# How much information is lost when sampling driving behavior data? Indicators to quantify the extent of information loss

*Jun Liu*
Department of Civil, Construction and Environmental Engineering, The University of Alabama, Tuscaloosa, Alabama, USA
*Asad Khattak and Lee Han*
Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, Tennessee, USA, and
*Quan Yuan*
State Key Laboratory of Automotive Safety and Energy, School of Vehicle and Mobility, Tsinghua University, Beijing, China

**Abstract**

**Purpose** – Individuals' driving behavior data are becoming available widely through Global Positioning System devices and on-board diagnostic systems. The incoming data can be sampled at rates ranging from one Hertz (or even lower) to hundreds of Hertz. Failing to capture substantial changes in vehicle movements over time by "undersampling" can cause loss of information and misinterpretations of the data, but "oversampling" can waste storage and processing resources. The purpose of this study is to empirically explore how micro-driving decisions to maintain speed, accelerate or decelerate, can be best captured, without substantial loss of information.

**Design/methodology/approach** – This study creates a set of indicators to quantify the magnitude of information loss (MIL). Each indicator is calculated as a percentage to index the extent of information loss (EIL) in different situations. An overall information loss index named EIL is created to combine the MIL indicators. Data from a driving simulator study collected at 20 Hertz are analyzed ($N$ = 718,481 data points from 35,924 s of driving tests). The study quantifies the relationship between information loss indicators and sampling rates.

**Findings** – The results show that marginally more information is lost as data are sampled down from 20 to 0.5 Hz, but the relationship is not linear. With four indicators of MILs, the overall EIL is 3.85 per cent for 1-Hz sampling rate driving behavior data. If sampling rates are higher than 2 Hz, all MILs are under 5 per cent for importation loss.

**Originality/value** – This study contributes by developing a framework for quantifying the relationship between sampling rates, and information loss and depending on the objective of their study, researchers can choose the appropriate sampling rate necessary to get the right amount of accuracy.

**Keywords** Driver behaviours and assistance, Sensor data processing, Information loss, Instantaneous driving decisions, Sampling rate, Undersampling

**Paper type** Research paper

## Introduction

In 2017, the National Highway Traffic Safety Administration of the United States announced its decision to move forward with the vehicle-to-vehicle (V2V) communication technology for all new light-duty vehicles (NHTSA, 2017). Newly manufactured vehicles will likely be equipped with dedicated short-range communication (DSRC) devices by regulation. As the roll-out of the V2V environment, diagnostic sensors will be installed on vehicles to collect data, and the data will be transmitted wirelessly between vehicles and nearby infrastructures. It would no longer have to rely on conventional data collection equipment, such as loop detector or video detections, and it collects much more information than the conventional ways (Liu, 2015; Liu and Khattak, 2016; Liu and

The current issue and full text archive of this journal is available on Emerald Insight at: **https://www.emerald.com/insight/2399-9802.htm**

Journal of Intelligent and Connected Vehicles
3/1 (2020) 17–29
Emerald Publishing Limited [ISSN 2399-9802]
[DOI 10.1108/JICV-10-2019-0010]



The authors would like to thank Department of Civil, Construction and Environmental Engineering at the University of Alabama, University Transportation Center for Alabama and Alabama Transportation Institute for funding support. The authors are also thankful for the funding provided by the Southeastern Transportation Center, the Region 4 University Transportation Center administered by the Research and Innovative Technology Administration as a part of the USDOT. The research was also supported through the TranLIVE University Transportation Center grant sponsored by the Research and Innovative Transportation Administration, of US DOT. The authors are grateful to anonymous reviewers who provided valuable comments that improved the manuscript.

Received 3 October 2019
Revised 14 December 2019
Accepted 22 December 2019

Khattak, 2019; Studer *et al.*, 2019). Measurements that are previously unknown are now available, which include but not be limited to vehicle speeds, positions, arrival rates, rates of acceleration and deceleration, queue lengths, stopped time and so on. With the increasing amount of data collected from DSRC-equipped vehicles, it is now made possible to explore micro-level driver behaviors. Instantaneous driving decisions are of particular interest, because they are the foundation of monitoring energy consumption, emissions and safety on a real-time basis. Driving decisions consists of a collection of maneuvers: accelerating, decelerating, maintaining speed, altering acceleration/deceleration, etc. Driving reflects a chain of instantaneous driving decisions made by drivers according to changes in surrounding circumstances, e.g. adjacent vehicles, roadway conditions and geometric changes in the roadway and weather conditions (Wang *et al.*, 2015).

Intuitively, higher rate sampled data can capture more information about the instantaneous driving decisions. Current data collection in industry can go as high as 800 MHz (Linear Technologies, 2014). However, driving data is not always necessarily sampled by such high rates in the transportation context. One problem of high sampling rates is cost, particularly under the context of the big data-driven intelligent transportation systems, in terms of requiring extra storage and processing time, which is called oversampling (Chawla, 2010). Another problem for data sampled by high sampling rates is the data accuracy. The Next Generation Simulation Program (NGSIM) collected detailed vehicle trajectory data in 10 Hz to develop behavioral algorithms in support of traffic simulation on microscopic modeling (Punzo *et al.*, 2011), as well as Safety Pilot Model Deployment (SPMD) sampling the safety messages (e.g. motion and location data) transmitted between connected vehicles and infrastructures at 10 Hz (Henclewood, 2014). The accuracy of NGSIM data is estimated at 2-4 ft (Kovvali *et al.*, 2007). For NGSIM data, in 0.1 s, the distance traveled by a 60 mph vehicle is about 8.8 ft but with a 2-4 ft error. Therefore, the accuracy of NGSIM data might be jeopardized with high sampling rates.

However, it does not mean low sampling rates are always desirable; undersampling/inadequate sampling may cause loss of critical information (Meade *et al.*, 1991). Jackson *et al.* (2005) discussed the validity of using in-vehicle GPS second-by-second (1 Hz) velocity data to track the 1-s driving operation modes, including acceleration and deceleration. Their results imply that the 1-s operation modes can be successfully measured by using GPS data sampled by 1 Hz (Jackson *et al.*, 2005), whereas the driving operation modes within 1 s are unknown. For example, if a driving command – "acceleration → deceleration → acceleration" occurs within 1 s, the 1 Hz sampled data may lose the information about the deceleration.

Current driving data are usually continuously sampled by rates from 0.2 to 10 Hz (Int Panis *et al.*, 2006; Ahn and Rakha, 2008; Campbell, 2012; Wang *et al.*, 2008; Hung *et al.*, 2007; Lyons *et al.*, 1986; Boriboonsomsin *et al.*, 2010; Simpson and Markel, 2012; TSDC Secure Transportation Data Project, 2014). Note that the continuous driving data are different from the traffic data collected by loop detectors (Bikowitz and Ross, 1985; Oh *et al.*, 2002). The focus of this study is the continuous driving data used to explore micro-driving behavior. The key question to be answered is what sampling rates are appropriate to capture micro-driving behavior without losing much information (i.e. by undersampling).

In the field of signal processing, Nyquist–Shannon sampling theorem gives the appropriate sampling rates for continuous signal. The Nyquist criterion for sampling rates is twice the bandwidth of a bandlimited signal or a bandlimited channel. The key question is to find out the bandwidth of a signal (Landau, 1967). However, the driving behavior does not fulfill the features of bandlimited signal. Driving behavior varies according to the decisions a driver makes to respond the instantaneous driving circumstances. This study aims to find out the appropriate sampling rates for driving behavior data through exploring the nature of driver's micro-driving behavior.

## Data description

Data used in this study comes from the University of Tennessee Driving Simulator Lab (DSL). This driving simulator, Drive Safety DS-600c, is fully integrated and immersive to driving test subjects with its visual and audio effects in the front half cab of a Ford Focus sedan and it provides 300° horizontal field-of-view via five projectors and back sight via three rear mirror liquid crystal displays (Yang *et al.*, 2013). The cab base is able to mimic pitch and 30 longitudinal motions. Since 2009, over 10 simulator studies have been conducted in DSL. The equipment has been recognized as a high-fidelity driving simulator and is qualified to be used to conduct driving behaviors-associated research. The data of driver responses (e.g. speed) gathered from simulator driving tests can be used as surrogate measures of driving behavior (Bédard *et al.*, 2010; Wang *et al.*, 2010). The driving data used in this study was collected from 24 subjects (13 males, 11 females, average licensed year – 17.6, standard deviation – 7.87). Note that, the scope of this study is to introduce the indicators to quantify the extent of information loss (EIL) when sampling driving behavior data. The influences of driving conditions on driving behavior are not examined in this study. Subjects were tested in a simulated driving scenario designed with various driving conditions, covering most possible driving conditions as a whole, including urban and rural environments, as well as freeways and local streets. Each subject completed the driving test in 22-29 min, depending on their travel speed and responses to traffic controls. The driving speed was sampled at 20 Hz. The final dataset used in this study includes 718,481 data points from 35,924 s (598 min) of driving tests.

