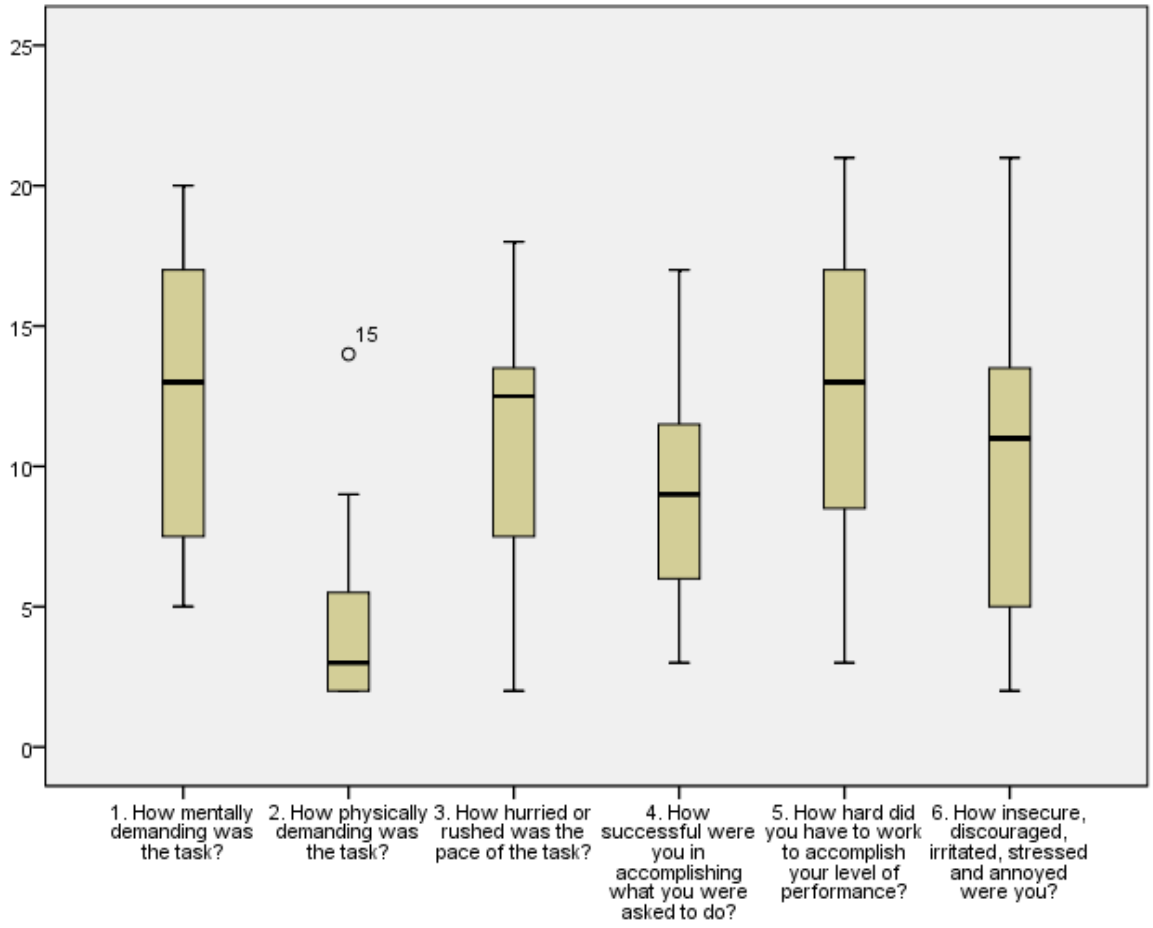

Figure 11: Median response to questions for UI1

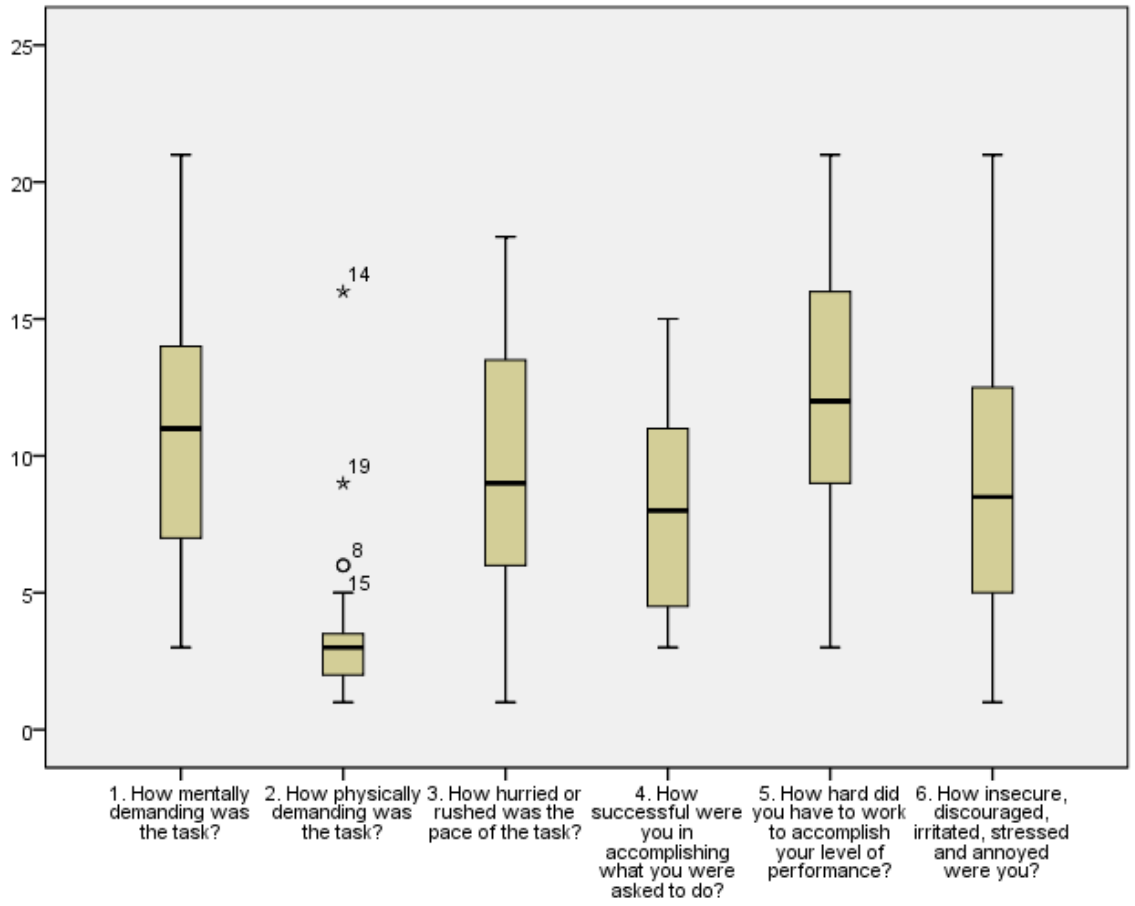

Figure 12: Median response to questions for UI2

There were significant differences in TLX scores between the UI1 and UI2 for two of the questions. Question 3: How hurried or rushed was the pace of the task? was rated significantly different ($p = 0.009$), with UI2 receiving lower rating. Question 4: How successful were you in accomplishing what you were asked to do? was rated significantly different ($p=0.047$), with UI2 receiving a lower rating.

4.5 Conclusions

Comparison of the initial and current UI designs for the SPEEDD Road Traffic prototype shows significant improvements in terms of usability and subjective rating of workload. The objective performance results will be presented in a later report. A point to note here is the ways in which the designs of the UI for the SPEEDD project have evolved. UI1 was designed on the basis of the data that would be available to operators and the reliance of CCTV in their current work activity. We consider this to be a design which reflects the data that operators access. UI2 is designed on the principles of Ecological User Interface Design which focus less of the data *per se* and more on the activities and judgements required for the tasks. The aim is to present information in a manner which, once learned, can be interpreted at a glance rather than needing to study and interpret the display. We believe that this could be one of the reasons for the differences, particularly in terms of workload. We also believe that the usability differences reflect superior ease of use for UI2.

5 Evaluation of the Event Recognition component

5.1 Introduction

This section is devoted to the evaluation of event recognition component. We first evaluate traffic congestion detection using both real and simulated data from the Grenoble south ring. Then, we compare the situations emitted from the Complex Event Processing (CEP) application with actual real congestions. In order to do so, we resort to annotated data, that is, timestamps for events during an elapsed time window that caused a sudden decrease in flow/sudden increase in density. The evaluation should answer the question: Can we forecast a congestion before it actually happens? In other words, is the inclusion of uncertainty aspects and the ability to predict a future event effective?

5.2 Evaluation of machine learnin for event recognition

5.2.1 Evaluation of with real data

In this task, the aim is to recognize traffic congestions which take place in the Grenoble South Ring, that links the city of Grenoble from the south-west to the north-east, by exploiting real time data collected from traffic sensors. The dataset comprises one month of data (≈ 3.3GiB of sensor readings), where each day is annotated by human traffic controllers for traffic congestions. The real data was collected from sensors placed in 19 collection points along a 12km stretch of the highway and each collection point has a sensor per lane. Sensor data are collected every 15 seconds, containing the total number of vehicles passing through a lane, the average speed and sensor occupancy. These readings constitute the simple derived events (SDEs) that concern activity on the highway.

