Yuri A. Vershinin
Fedor Pashchenko
Cristina Olaverri-Monreal *Editors*

# Technologies for Smart Cities

Springer

# Technologies for Smart Cities

Yuri A. Vershinin · Fedor Pashchenko ·
Cristina Olaverri-Monreal
Editors

# Technologies for Smart Cities

Springer

*Editors*
Yuri A. Vershinin
Faculty of Engineering, Environment
and Computing
Coventry University
Coventry, Warwickshire, UK

Fedor Pashchenko
Russian Academy of Sciences
Moscow, Russia

Cristina Olaverri-Monreal
Johannes Kepler University of Linz
Linz, Austria

ISBN 978-3-031-05515-7 ISBN 978-3-031-05516-4 (eBook)
https://doi.org/10.1007/978-3-031-05516-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Contents

# Safety and Security in Global Navigation Satellite Systems

**Yuri A. Vershinin and Georgios Antoniou**

**Abstract** The Global Positioning System (GPS) is currently the most widely spread and most used Global Navigation Satellite System (GNSS). Numerous industries and people rely daily on the system's ability to determine their position as well as to synchronize time with the atomic clock carried onboard the satellites. It is now more important than ever for GNSS to operate securely and reliably, to ensure the safety of its users. However, in this chapter, it is highlighted that this might not always be the case, as attackers constantly find new ways to exploit the GNSS features and pose an imminent threat to the system's reliability. GNSS receivers are especially vulnerable to three types of malicious attacks, in particular blocking, jamming and spoofing. In this chapter, a thorough research was conducted for the operation of both GNSS receivers and spoofing devices. A literature review based on the current available studies and research for detection and mitigation measures, was made. Then a new spoofing detection method is proposed and the principles and the research that was based on are analysed. Experimental data and results were gathered along with data extracted from simulated spoofed signals. The results from both the experiment and the simulation are reviewed, compared and discussed. Finally, based on those results conclusions are drawn.

## Nomenclature

| | |
|---|---|
| CRC | Cyclic Redundancy Check |
| GEDT | Goodness of fit empirical distribution equality test |
| GNSS | Global navigation satellite system |
| GOF | Goodness of fit |
| GPS | Global positioning system |
| MVDET | Mean vector distribution equality test |
| NMA | Navigation message authentication |

Y. A. Vershinin (✉) · G. Antoniou
Coventry University, Coventry, UK
e-mail: o292j@gmx.com

Y. A. Vershinin et al. (eds.), *Technologies for Smart Cities*,
https://doi.org/10.1007/978-3-031-05516-4_1

| | |
|---|---|
| NME | Navigation message encryption |
| NMEA | National marine electronics association |
| PRN | Pseudo-random noise |
| PVT | Position, Velocity and Time |
| SCA | Spreading code authentication |
| SCE | Spreading code encryption |
| SDR | Software defined radio |
| UAV | Unmanned arial vehicle |

## 1 Introduction

Global Navigation Satellite Systems (GNSS) make up the fabric of today's society as multitude industries completely rely on the unparalleled reliability and accuracy for positioning, navigation and time synchronization which GNSS offers. GNSS constantly find their way in new applications as the possibilities for implementation are seemingly endless. However, this rapid mass adaptation and increase in demand has left the security aspect of GNSS struggling to catch up, leaving users exposed to external threats and malicious attacks [18].

GNSS can provide highly accurate position, velocity and time (PVT) with the use of electromagnetic signals broadcasted from a consolation of satellites. The GNSS receiver calculates the PVT by applying the principle of trilateration. By measuring the time taken for the signal to arrive to the receiver, the in-between distance, which is known as the Pseudo range, can be calculated for every satellite that's connected to the receiver. Then the known distances from satellites to the receiver are taken as the radii of imaginary spheres with the satellite positions as the center of each sphere. The point where the spheres coincide on the earth's surface is taken as the true coordinates.

The GPS spatial segment operates as a satellite constellation consisting of six orbitual planes circling the earth. Each plane consists of four satellites as a baseline. Typically, the United States Defence Department keeps in orbit more than 24 satellites to ensure an unyielding coverage from any position on earth which allows for the user to have a connection with at least four satellites at any time. The GPS satellites transmit signals containing a navigation message and a Pseudorandom noise code, 50 times per second on three different carrier frequencies; the L1, the L2 and the L5 at 1575.42 MHz, 1227.60 MHz and 1176.45 MHz respectively [12]. The L1 and L5 frequency bands are currently used for civilian use while the L2 is reserved for military use. The L5 band is the newest civilian frequency and provides some updated features over the L1 signals, such as a higher transmission power of carrier signal, which means that the signal cannot be interrupted as easily and can be detected by the receiver faster compared to L1 signals [15].

The GNSS signal can often be blocked by tall buildings or tick metal structures. So, in order for the receiver to keep the connection with the satellites even when the direct signal is blocked, it connects with signals reflected from the ground or other surfaces. These signals as referred to as refracted or multipath signals. Electromagnetic waves can reflect from surfaces and find their way to the receiver's antenna; however, this reduces the accuracy as multipath signals take longer to reach the receiver and contain an increased amount of noise compared to direct signals. This phenomenon is known as the multipath error.

