UDRIVE

European Naturalistic Driving Study

EUROPEAN COMMISSION

SEVENTH FRAMEWORK PROGRAMME
FP7-SST-2012.4.1-3
GA No. 314050

eUropean naturalistic Driving and Riding for Infrastructure and Vehicle safety and Environment

Deliverable No.	UDRIVE D22.1	
Deliverable Title	Guidelines for data quality assurance	
Dissemination level	Public	
Written By	Ruth Welsh LBORO Steven Reed LBORO James Lenard LBORO Riku Kotiranta CHALMERS	31-03-2017
Checked by	Philipp Lindner (TUC)	01-06-2017
Approved by	Marika Hoedemaeker (TNO) Nicole van Nes (TNO)	04-06-2017 [dd-mm-yyyy]
Status	Final	

Please refer to this document as:

Welsh, R., Reed, S., Lenard, J., Kotiranta, R. (2017) UDRIVE deliverable D22.1 Guidelines for data quality assurance of the EU FP7 Project UDRIVE (www.udrive.eu)

Acknowledgement:

Disclaimer:

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 605170.

Executive Summary

The UDRIVE project aims to collect on the region of 100,000 hours of naturalistic driving data in order to support the analysis related to

- Crash causation, crash risk and normal driving
- Distraction and inattention
- Vulnerable road users
- Driving styles related to eco-driving

This document contains information relevant to data quality assurance for the UDRIVE project. Good quality data is a fundamental requirement for good quality analysis and data quality should be considered at all stages of the data processing chain:

- Data Acquisition System Installation
- During data collection
- Database management
    - Data pre-processing
    - Data post-processing

In order to deliver high quality data as an outcome from the UDRIVE project actions have been undertaken at each stage of the chain, following generic guidelines for data quality.

Before the full data collection commenced a number of pilot activities were undertaken. Prior to installing the DAS in the cars, trucks and mopeds bench testing and in-vehicle testing has been undertaken to assess the correct technical functioning of DAS systems in each of the vehicle types within the UDRIVE. Subsequently short (typically 2 week long) pilot studies were carried out at each of the operation sites (OS) in order to establish and remedy any difficulties with the installation and during data collection and transfer. Aspects of the database management process were also trialled.

Complete and comprehensive documentation relating to DAS configuration, installation and camera configuration have been provided within the project and this has been adhered to by all OS in order to collect good quality data. Additionally an on-line monitoring and tracking tool has been developed. The purpose of the OMT is to monitor the status of the DAS with corresponding data in the field and to keep a status log of collected data along its lifetime through the data flow. Essentially, the OMT has allowed each OS to check that vehicle based data are being recorded as expected for each vehicle and that the video data has not been disrupted. Correct use of the tracking feature by each OS ensures that no data is lost during the data transfer process.

All OS have been provided with guidelines related to data quality that covered:

- Recruitment
- Questionnaire data
- DAS Installation
- Field Data Collection and use of the on-line monitoring tool
- Data Storage and Transfer
- Feedback from Database quality
- Metadata
- Piloting

Hard disk drives have been exchanged and sent to Local Data Centres at regular intervals where de-encryption and data quality checks have been undertaken. These are primarily aimed at checking

- that fields existed, i.e. had been recorded and carried through the decoding process,
- that the range and distribution of field values were plausible,
- that different fields were consistent, where this could be cross-checked, e.g. by comparing independent measurements of velocity and distance travelled or by comparing independent measurements of velocity as provided by in-vehicle sensors or GPS location.

Finally, further processing of the data is undertaken by the data analysts aimed at the creation and integrity of parameters suitable for use and presentation as the results of naturalistic driving analysis. A template was created to record essential information about the origin and history of development of each script, including the name and affiliation of the first and any subsequent programmers, the date of revisions, the reason for revisions, the nature and outcome of checks made on the algorithms contained in the scripts, and notes on the scope and limitation of the scripts.

The measure undertaken during the course of the project have produced a good quality, high volume naturalistic driving database that can be further exploited beyond the life cycle of the UDRIVE project.

Table of contents

1. Introduction

The UDRIVE project will collect naturalistic data on passenger cars, trucks, and powered two-wheelers. All data - including video data showing the forward view of the vehicle and a view of the driver, as well as geographic information system (GIS) data - will be collected continuously to bring knowledge in the various research areas well beyond the current state-of-the-art. The UDRIVE project aims to collect on the region of 100,000 hours of naturalistic driving data in order to support the analysis related to

- Crash causation, crash risk and normal driving
- Distraction and inattention
- Vulnerable road users
- Driving styles related to eco-driving

Whilst the choice of operation sites (OS) was motivated by aiming at having a good spread over countries with different characteristics in terms of road safety records, road user behaviour, road infrastructure characteristics, the presence of vulnerable road users, climate, traffic density, etc., the quality of the recorded data and the resulting database available for analysis also have a direct bearing on the comparability and representativeness of data.

This document contains information relevant to data quality assurance for the UDRIVE project. An early draft has been used in order to develop data quality procedures specific to the UDRIVE project and to inform the video data annotation development. This final draft includes how generic principles related to data quality have been applied within UDRIVE.

The UDRIVE project follows a number of steps, namely

1. Defining research questions and hypotheses
2. Defining variables and data specification
3. Data acquisition
4. Data transfer
5. Data storage
6. Data base quality
7. Data analysis

The aim of this deliverable is to cover points 3 to 6 above though comments are made in the context of the other points.

The document begins by outlining the general principles of data quality elaborating on the areas identified in the PROLOGUE project (Welsh et al 2010). The next section shows how these principles have been applied within the UDRIVE project with reference to supporting UDRIVE deliverables. Finally, in an Appendix, comments are made on variables to be recorded during the data collection ppahse of UDRIVE including both their importance for the study i.e. how reliant the study is on the collection of the variable in question and the level of data quality issues associated with this variable. This appendix is of particular use for determing protoclos for coding the video data.

2. Principles of data quality

2.1 DAS installation

2.1.1 Physical considerations

Depending on the location and mounting of the Data Acquisition System (DAS), including logger, in the vehicle a number of quality issues may arise. These issues are outlined below and are not specific to particular vehicles or particular equipment but form some best practice.

Before considering these issues it is worth mentioning drop-out rates of sensors and the data loggers themselves; this can have a significant impact on data quality and reliability. It is true to say that almost anything within a vehicle can be measured; the only restriction to this is engineering and technical knowledge; it is however often imprudent to instrument a vehicle too extensively as drop-out, calibrations and technical issues will cause quality problems.

Another factor which can have a significant effect on data quality is in the physical installation of equipment. A review of system installation, whether it is a logging device or particular sensor, might not necessarily constitute state of the art techniques but should be considered good practice. As with all automotive applications there are numerous sources where data quality and reliability can be effected, these are diverse but include; interference (electrical, vibration), environmental (heat, moisture) or contaminants (fluids, dust), all of these can have a degenerative effect on data quality if left unchecked.

Using experience from recent Field Operational Tests (FOT), Naturalistic driving projects, experimental/test driving and from motorsport applications (where data loggers and sensor groups have been used extensively, and where, if anything, the strains on such equipment is greatly magnified) there is extensive reporting of good practice.

Equipment manufacturers often state how the devices should be fitted into a vehicle however this is often specific to a piece of equipment so a general review of installation good practice is included here under the headlines of the problem.

Vibration

All road vehicles vibrate (or resonate) at a certain frequency; this is due in part to their design, being manufactured predominantly of a metal shell/chassis containing within it a partially damped reciprocating engine. Different vehicles will vibrate at different frequencies; larger vehicles with longer stroke or slower revolving engines and softer damping will resonate slower than a very stiff vehicle with a high speed engine.

In general Road cars will have a resonance of approximately 1Hz; Motorcycles will have a resonance several orders above this (>10Hz) and trucks less so (~.5Hz). In addition to this some vehicles (particularly motorcycles) will have a natural resonance at certain engine speeds/road speeds. This vibration can increase the frequency considerably for short periods and may have significant effects on data quality or equipment life.

Vibration is an important issue which can affect the long term reliability of electronic components and the short term quality of logged data. Good practice for mounting data loggers suggests that if the item looks (or feels) as though it is vibrating then isolation should be considered. The method of isolation will depend on the situation but in areas of high vibration, very soft rubber mount may offer sufficient isolation, generally in these circumstances the softer the mount the higher the level of vibration isolation. If installation is temporary or the equipment needs to be removed on a regular basis then reusable adhesive putty or Velcro tape have sufficient security and damping qualities.

Temperature

Temperature specifications are, like vibration, equipment specific. However general good practice still exists in order to maintain high quality data. Environmental constraints in new car manufacture tends to push

engine efficiencies further and as a result greater temperature ranges can be experienced. For example; catalytic converters and exhaust systems can reach temperatures of over 900 ºC under hard use. Most commercial data logging systems can tolerate a temperature range of between 0 ºC and 80 ºC during use and -20 ºC to 85 ºC at idle; these can be exceeded easily depending on trial location and weather conditions during the trial. It is also not enough to ensure that peak temperatures are not exceeded; to maintain the best quality data it is necessary to ensure a (relatively) consistent operating temperature within these ranges. Additionally consideration has to be given to materials as plastic can deform as high temperatures and LCD screens become slow or even inoperable at lower temperatures, experience in the UK also shows that cameras can routinely become condensed up through swings in temperatures (from cold nights to warm days). This condensation will also be present on the electrical connectors and corrosion will occur unless insulated.

Interference

Interference can be caused by a number of factors; the two major forms outlined below are signal interference, caused by vehicle vibration, and electromagnetic interference caused by electronic components.

In addition to the reliability issues created by vehicle vibration outlined in the section above, quality issues can be present when recording in vehicles. Vibration in the vehicle structures, and therefore in the instrumentation, could interfere with sensor readings such as high sensitivity accelerometers. As previously mentioned in the vibration section this can be engineered out by careful isolation of components from this resonance or by applying filters to recorded data. Poor quality data signals (noise) are often attributable, simply to general vehicle resonance.

Routing cables inside a vehicle can also cause signal problems due to electromagnetic interference. This is caused by electric or magnetic fields, generated in parallel cables, coupling and creating ‘crosstalk’ in the signal. A method of controlling this is to use twisted pair cables, these will cancel out the interfering source providing that it is relatively uniform. The result is a ‘cleaner’ and higher quality signal at the data logger. Another method of reducing noise due to electric or magnetic interference is to reduce the length of the cabling. Careful logger positioning and routing of the cabling over the shortest distance can ensure a much higher quality signal.

Other on-board devices can also cause interference with the loggers and sensors, the effects of these are quite wide ranging but should be explored and controlled for where they occur. For example vehicles with OBD or CAN access can often produce spurious signals and vehicle ‘fault’ codes.

Moisture

Unless the test is to be conducted completely indoors then the vehicle will almost inevitably be exposed to water. This water combined with high sensitivity electronics can cause extensive quality problems. Protecting the loggers and sensors from moisture ingress is a major issue. Good practice dictates that, as far as possible, all equipment should be contained within the vehicle however this is not always possible, especially when considering instrumentation on a motorcycle or when considering wheel speed sensors, external cameras or radar sensors. Good quality electrical connectors (a good rule of thumb is to spend approximately 10% of the equipment budget on these) will help prevent water ingress and will make repairs easier when they do not. Avoiding areas of the vehicle where conditions are harsh is also worthwhile; areas such as wheel housings or chassis components will inevitably be subjected to more water contamination than others. Areas where temperature differentials are apparent, such as the engine bay, may cause condensation to gather in connectors or control boxes as they repeatedly cool or warm during trials.

Fluids

Water is only one type of fluid that is common with vehicle testing; most modern vehicles carry on board many different types of fluid. For example a modern road car will carry approximately 6 different fluid types with some up to 10. These include Engine oil and coolant, brake and power steering fluid, gearbox and

differential oils and specialist fluids such as for suspension systems or air conditioning. All equipment needs to be protected from these fluids as they tend to be more aggressive than water. Simple guidelines should apply when planning system design or installing equipment. Avoid areas where these contaminants are present, unless of course, it is what needs to be measured and shield wiring and connectors in case of a leak.

Dust and other contaminants

Equally important is to protect equipment from dust and other airborne particulates; these tend to be less invasive but can cause similar quality problems to water ingress. To maintain data quality equipment or loggers fitted with cooling fans should be sighted away from these particulates while careful routing of cables and connectors (unless absolutely necessary) will improve dropout rates.

Debris

Due to the nature of naturalistic and field operations trails the vehicle will more than likely be used on public roads, these are full of unexpected hazards that test tracks are generally not susceptible to. Debris in the carriageway is one such hazard that can cause sensor loss or damage and as such a complete loss of signal data with associated quality issues. Sensors mounted in vulnerable areas such as the vehicle under body, wheel housings or bumpers need to be protected so no damage can be caused to brackets, wiring or the sensors themselves. Sensor replacement, if the unexpected happens, can also be made easier if wiring and cable routing is carefully considered. Fig 2-1 shows a simple method of routing a concealed cable to, for example; a wheel speed sensor (shown in red). The connector plug (X) is mounted before the cable passes through the bulk head therefore protecting it from contaminants but making replacement of the sensor simpler with less cable to remove and refit.

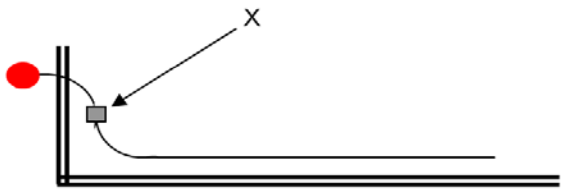

Figure 2-1: Schematic for simple service wiring

Cable routing

Most field operational or naturalistic driving trials require or suggest a level of concealment for the components; this has the advantage of making the driving as natural as possible as the vehicle remains standard looking however a disadvantage lies in the routing and subsequent servicing of wires and components.

With sensor failure being an excepted, but easily mitigated problem, access to equipment installed in the vehicle is of prime importance, this relates to the loggers, sensors or wiring equally.

Wiring should be routed with slack cable at the ends of connection, this relates to the sensor and the logger as these are likely to be frequently accessed. This slack cable acts as a cable full relief, relieving stress on the delicate connectors if the sensor or logger is moved.