We performed 10-fold cross validation over the entire dataset (172799 timepoints) using varying batch sizes. At each fold, an interval of 17280 timepoints was left out and used for testing. Figure 13 presents the evaluation results for OSLα and AdaGrad. In OSLα the predictive accuracy of the learned model increases initially, due to the increase in the number of learning iterations, and then decreases, due to the decreasing batch size. On the contrary, the accuracy of AdaGrad increases (almost) monotonically as the number of learning iterations increase.

OSLα achieves comparable predictive accuracy to the weighted manually constructed rules (AdaGrad), which is encouraging. Moreover, it can process data batches efficiently. For example, OLSa takes ≈ 11 seconds to process a 50 minute batch (1400 SDEs). As expected, AdaGrad is faster than OSLa. The predictive accuracy of the learned model, both for OSLa and AdaGrad, is low. This arises from the semi-supervised nature of the problem.

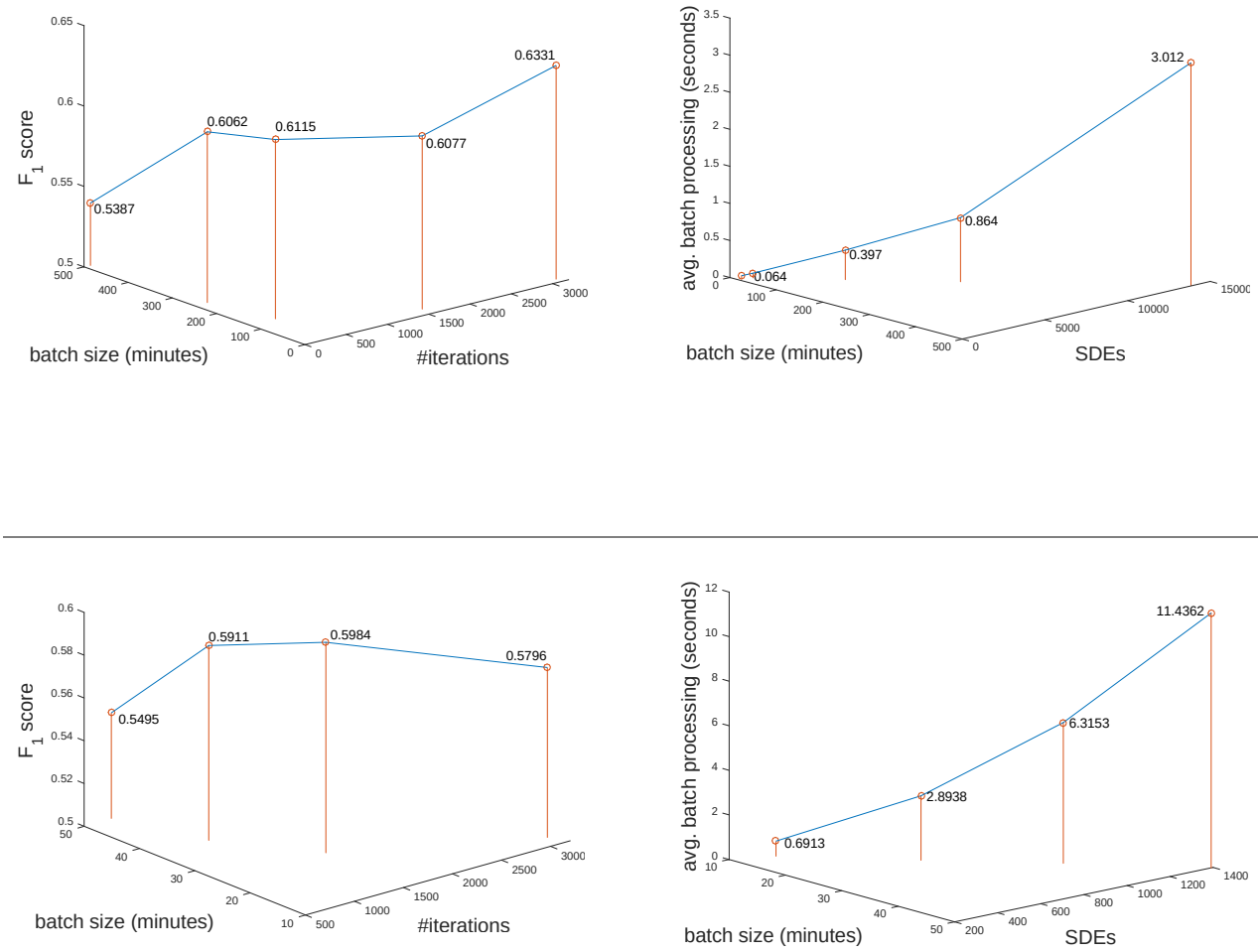

Figure 73: F1 score (left) and avg. batch processing time (right) for AdaGrad (top) and OSLα (bottom). In the left figures, the Y axis shows the number of learning iterations.