## Methodology

A fundamental question is "how much information is lost in going to lower sampling rates?" Driving can be volatile as drivers made driving decisions (e.g. accelerating and braking) according to the instantaneous changes of surrounding circumstances, e.g. adjacent vehicles, roadway conditions, geometric changes in the roadway and weather conditions (Wang *et al.*, 2015). Using the 20-Hz simulator driving data, this study creates a set of indicators to quantify the magnitude of information loss (MIL):

- $MIL_1$: instantaneous driving decision loss (based on combined direct and indirect "detectability" explained below) – equations (1)-(3);

- $MIL_2$: percentage of out-of-range observations during driving – equation (4);
- $MIL_3$: ratio of sampled to actual range in driving data – equation (5); and
- $MIL_4$: relative speed deviation from linear interpolation of undersampled data (based on observed speed deviation over the undersampled data) – equations (6) and (7).

An index, called Extent of Information Loss (EIL), is created for a sampling rate, as shown in equation (8). The overall methodological framework for this study is shown in Figure 1 and explained in more detail below. There are two groups of indicators: micro-driving decision indicators and magnitude-related indicators. The micro-driving decision indicators are used to capture the missing of micro-driving decisions when sampling data, and the magnitude-related indicators are to quantify the magnitude errors between the sampled values and ground truth values.

Each indicator is calculated as a percentage to index the EIL in different situations. The EIL is an overall indicator of information loss that combines the above indicators. The study quantifies the relationship between information loss indicators and sampling rates. A user can then select thresholds, e.g. 5 or 1 per cent of information loss may be acceptable, and find the appropriate sampling rate.

### Direct detectability of driving decisions

Driving decisions can be altered at any time and frequently when a vehicle is being operated. If the frequency of the driving decision alteration is considerably high and the data sampling rate is very low, then some driving decisions may be lost. As shown in Figure 2(a), the decision alteration – "acceleration to deceleration" between $n$ and $n+1$ s is missed by the 1-Hz sampled data (red points), as the speeds at $n$ and $n+1$ s are identical. In this case, undersampling causes information loss of micro-driving decisions. The information about going from "acceleration to deceleration" between $n$ and $n+1$ s is lost, whereas the information on "deceleration" or "no decision alternation" between $n+1$ and $n+2$ s is detected directly by the sampled data.

This study uses the 20-Hz simulator driving data to count the number of decisions made given a specific time interval, and then computes the possibility of no decision made cases,

**Figure 1** Study steps and indicators

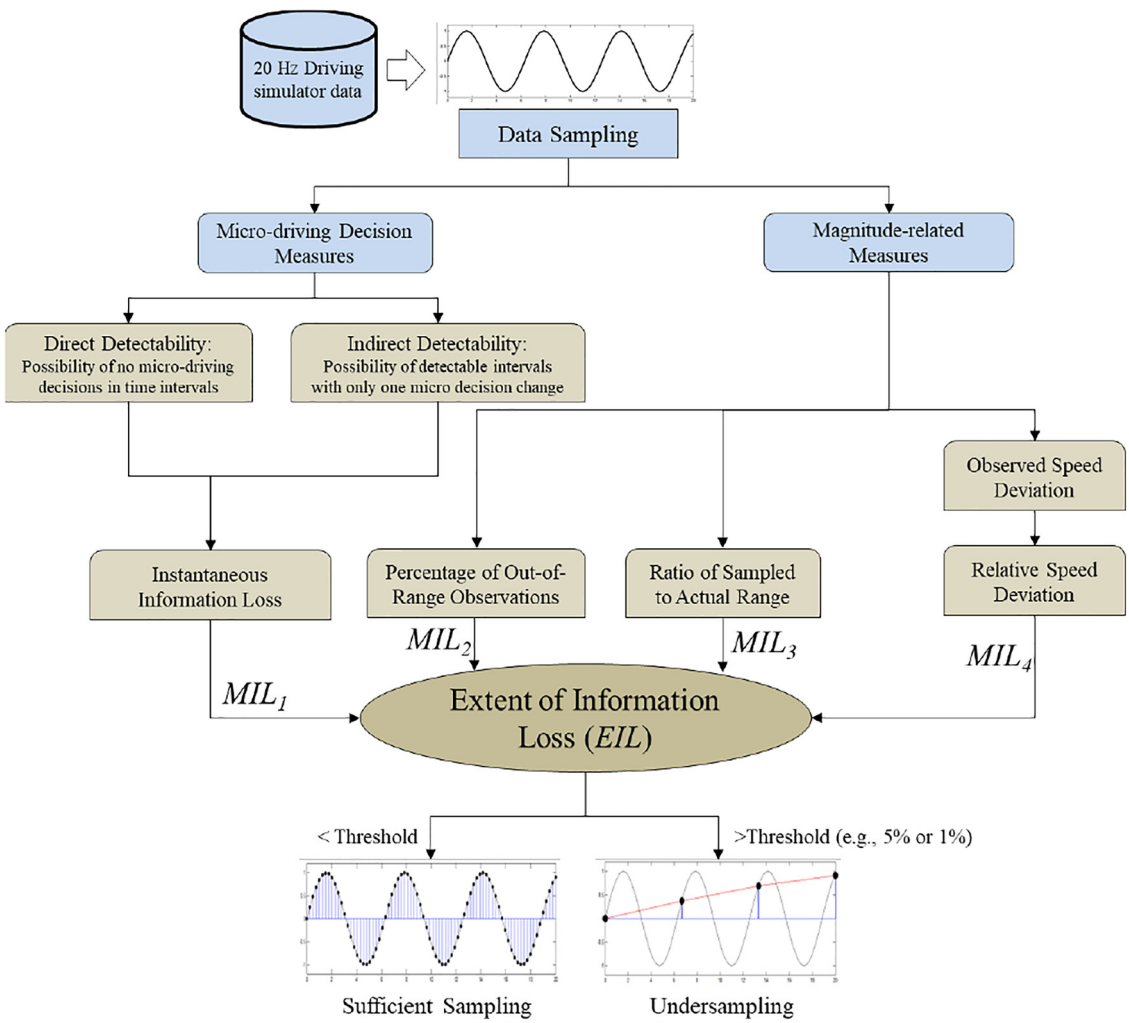
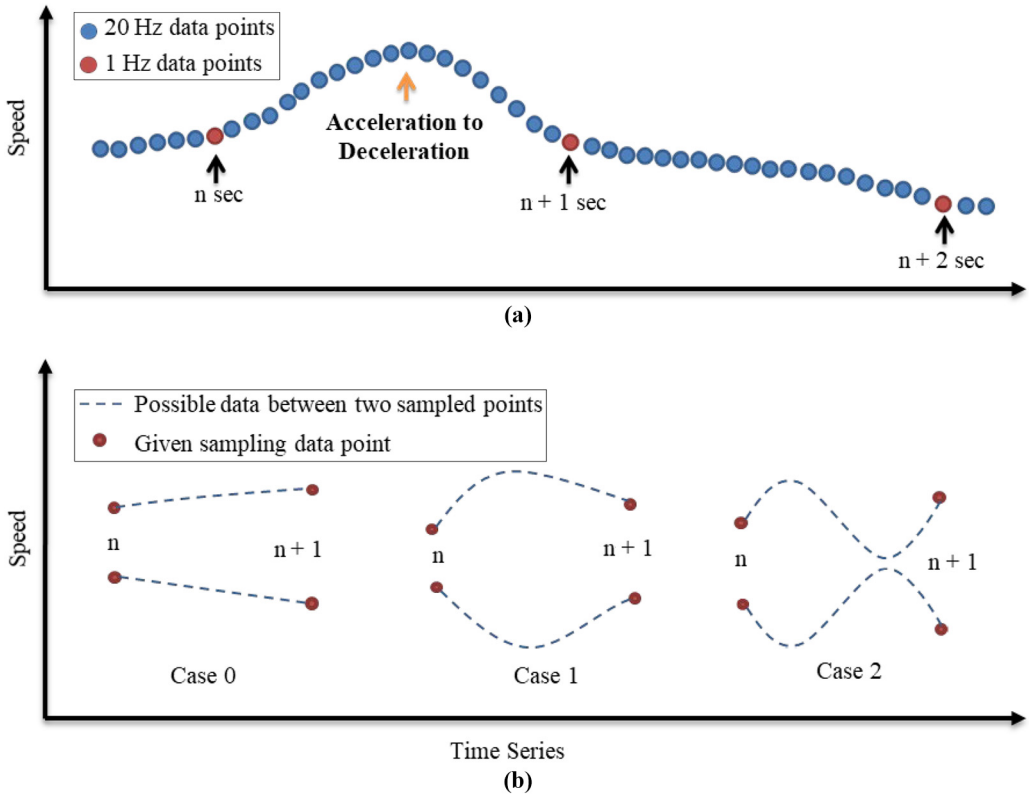