On unavoidably long cable runs secondary connectors should be used near the sensor end. This will allow simple maintenance as the majority of the cabling can be left, concealed in the vehicle while the sensor and short cable link is replaced or repaired. This is especially important if the cable routing passes through a bulkhead/firewall or exits the vehicle where servicing complexity is greatly increased.

2.1.2 Calibration

Orientation

A major consideration when installing data loggers in vehicles, especially in cases with an accelerometer is their orientation. Accelerometers are designed to work in defined planes and as such installing them correctly is vitally important.

In an ideal world a three degree of freedom accelerometer measuring force in X, Y and Z directions will have each axis aligned perfectly (as defined by a standardised coordinate system).

If the axis' are not aligned in parallel with an expected force (for example tilted slightly forwards) then this force could be seen on two axis' this is shown in figure 2-2; where x1 and z1 represent aligned axis and x2 and z2 the pitch of the accelerometer by the angle θ. In this example x2 and z2 could both record acceleration measures from the perfectly horizontal 'force'.

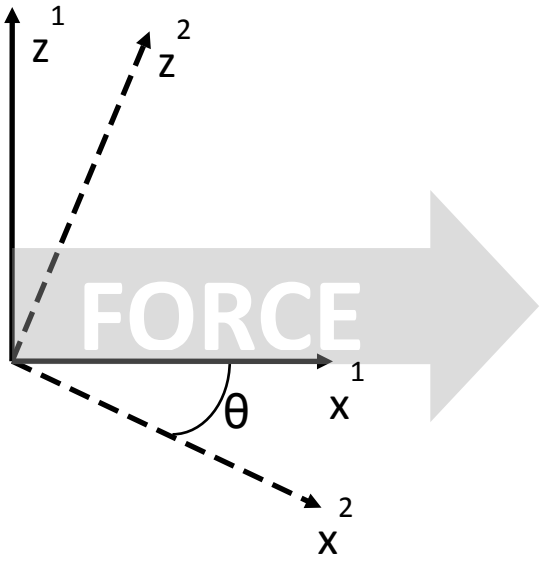

Figure 2-2: Mathematical coordinate system to demonstrate tilt effects

Mathematically, the above diagram and error in the accelerometer readings can be represented by:

g x sin(θ) = expected error in g

Where:

g is constant at 9.81 as this is always present in y

θ is the pitch angle – for this example 10°

This gives an expected error as:

1g x sin(10) = 0.17g

The angle of the accelerometer will also affect the sensitivity of the measurement to roughly the same degree; for this example (10° pitch) the sensitivity reduction could be approximately 2%. It is therefore possible to accurately calculate a range of error and sensitivity reductions for pitch changes in the accelerometer mounting.

This consideration is especially important when mounting data loggers in vehicles with non-destructive mountings, such as those fitted to participant's vehicles. Looking at existing systems fitted to windscreens it could be seen that a range of windscreen angles could give very large quality issues if not corrected for; for

example the MINI has a windscreen angle of approximately 44 degrees whereas the Ford Fiesta's is around 28 degrees, a difference of 16 degrees with associated large quality errors.

Issues with orientation are complicated even further when the vehicle in question is free to move in more than one direction. Instrumentation for motorcycles can be particularly difficult to install correctly as not only does the vehicle lean to turn but the dynamics of acceleration and braking are much more severe than those experienced in 4 wheeled road vehicles. It is unlikely that the same instrumentation in the same form will work equally well for both motorcycles and cars/trucks. For a motorcycle it may be necessary to understand rotational accelerations for cornering through roll sensors (in place of lateral acceleration), conversely it may not be necessary to record roll in cars but almost essential to record lateral acceleration.

Sensitivity

In addition to the sensitivity issues mentioned in the orientation section, brought on by an accelerometer not being ideally located, other sensitivity issues can be present in accelerometers.

Data quality can be very high from specific equipment designed to do the job; in instances where scientific grade equipment is used the output has a better chance of being of good quality (other requirement not withstanding). However these instruments are often prohibitively expensive and/or difficult to obtain in large numbers. To get around this, cheaper items can be used; they can however have a significant impact on quality.

Figures 2-3 to 2-5 show data from an accelerometer designed for industrial/scientific use (Blue line) compared to a cheaper system (similar to those fitted to smart phones – green line). The data is reversed for one device to make it easier to interpret. The data is for three braking events, all from 30mph to 0mph, and at 3 different severities. In each case it is clear that the budget device does represent the event well but close inspection of peak forces and smoothing show significant issues relating to quality.

Peak forces in the moderate braking test (fig 4) show significant differences in peak deceleration (40% reduction in peak force) which could have a major impact on quality if being used for event detection for example. Similarly the inbuilt smoothing of the data for the cheaper item shows very rough and variable data, this could also lead to many false positives or an unrealistic representation of a participants driving style.

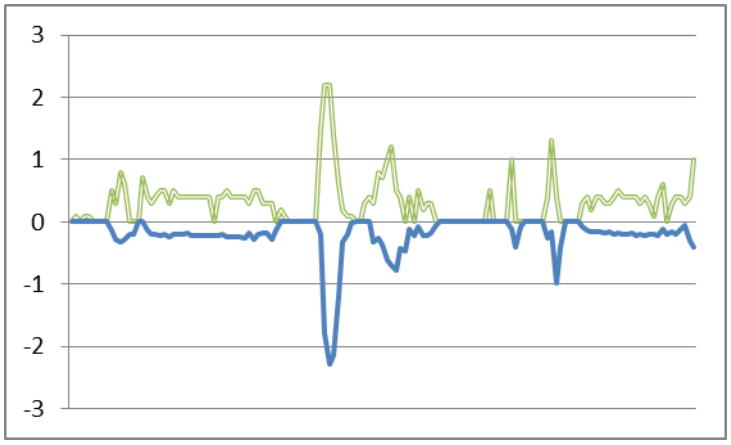

Figure 2-3: Low deceleration threshold top

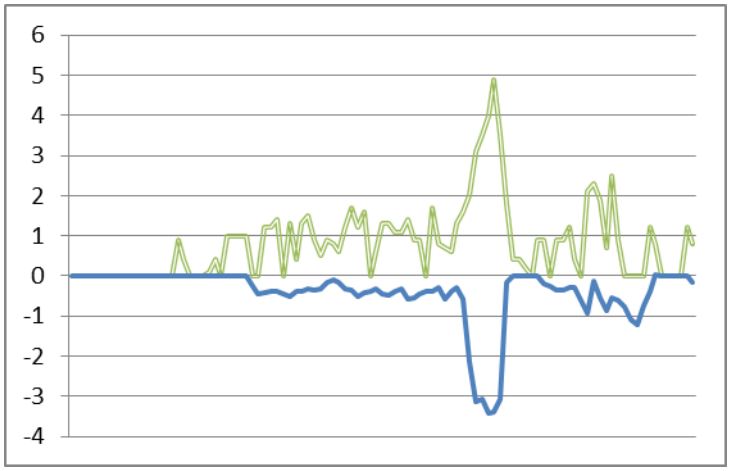

Figure 2-4: Moderate deceleration threshold stop

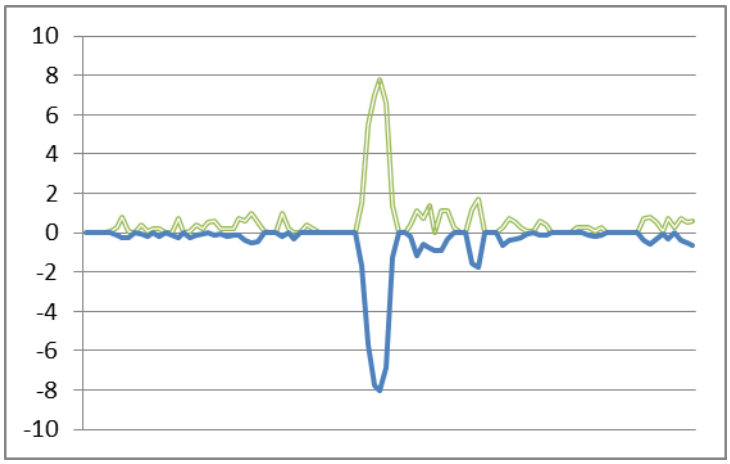

Figure 2-5: High deceleration threshold stop

Another factor that can effect quality and which is present in specifying cheaper instrumentation is that they may not all be equal. This issue is less critical if exploring a single vehicle in isolation but when combining large data sets with many (supposedly) identical instruments quality issues will become evident.

In a follow on experiment to that shown above a batch of identical data loggers with accelerometers fitted were installed into a test vehicle. Five devices in total were installed in a standard road car and data was collected and analysed. This data shows some differences between peak forces and sensitivity with some devices recording many more 'events' than others. To understand this issue it is sometimes necessary to perform a number of short but important tests to identify where issues may arise.

2.1.3 Check routines

A relatively simple step to ensure data quality is to engineer in some form of check routine. This step has a number of options with different associated outcomes but all will ensure data quality on one level or another.

The simplest form of check routine is to perform a short in vehicle calibration once the device has been installed within the vehicle. Most devices whether designed in-house or bought in from a commercial supplier will have completed some form of bench test/calibration phase, this does not mean however that the device will still work adequately once installed within a vehicle. The checks do not need to be particularly complicated or indeed comprehensive but may cover some basics such as the polarity of the accelerometer of the locational accuracy of the GPS receiver based on manufacturers (or project) tolerances.

It has been shown from other studies that a check at this stage does reduce the need to revisit the car shortly after instrumentation to fix problems that only become evident after the first data download. Particularly in motorsport applications a static 'shakedown' of the vehicle electronics is essential before the vehicle turns a wheel in competition.

In addition to the initial static calibration it may be necessary to complete repeated checks throughout the life of the data logger. This is covered in more detail in Chapter 2.3.2 – Off line checks.

2.1.4 Test data

Standard thresholds

As mentioned in the sensitivity discussion above it is not always clear when instrumenting vehicle in isolation that a device is not performing how it should. A simple test may be needed to identify instrumentation before the vehicle is returned to the participant.

In addition to the static 'check routine' type test it may be necessary to use the vehicle as it is intended (i.e. a short road trial) to ascertain whether the instrumentation is working and that it is in line with all other data collected; this step is especially important as it can help to identify instrumentation which is faulty in some way or which needs additional calibration.

In line with project goals a set of tests needs to be devised to test each installation within a vehicle; these tests may include physical driving such as an acceleration and braking event along with cornering tests alongside some more routine checks such as a 'cold' and 'hot' start to see if the equipment powers up and records data in a suitable time frame.

The physical road tests need to be repeatable so standardisation is very important. To keep things standard a general drive, such as running the vehicle around the block, may not be suitable. Instead it may be necessary to perform a number of repeatable tests such as performing a stop with ABS activation from 30kph and acceleration in second gear from 1000rpm with wide open throttle. These tests, although showing slightly different results from one vehicle to the next, have a set methodology and similar thresholds.

The vehicles should be tested to a set criteria ideally set down by the manufacturer and the project as a whole. There may be a need to perform some initial, pilot analysis to see if the planned on road tests give the correct data, this step will also help identify the tests in question.

2.2 Video quality assurance

A major component of any trial now, whether naturalistic or more trails based is the collection and analysis of video data. There are some significant quality considerations relating to this area which are covered in the next section.

2.2.1 Positioning

Positioning of the cameras should be considered equally important as getting good picture quality. There are now some commonly used camera locations in vehicles derived through fairly extensive work in previous projects, these split roughly into two groups; exterior (or contextual) cameras and interior (or behavioural) cameras

Exterior

Quality considerations when mounting cameras to capture an exterior view are similar to those highlighted in the physical consideration section. Some additional points to consider which can affect data quality considerably include how tamper proof the cameras are or how susceptible they are to becoming misaligned; cameras which are moved (either maliciously or accidentally) will have an impact on the data analysis that can be achieved. A quality issue relating to the mounting of cameras out of the way of drivers can be that they lie outside of the windscreen wiper swept area or that the area behind the rear view mirror

(a common mounting location) is often left un-cleaned if it does not hinder the drivers normal view (such as debris, oil or insects for example). Additionally areas such as behind the rear view mirror can commonly become condensed up or remain condensed up as HVAC does not always work efficiently right to the extremes of the windscreen.

Interior

Sighting the cameras in the vehicle cabin may prove the best view of the driver and other occupants and protect them from an aggressive environment, however, other quality issues could arise in the form of tampering and obscuring. This has been shown to occur in a number of studies, particularly early into the trial period, and unless addressed could lead to missing video channels for the whole trial period. Concealed cameras and/or familiarising the driver with the study aims could prove effective for such occurrences.

Other Quality considerations are perhaps more relevant to the research questions being asked. It can be common to record information about driver behaviour from the interior video channels and it can be seen that if critical information is not captured then the quality of the data will be compromised. In a recent UK FOT the camera focussed on the driver had to me moved once the analysis was started as the field of view missed critical information and would have had a huge impact on analysis quality. An early pilot analysis would have identified this issue and is therefore a very effective countermeasure for improving video quality.

In addition to the physical mounting consideration outlined above it is also important to ensure that the cameras can record data reliably and consistently; in this case how they cope with environmental conditions. This quality consideration is much more important if the vehicle is driven in low or variable light conditions as the camera may be unable to react quickly enough to lighting changes. An example of Spanish FOT trial data showed this issue clearly; in some cases a low sun 'bleached' the drivers face and made glance behaviour impossible to detect and in following frames the vehicle entered deep shade where the camera could not react fast enough and all video data was lost. In this case the lighting conditions alternated between direct lighting and deep shade and all video quality was reduced. Artificial lighting within the vehicle can provide good results with Infra-Red (IR) light sources and modified cameras showing very stable result, particularly in low light conditions. Unfortunately direct sun light also contains IR so care has to be taken if recording in these conditions.

2.2.2 Data extraction considerations

Like a lot of other quality issues it is worth considering how the data will be analysed to get results; in the case of video data it is not always straight forward as the data source itself is very rich in terms of information but very restrictive in terms of data extraction.