5.2.2 Evaluation with simulated data

In this task, similarly the aim is to recognize traffic congestions which take place in the Grenoble South Ring, but instead of the real data we are using the simulated dataset provided by the CNRS. The dataset compromises 10 simulations of 1 hour duration each and a specific location is annotated with traffic congestion. The annotated location for each simulation id is different and therefore one cannot use all the simulations for learning weighted patterns for one location. Therefore, we performed experiments for each simulation id separately. Because the training data provided for each annotated location are very little (121 timepoints), we performed training using a 60% of timepoints for training and the rest for testing. The pair of rules trained for location 1311 are presented below.

InitiatedAt(trafficjam(1311), t) ⇐
HappensAt(aggr(1311 , avgspd), t) ∧ 0 ≤ avgspd < 18

TerminatedAt(trafficjam(1311), t) ⇐
HappensAt(aggr(1311 , avgspd), t) ∧ avgspd > 45

Figure 14 presents the evaluation results for OSLα for simulation id 1 and location 1311. The purpose of the experiment is to present the accuracy change in a fully supervised dataset. The predictive accuracy of the learned model remains almost the same as the iterations increase and arrives at the best F1 score when batch sizes of 2.5 minutes are used. Note that the recognition results are much more accurate (highest F1 score) than the real data case presented in Section 4.1, which arise from the fully supervised nature of the data. Finally, as in the case of the real data, OSLα can process data batches efficiently. For example, it takes ≈ 3 seconds to process a 25 minute batch (6000 SDEs).

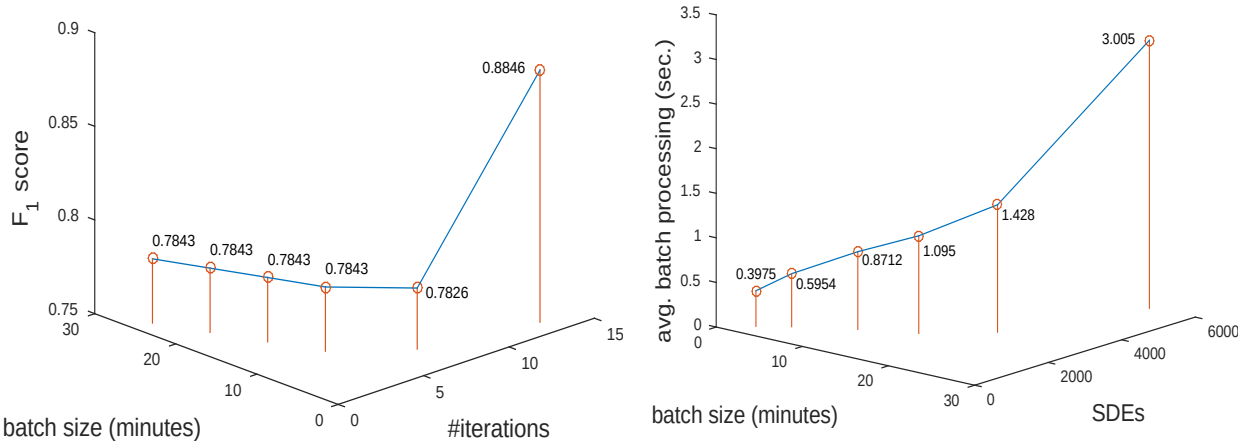

Figure 14: F1 score (left) and avg. batch processing time (right) for AdaGrad (top) and OSLα (bottom). In the left figures, the Y axis shows the number of learning iterations.

5.3 Evaluation of Event Processing in presence of uncertainties

As aforementioned, we aim at answering the question: is the inclusion of uncertainty aspects and the ability to predict an event effective? The way to address this is to have two applications or EPNs, once including uncertainty aspects and the other one without uncertainty, i.e. deterministic. This is a common approach in CEP engines dealing with uncertainty, see for example Cugola (2014).