**Figure 2** Example of information loss in instantaneous driving decisions




termed *direct detectability of driving decisions*. The formula is as follows:

$$Direct\ Detectability = \frac{1}{N}\sum_{i=1}^{N} w_i^0 \quad (1)$$

where:

$N = T \times f$, the number of time slices during total data duration $T$ in second;

$f$ = target sampling frequency/rates, e.g. 1 Hz;

$N = T \times f$, the number of time slices during total data duration $T$ in second;

$f$ = target sampling frequency/rates, e.g. 1 Hz;

$w_i^0 = \begin{cases} 1, & if \max\{v_{ij} - v_{i(j-1)}\} \times \min\{v_{ij} - v_{i(j-1)}\} \geq 0, \\ 0, & if \max\{v_{ij} - v_{i(j-1)}\} \times \min\{v_{ij} - v_{i(j-1)}\} < 0, \end{cases}$, indicator for micro-driving decision alternation during $i^{th}$ time interval; $t = \frac{1}{f}; i = 1, 2, 3, \ldots, N$;

$v_{ij}$ = speed at $j^{th}$ location in $i^{th}$ time interval, j = 1, 2, 3, . . ., $n$;

$n = \frac{T}{N} = \frac{F}{f}$, number of available data points in a given time interval; and

$F$ = sampling rate of original dataset, 20 Hz in this study.

In this study, time intervals without decisions made belong to Case 0 (this includes constant acceleration or deceleration), as shown in Figure 2(b), with one micro-decision made are referred to as Case 1 and with two decision alternations are referred to as Case 2. Case 1 will be further discussed below.

**Indirect detectability of driving decisions**

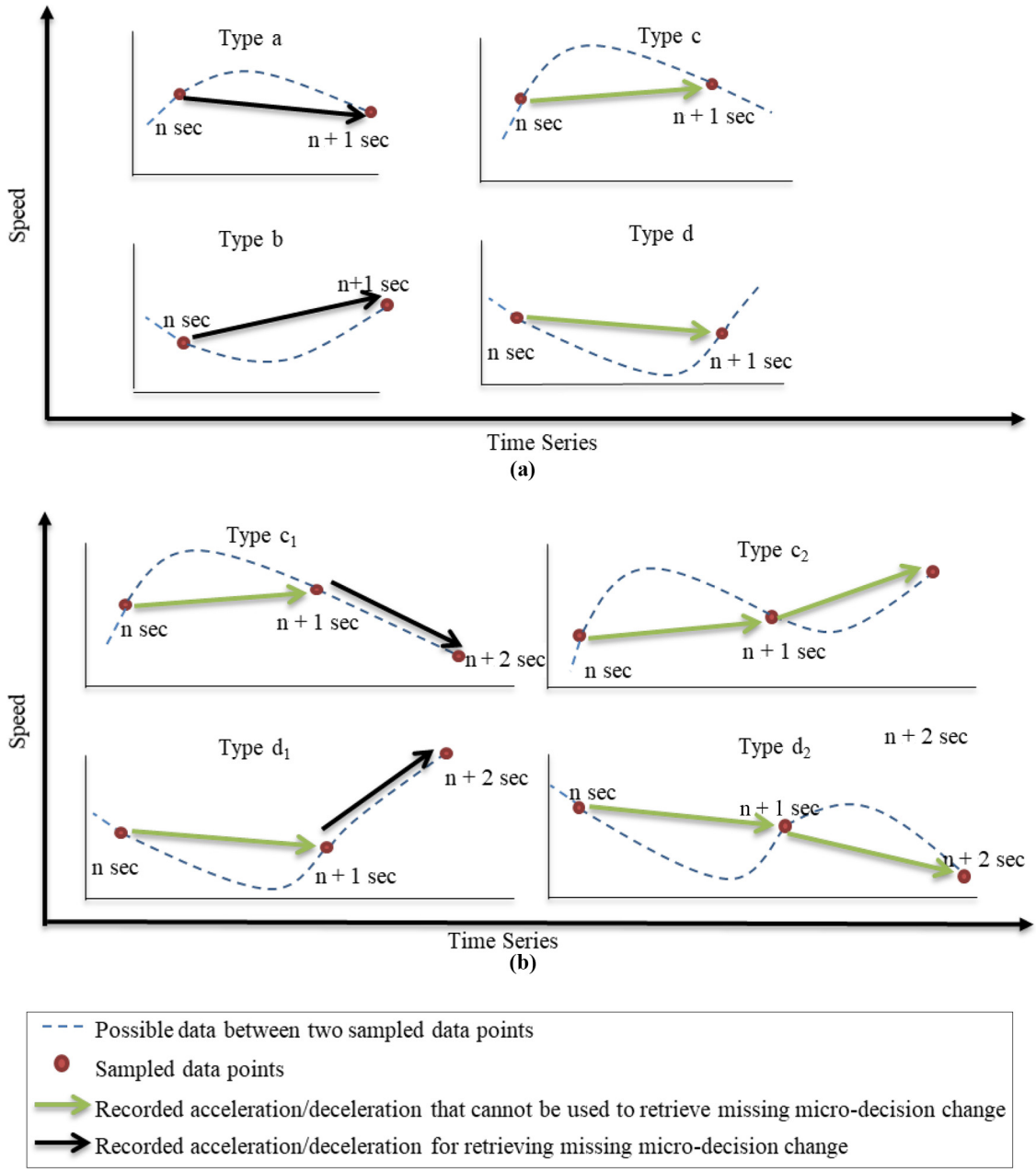
Direct detectability tells the chance of detecting micro-driving decisions directly with the sampled data. Next, this study discusses the chance of detecting driving decisions in Case 1. It is believed that driving speed can only continuously change without sharp changes. A sine wave illustrates the example of continuous changes, whereas square wave and sawtooth wave are examples of sharp changes (Elmore and Heald, 2012).

This study takes 1-s interval (corresponding to 1-Hz sampling rate) as the example for illustrating detection of driving decision alternation. Figure 3(a) presents six possible types of micro-driving behavior of Case 1 within 1 s. Types (a) and (c) show that there is a micro-decision made from accelerating to decelerating between $n$ and $n + 1$ s. Types (b) and (d) show that there is a micro-decision made from decelerating to accelerating between $n$ and $n + 1$ s.

For Type (a), there is a micro-decision made from accelerating to decelerating between $n$ and $n + 1$ s, whereas the speed measurement at $n$ and $n + 1$ s implies a deceleration during that second. Therefore, the missing micro-decision made within this second could be observed by using given sampling data points at $n$ and $n + 1$ s, though the amount/intensity of the driving decision change is not necessarily accurate. In the same fashion, Type (b) illustrates information detection for the micro-decision made from decelerating to accelerating. Therefore, for Types (a) and (b), the micro-decision change can be detected but with an error.

Types (c) and (d) do not meet the situations in Types (a) and (b), because the sampled data do not show the correct

**Figure 3** Examples of missing information when examining speed data over time




**Note:** Types (c) and (d) include the cases that speed at $n$ second is equal to speed at $n + 1$ second

micro-decision made between two sampled observations. Types (c) and (d) also include the cases that speed at $n$ second is equal to $n + 1$ s, as shown in Figure 2(a), because in these cases, the sampled observations cannot tell the micro-decision correctly.

Therefore, we move our sight to the next second, as shown in Figure 3(b). In Type ($c_1$), the sampled speeds at $n + 1$ and $n + 2$ s give a deceleration which uncovers the lost micro-decision made between $n$ and $n + 1$ s, but with a temporal error. The time stamped for the micro-decision using sampled data is at $n + 1$ s, but actually, it occurred between $n$ and $n + 1$ s. Type ($d_1$) is similar to Type ($c_1$), but for detecting a micro-decision from decelerating to accelerating.