It is not always enough just to place video cameras within a vehicle and point them at the subject to be recorded. With basic information extraction such as weather/lighting conditions or as a count of passengers this may well be the case but as research questions become more complex so does the installation and data collection from video.

Data quality issues arise when information on distances, headway or road positioning are included in the research question, this data collection will involve some form of scaling off the video image which is not always possible and is in most cases highly variable. As an example data on distance to an object/vehicle ahead can be affected considerably by (by not exclusive to): road inclination, size of object ahead, weather condition, camera mounting position/angle, camera configuration and lens construction. Getting a camera to record this information reliably and consistently to allow for data extraction will require good calibration.

The effect of this is easy to appreciate when comparing video from a forward mounted camera mounted behind a rear view mirror (Fig2-6) with a photograph taken from roughly the same location (Fig2-7). This image shows the typical case of the forward road scene in the video being foreshortened towards the foreground and lengthened in the distance due, in part, to the inclination of the camera and the carriageway. In both these images scaling for distances could be unreliable but as an example illustrate

where video scaling may create data quality issues. Any research question that requires complex video analysis can have major quality issues associated

Figure 2-6: Forward facing video

Figure 2-7: Photograph of road scene

Connected to the issue covered above about recording information from the video frame, is the necessity to have adequate resolution. Video resolution can be described in a number of ways but always refers to how much detail the image (or video frame) holds. The more information in the image (frame) then, theoretically, the more information can be coded for analysis.

When considering the previous issue, that of scaling, it can be seen that information relating to a distant object (pedestrian, cyclist etc.) will be easier to code if the image has sufficient detail. This is especially important if the research question requires information about the behaviour of distant objects; lower resolution will provide considerably less data quality compared to higher resolution images. Even when filming subjects in close proximity a high resolution will provide the best data quality as more detailed information can be collected. A common example of this is the recording of glance behaviour from a driver where a camera may be focussed on the face but eye behaviour is still difficult to collect due to insufficient resolution.

Camera specification and set up will be a compromise as it will be impossible to film every angle in high resolution, partly because of technical issues (data size) and partly because of cost implications; however it is essential that where a research question depends on certain video data to be recorded then a camera of sufficient resolution is supplied to capture this else data quality will suffer.

2.2.3 Limitations of cameras

Exposure

The very nature of filming interior and exterior scenes in real world conditions will have an impact on data quality. Some of the issues that can be experienced are fundamental to the operation of video cameras but can be mitigated or optimised to provide good quality data for the analysis phase.

A common issue with filming on road is the changing light conditions; these are pretty straight forward to understand but come in two main forms, the first being a macro change in lighting conditions (from daylight into darkness) and the second being micro changes (moving from direct sunlight into shade for example).

The first example of changing or different lighting conditions (daylight and darkness) is easier to control for in terms of quality assurance as the change in conditions is slower or constant i.e. continuous driving in darkness/daylight or a gradual change in lighting conditions. In this case a camera with suitable Lux rating will be sufficient to capture good data in these conditions. In addition and depending on the research question it may be necessary to provide additional lighting into vehicle cabins for driving in low light conditions; Infra-Red lighting is ideal for this purpose but will necessitate filters for the cameras in the vehicle.

In conditions with variable and rapidly changing lighting conditions such as those created by roadside objects, bridges, buildings etc. it is more difficult to control for the changes. Figure 2-8 shows a series of video frames taken from a monochrome camera directed at the driver, these show a range of exposure issues with no single frame providing suitable data. A camera with poor automatic exposure control such as that illustrated in the images will also have some latency built in so data can be of poor quality for a number of seconds after the lighting change has occurred, this can impact the data quality considerably if the vehicle is driven in such an environment.

Figure 2-8: Video frames showing under/over exposure

In addition it is also common to experience some degree of image pixilation when a camera experiences a sudden change in lighting (normally from a well-lit environment to a dark environment) as the camera auto exposure struggles to react quickly, this can also have a significant impact on data quality especially if gaze behaviour or other detailed parameter is to be coded. Figure 2-9 shows how this small effect of pixilation can have significant effect on determining the gaze direction.

Figure 2-9: Video pixilation during short periods of low light driving

A difficulty when filming in and around vehicles is the mix of focus and zoom levels that are required. This is evident when filming both near objects; the drivers face, and far objects; the view of the road ahead.

2.3 Tool development

2.3.1 On-line checks

On-line checks in this context refer mainly to 'functional' checks. The definition of functional for this section will be, in accordance with the dictionary "capable of functioning; working". The basis of the quality checks in this sections therefore will be devoted to the working of the DAS in the vehicle and will not (unless absolutely essential) involve other quality issues such as processing for analysis etc.

The major concern in terms of data quality in this section will be faults. These can occur in a number of subtly different ways but all should be covered in the on-line checks to ensure quality.

The on-line checking system should be prioritised for fault finding and automatic reporting of these faults. It is extremely difficult to outline what faults may be encountered during the duration of a trial however there are some basic parameters which the system (and data) should expect.

The first category of these parameters will be focussed on the functional element; is the device working as it should? This will have some very easily definable conditions based on the operational constraints developed during DAS development. Checks may be concerned with checking power supply (Y/N), voltage (within operating range), and temperature (within operating range) among a number of other checks deemed critical to the full operation of the DAS. These issues can have a major quality impact and will need to be monitored throughout the trials and tested during prototyping and piloting.

Beyond the pure operational checks will be a range of other data critical checks. These checks will not necessarily analyse the data to ensure high quality at this stage (this is left for the off-line checks) but may provide an alert if a variable is constantly recording out of range or is not recording at all. The quality check at this point is more likely to form a GO/NO GO check rather than providing a solution or processing any data.

The on-line checks for the data should aim to cover the largest range of variables possible; this will be based on what the research questions require. The data checks will almost certainly form a hierarchical relationship. Some variables (depending on the final set of research questions) will be hugely important for a large number of questions; these will form the basis of the GO/NO GO on data quality at this stage. Data to be checked should also go beyond the pure digital data and will certainly include the video stream; this may be in the form of a small video 'clip' or just one frame as a still image but the importance of checking video quality across the trial durations is hugely important.

The data checks in the on-line phase may, for reasons or practicality, be restricted to routine or expected errors based on the expected data range. Unfortunately data acquisition systems, sensors, and video cameras do not always create errors in a predictable or even uniform way. In these cases, where an error may only affect one data line out of 100,000, it is likely that an error can go unnoticed for some time. A further on-line check should be employed to spot these and collate the errors to understand their scale and effect on data quality.

2.3.2 Off-line checks

Logger conformity

Despite data 'passing' a quality check such as that outlined above it is also necessary to complete a number of other quality checks. Not all loggers in all vehicles will record data equally and it would be unwise during a trial of long duration to assume they will return comparable results. In some cases the data recorded will vary by only a small degree; this may be deemed acceptable but variations above this set level will produce data quality issues.

It is important to identify loggers that do not fit the required standards early. In the first instance these loggers will be detected during the installation and calibration phase where it is specified that a number of simple tests are conducted to identify loggers that return unusual results. Further to this it is essential to replicate this test while the trials are underway and this is best achieved during regular off-line checks.

An off-line check can be of any description but it is essential to compare data from all the vehicles in the trial in order to identify issues. This check can be automated on data upload so to give the fastest possible response time to erroneous data and loggers.

If, for example, an isolated logger drifts from the calibrated settings it may not be enough to identify this particular device under the weight of other data; it may not stand out as being particularly different, especially considering differing driving styles, but it may be significant in that it is not reflecting what is actually happening to that vehicle.

Combining all the vehicle data and comparing each vehicle to this combined group can provide a very clear indication of when a logger begins to fail and the data 'creeps' out of calibration. Minor errors, which in isolation may look insignificant, will become clear as the single vehicle trend moves away from the group as a whole.

This technique was used to good effect in a UK study as part of the TeleFOT project where 80 identical loggers were fitted to passenger vehicles. Even taking into consideration different dynamic characteristics of the vehicle fleet and different driving styles of the participants it was clear from an early stage that some reporting, particularly of accelerations, was either particularly pessimistic or overly optimistic.

The countermeasure to this data quality issue is not always to remove and replace instrumentation. In some cases it is possible to correct for the errors in post processing or to exclude this from the analysis depending on the variance seen. In cases where this technique is not applied and correspondingly the data errors are not identified then this quality issue will remain right through to the analysis stage.

Sensor States

Although not directly a quality issue the categorisation or harmonisation of sensor states is an important issue when it comes to the analysis stage. Sensor states refers to the number of different and unique

outputs a sensor can record for a common variable, take for example something as simple as windscreen wiper activation (something which can have implications for 'weather' detection); The most basic form of this in analysis terms is Y or N, whether the windscreen wipers were on or not. This provides adequate information for the type of weather the vehicle is experiencing (windscreen washing excepted) and can be used as a trigger for analysis (for example checking to see if vehicle speed reduces or headlights are turned on in wet conditions).

There are slight complications to this variable however if some vehicles are equipped with multi speed, intermittent (with multi speed) or automatic wipers – as a lot of modern cars are – as this can in, effect confuse, the raw data. What is needed in terms of quality and robustness of results is a common minimum standard which is applied throughout the vehicle fleet. If for example some of the fleet can only record wiper activation as Y / N then the other more complex coding (possibly 20 different wiper settings) will also need to be reduced to this level; the raw data (those 20 settings) can be kept for further analysis but by harmonising the data it will be possible to conduct a more thorough analysis of all available data.

2.4 Database quality control

Almost irrespective of the study size it will be impossible to check all the data manually to see if it all makes sense. In most cases and in most previous studies the first time a longstanding error is spotted is in the analysis stage; in some cases it is then too late to improve the data quality and the results are affected irrevocably.

The importance of conducting an off-line quality or 'Sanity' check is well understood in road trials and can provide an on-going 'health check' for the systems within the vehicle and the data recorded.

Unlike experimental driving the vehicles in naturalistic driving studies have no limitations applied to their travel. If an issue is detected with a data logger in may only be when the data has been uploaded and analysed or that the driver 'self-reports' a problem. In both these cases a time delay is almost inevitable and repair could be difficult due to location. Automated quality checks can be a very important way of identifying a problem early.

In order to complete a quality check it is important to think of the analysis to be conducted; this may seem the wrong place to start but it will help identify the nature of the check to be conducted. It would be unwise to complete road trials first only to discover that a critical variable in the analysis has been recording erroneous data.

In terms of time scale, particularly in the case of a new study, the first check should be the first data download from the pilot trials or at least the first stage of the pilot analysis (ideally before full scale trial implementation). Data checks however do not stop at the Pilot phase and in most instances during a running trial a routine automated check is required. As mentioned at the start of this section it is not always enough to just look at the data; after all data looks like data whether it is correct or incorrect.

One of the first lines of defence against data errors is a basic sanity check. This process, normally conduced off-line, makes sure some of the most basic but most critical variables are recorded reliably and accurately. For example data which is recorded correctly but has missing time data or missing participant identification information will be useless in the final analysis. Basic quality checks can spot these issues and help reduce missing or unreliable data.

These data checks can be of any form but are mainly focussed on the major data variables as these underpin almost every analysis. For example missing information on time headway for one participant from one country is unfortunate but will only impact a few research questions whereas missing date/time information for one participant could eliminate all their data for all research questions. It is important to identify the critical variables and base the checks around these;

The data checks will almost certainly form a hierarchical relationship. Some variables (depending on the final set of research questions) will be hugely important for a large number of questions – these will form the first

and most important checks. After this the research questions and associated variables used to answer them can be ranked in order of dependence; i.e. How dependent is the success of the project/research question on this variable recording correctly. It will not be practicable to check every variable in great detail in a large scale naturalistic driving study so this 'ranking' will form the basis of the quality check determination.

In addition, the quality of the database depends on good management of the case materials and a well-designed user interface for data input and data downloads. Oracle, SQL Server and MS-Access are examples of database applications available to serve as the central database. A logical hierarchy of relationships between the component data tables, if applicable, is an important foundation and a user-friendly data input system with the capability for validation checks is necessary to create a good quality database.

The trials may generate digital case materials that cannot be incorporated into the central database. This could include images, video, sensor output, and time-location data-streams. It is important that these files are systematically named according to rigorous protocols so that they can be identified by computer logic. This also applies to case directories and folders.

The management of data records - creation (especially), modification and deletion - is the first general requirement. The system should also respond interactively to data input by only showing relevant sections of the forms, hiding those that are irrelevant or not applicable to the case at hand.

2.4.1 Data quality assessment framework

As good practice, a Data Quality Assessment Framework should be followed (Sebastian-Coleman 2013). The purpose of the framework is to define a set of measures that enable basic stewardship of the data based upon objective aspects of the dimensions of quality. A very simple but critical aspect of the design of the database is to ensure that a value has been entered for each field (even if is 'Not Known' or 'Not Applicable') and the data that are entered are valid. A warning should be issued at data entry for values that are valid but extreme, rare or otherwise improbable. There for it is imperative that the expected and valid range for each variable is defined by the analyst in advance to aid the database management. This equally applies for variables derived from the raw data being collected. These definitions form the basis of the data model. Table 2.1 below shows an example.

Table 2-1: Data quality assessment framework - Speed

	Value	Notes
Units	Km/h, m/s	The units the variable will be recorded in
Format	Numeric	Numeric values only, text values will automatically be recorded as -999
Resolution	000.000	To 3 decimal places
Max value	250	Max speed possible, values above this should be recorded as -999
Min value	0	Reversing speeds normally recorded as positive speed. Minus values automatically recorded as -888
Max value (norm')	160	Maximum expected speed = motorway speed + 20% (values above this but below 250kph should be questioned)
Max change between consecutive data points	$8m/s^2$	Repeated changes in speed over this threshold should be recorded as -888
Not known/null	-999	This value should fall outside of the data value range
Not applicable	-888	This value should fall outside of the data value range*

2.4.2 Dimensions of data quality

Dimensions of quality can be quantified as

- Completeness
- Timeliness
- Validity
- Consistency
- Integrity

These dimensions can be assessed by undertaking structured checks on the data. Some checks require an initial one-time assessment such as gaining an understanding of the data and the data environment. Others require periodic measurement and utilise automated processes and in-line database assessments. Figure 2-10 below illustrates the database quality control.