5.3..1 Sample data

Data encompassing 3 hours (from 4PM-7PM) has been simulated in Aimsun. A stream is received every 15 seconds as in the real physical sensors. The data from the Aimsun includes the following attributes:

did - Replication ID for random seed
oid - Detector ID

sid - Vehicle type(0 = for all vehicles, 1 = Car , 2 = Truck)
ent - Time interval, from 1 to N, where N is the number of time intervals, and 0 with the aggregation of all the intervals. As the data covers 3 hours, we have a total of 10,800 sec. Having intervals of 15 sec gives 720 intervals for each simulation (10800/15=720).
countveh - Vehicle count
speed - Speed [km/h]
occupancy - Occupancy
density - Density at the specific location [#of vehicles/km]
A total of 10 simulations have been created with random seeds for all the simulations. For each of the simulations a csv file was produced. In addition, another annotated file has been produced containing information about incidents that had been created during the simulations. For each simulation (csv file) an intended incident was created and annotated. An incident causes a rapid build-up of congestion and helps us evaluating our application. Only two out of the 10 incidents occur in the main road, namely in simulation #6 and #10. All other incidents appear in off or on ramps.
The annotated file includes the following fields:
Aimsun Section ID - Where incident occurred
Duration of incident – the time window for the incident [hh:mm:ss:]
Start Time – of the incident [hh:mm:ss:]
Length of Effected Area – in [meters]
In all the ten simulations there is a congestion building up in all sensors close to the end of the road around 18:40 that lasts until the simulation end, with no clearing up of the congestion.

Details of data preprocessing can be found in deliverable D3.2.

5.3.2 Summary of results

Our findings are consistent among all the simulations. First, we were able to detect all congestions resulting from the simulated (annotated) incidents. Furthermore, we were able to detect more congestions as they happened in the simulations and indicated by the sudden drops of speed and high increase in density values. Moreover, *PredictedTrend* situations were detected and emitted which even caused *Congestion* situations a few minutes later. Note that we are running the CEP module in isolation so we don't have any feedback from the decision making module or any action taken that can help in alleviating the potential congestion.
Congestions or even predicted congestions have an impact upstream, and therefore forecasting a congestion in upstream locations when a congestion is detected in a downstream location can help in clearing up the whole area.
However, we still need to refine and validate the level of certainty that actually indicates a congestion, i.e. from which certainty value we should take an action (a congestion is very likely to happen). One time we get a congestion after certainty value of 0.763 and one time after 0.474 while we didn't get a congestion after certainty value of 0.88. There are some fluctuations in the numbers which should be further analyzed and comprehended. To this end, we ran and analyzed a second set of simulations as detailed in the next section.

5.3.2.1 Recall and precision

In order to explore the quality of our results we ran a second set of simulations which comprised of:

- 20 simulations with annotations of congestions. All simulations last an hour.
- The annotations of congestions include the location and the time the congestion is detected.
- Other characteristics as file format and pre-processing of data remained the same, apart of the additional field for annotation of congestion (a Boolean field, having 1 for a congestions and 0 otherwise).

First, we checked the quality of our *Congestion* pattern against the annotated data. First, we checked the proportion of detections by our EPA that were annotated in the data as congestions (*precision*) and second, the proportion of congestions we were able to detect out of all the annotated congestions (*recall*). In all our simulations our precision was 100%, while the average recall over all the simulations was 72%. This can be easily explained: the rule implemented has been given to us by the domain expert, who is the one to identify the congestions in the simulations, thus giving a perfect precision. However, when implementing the pattern we applied a "stricter" criterion for the rule: we took into account not just the average *speed* critical thresholds , as was done in the simulations, but also *density thresholds*, therefore we have a less success rate in the recall of the results, i.e., there were annotations of congestion in the data that we "missed". When we "relaxed" the pattern and run the same rule as in the simulations we were able to detect all congestions with a perfect score in both precision and recall.

As a second step (as in our previous series of tests), we aimed at checking a more interesting question, that is, whether the inclusion of uncertainty aspects enables us to predict a congestion in the high way before it reaches critical thresholds, as opposed to detecting it once it happens. As before, we addressed this question by having two EPNs, once including uncertainty aspects and the other one without uncertainty, i.e. deterministic; and running the tests twice, one time for each EPN (with and without uncertainty). The deterministic case served as the "ground truth", as we knew at this stage that all our congestions have been detected correctly. The precision of our results indicates the proportion of congestions we were able to predict (in other words, *PredictedTrend* pointed out correctly to a congestion), whereas the recall indicates the proportion of congestions we were able to detect out of all the annotated congestions (in other words, *PredictedTrend* pointed out correctly out of all congestions). Important to note that we used a threshold of 0.6 in the *certainty* attribute to determine whether to consider *PredictedTrend* as a congestion. In other words, only *PredictedTrend* alerts with a certainty value larger than 0.6 were considered in our calculations of precision and recall.