Types ($c_2$) and ($d_2$) illustrate two types of micro-decisions which cannot be easily detected, because there are two distinct micro-decisions (acceleration and acceleration) made in two sequential sampling intervals. Besides, for cases with two or more micro-decisions made within one particular time interval, there is no way to detect them by the above methods. This study mainly discusses Case 1 with one micro-decision made and tries to find the possibilities of having Types (a), (b), ($c_1$) and ($d_1$) given a time interval. The indicator, *indirect*

*detectability of driving decisions*, is the sum of the possibilities of having Types (a), (b), ($c_1$) and ($d_1$).

The formula is as follows:

$$Indirect\ Detectability = \frac{1}{N}\sum_{i=1}^{N}\left(w_i^a + w_i^b + w_i^{c_1} + w_i^{d_1}\right) \quad (2)$$

where

$N = T \times f$ is the number of time slices during the total data duration $T$ in second;

$f$ = target sampling frequency/rates, e.g. 1 Hz;

$w_i^1 = \begin{cases} 1, \text{ if } \sum_{j=1}^{n-1} z_j = 1, \\ 0, \text{ if } \sum_{j=1}^{n-1} z_j \neq 1 \end{cases}$ indicator for whether there is only one decision change during $i^{th}$ time interval $t = \frac{1}{f}, i = 1, 2, 3, \ldots, N$;

$z_j = \begin{cases} 1, \text{ if } (v_{ij} - v_{i(j-1)}) \times (v_{i(j+1)} - v_{ij}) < 0 \\ 0, \text{ if } (v_{ij} - v_{i(j-1)}) \times (v_{i(j+1)} - v_{ij}) \geq 0 \end{cases}$, indicator for whether two consecutive micro-decisions are the same (either acceleration or deceleration);

$v_{ij}$ = speed at $j^{th}$ location in $i^{th}$ time interval, $j = 1, 2, 3, [\ldots], n$;

$n = \frac{T}{N} = \frac{F}{f}$, the number of available data points in a given time interval;

$F$ = sampling rate of the original dataset, 20 Hz in this study.

$$w_i^a = \begin{cases} 1, \text{ if } (v_{ij} - v_{i(j-1)}) > 0 \text{ and } (v_{i(j+n)} - v_{i(j+n-1)}) \\ \quad < 0 \text{ and } v_{ij} > v_{i(j+n)}, \text{ indicator Type for } (a) \text{ error;} \\ 0 \end{cases}$$

$$w_i^b = \begin{cases} 1, \text{ if } (v_{ij} - v_{i(j-1)}) < 0 \text{ and } (v_{i(j+n)} - v_{i(j+n-1)}) \\ \quad > 0 \text{ and } v_{ij} < v_{i(j+n)}, \text{indicator for Type } (b) \text{ error;} \\ 0 \end{cases}$$

$$w_i^c = \begin{cases} 1, \text{ if } (v_{ij} - v_{i(j-1)}) > 0 \text{ and } (v_{i(j+n)} - v_{i(j+n-1)}) \\ \quad < 0 \text{ and } v_{ij} < v_{i(j+n)}, \text{indicator for Type } (c_1) \text{ error;} \\ 0 \end{cases}$$

$$w_i^d = \begin{cases} 1, \text{ if } (v_{ij} - v_{i(j-1)}) < 0 \text{ and } (v_{i(j+n)} - v_{i(j+n-1)}) \\ \quad > 0 \text{ and } v_{ij} > v_{i(j+n)}, \text{indicator for Type } (d_1) \text{ error;} \\ 0 \end{cases}$$

## Instantaneous driving decision loss

With the direct and indirect detectability of driving decisions, we can detect micro-driving decision made given a particular sampling rate. The formula for instantaneous driving decision loss ($MIL_1$) is as follows:

$$Decision\ Loss = 1 - (Direct\ Detectability + Indirect\ Dectetability) \quad (3)$$

Empirical results are shown later. Theoretically, higher sampling rates lower the possibility of missing critical decisions, but they increase the possibility of "noise" in the data and the data storage and processing requirements. The challenge is to not lose decision information while reducing the noise in the data.

## Indicators concerning magnitudes

It is important to know whether sampled values represent the population and the magnitude of errors, if any. In other words, whether the one point (e.g. 1 Hz data) can represent the 20 data points (20 Hz data) during the same second? If the 20 data points provide only marginally more information (such as constant speed during 1 s), one data point might be sufficient for sampling this second.

Figure 4(i) shows an example using 20 Hz simulator data, along with two 1-Hz sampled points at the $n$ and $n + 1$ s. The speed is 10 mph at $n$ second and 12 mph at $n + 1$ s. The question would be whether all speed values between $n$ and $n + 1$ s are within the micro-speed range 10-12 mph. The example shows that given a 1-s time interval, there are six data points, or 30 per cent (6 out of 20) data points with speed values out of range 10-12 mph. In this case, two data points with records of 10 and 12 mph cannot fairly represent the driving behavior from $n$ to $n + 1$ s. The *percentage of out-of-range observation* ($MIL_2$) is an indicator that captures how many data points are out of the sampled micro-speed range. Theoretically, the value of $MIL_2$ can be from zero to extremely close to 100 per cent.

The formula for *percentage of out-of-range observation* ($MIL_2$) is as follows:

$$Percentage\ of\ Out\ Range\ Observations = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{n} OR_{ij}}{n} \quad (4)$$

where

$OR_{ij} = \begin{cases} 1, \text{ if } v_{ij} > \max\{v_{i1}, v_{in}\} \text{ or } v_{ij} < \min\{v_{i1}, v_{in}\} \\ 0 \end{cases}$, indicator for out-of-range observation.

The ratio of sampled micro-speed range over actual micro-speed range during the same second is another indicator of information loss and it is termed *ratio of sampled to actual range* ($MIL_3$). In the example, the sampled micro-speed range is 12 − 10 = 2 mph, whereas the actual micro-speed range is 12.3 − 9.6 = 2.7 mph. The ratio is 2/2.7 = 0.74, or 74 per cent. The formula is as follows:

$$Ratio\ of\ Sampled\ to\ Actual\ Range = \frac{1}{N}\sum_{i=1}^{N}\frac{R_i^{Sampled}}{R_i^{Actual}} \quad (5)$$

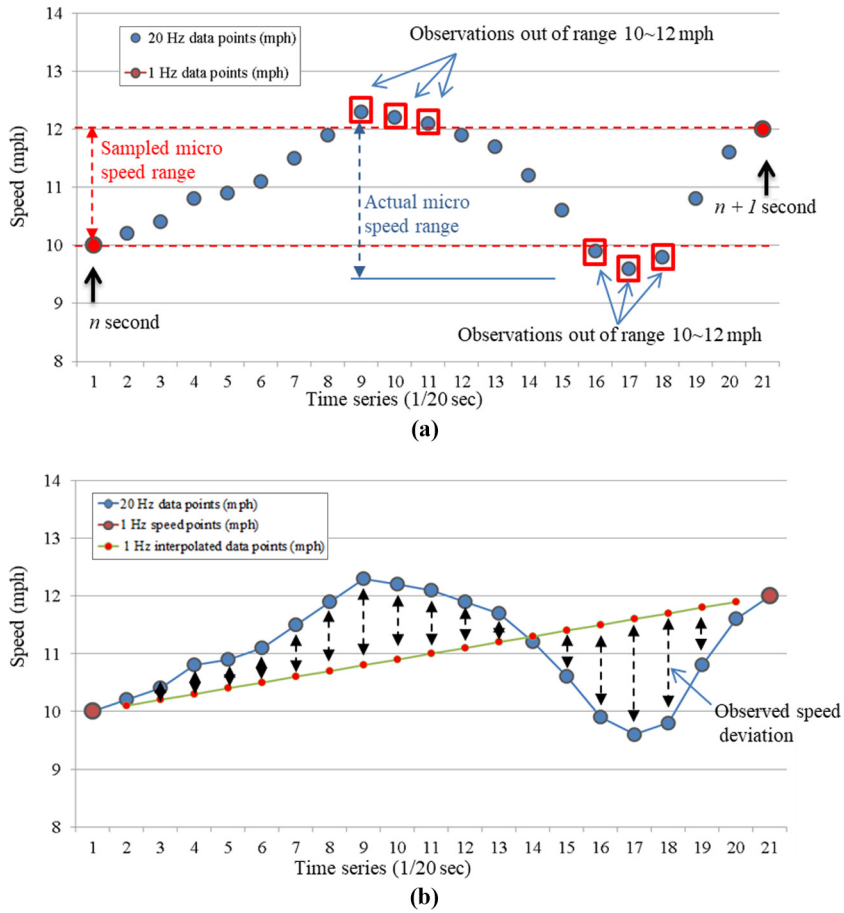
where

$R_i^{Sampled} = |v_{i1} - v_{(i+1)1}|$, sampled speed range for $i^{th}$ time slice;

$R_i^{Actual} = \max\{v_{ij}\} - \min\{v_{ij}\}$, actual speed range for $i^{th}$ time slice.