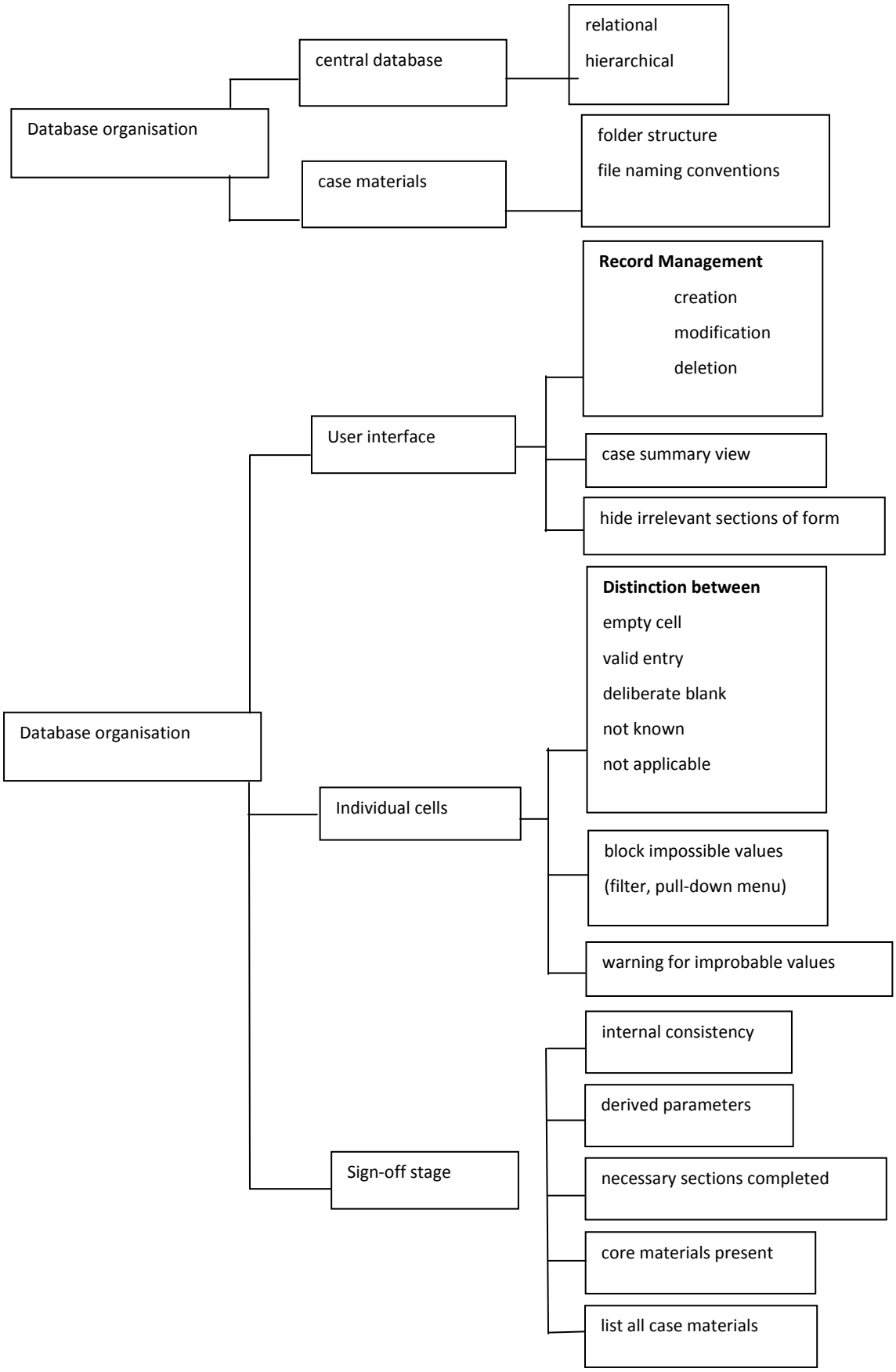

Figure 2-10: Database quality control

2.4.3 Data quality measurements

Formally, the data quality can be measured by taking 'Measurements' against each of the objective dimensions listed above. Examples are given in the table 2.2 below.

Table 2-2: Data quality measurements

Dimension	Measurement Type	Description	Object of Measurement	Assessment category
Consistency	Consistent use of default values in a field	Assess column properties and data for default values assigned for each field	Data model	Initial one-time assessment
Timeliness	Timely delivery of data for data processing	Compare actual time of delivery to scheduled time	Operation site adherence to defined schedules	In-line measurement
Completeness	Filed completeness, non-nullable fields	Ensure all non-nullable fields are completed	Condition of data upon receipt	Process control
Validity	Single field validity	Compare values on incoming data to data model definitions	Condition of data upon receipt	In-line measurement
Integrity	Data-set integrity - duplication	Identify and remove duplicate records	Conditions of data upon receipt	Process control

The Database quality should be built around a logical structured data quality assessment frame work,

The database quality hinges upon receipt of a data model that specifies as a minimum the expected and valid ranges for each variable. Once the data model is available a data base template can be produced. it is recommended that the database quality checks should take into account the following:

Ensuring correct receipt of the data for processing

Ensuring that the data are in their expected condition upon delivery reduces risks associated with data processing. The simplest checks are to ensure that the data are complete. These checks could include;

- Confirm that all required files are available for processing
- Compare record counts in a file to a documented control file
- Compare summarised data to summaries in a control record
- Compare size of input file to typical past input file for individual participants

Inspecting the condition of the data upon receipt

The purpose of these checks are to ensure that the initial condition of the data conform to the expectations described in the specification for each variable, for example;

- Record completeness – length of records matches a defined expectation
- Field completeness - All non-nullable fields are populated
- Data set integrity

Any issues identified during the Database quality check must be reported back to the relevant OS as quickly as possible in order to minimise the amount of data that are compromised. It is also therefore essential that data are uploaded to the data base and data quality checks are applied within a quick time frame. Whilst

issues due to post processing can be rectified relatively easily, problems with raw logged data can result in data loss.

- Duplicate records are identified and removed

In addition to these technical aspects, reasonability or sanity checks can also be made;

- Number and percentage of records defaulted follows a historical pattern
- Ratio of duplicate records matches historical pattern

These sanity checks are most applicable when there is the expectation that the data would have a relatively consistent content upon each receipt.

Checking the results of data processing

Sometimes processing data can produce unexpected results. It is therefore important to check the results of data processing (for example to calculate derived variables from raw data) in a similar way that the initial condition of the data is assessed. In particular, the processed data should adhere to the data specification documentation in relation to completeness and integrity discussed above.

Checking the validity of data content

Before undertaking analysis of a dataset, it is important to determine how much of the dataset is valid. The basis for validity is to make comparisons to a standard or rule that defines the domain of valid values. The two fundamental validity checks are

- Basic validity check – comparison between incoming values and valid values defined in the data specification, including a data range
- Validity check based upon an algorithm – e.g. time and data are converted correctly into an expected value, the interval between consecutive data points is as expected

Checking the consistency of data content

Similarly to validity, consistency focuses on the actual content of the data fields. In order to assess consistency, there is the expectation that data points or data sets will be comparable in definable ways since the consistency of the data is determined through comparisons with previous examples of the same measurements. Checks for consistency include;

- Consistent content of an individual field – measured through a record count distribution of values known as a column profile, or the stability of patterns within the data field
- Consistent content across fields – relationship profile of values in two or more fields

Assessing the overall database content

At certain stages of data input it is recommended to have the user "sign-off". This signals the completion of part or all of the data input stage, including compilation of digital case materials. The data input system should then make cross-checks to ensure that the whole database is internally consistent. Such checks include

- Calculation of derived values;
- Checking that all necessary parts of the database have been filled in;
- Reading the case folders to ensure that all core, required materials are present;
- Listing of all case materials.

This relies on the folder structure and file-naming protocols mentioned above. If problems are detected, the program should block the case from being marked as completed.

The data quality checks should also monitor the vehicle usage in the context of information received from questionnaire / metadata. (e.g. idle period due to holidays, change in driving patterns due to work relocation etc).

2.4.4 Normal Use

Other checks that can be made with the off-line data are based around the predicted use of the vehicles and the participants' background questionnaires. This quality check can maintain good data by identifying vehicles and drivers who are driving out of a predicted range.

This step can be seen as prying, particularly with the participants, but it has been shown to identify issues that cannot be predicted during the project planning, the test site set up and the trial running.

The following are just a few examples from previous projects that have worked in identifying quality issues in the data:

Change of home address/work address

These pieces of information are normally (although not exclusively) collected in the participant background questionnaire and form a framework to the understanding of the driving data; especially if any travel based analysis is to be conducted. During long trials with lots of participants it will be quite common for people to move house or change job and this is normally something that the project would need to know. The associated changes in driving distances, durations and area type can have a large quality impact on the data if this continues unknown. Of course other more significant issues can also arise if a change of travel mode is associated with a change of home or work location – this will have a major impact on the data volume collected and a much lower cost benefit of installing the DAS into a particular vehicle.

Change of use of vehicle

It is not always known when a participant changes their vehicle; in fact numerous stories from trials conducted worldwide suggest that participants prioritise buying and selling vehicles (understandably) higher than maintaining consistent data for the project. Even if the vehicle is not sold then it could change its use; perhaps moving from a 'first' car to a second vehicle in the family. Examining the travel data can identify these issues early and therefore identify whether a participant can still be included in the trial. It is worth understanding that participant annual mileage recorded in background questionnaire is often overestimated and as such a lower than predicted mileage total (per week/month) in the collected data is not necessarily indicative of long term data quality problems.

Idle periods/holidays

It is highly unlikely that, over the course of a trial, the vehicle will be used every day; in these cases stationary data (or more likely no data) will be recorded. This situation is perfectly normal and can be as a result of some of the issues outlined above; however there are situations when long stationary periods can be detrimental to data quality. Intentional non-use is quite common and is normally evident in long vacations; however it is essential to understand that this is a vacation and not a reoccurring logger problem. In order not to interfere with each participant in these situations it is essential to build in a time delay to the checks, this delay will also allow the check to detect the end of the vacation if this is the case. If not then contact with the participant may be needed in order to establish the true nature of the idle period.

3. Data Quality application within UDRIVE

In this section, the general principles from the previous section are applied to the UDRIVE project and the activities undertaken within the project in relation to data quality are described. Data quality within UDRIVE is reported under three headings:

- Pilot testing including piloting the in-vehicle DAS, test of the data management chain and OS piloting
- Data process chain including vehicle installation, on-line monitoring during data collection, data tracking, data pre-processing and data post-processing
- Operation site guidelines providing specific actions to be undertaken by each OS

3.1 Pilot Testing

In order to ensure quality of the data upon full scale operations within UDRIVE, pilot testing has been undertaken in a number of areas. These include piloting the in-vehicle DAS, tests of the data management chain and also piloting at each of the Operation Sites. A brief summary of the objective of each of the pilot activities is given here, full reports on the activities are available in the UDRIVE deliverable D25.1 Validation report in technical piloting of DAS and data management chain (Restricted), D33.1 Overview of OS preparation, sample characteristics & piloting (Public) and D33.2 Overview of OS preparation, sample characteristic & piloting update (Public).

3.1.1 Piloting the in-vehicle DAS

The objective of these pilot tests are to assess the correct technical functioning of the data acquisition systems in each of the vehicle types within the UDRIVE project, cars, trucks and PTWs. The tests address the specificity of these different vehicle types to ensure the correct adaptation of equipment as well as the quality of the data acquired in driving conditions. Both bench testing (laboratory) and in-vehicle testing has been undertaken.

The main functions tested include:

- Power management
- CAN logging
- Video logging
- Storage
- Monitoring / remoting
- Configuration / update

3.1.2 Test of the data management chain

After the finalisation of the pre-test implementation of the database and the respective analysis tools, these aspects of the data management chain have been tested to provide input for potential areas for improvement prior to final implementation. The tests include upload and download of the data, the access to the data with the developed tools and the enrichment of the data using these tools. The steps identified in the data processing section below are also tested. Further elaboration on the data management chain is provided in sections 3.2.3 and 3.2.4.

3.1.3 Operation Site Piloting

In order to be confident that good quality data will be collected at by each OS, a small scale pilot (typically 2 weeks long) has been carried out by every OS. This builds on the technical piloting by validating the data collection and management under actual study conditions.

The pilot is a preliminary study that is also representative of the main study. The aim is to validate the complete tool chain (from participant and vehicle reception, Data Acquisition System installation to data collection) and the corresponding procedures in each OS specific context.

This preliminary study has been divided into several successive steps each of which are validated before progressing. The steps undertaken in the pilot process are listed below. Specific data quality checks related to the OS are elaborated in section 3.3.

Prerequisites

Before starting piloting, each OS should check that:

- Operation Site responsible and teams are identified
- Documents needed by the OS before the piloting are ready
- On-line Monitoring Tool (OMT) is configured for the OS
- Suitable suppliers have been selected for:
    - Data Acquisition System Installation team
    - GSM data transfer (SIM cards)
    - Hard drive shipping to Local Data Centre (LDC)
- Pilot vehicle and participants are selected
- Suitable facilities to receive participants are prepared

Description of PILOT TESTS PROCEDURES to be followed by the OS

- Registration of the participants
- Reception of the vehicle and instrumentation by the DAS installation team
- On-line Monitoring Tool connection
- Data collection
- Data quality check
- Data transfer to LDC/CDC
- DAS removal by DAS installation team

Report on pilot testing

Each OS reports the results of the piloting, including feedback on the difficulties encountered, the solutions which were applied to solve each problem, and fills-in the checklist and tables which are given as guidance.

Based upon the individual and collective experiences during the piloting, adjustments can be made to ensure that the quality of the data is not compromised during the main study.

3.2 Data process chain

This outlines the core activities that have been undertaken with the UDRIVE project throughout the data process chain. These are identified as :

- Installation

- On-line Monitoring
- Data pre-processing
- Data post-processing.