Table 2 summarizes our findings. As can be seen the average precision is 90.75% and the average recall is 75.05%. In addition we can see that sometimes a very high score in precision comes along with a lower score in recall (e.g., simulation #17, precision = 96.15 and recall = 51.85).

In general, the results indicate that the *PredictedTrend* pattern is a very good estimator of congestions to occur a few minutes ahead in time, thus enabling the system to take proactive actions in order to alleviate these congestions. However, lower scores in recall indicate that there are other situations that cause congestions which are not detected by our pattern. Further analysis on both real and generated data of these congestions that have not been "caught" by our pattern shows that these situations are characterized by "jumping data", meaning, the values of speed and density tend to jump thus not satisfying the increasing build-up which is required in our pattern. We are currently investigating these "jumping" cases to see if we can identify some common behavior/pattern.

Table 2: Summary of recall and precision results

Simulation	Precision	Recall
1	97.61	90.90
2	80	93.15
3	81.6	95.2
4	87.88	85.56
5	84.88	83.62
6	83.33	92.22
7	76.12	81.72
8	89.25	93.22
9	88.61	90.27
10	98	68.99
11	96.77	67.11
12	85.42	89.39
13	96	57.33
14	88.46	56.94
15	100	51.25
16	96.43	61.11
17	96.15	51.85
18	96.77	60.87
19	96.67	58.54
20	95	71.76
Average	90.75	75.05

6 Evaluation of the Decision making Component

Decision making (DM) is an important component of proactive traffic management. We evaluate the performance, in terms of two main aspects, of the decision making component for the traffic use case: evaluation of Real-time capability and comparison with state-of-the-art.

6.1 Evaluation of Real-Time Capability

A detailed evaluation of the time needed to process individual events by DM has been conducted and presented in the Decision Making Deliverable D4.2. To ensure perfect control over the type of events processed as well as the attribute values, this test is run outside of the SPEEDD cluster architecture on a local computer (MacBook Pro with OSX El Capitan, 2.3 GHz Intel Core i7).

In total, 10^6 events are processed in 35s. Note that this time and all the other times reported in this section, only account for the computation time of DM **once the event has been received by DM**, as we only intend to show **real time capability of the DM algorithms**. For example, the overhead incurred to send, transmit and emit events is not included.

The computation times for the test events exhibit a bimodal distribution, with many events being handled in less than 1μs but others taking between 2μs and 0.1ms to process. This is because for certain events, no immediate action by DM is necessary or advisable. The processing of such events can be as simple as updating certain variables, and is often accomplished in less than 1μs. In contrast, other events require immediate action by DM. For example, a reevaluation of the ramp metering policy requires at least a recomputation of the metering rates. These events then take longer to process. We report the computation times for various DM events in Table 1 numerically. To account for the bimodal nature of the distribution, we first report mean computation time, and then also report the 90%, 99% and 99.9% quantiles. The mean computation times can only be interpreted along with the percentage of events that do not require immediate processing (presented in Deliverable D4.2). The 90%, 99% and 99.9% quantiles more accurately capture the tail of the distribution of computation times.

Table 3: Statistics of computation times for individual events.

Event	Mean	90%	99%	99.9%
PredictedCongestion	15μs	8.2μs	18μs	36μs
Congestion	13μs	8.2μs	19μs	42μs
ClearCongestion	15μs	8.2μs	18μs	44μs
setMeteringRateLimits	19μs	9.6μs	21μs	48μs
PredictedRampOverflow	19μs	9.3μs	21μs	49μs
ClearRampOverflow	19μs	9.0μs	21μs	47μs
AverageOnRampValuesOverInterval	49μs	42μs	85μs	212μs
AverageDensityAndSpeedPerLocation	50μs	42μs	84μs	195μs

6.2 Comparison with state-of-the-art

We have presented a number of algorithms for decision making for the traffic use case in D4.2. For each of these algorithms, we present the quantitative improvements against a *benchmark* and the *optimal* solution, if available, in Table 4. The choices for the benchmarks are given in the table itself. Note that the *optimum* refers to the global optimal solution under perfect model knowledge and perfect traffic demand prediction. This performance will be unobtainable for any real-world controller and it merely provides an estimate of the potential for further improvement. Depending on the scenario, we compare algorithms using the more appropriate of the following traffic metrics:

- Total Time Spent (TTS) and Total Waiting Time (TWT): These are defined in Section 2.2 of D4.2. Note we seek to *minimize* both quantities, therefore negative values in the table indicate an improvement of the proposed algorithms over the benchmark performance.
- Total Travel Distance (TTD) and the Service of Demand (SoD): These are defined in Section 2.4 of D4.2. Note that we seek to maximize these quantities and hence, positive values in the table indicate an improvement in performance of the proposed algorithms over the benchmark.