An indicator of information loss is through speed deviations. The deviations are measured based on the linear distance between observed speeds and sampled speeds. Sampled data can be used to linearly interpolate the data points in between two time stamps. This can be compared with observed data at a higher frequency (20 Hz in this case). Figure 4(b) uses 20-Hz driving simulator data and measures *observed speed deviation*, which is the mean of

**Figure 4** Quantifying magnitude errors in sampled data




absolute deviations within time intervals. Another indicator is *relative speed deviation* ($MIL_4$), which is the average deviations over interpolated speed values, providing the extent of deviations. The formulas are as follows:

$$Observed\ Speed\ Deviation = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n}\sum_{j=1}^{n}\left|v_{ij} - j \times \frac{v_{i1} - v_{i(n+1)}}{n}\right|\right) \quad (6)$$

$$Relative\ Speed\ Deviation = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\left|v_{ij} - j \times \frac{v_{i1} - v_{i(n+1)}}{n}\right|}{v_{ij}}\right) \quad (7)$$

### Index for magnitude of information loss

The *instantaneous driving decision loss*, *percentage of out-of-range observation*, *ratio of sampled to actual range* and *relative speed deviation* quantify the MIL from different angles. All these indicators are finally calculated in terms of percentage of information loss. Then, these indicators can be combined (weighted equally) to create an index capturing the *EIL index*, given a sampling rate. The formula is as follows:

$$Extent\ of\ Information\ Loss\ Index = \frac{MIL_1 + MIL_2 + (1 - MIL_3) + MIL_4}{4} \quad (8)$$

where

$MIL_1$ = instantaneous driving decision loss;
$MIL_2$ = percentage of out-of-range observations;
$MIL_3$ = ratio of sampled to actual range; and
$MIL_4$ = relative speed deviation.

Users of data in the transportation context can either choose a threshold for information loss and find the appropriate sampling rate or vice versa.

## Results

### Direct detectability of driving decisions

To capture alternations between acceleration and deceleration within the given time interval (e.g. 1 s) corresponding to a sampling rate (e.g. 1 Hz), the number of alternations was counted by using 20 Hz data. All possible alternations within the data, given different time intervals and starting locations, were counted. If all decisions made occur exactly at the sampled points, no information will be lost. For example, in Figure 2, if the data was just sampled at $n + 0.5$ s and $n + 1.5$ s

instead of $n$ and $n+1$ s, then the driving decisions from accelerating to decelerating can be detected accurately, even if the data are still sampled at 1 Hz. The example in Figure 2 shows that there are 20 possible locations to start sampling the 1 Hz data.

Figure 5(a) presents the direct detectability and possibility of no decision made (Case 0), given a specific time interval, and Figure 5(b) presents the distribution of the possibilities of the three cases (discussed above) in different time intervals. For short time intervals, the location does not have a significant influence on the data sampling. Specifically, for time interval of 1 s (1 Hz sampling rate), the direct detectability is around 89.9 per cent, i.e. Case 0 or no micro-decision made during 1-s intervals. The reason is probably related to the driver reaction time, which is usually more than 1 s (AASHTO, 2011).

In Figure 5(b), the percentages of possibilities of the three cases (i.e. no decision, one decision and two and more decisions made within the sample interval) are provided. Shorter time intervals (higher sampling rates) are related to the lower information loss in terms of instantaneous driving decisions, as expected. For time interval of 1 s (1 Hz sampling rate), Case 1 accounts for 9.2 per cent and Case 2 accounts for 0.9 per cent of sampling intervals (1 s).

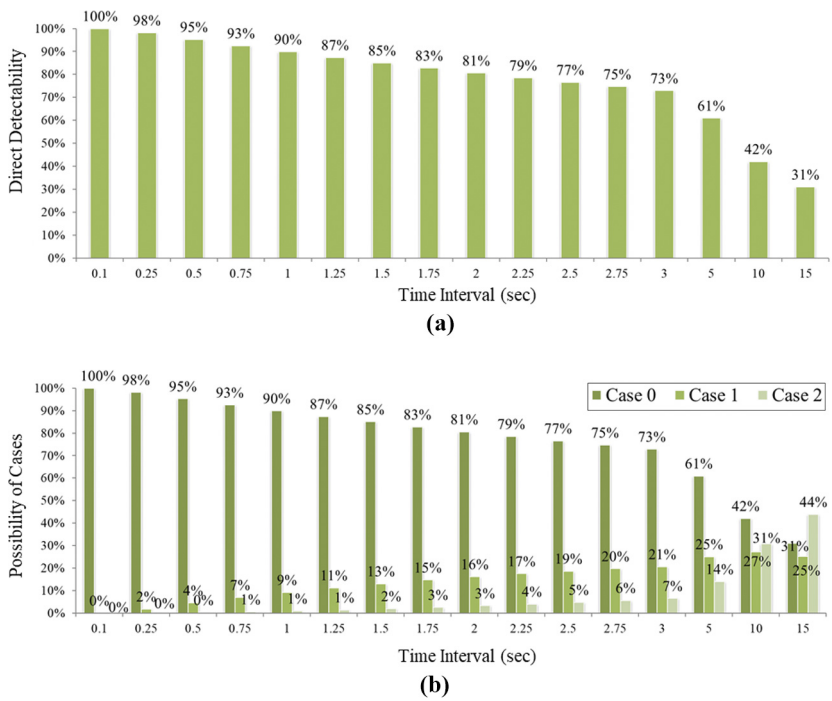
### Indirect detectability of driving decisions

Figure 6(a) shows percentages of Types (a), (b), ($c_1$) and ($d_1$) in Case 1 (one decision change). Specifically, given a 1-s time interval (or 1-Hz sampling rate), Types (a), (b), ($c_1$) and ($d_1$) constitute 31, 25.37, 21.42 and 16.14 per cent of Case 1, where only one micro-decision is made between two sampled data points. These four types of patterns contain detectable driving information. The indirect detectability is the sum of these possibilities, shown in Figure 6(b). For 1-s time interval (or 1-Hz sampling rate), the indirect detectability is around 31 per cent + 25.37 per cent + 21.42 per cent + 16.14 per cent = 93.92 per cent. With the time interval getting longer, this indirect detectability decreases.