Figure 3-1 below illustrates this data process chain.

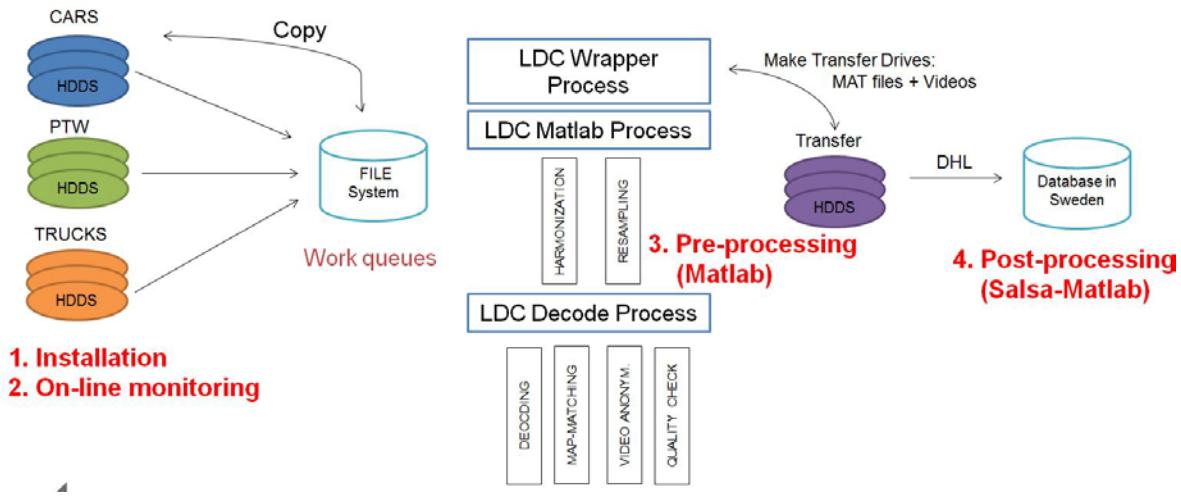

Figure 3-1: Data processing chain overview

3.2.1 Installation

The following have been delivered and undertaken within the project to ensure quality in the DAS installation:

- DAS adapted for each vehicle type
- Installation manuals provided for each vehicle type
- DAS configuration training provided for OS managers
- DAS configuration instructions provided to OS
- Camera configuration manual provided
- Training sessions provided for each OS intallation team
- Support provided upon request at all stages of the installation

Only approved installers who have undertaken the training will be used for installation and de-installation of the DAS.

3.2.2 On-line Monitoring Tool and data tracking

UDRIVE has developed an On-line Monitoring Tool. The purpose of the OMT is to monitor the status of the DAS with corresponding data in the field and to keep a status log of collected data along its lifetime through the data flow. Essentially, the OMT allows the OS to check that vehicle based data are being recorded as expected for each vehicle and that the video data has not been disrupted. The OMT is accessible through a website with restricted access. Daily usage of the OMT includes monitoring the status reports received from the instrumented vehicles. A heartbeat signal is sent from each DAS at start-up and indicates the start of a new record. At the end of a session a status report is then created for the record. At the next DAS start-up, the status report is then then sent via GPRS to OMT server where it is parsed and checked and then entered into the database. As soon as the information has been entered it can be seen on the OMT website thus

completing the information about the record. The OMT contains website contains overview pages for different 'entities' these being vehicles, records, DAS or hard drives.

By monitoring the different record report entries, the OS user can assess whether the DAS is in a healthy state or not. For data protection, the video snapshots can only be reviewed by authorised OS personnel and the snapshots are hidden by an overlay to minimise accidental display. The mouse pointer must be moved over the snapshot in order for it to be shown. The record details page provides the possibility to mark a record for deletion. This function is for cases where the vehicle owner requests that a certain part of the data should be excluded from processing. Subsequent data handling at the LDC will make sure that records marked for deletion are not processed any further. Table 3-1 shows the report entries are available for review and used for monitoring the DAS status.

Table 3-1:OMT report entries

Entry name	Unit	Description
Storage	%	Percentage of free space on the HDD
Last Seen	yy/mm/dd hr:min:sec	The date and time of the last heartbeat /full record
Corrupt	True/False	Indicates a corrupt status report
DriveHealth	0 (normal) 1 (warning) 2 (error)	Based on SMART data received by the DAS from the HDD, this indicates the state of the HDD
GPSCount	-	Number of GPS packets received
GPSCountR	Packets/time	Ratio of number of recorded GPS packets to total recording time
RecordRatio	Byte/ms	Ratio of record size and total recording time
Has Excerpts	True/False	Indicates whether the report contains data excerpts and video snapshots or not
SnapshotNr	-	Number of video snapshots included
CANTotal		Total number of CAN frames received by the data logger
CANTotalIR	Frames/minute	Frame collection rate

Guidelines for use of the OMT

Responsibility

The following should be observed when possible:

- At least one primary and one secondary person should have access to and responsibility for checking the OMT. It is the main responsibility of the primary person but the secondary will take over in cases of illness or holiday.
- Ideally those responsible should refrain from taking holiday at the same time. Should this prove unavoidable contingency measures will be required in order to maintain the data quality such as a third person access to the OMT.

Frequency

- The OMT should ideally be checked daily during the course of the working week.
- The OMT should not be checked remotely (i.e. from home across the weekend) if this contravenes the DPC for the OS

Record Keeping

Record keeping is appropriate for the following

- To record meta data relating to the participants travel
- To record the status of parameters when the OMT is checked

Meta data

A record should be kept of any information relating anticipated interruptions to data logging. These would be for example

- Participant holidays
- Participant illness
- Vehicle in garage

These can then be cross referenced against any breaks in data logging observed in the OMT. In an ND study it is important to maintain as little contact as possible with the participants in order to preserve ecological validity. Participants can however be reminded, for example when there is a HDD exchange, that they should let the OS know of any periods when the vehicle is unlikely to be on the road.

OMT check LOG

Certain key information should be recorded and updated each time the OMT is checked. This helps to easily identify key irregularities and also serves as proof that the DQ activity has been undertaken. The following very simple template is suggested (table 3-2):

Table 3-2: OMT check LOG

OMT DAILY LOG DATE: 24/11/2105 Checked by: Martyn						
ID	Last Seen	Last good record	Last snap shot	Snap shots ok?	% HDD free	Notes
UK_001	24/11/15	24/11/15	23/11/15	Yes	78	
UK_002	23/11/15	23/11/15	23/11/15	No	84	Driver action camera appears dislodged
UK_003	24/11/15	24/11/15	24/11/15	Yes	76	
UK_004	20/11/15	19/11/15	19/11/15	Yes	85	On holiday 20/11 till 28/11

This record can be kept and updated for example in excel as a spread sheet or more simply as a wipe clean board. Whichever way is chosen, the purpose is to highlight any actions that need to be taken. These are discussed in more detail below.

Actions based upon OMT checks

Last saw the vehicle

If there is a period (a couple of days) where the vehicle has no record

- Based upon a review of records for this vehicle, establish whether it is normal to see no activity for a few days
- Check against any meta data for the participant, are they on holiday for example?
- If needs be, contact the participant for an explanation

If there is no reason why the vehicle should not have logged with the OMT then try and establish where the problem may be.

- Call the vehicle in for checking
- Check no cables are loose at the DAS
- If possible check the SIM activity through the mobile provider (is it the DAS or the SIM at fault?)

- Reconfigure the DAS
- If none of the above resolve the problem then contact project technical support

Last snapshot

Typically snapshots are sent roughly every 48 hours. If there are records on the OMT, so the vehicle is 'talking' to the OMT but no snap shot for 3 or 4 days then

- Call the vehicle in to be checked
- Check the camera cable into the DAS
- Reconfigure the DAS (this normally rectifies the problem)

If none of the above solve the problem then contact project technical support

Check camera views

If there are snap shots on the OMT but they are incorrect or some are missing then

- If camera view has moved i.e. camera dislodged, adjust the camera and check using the camera settings procedure (installation)
- If camera is not working, check cables and if this does not solve the problem contact project technical support

HDD space

The HDD should be exchanged frequently enough to ensure that the data flows sufficiently throughout the course of the project. Ideally a HDD exchange should be undertaken every 2 months.

- It is unlikely that a HDD will become full in 2 months
- However, monitor the storage capacity and ensure that, if need be, sufficient time is set aside to arrange a HDD exchange with the participant before the disk is full.
- As a guideline, arrange a HDD once there is only 20% capacity left.

OMT Tracker

The OMT tracker works exclusively with the use of pre-assigned QR codes. Each Operation Site, vehicle, DAS and HDD has its own ID on the form of a QR code. The QR codes are printed and attached to the relevant unit so as to be easily accessible but out of sight of non-authorised personnel

Using a smartphone, a web link can be accessed through the QR code. The website provides different actions for assigning, attaching/detaching, storing and transporting units. The tracker allows information about the present set up of a vehicle/DAS/HDD combination to be easily entered and stored for reference as well as the location of any external HDD to be identified. For the tracker to function properly if is very important that the QR codes and web links are used for each of the different actions listed below:

- Installing a DAS in to a vehicle
- Attaching an external drive to a DAS for the first time
- Detaching an external drive from a DAS
- Replacing an external drive
- Transporting an external drive for temporary storage at the Operation Site
- Receiving / storing an external drive at an Operation Site or a Local Data Centre
- Transporting an external drive to a Local Data Centre

- Removing a DAS from a vehicle

Detailed instructions for each operation in relation to the actions listed above have been distributed to the Operation Sites and are reported in section 3.3.

3.2.3 Data pre-processing

Following the data quality checks at installation and through the on-line monitoring tool as described above, a third point of intervention has been at pre-processing, i.e. after vehicle instrument readings have been decoded in the LDC Wrapper Process, as represented in Figure 3-1, and imported into Matlab. Prior to this, most of the data was encrypted according to the vehicle manufacturer's proprietary format. Scripts were written in Matlab to scan the data files, either individually or in batch mode. Three main types of check were performed:

- that fields existed, i.e. had been recorded and carried through the decoding process,
- that the range and distribution of field values were plausible,
- that different fields were consistent, where this could be cross-checked, e.g. by comparing independent measurements of velocity and distance travelled or by comparing independent measurements of velocity as provided by in-vehicle sensors or GPS location.

Alerts were raised where anomalies were detected. The threshold values within the scripts that determined the issuing of alerts were programmed to be easily adjustable as experience with the naturalistic driving data increased.

3.2.4 Data post-processing

Following pre-processing, UDRIVE data was made available for analysis within the Salsa interface. It was necessary to derive many of the parameters used in analysis from the source data fields. Two of the main ways in which this was achieved were (a) through annotations based on manual video review and (b) computational processing of the vehicle instrument readings, e.g. by searching for high acceleration values. The processing of the data required at this stage to derive reliable parameters for direct analysis is shown in Figure 3-1 (4. Post-processing).

The scripts created at the post-processing stage were written in Matlab within the Salsa environment. Whereas pre-processing was aimed at ensuring the existence and integrity of the source data, post-processing was primarily aimed at the creation and integrity of parameters suitable for use and presentation as the results of naturalistic driving analysis. A template was created to record essential information about the origin and history of development of each script, including the name and affiliation of the first and any subsequent programmers, the date of revisions, the reason for revisions, the nature and outcome of checks made on the algorithms contained in the scripts, and notes on the scope and limitation of the scripts.

3.3 Operation site guidelines

The following guidelines have been provided to each OS.

It is the responsibility of each Operation Site to ensure that the data delivered to the local data centres is as complete and reliable as possible. In order to achieve this, a number of data quality tasks should be undertaken and a team member be named as responsible for the data quality (this may be someone with other OS responsibilities such as the OS leader or the OMT user). The data quality responsible person should monitor the completion of the recommendations made below. These relate to all stages of the data collection process; Piloting, Recruitment, Questionnaire data, DAS Installation, Field Data Collection. OS Data Storage, Data transfer, Feedback from Database Quality Checks, Metadata Collection. The project will provide tools to enable some of the tasks required, others will require a protocol to be established as part of the OS management.

3.3.1 Recruitment

The outline data analysis will be developed based upon data collected according to the sampling plan outlined in Deliverable 1.2.1 Study Plan. It is therefore important that this sampling plan is followed in order that the later analysis is not compromised.

- All efforts should be made to meet the sampling requirements outlined in Deliverable 1.2.1
- OS should outline a suitable advertising strategy which gives confidence that recruitment targets will be met.
- OS must inform SP3 leader in the event that it becomes clear that targets will not be met.
- OS should feel confident that potential participants will engage with the project throughout the course of the entire data collection period. Some issues that may affect dropout can be explored in the OS recruitment questionnaire (e.g. likely job change, change of vehicle). Any foreseeable potential issues should be identified prior to confirming recruitment into the study.
- It is not anticipated that there will be budget to enable 'reserve' participants to take the place of any drop outs.
- It is therefore important that participants are inconvenienced as little as possible in order to further minimise the risk of dropout.

3.3.2 Questionnaire data

Once confirmed as a participant, each participant will be required complete a participant questionnaire that should be available in each local language. It is important that this data is complete and reliable. The following recommendations are made.

- The questionnaire should be piloted among non-project related volunteers in order to establish any ambiguities / difficulties with completion.
- Any issues identified from the piloting should be addressed locally and also shared for the benefit of other OS.
- If possible, the participant questionnaire should be completed by participants in the presence of a member of the OS team (e.g. when also signing the participant agreement), This will allow any questions to be addressed directly to the OS staff member and reduce the risk that guesswork may take place in the event of uncertainty.
- If the questionnaire is to completed in written form, a procedure should be in place to ensure the quality of the data entry into electronic format. The standard approach is for a random sample to be double checked for accuracy.

3.3.3 DAS Installation

The DAS installation is fundamental to the subsequent data collection, if this is not undertaken with due care then the data could be compromised. This has been covered in detail above but the following guidelines should be considered by the OS manager;

- The DAS installer should be competent and trained in respect of the OS DAS
- The OS should provide a location for installation that allows the installer to undertake the work as stipulated in the installation guide
- The installation guide must be followed without exception
- Efforts should be made (within reasonable constraints) to undertake installation at the convenience of the participant.