Table 4: Performance comparisons between the algorithms proposed in this deliverable and various benchmarks. Improvements against the benchmarks are indicated in green and suboptimal values are indicated in red. Note that we minimize the TTS and TWT metrics, whereas we maximise TTD and SoD.

	Benchmark		**Optimum**	
	TTS	TWT	TTS	TWT
D4.2 Sec. 2.2 Local feedback, benchmark: No control	-5.33%	-17.53%	+0.054%	+0.178%
D4.2 Sec. 2.2 Local feedback, benchmark: ALINEA	-0.43%	-1.42%	+0.054%	+0.178%
D4.2 Sec. 2.3 Coordinated metering, benchmark: No coordination		Up to -31%	Unknown	Unknown
	TTD	SoD	TTD	SoD
D4.2 Sec. 2.4 Dec-MILP, Network 2.18a	18%	14%	Unknown	Unknown
D4.2 Sec. 2.4 Dec-MILP, Network 2.18b	17%	15%	Unknown	Unknown

For some control problems, there is no efficient method to compute the optimal solution because these problems are nonconvex and far too large to be solvable by available tools. Therefore, the relation of the obtained results to the optimal solution are reported as “unknown”.

In the traffic use case, existing solutions for freeways such as *Alinea* are proprietary systems and the decision making algorithm used in these approaches are not publicly available. To circumvent this problem, we suggest the use of the first version of the SPEEDD prototype as the state-of-the-art.

For inner city traffic, this problem is only more acute. Design principles for existing traffic controls in the city of Geneva are opaque. In fact, many of the existing urban traffic control solutions around the world have been hand-tuned over several years, and the design principles are absent or lost. Also, these systems have been designed at some point for existing traffic conditions and have not been thoroughly or rigorously updated since. In this context, an important contribution of the decision making component in SPEEDD are the systematic design approaches developed in this project.

7 Changes in Road Traffic Activity at DIRCE

7.1 Introduction

The initial visit to the DIRCE control room took place in February 2015. During this visit we noted that the operators did not seem to make effective use of the large bank of CCTV screens on the back wall. Our eyetracking studies showed that operators rarely looked at these. We also noted that there was no common view of the road system for the operators, which meant that situation awareness could be limited to individual rather than a collective view of the state of the road network. Over the course of the intervening 12 months, the control room has undergone several changes.

7.2 Changes in room layout and use of large screens

Figure 15: Comparing the layout of the control room in February 2015 and February 2016

As the image on the right of figure 15 shows, the control room now has a new bank of screens. The bank of screens on the right shows a map of the ring road, together with the location of CCTV and controlled signs. This map is similar to the version that was available to operators on their individual monitors, but is now placed prominently on the large display. It is also apparent that the images of the two banks of CCTV screens (in the right-hand image) are showing different scenes. One bank is showing snow covered roads and is monitored by the inter-urban traffic monitor, and the other is showing the ring road and is monitored by the Rocade Sud operators. In addition to having an additional bank of CCTV screens, the control has also had the desktop monitors repositioned; they have been lowered in order to allow the operators to attend to the large bank of screens.

7.3 Changes in use of CCTV

In February 2015, the operators described how the automated incident detection system could be applied to the CCTV images. We did not see this in operation during the initial visit. In February 2016, this system is operational and an integral aspect of the monitoring of the large bank of CCTV screens. Figure 15 shows the system in operation. On the left is the ring road running normally. On the right, one of the CCTV screens has blacked out (indicated by the circle). This blacked out screen has a message stating that the automated detection system is running.