### Instantaneous driving decision information loss

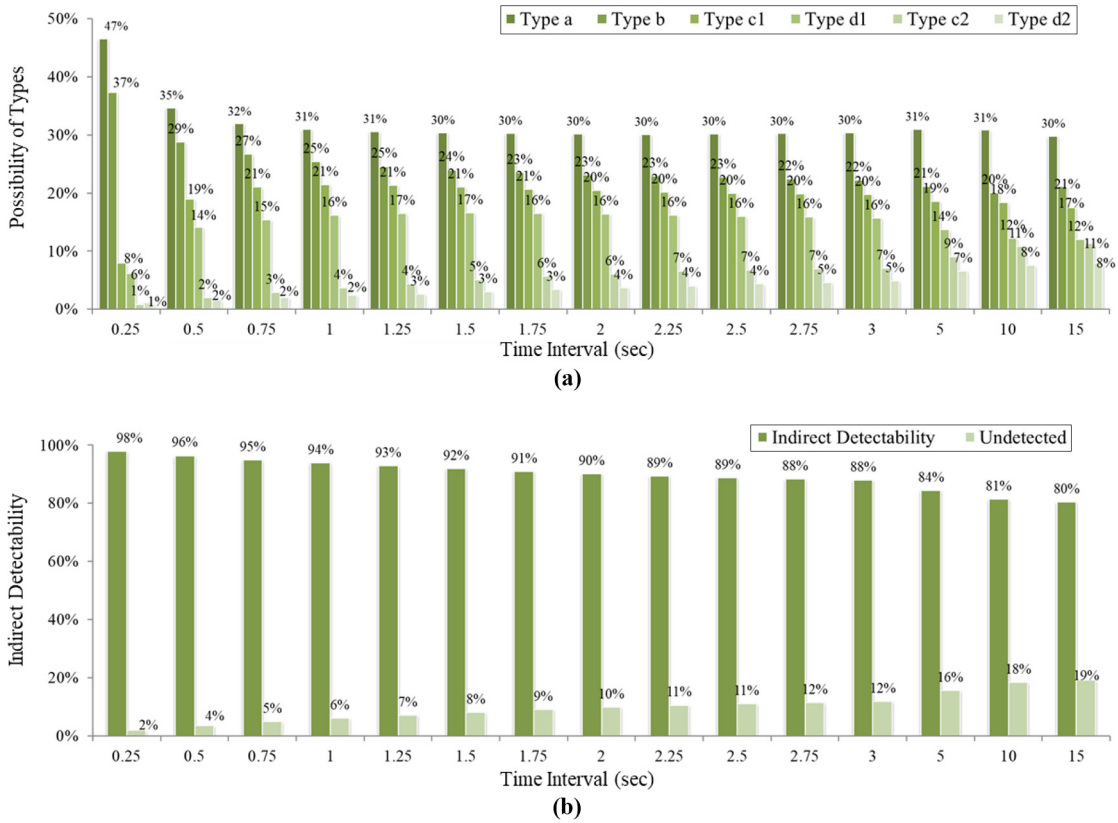
The combined results of instantaneous driving decision loss are shown in Table I. There is an 89.90 per cent chance that there is no micro-decision (Case 0) within 1 s (1-Hz sampling data, highlighted in Table I) and 9.20 per cent chance that there is one micro-decision (Case 1). For Case 1 with only one micro-decision, there is a 30.99 per cent chance that the Type (a) decision pattern would occur, and 25.37, 24.42 and 16.14 per cent for Types (b), (c) and (d), respectively. These four types include micro-decisions that can be detected. Therefore, in summary, the feasibility of detecting micro-driving decisions for 1 Hz sampling data are 89.90 per cent + 9.20 per cent × (30.99 per cent + 25.37 per cent + 24.42 per cent + 16.14 per cent) = 98.54 per cent, and 1.46 per cent of information about micro-decisions would be lost. Data sampled by rates higher than 0.5 Hz can reflect more than 95 per cent of micro-decisions and the instantaneous driving decision loss is less than 5 per cent.

**Figure 5** "Direct detectability" for data sampled between different time intervals




**Notes:** Case 0 = No decisions; Case 1 = One decision; Case 2 = Two and more decisions; Decision = Driving from "accelerating to decelerating" or "decelerating to accelerating"

**Figure 6** Indirect detectability in different time intervals




### Indicators concerning magnitudes

Results in Table II show that lower sampling rates (or longer time intervals) are associated with larger percentages of out-of-range points, smaller ratio of sampled-to-actual range, larger speed deviations and relative speed deviations, as expected. Percentage of out-of-range points concerns the sampled micro-speed range within a time interval. The sampled micro-speed range is determined by two sequential recorded data points, as shown in Figure 4. The results show that, on average, 1.75 points (or 8.75 per cent) are out of the sampled micro-speed range for 1-s time interval (or 1-Hz data), because there is a large possibility that there is no micro-decision changes during 1 s. It is consistent with the above finding that for the time interval of 1 s, the average possibility of no micro-decision change is 88.90 per cent (see Figure 5). For 1-Hz data, the ratio of sampled to actual micro range is 0.957, which means the extent of representativeness of the 1-Hz data to 20-Hz data is about 95.7 per cent in terms of magnitude. Though some data points are possibly out of the recorded micro ranges, these points do not deviate broadly. Further, 1-Hz data have an observed speed deviation of about 0.076 mph. Note that 1 per cent percentile of 718,481 20-Hz speed records is 0.493 mph, and thus the deviation of 0.076 mph is not substantial in the distribution of speed records. This is consistent with EPA drive cycle data, which is based on 10-Hz (EPA, 2013). Further, the relative speed deviation, ratio of deviation over interpolated speeds, shows that 1-Hz data has a relative speed deviation to 20-Hz speed records at 0.87 per cent, substantially lower than the 5 per cent threshold.

### Extent of information loss

The overall EIL is an equally weighted indicator, calculated using equation (8). The results are shown in Table II. We know if the sampling rate is 1 Hz, the percentage of out-of-range points is 8.77 per cent, ratio of sampled to actual range is 95.71 per cent, relative speed deviation is about 0.87 per cent and the instantaneous driving decision loss is about 1.46 per cent. So, the overall EIL is (8.77 per cent + (100 per cent − 95.71 per cent) + 0.87 per cent + 1.46 per cent)/4 = 3.85 per cent. Thus, overall, about 3.85 per cent of the driving information, including the micro-driving decisions and speed magnitude, might be lost if the sampling rate is 1 Hz instead of 20 Hz. If 5 per cent of information loss is the threshold, a sampling rate higher than 0.8 Hz can be acceptable, if EIL is considered. If all MILs need to be under 5 per cent for importation loss, then the 2 Hz sampling rate might be the lowest sampling rate to meet the information loss threshold.

Figure 7 presents the final results quantifying various information loss indicators and different sampling rates. The results show that different indicators have different levels of information loss at a given sampling rate and the relationship is nonlinear. At sampling rates higher than 2 Hz, all MILs are under 5 per cent for importation loss. The indicator of $MIL_2$, percentage of out-of-range observations, seems to be with higher values than other MILs across sampling rates. This

**Table I** Instantaneous driving decisions information loss

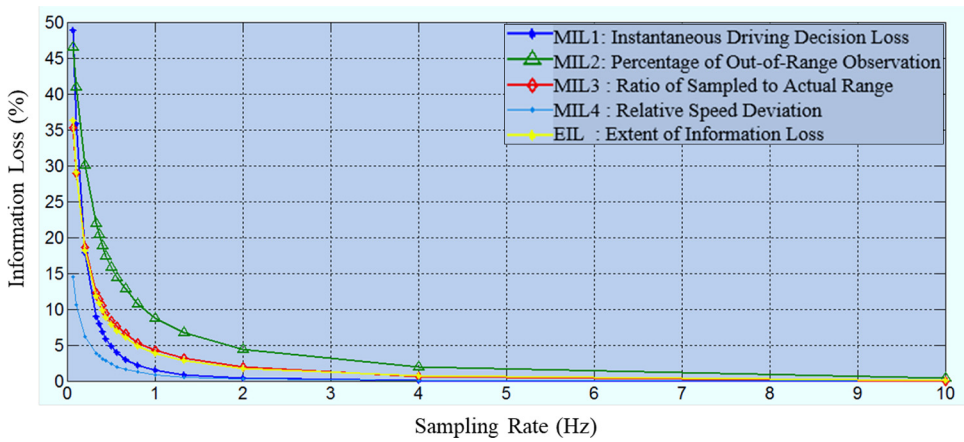
| Sampling rate (Hz) | Time interval (s) | Percentage of total sample | | | Percentage of Case 1 | | | | | Feasibility of detecting micro-decisions (%) | Instantaneous driving decision lost (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Case 0 (%) | Case 1 (%) | Case 2 (%) | Type a (%) | Type b (%) | Type $c_1$ (%) | Type $d_1$ (%) | Type $c_2$ (%) | Type $d_2$ (%) | |
| **10** | 0.1 | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100.00 | 0.00 ^ |
| **4** | 0.25 | 98.16 | 1.78 | 0.05 | 46.53 | 37.28 | 7.98 | 6.16 | 0.88 | 1.17 | 99.91 | 0.09 |
| **2** | 0.5 | 95.27 | 4.49 | 0.24 | 34.60 | 28.79 | 18.96 | 14.06 | 1.99 | 1.60 | 99.60 | 0.40 |
| **1.333** | 0.75 | 92.53 | 6.95 | 0.52 | 31.91 | 26.65 | 21.04 | 15.40 | 2.94 | 2.06 | 99.13 | 0.87 |
| **1** | 1 | 89.90 | 9.20 | 0.90 | 30.99 | 25.37 | 21.42 | 16.14 | 3.68 | 2.40 | 98.54 | 1.46 |
| **0.8** | 1.25 | 87.40 | 11.22 | 1.38 | 30.55 | 24.55 | 21.29 | 16.52 | 4.44 | 2.65 | 97.83 | 2.17 |
| **0.667** | 1.5 | 85.03 | 13.03 | 1.94 | 30.36 | 23.96 | 21.00 | 16.58 | 5.11 | 2.99 | 97.01 | 2.99 |
| **0.571** | 1.75 | 82.77 | 14.68 | 2.55 | 30.28 | 23.48 | 20.64 | 16.50 | 5.69 | 3.41 | 96.11 | 3.89 |
| **0.5** | 2 | 80.61 | 16.16 | 3.24 | 30.16 | 23.16 | 20.42 | 16.40 | 6.12 | 3.74 | 95.17 | 4.83 |
| **0.444** | 2.25 | 78.54 | 17.47 | 3.99 | 30.09 | 22.95 | 20.14 | 16.20 | 6.57 | 4.05 | 94.16 | 5.84 |
| **0.4** | 2.5 | 76.58 | 18.63 | 4.79 | 30.14 | 22.69 | 19.98 | 16.02 | 6.81 | 4.36 | 93.13 | 6.87 |
| **0.364** | 2.75 | 74.70 | 19.68 | 5.63 | 30.22 | 22.42 | 19.89 | 15.90 | 6.96 | 4.62 | 92.10 | 7.90 |
| **0.333** | 3 | 72.90 | 20.59 | 6.50 | 30.35 | 22.20 | 19.76 | 15.71 | 7.10 | 4.88 | 91.03 | 8.97 |
| **0.2** | 5 | 60.97 | 25.07 | 13.96 | 30.98 | 21.15 | 18.60 | 13.68 | 9.02 | 6.57 | 82.13 | 17.87 |
| **0.1** | 10 | 42.04 | 27.13 | 30.83 | 30.82 | 20.06 | 18.36 | 12.20 | 10.88 | 7.58 | 64.14 | 35.86 |
| **0.0667** | 15 | 30.98 | 25.15 | 43.88 | 29.79 | 21.14 | 17.47 | 12.01 | 11.30 | 7.96 | 51.20 | 48.80 |