3.3.4 Field Data Collection

Complete and accurate field data is crucial to the later analysis. The On-line Monitoring tool has been developed to enable the OS to see that data are streaming. (Guidelines for using the OMT are elaborated fully in section 3.2.2)The OS should consider the following in order to minimise any compromise to the data quality.

- The OS should ensure good communication channels are available for the participants to contact the OS throughout the course of the data collection phase. This will help to reduce the risk of participant dropout.
- The OS OMT user and / (or if the same person) the data quality responsible person should be trained in the functionality and use of the On-line Monitoring Tool (OMT).
- There should be another OS team member capable of using the OMT in order to cover the absence of the DQ responsible person.
- The output from the OMT should be reviewed daily in order that any issues are discovered quickly. This includes communication between the DAS and the OMT (does DAS send signal of communication even if vehicle doesn't move?)
- If problems are highlighted by the OMT then immediate action should be taken to investigate further and solve the issue (liaison with the DAS technical support and the participant)
- An initial DAS hard drive swap should be made early in the field data collection phase in order that data can be uploaded to the central data centre and database quality checks performed (e.g. after 2 weeks of live data collection). This will highlight any issues with the data not covered by the OMT that can then be addressed early on in the data collection phase.
- Adequate time should be allowed for collection and exchange of DAS hard drive. A possible guideline would be that contact is made with the participant to arrange a suitable time when there is 2 weeks drive time storage capacity remaining. The amount of capacity will vary from individual to individual but can be estimated for each participant based upon historical usage over time.

3.3.5 Data Storage and Transfer

It is important to ensure that no data are lost during the process of transferring data from the OS site to the LDC.

- A reputable and reliable data courier should be used for data transfer.
- The correct data tracking procedures should be followed (see section 3.2.2)

3.3.6 Feedback from Database quality

Once received by the LDC / CDC, data quality checks will be applied to check the integrity of the data. It is anticipated that the OMT will show whether or not data are streaming, but it will not check for the validity of these data. This will be dome at a database level. Feedback should be provided by the database manager to the OS data quality person so that any problems with the validity of the data can be managed at the OS.

- Communication should be established between the OS data quality person and the DB quality person.
- Issues relating to the validity of the data should be logged dealt with as soon as possible after alert from the DB quality checks.
- The OS DQ person should direct the alert to the most appropriate person dependent upon the nature of the problem.
- Upon rectification, sample data from the effected DAS should be sent to the DB for further checks to ensure the data are now valid.

3.3.7 Metadata

Metadata is a set of data that describes and gives information about other data. It enables the user of the data to understand more about the data. In the context of UDRIVE, metadata is important as it can help to explain unusual patterns in the data and deviations from the normal for participants. It represents non-driver decision related factors that could have an influence on participant's driving behaviour/ pattern.

This type of data will not be logged by the DAS and will not be included in the participant questionnaire since is represents changes that occur during the course of the data collection period. The OS should log metadata so that it can be added to the data base for data clarification at a later date.

Examples of metadata include extremes of weather, national road policy, change of work place. Clearly metadata in the context of UDRIVE is broad.

3.3.8 Piloting

There is a requirement that all OS pilot the data collection procedure (see section 3.1.3). The above guidelines should be followed during the piloting. It is essential that any issues relating to data quality that become apparent during the piloting are logged and reported back to the authors of this deliverable so that amendments / additions to the data quality guidelines can be made and shared among all of the OS.

- The DQ responsible person should log any issues relating to DQ during the piloting
- The DQ responsible person should report these issues to the DQ deliverable authors

4. Conclusions

This deliverable presents the concepts and application of data quality in the context of data collection, storage and database management. The processes described are those applicable to collecting field based driving data.

Principles of data quality have been discussed in relation to:

- DAS installation
- Video data quality assurance
- Tool development in relation to on-line and off-line checks
- Data base quality control

These generic principles have been applied within the UDRIVE project in order to assure good quality data are available for analysis and these have been described within this deliverable. Full details for some of the processes are available in other reference UDRIVE deliverables. Where these are restricted deliverables, only aims of the work undertaken has been presented in this deliverable.

Within UDRIVE, piloting has been carried out that tests the data management chain and also the all aspects of the data collection by pilots undertaken at each operational site.

Measures have also been put in place though the project that demonstrates data quality consideration throughout the entire data processing chain within the UDRIVE project:

- Vehicle installation with consideration to cars, trucks and scooters
  - Guides developed for installation teams
  - Manual provided for DAS configuration and camera configuration
  - Training provided for installation and configuration
- The development of an On-line Monitoring Tool for monitoring the on-going quality of the data by the operation sites
- Instigation of a QR controlled data tracking system
- Data base quality checks (pre-processing)
- Data post processing logging

Additionally, full guidelines have been established for adherence by each operation site in order to reduce the risk of data loss during the course of the data collection.

Thus, the objectives and aims of the deliverable have been met.

5. References

Sebastian-Coleman, L (2013) Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework (The Morgan Kaufmann Series on Business Intelligence)

Lai, F., Carsten, O., Schmidt, E., Petzoldt, T., Pereira, M., Alonso, M., Perez, O., Utesch, F. and Baumann, M. (2013). Study Plan. Deliverable 12.1 of the EU FP7 Project UDRIVE

UDRIVE deliverable [25.1] [Validation report on technical piloting of DAS and data management chain] of the EU FP7 Project UDRIVE

Quintero, K., Val, C., (2016) Overview of OS preparation, sample characteristics and piloting. Deliverable D33.1 of the EU FP7 Project UDRIVE

Welsh, Reed, Talbot, Morris 2010 Prologue D2.1 Data collection, analysis methods and equipment for naturalistic studies and requirements for the different application areas https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/9322/5/AR2649%20Data%20collection,%20analysis%20methods.pdf

6. List of abbreviations

Abbreviation	Meaning
DB	Data Base
CAN	Controller Area Network
CDC	Central Data Centre
DAS	Data Acquisition System
DPC	Data Protection Concept
DQ	Data Quality
FOT	Field Operational Test
GIS	Geographic Information System
GPRS	General Packet Radio Service
GPS	Global Positioning System
GSM	Global System for Mobile Communication
HDD	Hard Disk Drive
HVAC	Heating, Ventilation and Air Conditioning
HGV	Heavy Goods Vehicle
IR	Infra-Red
LCD	Liquid-Crystal Display
LDC	Local Data Centre
OBD	On-Board Diagnostics
OMT	On-line Monitoring Tool
OS	Operation Site
POV	Principle Other Vehicle
PV	Participant Vehicle
PTW	Powered Two Wheeler
SCE	Safety Critical Event
SP	Sub-Project

7. List of Figures

8. List of Tables

Appendix A Consideration of UDRIVE Variables

Throughout this section the variables to be recorded are assessed for both their importance for the study i.e. how reliant the study is on the collection of the variable in question and the level of data quality issues associated with this variable.

Each section below begins with the title of the variable drawn from the variable list supplied within the project and is followed by a discussion of the data quality issues identified for collecting this variable successfully and with high quality.

Following the text discussion is a table whose main aim is to outline the steps that need to be taken to improve or ensure high data quality.

RQ reliance
Identifies how reliant the study is on this variable – rated between high, medium and low
Test required
Step 1 to ensure/improve data quality
Step 2 to ensure/improve data quality
Step 3 to ensure/improve data quality……..

Colour coding for reliance and quality issues

Green: Low

Amber: Moderate

Red: High

Reliance	Quality

A.1 Time

Normally GPS time – format is well defined and will not present many if any data quality issues

N.B if synchronising time then the GPS time definition should be used as the root as it will be more stable than other computer or device times.

RQ reliance		
High		
Test required	Reliance	Quality
Independent time comparison with pilot data logger		

A.2 Time of day

A clear definition needs to be provided for this variable. Time of day can be classed in a number of ways; for example, what times constitute Morning, Noon, Afternoon, Evening and Night need to be tightly defined in the coding algorithm.

RQ reliance		
High		
Test required		
Independent time comparison with pilot data logger	Reliance	Quality
Clear definitions for time of day		

A.3 Weather

Defined from video coding – coding ambiguity the biggest challenge. Difficulties, and associated data quality issues, can arise if the coding manual does not provide clear enough instructions or if the video data is not sensitive enough to meet the coding manual requirements. For example if the coding manual is very detailed then the video data needs to be of a high enough quality to determine between dense, light rain, Fog and Mist.

A coding manual with examples of the weather should be produced as a reference to all video coding centres to ensure consistent and high quality data.

Some variables such as 'sunny' or 'raining' will be more easily defined than others; care needs to be taken that these do not become 'default' variables i.e. it is easier just to code 'rain' than to determine between light rain and heavy rain.

RQ reliance		
Moderate		
Test required		
Coding manual to be produced and followed to illustrate case		
Independent weather coding of short video sections in Pilot – cross centre comparison to check quality.		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Spot checks and coding committee scheduled during full scale data analysis		

A.4 Road condition

Defined from video coding – coding ambiguity will be the biggest challenge. These variables need defining to the same level as weather (above)

It is often difficult to determine detail of road condition from video as low quality images tends to make all road surfaces look better than they are. The higher the resolution of video camera the more detail will be recorded from the road surface; clearly the forward (roadway) camera will be compromised and as such this detail will be lost.

Coding definitions should be designed to promote positive coding and to avoid ambiguous values. The danger with creating a generic coding manual for road surfaces is that all video will be coded 'good' by default as this is easy, by developing the coding manual (more and clearer definitions) it may be possible to encourage positive coding and increase data quality.

Values such as 'slippery' or 'dangerous' should be avoided as these are intangible and are a by-product of the road surface, not a description of the road surface itself. It will also be very difficult to determine with any certainty if a road surface is slippery from a video channel.

RQ reliance		
Moderate		
Test required		
Coding manual to be produced and followed to illustrate case		
Independent road condition coding of short video sections in Pilot – cross centre comparison to check quality.		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Spot checks and coding committee scheduled during full scale data analysis		

A.5 Number of lanes

Number of lanes is relatively easy to define from video analysis providing a clear definition of what constitutes or divides a lane is provided. On Motorways the coding is relatively straight forward and therefore quality of coded data should be high, however in towns, with bus lanes, restricted lanes and cycle lanes it is more difficult to clearly determine this.

It will be necessary to determine clearly in the coding manual as to what constitutes a lane. For example if a cycle lane is painted (but not physically divided) to the side of an unrestricted lane it may not be relevant until there is a cyclist using it, this may be particularly relevant in some of the research questions which refer to cyclists; how this is managed will have an effect on overall data quality.

RQ reliance		
Moderate		
Test required		
Coding manual to be produced and followed to illustrate case		
Independent lane coding of short video sections in Pilot – cross centre comparison to check quality.		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Spot checks and coding committee scheduled during full scale data analysis		

A.6 Visibility conditions

A clear definition of visibility conditions needs to be provided to achieve good quality data; it is important not to confuse weather conditions with visibility (there is, after all, a weather conditions variable) so the coding needs to convey the visibility and not the environmental conditions i.e. the code should be independent of lighting or weather.

Possible but extensive quality issues may arise with this variable as visibility is a pretty subjective matter, or rather that there is almost always a discrepancy between what a driver can see and how the video shows what a driver can see.

Other issues may arise in the case of camera fogging or windscreen misting as these can give an overly negative report of visibility.

RQ reliance		
Low		
Test required		
Pilot tests for different visibility conditions such as darkness or low sun – this should provide data for coding manual		
Coding manual (and pictorial examples) to be produced based on findings from pilot tests		
Independent Visibility coding of short video sections in Pilot – cross centre comparison to check quality.		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Spot checks and coding committee scheduled during full scale data analysis		

A.7 Occlusion of sight (inside)

Typically easy to identify objects in the region of the windscreen, can be very difficult to determine actual occlusion of sight.

Windscreen mounted devices (such as that proposed) will have a very limited and possibly distorted view of the windscreen making a completely reliable coding methodology, and therefore high quality data collection process impossible.

Due to the focal length of cameras filming through glass (front and rear cameras for example) it is often easy to record misleading data; small occlusions (dirt/ice) can seem very significant at times and insignificant at others depending on where they fall in the camera view.

Fogging of side windows is more reliable but can be susceptible to lighting conditions as this can make the occlusion seem worse (bright sunlight) or better (shade) than it is in reality.

RQ reliance		
Low		
Test required		
Artificial occlusions should be applied with pilot instrumentation to understand scope		
Coding manual (and pictorial examples)to be produced and followed based on above test	Reliance	Quality
Independent Occlusion coding of short video sections (or still images) in Pilot – cross centre comparison to check quality.		

A.8 GPS coordinates

By utilising the standard format for GPS location coordinates it will be possible to achieve good data quality. This quality will be dependent on the quality of the GPS equipment built into the logger device

Quality issues with the data can occur through a number of channels; the most common of these are if the GPS signal is poor or if the data is processed before being used. These quality issues may not be evident when viewing or analysing the raw data; visualising the data may provide a good representation of the route taken, however these quality issues may have a knock on effect when map matching is attempted with this data as the GPS positions may return spurious data or match to adjoining or adjacent roads.

A clear understanding of where GPS reliability changes from reliable to unreliable needs to be developed. Commonly this is derived from the number of satellites used to determine the position; the more satellites the more accurate the location. As satellite numbers tend towards zero then a cut-off point needs to be identified from where no further data is used due to these inherent inaccuracies.

If data is converted, rounded, imported in other formats it is important to check that the data quality is carried over – data points can shift if data is not directly replicated causing extensive data quality issues if other variables are reliant – such as map matching, speed limits, area type etc.

RQ reliance		
High		
Test required		
Data quality check on GPS point accuracy for pilot analysis		
Quality check the processing procedure to determine the accuracy after any transfer or processing		
Spot check data during data collection	Reliance	Quality
Determination of cut-off data to ensure only high quality data is collected.		

A.9 Traffic light status

A clear definition of expected traffic light states needs to be completed prior to analysis; traffic lights differ across European countries and while these are minor could have a bearing on coded data quality.