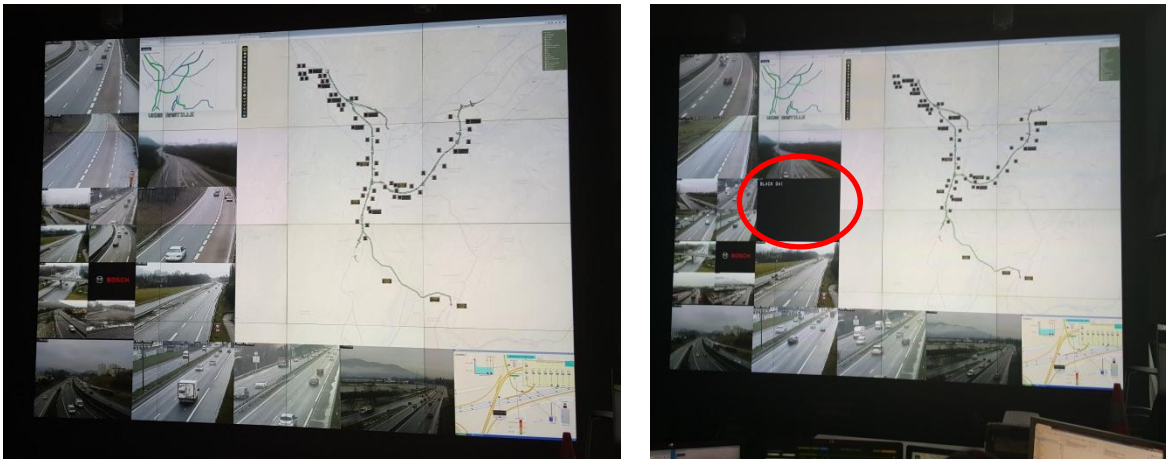

Figure 8: Indicating Automatic Incident Detection

In this instance, the automated incident detection system had indicated stationary vehicles on the side of the road. This could be a hazard that requires managing. However, in this instance, a pair of road maintenance vehicles had stopped on hard shoulder. The Control Room was aware that maintenance was planned but not the location. The Controllers decided this was not an 'event' and waited until the vehicles moved on (Figure 15).

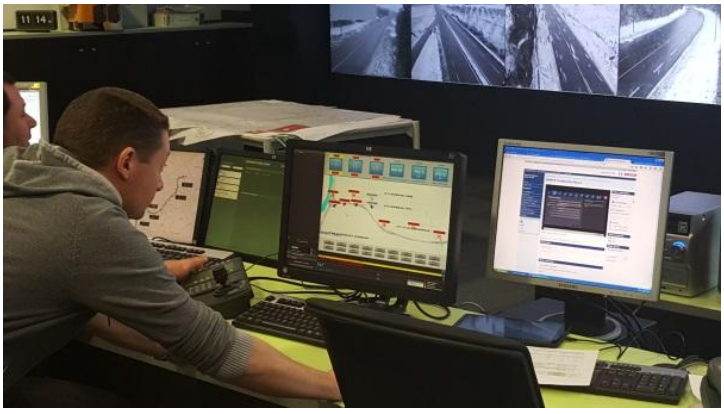

Figure 16: Determining whether an incident (from the automated incident detection system) should be classified as an event to be managed

In addition to the use of the automated incident detection system, the operators said that they used the CCTV images to monitor the build up of congestion during rush hour. The images on the CCTV screens change every 20 seconds. As there are over 100 cameras on the ring road, most of the road can be monitored in a few screen changes.

7.4 SPEEDD contribution to current traffic operation centers

In most road traffic control centers, operators are commited to monitor the current traffic status and to take decisions if an event occurs, typically an accident. Detection of such events is usually achieved using automated incident detection systems, which can raise false alarms mitigated by opertaors through CCTV screening. Congestion can be detected through a speed heatmap with 3 to 4 different colors and also by CCTV screening. However, the only action that can be taken in case of congestion is to inform drivers, while in case of accident, the traffic center not only inform drivers but also alerts police and emergency services.

To handle congestion, some cities are now implementing ramp metering. Grenoble will implement such system within a year. The role of the operators is not yet well defined. The system will not be fully automatized and will not be proactive. In the loop, the operators will be restricted to switch off/on the system, typically ALINEA (Papageorgiou et al., 1991; 1997).

SPEEDD prototype intends to bring to the community proactive policies. Indeed, based on current traffic status and predictions for the future, SPEEDD prototype can suggest to the operators the best decision to be taken. For this purpose, the user interface provided by SPEEDD will display the necessary information to the operators in order to help them to take the right decisions. Instead of reacting to events, SPEEDD will help to act for mitigating future impacts of some events.

8 References

Brooke, J., (1996), SUS: a quick and dirty usability scale. In: P.W. Jordan, B. Weerdmeester, B.A. Thomas and I.L. McLelland (eds) *Usability Evaluation in Industry*, London: Taylor and Francis, 189–194.

G. Cugola, A. Margara, M. Matteucci, and G. Tamburrelli (2014). Introducing uncertainty in complex event processing: model, implementation, and validation. Computing, pages 1–42, 2014.

Hart, S. G., & Staveland, L. E. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, *Advances in psychology*, *52*, 139-183.

M. Papageorgiou, H. Hadj Salem, and J-M. Blosseville (1991) "ALINEA: A Local Feedback Control Law for On-Ramp Metering, " Transportation Research Record , vol. 1320, pp. 58-64.

M. Papageorgiou, H. Hadj Salem and F. Middleham (1997) “ALINEA Local Ramp Metering: Summary of Field Results,” Transportation Research Record, vol. 1603, pp.90-98.