**Note:** ^Extremely close to 0%

**Table II** Overall magnitude of information loss

| Sampling rate (Hz) | Time interval (s) | Count of out-of-range observations | $MIL_2$ – percentage of out-of-range observations (%) | $MIL_3$ – ratio of sampled to actual range (%) | Observed speed deviation (mph) (%) | $MIL_4$ – relative speed deviation (%) | $MIL_1$ – instantaneous driving decision loss (from Table I) (%) | EIL (%) |
|---|---|---|---|---|---|---|---|---|
| **10** | 0.1 | 0.008 | 0.42 | 100.00 | 0.001 | 0.01 | 0.00 | 0.11 |
| **4** | 0.25 | 0.100 | 2.00 | 99.37 | 0.005 | 0.05 | 0.09 | 0.69 |
| **2** | 0.5 | 0.442 | 4.42 | 98.11 | 0.020 | 0.23 | 0.40 | 1.73 |
| **1.3333333** | 0.75 | 1.010 | 6.73 | 96.87 | 0.045 | 0.52 | 0.87 | 2.81 |
| **1** | 1 | 1.754 | 8.77 | 95.71 | 0.076 | 0.87 | 1.46 | 3.85 |
| **0.8** | 1.25 | 2.677 | 10.71 | 94.68 | 0.115 | 1.24 | 2.17 | 4.86 |
| **0.6666667** | 1.5 | 3.847 | 12.82 | 93.38 | 0.160 | 1.66 | 2.99 | 6.02 |
| **0.5714286** | 1.75 | 5.050 | 14.43 | 92.40 | 0.208 | 2.00 | 3.89 | 6.98 |
| **0.5** | 2 | 6.345 | 15.86 | 91.66 | 0.258 | 2.35 | 4.83 | 7.85 |
| **0.4444444** | 2.25 | 7.848 | 17.44 | 90.65 | 0.316 | 2.78 | 5.84 | 8.85 |
| **0.4** | 2.5 | 9.441 | 18.88 | 89.53 | 0.371 | 3.11 | 6.87 | 9.83 |
| **0.3636364** | 2.75 | 11.216 | 20.39 | 88.63 | 0.426 | 3.45 | 7.90 | 10.78 |
| **0.3333333** | 3 | 13.172 | 21.95 | 87.70 | 0.491 | 3.88 | 8.97 | 11.78 |
| **0.2** | 5 | 30.058 | 30.06 | 81.42 | 0.974 | 6.15 | 17.87 | 18.17 |
| **0.1** | 10 | 81.855 | 40.93 | 71.10 | 2.088 | 10.57 | 35.86 | 29.07 |
| **0.0666667** | 15 | 139.545 | 46.51 | 64.73 | 3.131 | 14.52 | 48.80 | 36.28 |

**Figure 7** Extent of information loss with different sampling rates




indicator may be critical for some purposes, e.g. crash reconstruction and reporting. Therefore, for studies dealing with crashes, especially crash reconstruction studies that are highly sensitive to speed magnitude, higher sampling rates can be beneficial. The curves, including the overall information loss indicator, show that information loss becomes rather high between at 1- and 2-Hz level.

## Limitations

The data used in this study comes from a simulator driving test, i.e. they are from a hypothetical but controlled test environment. Having few test subjects is recognized as a limitation, though it is not very germane to this study. The data was sampled by 20 Hz. It is possible that micro-driving decisions between the 20-Hz time-stamp data points were lost. This study assumes the chance of having micro-decision changes within 0.05 s is very small, given a perception reaction time of about 1 s. In the future, driving data sampled at even higher sampling rates can be used to verify the results of this study. The proposed indicators can be used for analysis of information loss with any range of sampling frequency.

The scope of this study is to develop the concept of MILs or EILs that can be used to quantify the EIL when sampling driving behavior data. This study introduced a limited number of indicators, and more indicators can be developed to quantify the information loss. In addition, the results of quantified information loss may vary significantly across different traffic conditions, e.g. urban and rural environments. The road configurations would also have a significant impact on driving behavior. Therefore, the recommended sampling rates for collecting driving behavior data may need to be specified for particular driving conditions of interest.

## Conclusions

The key question investigated in this study is: what sampling rates are appropriate to capture micro-or short-term driving decisions? Oversampling can result in noisy data, and waste storage and processing resources. Undersampling can result in loss of information about important instantaneous driving decisions. This study developed indicators of information loss and quantified their relationship with sampling rates. It discussed driving behavior information from two angles: instantaneous driving decisions and speed magnitudes. Four main indicators were created to quantify the magnitudes of driving behavior information loss:

- $MIL_1$ – instantaneous driving decision loss (combined direct and indirect "detectability");
- $MIL_2$ – percentage of out-of-range observations;
- $MIL_3$ – ratio of sampled-to-actual range; and
- $MIL_4$ – relative speed deviation from linear interpolation of sampled data (based on observed speed deviation over interpolated speed).

These indicators quantify the EIL. With these four indicators, the overall MIL index was generated by equally weighting them. The index, termed by EIL, simply tells us how much information might be lost, given a sampling rate.

The results show that shorter time intervals (i.e. higher sampling rates) are associated with larger direct detectability of instantaneous driving decisions. In other words, there is a smaller chance of having cases with micro-driving decisions between two sampled data points. Drivers typically keep constant acceleration/deceleration rates during a short time. Specifically, for the time interval 1 s (i.e. 1-Hz sampling rate), the direct detectability is 88.90 per cent. The large possibility of no micro-decision in 1 s may be because of the driver reaction time. The reaction time includes the time for driver perception, identification, judgment and reaction (TRB, 1998). The whole process usually takes more than 1 s (AASHTO, 2011). This study further observed cases of one micro-driving decision made within a particular time interval and discussed the possibility of detecting such micro-driving decisions. Through defining the six possible micro-driving decision patterns, the study found the four of six patterns include the micro-driving decisions that can be detected indirectly by using the sampled data points. These four patterns dominate the cases in short time intervals (less than 3 s). Specifically, the indirect detectability for 1-s time interval (or 1-Hz sampling rate) is around 93.92 per cent. The feasibility of detecting micro-driving decisions combines direct detectability and indirect detectability. Thus, the feasibility of detecting micro-driving decisions by 1-Hz data are 89.90 per cent + 9.20 per cent × 93.92 per cent = 98.54 per cent, and 100 per cent − 98.54 per cent = 1.46 per cent of information about micro-decisions ($MIL_1$) will be lost by 1-Hz data.