It may be necessary to define all the expected states (static light phases as well as flashing light phases) alongside other instructions such as turn filter lights, cyclist/pedestrian phases and countdown lights.

There are of course other traffic light signals that offer instructions; these are predominantly found on major trunk roads and motorways but can sometimes be found in urban settings. Lights such as temporary speed limits on motorways and lane closure information for multi-lane roads/tunnels/bridges are some of the most common. It may be necessary to include these in this variable as they will almost certainly have an effect on the driving behaviour of the participant.

Not all traffic lights will be visible from the viewpoint of the forward facing camera, especially if the vehicle is first in line or behind a large vehicle, there needs to be allowances for unknown codes to aid data quality. Positioning of traffic lights can also make then difficult to determine from video.

RQ reliance		
Low		
Test required		
Coding manual (and pictorial examples) to be produced and followed to illustrate case		
Independent traffic light coding of short video sections (or still images) in Pilot – cross centre comparison to check quality.		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Spot checks and coding committee scheduled during full scale data analysis		

A.10 Presence of road works

Depending on the coding strategy this variable is relatively easy to ensure good quality data; potential issues relate to whether the road works are relevant or not to the road scene (isolated works on the footway for example) or whether different types of road works need to be coded for clarity (traffic light controlled and segregated works compared to ad hoc repairs).

RQ reliance		
Low		
Test required		
Coding manual (and pictorial examples) to be produced and followed to illustrate case		
Independent road work coding of short video sections (or still images) in Pilot – cross centre comparison to check quality.		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Spot checks and coding committee scheduled during full scale data analysis		

A.11 Traffic density

Potential serious data quality issues can arise if coding scheme is not defined carefully and followed by coding centres. Definition of traffic density needs to be defined based on the requirements of the research question.

Congestion (or breakdown flow) has predefined definitions (Level of Service for example), likewise heavy traffic (unstable flow); these are not always based on the presence of other vehicles but on other factors such as the travel speed of the vehicle(s) compared to the posted speed limit.

In addition traffic density (other vehicles) may only be present in front of the SV, for example if the SV meets queuing traffic – it could be that the opposite is true and that other traffic is only visible behind the SV. The coding needs to be flexible enough to cope with these differences but still allow traffic density to be

accurately coded. Another condition could be that traffic density is present in an adjacent lane and the SV lane is clear (figure A-1); advanced driver training would suggest that the SV speed should be slow(er) at these points so coding 'low traffic density' but recording high speed may be misleading in terms of accident prevention.

Figure A-8-1: Differing traffic density

RQ reliance		
Low		
Test required		
Coding manual (and pictorial examples) to be produced and followed to illustrate case		
Independent traffic density coding of short video sections (or still images) in Pilot – cross centre comparison to check quality		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Spot checks and coding committee scheduled during full scale data analysis		

A.12 Traffic control

Traffic control can mean a number of very different things in different countries and at different locations; it is important for data quality that these are all understood and clearly defined.

For example there are a number of different locations that traffic controls can be applied; at junctions, on normal road sections, through restricted areas and at special scenes (such as road works or collision sites). These controls can also occur in a number of different formats; automated controls, cooperative signs (give way, width restrictions), signs giving orders (stop sign, no entry) and even signals by authorised persons (police officers, school crossing patrols)

For data quality a clear definition as to what each traffic control means and where it should be coded should be provided.

N.B. clarification as to what traffic control means should be provided as it may also be necessary to see whether unauthorised persons give traffic signals which are followed by participants such as an oncoming driver flashing their headlights or a cyclist/horse rider waving a vehicle past.

RQ reliance		
Low		
Test required		
Definition as to what constitutes a traffic signal (based on what is required for the research questions)		
Coding manual (and pictorial examples) to be produced and followed to illustrate case		
Independent traffic control coding of short video sections (or still images) in Pilot – cross centre comparison to check quality.		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Spot checks and coding committee scheduled during full scale data analysis		

A.13 Driver state

Data quality issues will always be more prevalent when coding subjective data from a video stream; depending on the coding scheme to be used it is not always possible to accurately define the driver state. For example when does a drowsy driver become sleepy? How is the crossover from awake to drowsy identified and accurately coded for all participants?

A clear and unambiguous coding manual needs to be developed that on one hand allows positive coding of the perceived driver state and on the other allows a bit of flexibility as the driver state will rarely be clearly identifiable.

For an important variable such as this with a lot of dependent research questions it is essential to be as accurate and repeatable in the coding – in these cases it may be necessary to peer review all sections of interest to ensure that the data is as reliable as possible.

RQ reliance		
High		
Test required		
Definition as to what constitutes driver state (based on what is required for the research questions)		
Coding manual (and pictorial examples) to be produced and followed to illustrate case		
Independent traffic control coding of short video sections (or still images) in Pilot – cross centre comparison to check quality.		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Spot checks and coding committee scheduled during full scale data analysis		

A.14 Driver activity

It is often easier to define what a driver is doing in a vehicle than it is to determine what they are looking at; for this reason the coding of driver activity is relatively simpler to code than gaze related variables.

The resultant data quality from this variable relies on a few simple and clear steps; the first relate to the physical instrumentation within the vehicle whereby the cameras should be positioned to cover a good view of the driver areas and have a high enough resolution and quality to pick up detail of any task the driver is conducting (bearing in mind some of these may be detailed tasks and therefore more difficult to spot and identify with poorer camera equipment).

The second part relates to the analysis; without a clear understanding of the types of task that may be expected it will be impossible to code clearly and most importantly with high quality. Drivers will do a lot of

different tasks within a vehicle, some more expected than others, and it is important to define what ones are relevant to the research questions and which ones are not, for example reading or composing a text message maybe critical to know whereas adjusting clothing or seatbelt maybe not; the quality of the data will be dependent on how these are defined.

RQ reliance		
Moderate		
Test required		
Development of pilot coding taxonomy to suit the research questions and capability of instrumentation		
Pilot coding taxonomy tested with pilot data		
Definitions reviewed and modified if needed based on results from Pilot analysis		
Independent driver activity coding of short video sections in Pilot – cross centre comparison to check quality.	Reliance	Quality
Spot checks and coding committee scheduled during full scale data analysis		

A.15 Gaze coding

In terms of data quality, gaze coding is one of the most difficult variables to code. Due to the large number of different surfaces/objects/areas that a driver may look at while driving it can be very difficult to determine detailed and accurate gaze location from video; certainly, coding specific interior surfaces such as different gauges on the instrument panel is nearly impossible.

Coding general areas of gaze is much more reliable and in turn data quality will improve. General surfaces of the vehicle interior (and areas outside the vehicle) will make the determinations much easier, for example substituting individual surfaces such as speedo, rev counter, fuel gauge, temp gauge etc. for a single code of 'instrument panel' will greatly improve data quality.

Different vehicles will present different challenges for coding and it is here where a more holistic view of the project is needed with a reduced number of vehicle makes and models increasing data quality. In analysis it takes time and experience to identify different glance locations; the additional complexity of different vehicle geometries will only have a detrimental effect on data quality.In terms of vehicles it is really only practical to code reliable gaze coding for closed vehicles such as passenger cars or trucks – motorcyclists present a special case which will be particularly dependent on the instrumentation (camera location and quality) and the type of research question to be asked (general gaze info such as 'blind spot check' as opposed to exact gaze locations).

In general an open coding taxonomy will always result in higher data quality. Coding a glance as 'Exterior object – not further specified' will provide more high quality data compared to individual codes describing a multitude of objects. Of course the effect of this will be lower data detail; it will not be possible for instance to see how often a driver looked at a street sign, however it will give much more robust and high quality data.

RQ reliance		
High		
Test required		
Development of pilot coding taxonomy to suit the research questions and capability of instrumentation		
Pilot coding taxonomy tested with pilot data		
Definitions reviewed and modified if needed based on results from Pilot analysis		
Independent driver activity coding of short video sections in Pilot – cross centre comparison to check quality.	Reliance	Quality
Spot checks and coding committee scheduled during full scale data analysis		

A.16 Long eye closure coding

On paper, an easy variable to code providing a good camera view of the drivers' eyes. Clarification needed to determine what 'long' really means and whether this is accurately timed to record duration or banded (short eye closure, long eye closure).

This variable will need to be processed with video data with a known frame rate as this can be used to determine the eye closure duration.

This variable is unlikely to be easily recorded for PTW rides depending on camera position and helmet design (an open face helmet may provide adequate view)

N.B. analysis of data from previous driving trials suggests that long eye closures are very rare. Large amounts of video data and long durations will need to be viewed to identify these events. Targeting may provide a short cut to finding these – for example looking at drivers who have driven for more than 2hrs or driving late at night or early in the morning.

RQ reliance		
Moderate		
Test required		
Determination of proposed recording scheme based on definition of a 'long eye closure'		
Pilot coding taxonomy tested with pilot data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent long eye closure coding of short video sections in Pilot – cross centre comparison to check quality.		

A.17 Gaze eccentricity

In a similar vein to gaze coding this variable provides a number of data quality issues. In some ways the coding is simplified with only the eccentricity from the forward view recorded and not the location the driver was looking at however this does not remove all possible causes of data quality issues.

Depending on camera location it can be difficult to accurately determine head rotation as it is likely the camera is off centre (i.e. not directly in front of the drivers head); this can lead to miscoding of gaze eccentricity.

The way a driver, and more pertinently, a PTW rider looks may not be represented easily by what can be seen in the video image – the difference can be seen when comparing the view from a head/helmet mounted camera with supposed intention; in this case a blind spot check.

The view from the camera in this case will never show what the driver/rider is looking at; this is made even more difficult when assessing from a camera mounted in the vehicle as a blind spot check (or over shoulder glance) will in most cases look similar to a glance out of the side window. Only the drivers intention (such as a subsequent lane change or change in road position) will give the nature of the gaze eccentricity away.

In some cases gaze eccentricity will vary as the driver/rider follows a target object from a moving vehicle (or vice versa), this can create a change in eccentricity over time as the object changes position; this adds immeasurably to data detail but can further reduce data quality as more unknown or unquantifiable head positions are coded.

Head position is also rarely in one plane; despite the road environment being, on a macro level, a two dimensional world (i.e. roads never extend sharply upwards or downwards in relation to the participants vehicle) head movement will also extend into vertical movements and more often than not in conjunction with horizontal movements

Despite this data quality can be relatively high if coding to general areas such as 'ahead' (normal scanning of the road scene ahead), 'off centre ahead' (looking through the windscreen but to the sides of the road) or 'side L/R' (a glance out of the side window on either side). Finer measures than this will always reduce data quality as it becomes increasingly difficult to code positively.

RQ reliance		
Moderate		
Test required		
Development of pilot coding taxonomy to suit the research questions and capability/fitment of instrumentation		
Pilot coding taxonomy tested with pilot data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent gaze eccentricity of short video sections in Pilot – cross centre comparison to check quality.		

A.18 Driver Reaction

Data quality issues relating to this variable predominantly arise from identifying driver reaction and coding accurately. Although in some cases driver reaction can be quite apparent, such as stiffening in the seat or through facial expressions, it will not always be possible to spot in some cases.

Using SCE data (i.e. periods of video data where an event has been already spotted) may help in identifying these reactions however they are, as with all types of human behaviour, highly variable both between participants and for each individual.

RQ reliance		
Low		
Test required		
Determination of proposed recording scheme based on definition of a 'driver reaction'		
Pilot coding taxonomy tested with pilot data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent driver reaction coding of short video sections in Pilot – cross centre comparison to check quality.		

A.19 Helmet use

Helmet use is unlikely to cause any significant data quality issues. Coding will not need to be of a particularly high resolution, perhaps a few codes per journey. Unless the helmet type or other detailed information is needed the variable should contain relatively high quality data.

RQ reliance		
Low		
Test required		
Definition of 'Helmet use' derived from analysis of Pilot data	Reliance	Quality
Independent helmet use coding of short video sections in Pilot – cross centre comparison to check quality.		

A.20 Number of passengers in vehicle

Straight forward coding in terms of data quality; the only issue relates to the positioning of the interior camera for data capture and the layout of the vehicle which may obstruct vision to smaller passengers – particularly for the rear seats.

RQ reliance		
Moderate		
Test required	Reliance	Quality
Analysis of camera positions in vehicle – check with different sized occupants in different seating positions		

A.21 Headlight activity

In essence a very simple variable with clear definition but with associated recording and coding quality issues. To record this variable with high quality throughout then an approach other than video may be more appropriate; however it is possible.

Ideally, to record high quality data using video, a dedicated camera should be used, this may need to be directed at either the light switch or the instrument panel (to record the light reminder). This solution, despite providing high quality data, will most likely be impractical.

Identifying light activation as described above but using an interior video stream will prove extremely difficult depending on camera location and field of view, in most cases it is only evident when a participant switches on/off the lights; something which will be unlikely to see.

RQ reliance		
Low		
Test required		
Analysis of camera positions in vehicle – check view of light switches and instrument lights		
Determination of coding scheme based on positive result of above step		
Re-evaluate camera quality or positions based on a negative result of step 1		
Development of coding taxonomy and test with pilot data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent headlight activity coding of short video sections in Pilot – cross centre comparison to check quality.		

A.22 Optical size of POV (principal other vehicle)

A number of data quality issues will arise from this variable depending on how it is defined; these issues will impact the data collected and the analysis conducted unless fully understood and controlled for.

It is essential to have the video camera predominantly used for this determination (normally the front view) installed comparably between all vehicles in the fleet; this is important if this variable is to be collected for all modes (passenger vehicle, PTW and truck) as any difference in zoom level, orientation or quality could affect the data collected.

Objects filmed through the windscreen invariably look further away than they do from the driver seat, this is essentially due to the angle the scene is viewed at; the higher up the more 'road' is visible between the participant vehicle (PV) and the POV. This effect is not apparent when calculated mathematically however it may have effects on the perceived optical size for coding purposes between video from a low passenger vehicle (camera position at approximately 1450mm from ground) and a truck (nearly 3000mm).