The indicators of information loss magnitude reveal that smaller sampling rates or longer time intervals are related to more missing data points because of their too large or too small values. Though there are some data points out of the micro-speed ranges (about 8.77 per cent of points out of the micro-ranges for 1-Hz data, $MIL_2$), these points do not deviate broadly when sampling rates are equal to or higher than 1 Hz. Specifically, the ratio of sampled to actual ranges ($MIL_3$) is 95.7 per cent for 1-Hz data. And 1-Hz data has an average speed deviation of about 0.076 mph. The small deviation supports the assumption that driving behavior within 1 s shows nearly constant acceleration (EPA, 2013). Further, the relative speed deviation ($MIL_4$) of 1-Hz data to 20 Hz is around 0.87 per cent. With four indicators of MILs, the overall EIL can be calculated. For 1-Hz sampling rate, the EIL is about 3.85 per cent.

This study proposed indicators to quantify the MIL regarding the longitudinal driving behavior. The indicators can be used individually or combined to create an index. The calculation results are not intended to be directly used by all other driving behavior studies, as the results may vary significantly across different traffic conditions and driver behaviors. The trends of MILs and EIL may be useful to researchers to understand how information might be lost because of the low sampling rates. The calculation process can be easily replicated by other researchers who aim to determine an appropriate sampling rates for their study data collection, or to evaluate the extent of information loss for driving behavior data that have been collected at a known sampling rate. The results show that lower sampling rates are associated with greater information loss, but the relationship is nonlinear. This study contributes by developing a methodology to quantify the relationship between sampling rates and information loss. Depending on the objective of their study, researchers can choose the appropriate sampling rate necessary to get the right amount of accuracy. For some studies, e.g. quantifying energy consumption or emissions, 2-Hz sampling rate may be sufficient, whereas for safety studies, higher sampling rates may be required. In addition, different indicators may capture different aspects of the information loss while sampling data to study driving behavior. The indicators introduced in this study are for longitudinal driving behavior. Indicators for lateral behavior such as steering angle need to be developed in future research.

## References

AASHTO (2011), *A Policy on Geometric Design of Highways and Streets*, 6th ed., American Association of State Highway and Transportation Officials, Washington, DC.

Ahn, K. and Rakha, H. (2008), "The effects of route choice decisions on vehicle energy consumption and emissions", *Transportation Research Part D: Transport and Environment*, Vol. 13 No. 3, pp. 151-167.

Bédard, M., Parkkari, M., Weaver, B., Riendeau, J. and Dahlquist, M. (2010), "Assessment of driving performance using a simulator protocol: validity and reproducibility", *American Journal of Occupational Therapy*, Vol. 64, pp. 336-340.

Bikowitz, E.W. and Ross, S.P. (1985), Evaluation and improvement of inductive loop traffic detectors.

Boriboonsomsin, K. Vu, A. and Barth, M. (2010), Eco-driving: pilot evaluation of driving behavior changes among US drivers.

Campbell, K.L. (2012), *The SHRP 2 Naturalistic Driving Study: Addressing Driver Performance and Behavior in Traffic Safety*, *TR News*.

Chawla, N.V. (2010), *Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook*, Springer.

Elmore, W.C. and Heald, M.A. (2012), *Physics of Waves*, Courier Dover Publications.

EPA (2013), "Dynamometer drive schedules", United States Environmental Protection Agency, available at: www.epa.gov/nvfel/testing/dynamometer.htm (accessed 3 March 2014).

Henclewood, D. (2014), *Safety Pilot Model Deployment – One Day Sample Data Environment Data Handbook*, Research and Technology Innovation Administration, US Department of Transportation, McLean, VA.

Hung, W., Tong, H., Lee, C., Ha, K. and Pao, L. (2007), "Development of a practical driving cycle construction methodology: a case study in Hong Kong", *Transportation Research Part D: Transport and Environment*, Vol. 12 No. 2, pp. 115-128.

Int Panis, L., Broekx, S. and Liu, R. (2006), "Modelling instantaneous traffic emission and the influence of traffic speed limits", *Science of the Total Environment*, Vol. 371 Nos 1/3, pp. 270-285.

Jackson, E., Aultman-Hall, L., Holmén, B.A. and Du, J. (2005), "Evaluating the ability of global positioning system receivers to measure a real-world operating mode for emissions research", *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1941 No. 1, pp. 43-50.

Kovvali, V.G., Alexiadis, V. and Zhang, P.E. (2007), "Video-based vehicle trajectory data collection (No. 07-0528)" available at: http://sites.poli.usp.br/ptr/lemt/documents/07-0528.pdf

Landau, H. (1967), "Sampling, data transmission, and the nyquist rate", *Proceedings of the Ieee*, Vol. 55 No. 10, pp. 1701-1706.

Linear Technologies (2014), *LTC6412 – 800MHz, 31dB Range Analog-Controlled VGA*, available at: www.analog.com/media/en/technical-documentation/data-sheets/6412fa.pdf

Liu, J. and Khattak, A. (2019), "Informed decision-making by integrating historical on-road driving performance data in high-resolution maps for connected and automated vehicles", *Journal of Intelligent Transportation Systems*, pp. 1-13, doi: 10.1080/15472450.2019.1699076.

Liu, J. and Khattak, A.J. (2016), "Delivering improved alerts, warnings, and control assistance using basic safety messages transmitted between connected vehicles", *Transportation Research Part C: emerging Technologies*, Vol. 68, pp. 83-100.

Liu, J. (2015), "Driving volatility in instantaneous driving behaviors: Studies using Large-Scale trajectory data", PhD dissertation, University of Tennessee, available at: https://trace.tennessee.edu/utk_graddiss/3308.

Lyons, T., Kenworthy, J., Austin, P. and Newman, P. (1986), "The development of a driving cycle for fuel consumption and emissions evaluation", *Transportation Research Part A: General*, Vol. 20 No. 6, pp. 447-462.

Meade, M.L., Dillon, C.R. and Dillon, C.R. (1991), *Signals and Systems*, Springer.

NHTSA (2017), *V2V Statement, National Highway Traffic Safety Administration*", available at: www.nhtsa.gov/press-releases/v2v-statement

Oh, S., Ritchie, S.G. and Oh, C. (2002), "Real-time traffic measurement from single loop inductive signatures", *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1804 No. 1, pp. 98-106.

Punzo, V., Borzacchiello, M.T. and Ciuffo, B. (2011), "On the assessment of vehicle trajectory data accuracy and application to the next generation SIMulation (NGSIM) program data", *Transportation Research Part C: Emerging Technologies*, Vol. 19 No. 6, pp. 1243-1262.

Simpson, M. and Markel, T. (2012), "Plug-in electric vehicle fast charge station operational analysis with integrated renewables", *EVS26 (Electric Vehicle Symposium)*.

Studer, L., Agriesti, S., Gandini, P., Marchionni, G. and Ponti, M. (2019), "Impact assessment of cooperative and automated vehicles", in Lu, M. (Ed.), *Cooperative Intelligent Transport Systems: Towards High-Level Automated Driving*, IET (Institution of Engineering and Technology), London. ISBN: 978-183953-012-8 (Print)/978-183953-013-5.

Transportation Research Board (1998), *Managing Speed: Review of Current Practice for Setting and Enforcing Speed Limits*, *Transportation Research Board*, National Research Council, Washington, DC.

TSDC Secure Transportation Data Project (2014), "Transportation secure data center, national renewable energy laboratory", available at: www.nrel.gov/transportation/secure-transportation-data/tsdc-about.html

Wang, H., Fu, L., Zhou, Y. and Li, H. (2008), "Modelling of the fuel consumption for passenger cars regarding driving characteristics", *Transportation Research Part D: Transport and Environment*, Vol. 13 No. 7, pp. 479-482.

Wang, X., Khattak, A., Liu, J., Masghati-Amoli, G. and Son, S. (2015), "What is the level of volatility in instantaneous driving decisions?", *Transportation Research Part C: Emerging Technologies*,.

Wang, Y., Mehler, B., Reimer, B., Lammers, V., D'Ambrosio, L.A. and Coughlin, J.F. (2010), "The validity of driving simulation for assessing differences between in-vehicle informational interfaces: a comparison with field testing", *Ergonomics*, Vol. 53 No. 3, pp. 404-420.

Yang, Q., Overton, R., Han, L.D., Yan, X. and Richards, S.H. (2013), "Driver behaviours on rural highways with and without curbs – a driving simulator based study", *International Journal of Injury Control and Safety Promotion*, pp. 1-12.

## Corresponding author

**Jun Liu** can be contacted at: jliu@eng.ua.edu