RQ reliance		
Moderate		
Test required		
Review of camera positions between vehicles – check with different vehicles in fleet		
Pilot coding based on video (or stills) from positional check		
Alterations to camera positions or coding as required	Reliance	Quality
Coding definitions reviewed and modified if needed based on results from Pilot analysis		

A.23 POV eccentricity angle

A useful addition to the above variable as a vehicle which is angled in relation to the subject vehicle will have a perceived larger optical size or be presenting a different part of their vehicle to the subject vehicles camera view.

Apparently straight forward, this variable may have a some data quality concerns with coding as not all vehicle to vehicle situations will be easy to code; how for example is the eccentricity angle of a pedestrian coded – it may be easier to determine vehicle angles.

It is also clear that as the subject vehicle is driven along then the angle of a crossing vehicle or turning vehicle will appear to change; this needs to be understood as it could be misleading to describe a POV eccentricity angle to be changing when in fact it is not; this data quality issue will arise if coding from still images as video data will show vehicle movement more clearly.

Other small data quality issues arise when coding different vehicles on the road as an articulated vehicle, for example, may have two eccentricity angles at the same time as the trailer may be straight ahead and the tractor unit may be turned.

A clear definition of the coding scheme needs to be developed to indicate how angles are recorded; for example, are positive and negative eccentricity angles to be used to describe right and left turning vehicles respectively and are angles to be coded in bands (<10°, 10°- 20° etc.) or measured somehow.

RQ reliance		
Moderate		
Test required		
Determination of proposed coding taxonomy based on requirements of research questions		
Pilot coding taxonomy tested with pilot video data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent POV eccentricity angle coding of short video sections in Pilot – cross centre comparison to check quality.		

A.24 POV type

A relatively straight forward variable to code from video data; it would seem easy to differentiate between a passenger vehicle and a truck. However, and depending on what is needed to be known for the research questions, it may be difficult to determine between truck classes (HGV, 7.5t etc.) or between car derived vans and passenger cars (Figure A2).

Figure A-8-2: HGV recognition

This variable depends very much on how clear the 'type' needs to be defined – high data quality can easily be achieved if the classification is quite course (such as Car, Van, Truck, PTW, cyclist, pedestrian) but a lot more difficult to retain data quality if the classification is finer (different truck classes, different van classes and different PTW classes etc.)

RQ reliance		
Moderate		
Test required		
Determination of proposed coding taxonomy based on requirements of research questions		
Pilot coding taxonomy tested with pilot video data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent POV type coding of short video sections in Pilot – cross centre comparison to check quality.		

A.25 Brake light onset of POV

Distinctions between brake light phases are easy to determine from video in most cases with a POV at low eccentricity angles (i.e. there is a clear view of the brake lights).

The addition of third (high level) brake lights increases ability to spot brake activation rather than other light activation. Strong sunlight on the rear of the POV can cause issues with identifying light onset but rarely causes an issue.

Higher eccentricity angles and issues with offset vehicles (such as those moving over before turning) can effect data quality however this effect should be small.

RQ reliance		
Moderate		
Test required		
Sample data reviewed to identify limits of scope	Reliance	Quality
Pilot coding taxonomy developed and tested with pilot video data		

A.26 Occlusion of objects (outside)

Similar data quality issues exist for this as well as other variables above concerned with reading data from the video channel; the most serious of which being the accurate and repeatable coding of data from the video stream.

For this variable to be coded with high quality it is essential that the video cameras installed in the vehicles are correctly positioned. This will need to be checked before trials begin and reviewed through a pilot process to see whether the video view through the windscreen (where most occlusion data will be evident) is reflective of the drivers/riders view point. From reviewing video data it is clear that some occlusions only exist from the drivers view point, a camera positioned higher can often give a much better view of the road ahead, conversely a high camera position can sometimes hide information that a driver has seen and reacted to, such as the view of a pedestrian through or under a vehicle which the camera position can mask. Occlusion can be obvious but in a lot of cases where a reaction from the driver is recorded the occluded object is much more subtle.

A secondary issue that can occur with the availability of video data is that it is possible to review a section of driving many times; this sometimes can give the impression - from a reviewers point of view - that the occlusion was much more or less severe than it appeared to the driver. It is necessary to understand that the driver has only fractions of a second to see, understand and react to an occluded object emerging whereas an analyst may have many minutes and in slow motion.

There is a need to develop a clear glossary of the types of occlusion expected. This will need to cover detail on how an object is defined as occluded and when it is partially occluded, for example does the smallest glimpse of an object emerging qualify as partially occluded or does it need to be a quarter visible or half visible; this may also be defined on time, i.e. timing an average reaction time from fully occluded before coding partially occluded.

RQ reliance		
Low		
Test required		
Camera position check to identify whether driver occlusions are reflected by the video data		
Determination of proposed coding taxonomy based on requirements of research questions and camera position check		
Pilot coding taxonomy tested with pilot video data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent occlusion coding of short video sections in Pilot – cross centre comparison to check quality.		

A.27 Presence and position of other vehicles

There needs to be provided a definition of what constitutes ‘other vehicles’. Should this be expanded to include all other vehicles in the visible scene or just the ones adjacent to the subject vehicle? This definition needs to be clear and unambiguous as any vagueness will encourage a lack of data quality; it may be necessary to identify parked vehicles and other road users too as these can both influence how the driver reacts to the scene.

Coding positions of other vehicles will be fraught with data quality issues as this will be difficult to do accurately unless the coding is very general. For example, the amount of data processing involved in recording the location (ahead, behind etc.), distance (10m, 20m etc) and any lateral movement (to the side, directly ahead, parallel) will only introduce more data quality issues. However if the coding was more generalised (ahead, beside, parked, behind, waiting to turn in/out) then the coding quality (but not data detail) will improve.

RQ reliance		
Low		
Test required		
Determination of proposed coding taxonomy based on requirements of research questions.		
Pilot coding taxonomy tested with pilot video data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent coding of short video sections in Pilot – cross centre comparison to check quality.		

A.28 Pedestrian/Cyclist head direction

This variable could present a number of very serious data quality issues as the information to be recorded from video can, and invariable is, extremely subtle. The success of this variable is highly depended on the camera quality and proximity to the pedestrian/cyclist in question.

An example of the issues that may arise could be a glance from a travelling cyclist (i.e. not one waiting to cross the road). This glance may be quite quick and quite small, combine this with the fact that the cyclist may be a long way ahead of the SV and the challenges in reading data from the video stream are quite apparent.

The following images extracted from driving data show the situation of camera quality very clearly. This particular section of road has been chosen as each painted line on the approach to the crossing is approximately 2m in length so a clear measure of distance can be inferred. In the first image (Figure A-3) recorded 16 meters from the crossing the pedestrians waiting for the crossing are clearly visible however it is difficult to determine direction of head or even, for that matter, body position. The second image (Figure A-4) shows head position clearer but still errors in recording accurate data could occur; this image is only approximately 6 meters from the first pedestrian.

This section of video is also simpler to analyse due to the fact the car is slowing. There are 50 available video frames to analyse head position. If however the vehicle was travelling at the speed limit at this location (30mph) then this data reduces to around 20 frames, or less than 1 second of data.

Figure A-8-3: Video impage 16m from pedestrian

Figure A-8-4: Video image 6m from pedestrian

The challenges become less critical when considering a pedestrian waiting to cross a road in close proximity, however it may still be difficult to extract information accurately depending on how the coding taxonomy is defined.

An example of this is the importance of defining clearly whether head direction is in relation to the SV (looking at SV, looking away from SV), the roadway (looking down side road, looking across lanes) or the pedestrian/cyclist (looking forwards, looking over shoulder) – each of these has a slightly different level of difficulty and therefore expected data quality.

An essential factor in collecting high quality data is an extremely high quality camera; only the best image quality will give sufficient detail to allow high quality coding.

RQ reliance		
Moderate		
Test required		
Development of proposed pilot coding taxonomy to suit the research questions and capability of instrumentation		
proposed pilot coding taxonomy tested with pilot video data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent coding of short video sections in Pilot – cross centre comparison to check quality.		

A.29 Pedestrian/Cyclist density

A variable that could be highly dependent on a clear coding taxonomy; it is possible to record the information needed for this variable from video footage but, in order to reach suitable data quality levels, a clear definition of coding needs to be developed through the pilot phase.

Items to define are; what constitutes the area from which density is calculated. This could be defined in a number of ways but each will be related to the output data quality. A very long field of view (as far as is visible for example) may need a much higher quality camera to capture the information and a shorter distance will need a definable distance which is repeatable for all SV types (see comment on camera height

in **Optical size of POV**). In addition coding information from a snap shot of data (i.e. one still image from a video stream) may prove misleading as pedestrian/cyclist density may have been distinctly different only a few camera frames previously.

It will also be necessary to identify what pedestrians/cyclists are included or excluded in the density calculation. For example it may not be necessary depending on the research question aims to count young children in push chairs or accompanied by adults. Similarly what is the coding convention if there are cyclists on a segregated cycle path and some in the carriageway, likewise if cyclists are waiting at a stop sign or crossing – should these also be counted in the density coding or should these be deemed as irrelevant to the traffic scene?

From viewing previous examples of driving video it is clear that it may not always be possible to determine the exact number of cyclists or pedestrians in the view; this is particularly apparent when there is bunching (i.e. a group of pedestrians waiting at a crossing) as the distinction between individuals becomes less clear. In these cases it may be necessary to define a coding taxonomy that allows for this, for example, the inclusion of a value that states '10+' or 'many'.

RQ reliance		
Moderate		
Test required		
Development of proposed pilot coding taxonomy to suit the research questions and capability of instrumentation		
proposed pilot coding taxonomy tested with pilot video data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent coding of short video sections in Pilot – cross centre comparison to check quality.		

A.30 Pedestrian/Cyclist age

This variable should, on the face of it, provide very high quality data; however in some cases it could be very difficult to determine ages for all relevant pedestrians. The coding taxonomy may need to be designed to collect data in a very course manner as this will only influence data resolution rather than quality. For example it may prove very difficult to determine ages even within 10 year bands, this becomes more problematic in winter or low light conditions where there is less available information from faces/clothing to 'guess' an age accurately. In this case it may be necessary to keep the age classification very open with codes as simple as Child, Adult or Elderly. By implementing this coding scheme it will be possible in almost all cases to positively assign a value, whereas with finer classification a level of ambiguity will always be present.

As far as instrumentation is concerned the same concerns listed in other subjective variables exist here; the camera used for the data collection should be of a high enough quality to give analyst every chance of identifying the correct age category for pedestrians.

RQ reliance		
Low		
Test required		
Development of proposed pilot coding taxonomy to suit the research questions and capability of instrumentation		
proposed pilot coding taxonomy tested with pilot video data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent coding of short video sections in Pilot – cross centre comparison to check quality.		

A.31 Pedestrian/Cyclist gender

Gender of cyclists and pedestrians is perhaps marginally easier than age but still difficult. In most situations there will be suitable information contained in the video to determine this positively however, in others data quality issues will occur unless considered.

For example it may be more problematic to identify gender of a cyclist travelling in the same direction as the SV especially if helmet use and non-gender specific cycling clothing is present. Other issues as identified in age may also be present if the pedestrian, for example, is wearing a lot of cold weather clothing including hat and scarf.

Gender specific identifiers recorded by the video are very valuable; these include but are not restricted to, clothing style, clothing colour, hair styles/lengths, walking gait etc. This method is also fallible; with a small but possibly significant number of unknowns or miss-codes present.

Analysis of video data will provide the best data quality as gender cues contained in this are much more powerful compared to one isolated video frame.

RQ reliance		
Low		
Test required		
Development of proposed pilot coding taxonomy to suit the research questions and capability of instrumentation		
proposed pilot coding taxonomy tested with pilot video data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent coding of short video sections in Pilot – cross centre comparison to check quality.		

A.32 Pedestrian/Cyclist activity

Following on from the previous variable (Pedestrian/cyclist gender) the same format of coding too achieve high quality data is also relevant here. In this case it will be necessary to either create a very comprehensive and clear coding taxonomy covering all the anticipated pedestrian/cyclist activity or have some very general codes which enable and encourage positive coding.

This first approach brings with it coding quality issues as it may not always be possible to identify the type of activity being undertaken and therefore a level of coding ambiguity will be inevitable in the final data; conversely using a very course coding scheme (such as activity: Yes/No) will greatly increase data quality for analysis; at the expense of data resolution.

RQ reliance		
Moderate		
Test required		
Development of proposed pilot coding taxonomy to suit the research questions and capability of instrumentation		
proposed pilot coding taxonomy tested with pilot video data		
Definitions reviewed and modified if needed based on results from Pilot analysis	Reliance	Quality
Independent coding of short video sections in Pilot – cross centre comparison to check quality.		

Appendix B Review report template; checklist for reviewers

B.1 Overall judgement: readibility, structure and format

		Yes	No	N/A
	Does the deliverable reflect the content described in the Description of Work?			
Comments				
	Is the deliverable sufficiently understandable: did you fully understand it (even if slightly off topic for you)?			
Comments				
	Does the deliverable include learning from mistakes/challenges encountered and does it stimulate to further research?			
Comments				
	Is the document template applied properly?			
Comments				
	Is the structure of the deliverable easy to follow? Do you suggest any changes to the structure to make the deliverable more accessible?			
Comments				
	Is the English in the deliverable good? Is it clear and accessible?			
Comments				
	Are the figures and tables understandable and refered to in the text?			
Comments				

B.2 Scientific judgement

		Yes	No	N/A
	Is the issue which is being researched clearly and simply stated?			
Comments				
	Are the objectives as described in the deliverable in line with the Description of Work (description of the Task)?			
Comments				
	Is the quality of the study design sufficient, are the methods/procedures as well as their actual application appropriate/correct?			
Comments				
	Do the results match the objectives as described in the Description of Work?			
Comments				
	How are the findings and results of the work described in the deliverable? Does the conclusion chapter reflect all described main important issues in the report and are the conclusion well based? Are the conclusions clearly stated? Are the conclusions relevant and applicable?			
Comments				
	Does the report include the relevant and necessary references? If relevant, is the			

	necessary wider view on the field of work properly given?			
Comments				
	Other comments			