The medium that the signal travels through, can as well compromise the accuracy, as the earth's atmosphere refracts the GNSS signals, causing what is known as the ionospheric effect and the tropospheric delay, which cause errors in the calculation of Pseudo range and as a result the range appears greater than in reality [8]. These effects are amplified when the satellite is positioned near the horizon relative to the receiver, as the signal has to travel for longer through the medium, whereas the effects are reduced when the satellite is positioned right above the receiver at its zenith.

Beside the physical challenges that the GNSS signals must endure, there are also artificial and malicious attacks to which the GNSS users are exposed to. Most common types of GNSS attacks are the jamming attacks during which the attacker transmits a signal filled with white noise at the same frequency as the GNSS signals but with a higher signal strength which interrupts the connection between the GNSS receiver and the satellite. As the signal is weakened by the time it reaches the earth's surface, it is quite easy for attackers to jam the signal [18].

Another common and perhaps far more threatening type of attack is spoofing which not only blocks the authentic GNSS satellite signals but also replaces them with fake GNSS signals which contain false navigation data and are almost indistinguishable from the authentic signals [18]. A software defined radio (SDR) is the device of choice for spoofers as the device can both receive and transmit signals at a wide range of frequencies defined by the user. Typically, in spoofing attacks SDR's are programmed to first jam a GNSS signal and then transmit counterfeit GNSS signals at a higher power than the authentic ones in order to mislead the receiver into tracking the fake GNSS signals which could contain false positioning, navigation and timing information [18].

Spoofing attacks can be separated in three main categories, namely the asynchronous attack, the synchronous attack and synchronous attack with multiple transmitters. An asynchronous attack can be done cheaply with any GNSS signal imitator or an RF transmitter programmed to recreate a GNSS signal. A Synchronous is more expensive and requires more sophisticated equipment, making it harder to detect. Synchronous attack with multiple transmitters is the most expensive from all, as it uses a number of different transmitters at different locations. This technique makes the attack very hard to detect as its more sophisticated than any other.

The aim of this research is to study further into the current safety and security measures of GNSS against malicious attacks which present a major threat to GNSS users. A review recent published literature is conducted to gain a deeper understanding of the techniques that are currently used both spoofing and anti-spoofing which will lead to the development of a critical thinking around the topic and new ideas for overcoming the issues. The ideas are proposed later in the report and with the use of u-center and MATLAB, measurements are gathered from experiments and are analysed to detect the presence of a spoofer.

## 2 Literature Review

Various methods for dealing with jamming and spoofing attacks have been developed so far. Some research, such as [1], focus on some characteristic features of malicious attacks and search for those in the received signals to determine whether a GNSS signal can be considered safe or dangerous. Based on pairwise correlation, [1] have developed a method for categorising signals into two different groups based on how closely correlated the signals are to each other. As the spoofer can generate multiple signals from the same source, the spoofed signals are highly correlated in terms of their spatial signature. The first group contains the signals with high spatial correlation, labelled as “spoofed signals”, and are separated from the second group with low correlation, which is labelled as “authentic signals”. However, it is mentioned in the paper that there is the possibility of accidentally placing spoofed signals in the authentic group, if the authentic GNSS signal is blocked or ignored by the receiver, which can occur if the signal is too weak. Although, the spoof signal will soon be pulled out of the authentic group as it will be correlated with other signals and relocated to the spoofing group. However, attackers can easily get around this technique, by using a sophisticated spoofer which can first jam the authentic GNSS signals with a low power jammer which will be just enough to interrupt the weak GNSS signals and then transmit spoofed signals at high power level so that they do not get blocked by the jammer. Furthermore, the receiver can be deceived by a set of spoofers at different locations, each transmitting a single GNSS signal, so that the signals would not be correlated, although such an attack would be unlikely as it requires multiple expensive equipment located at different positions and operated by multiple people.

In another recent study, [7] took a more statistical approach as they investigated two new methods for early detection of GNSS signal jamming of unmanned aerial vehicles (UAVs). They have used limiting distribution of random matrices obtained from samples of the received signals, to define a test statistic, which rejects the signals which are considered to be jamming. Their first method presented is the Mean vector distribution equality test (MVDET), during which radio signal samples from a jamming-free environment are collected and used to create a computational threshold of probability distribution which can be used as a baseline to reject signals which exceed that threshold and are considered to be jamming. The second method is called the Goodness of fit empirical distribution equality test (GEDT) and it obtains samples of jamming-free signals and use them as reference. Those samples always follow chi-square distribution, so the signals are compared to the reference, by using the goodness of fit test (GOF). If the GOF test shows a significant variation, then the signal is rejected, otherwise it is accepted. These methods for jamming detection have the advantage that they do not require heavy computational complexity in comparison to other methods of jamming detection. Although, these methods to be effective require to obtain baseline measurements of satellite signals in a jamming-free environment, which might not always be possible.

Qi et al. [5] propose in their paper another method for spoofing detection based on the Doppler effect and the relative frequency of the signal's carrier. Based on the amount of Doppler shift, the algorithm can determine if the signals originate from a single source, which is a suggestive sign of a spoofing attack, as typically one spoofing source contains several signals and therefore the signals will have the same amount of Doppler shift. On the contrary, authentic GNSS signals originate from multiple different satellites at different locations, so each of the signals will appear with a different amount of Doppler shift, as each satellite is moving in a different direction relative to the receiver. The Doppler shift change is observed from the change in frequency of the Pseudo-random noise (PRN) for each signal. In authentic signals, when the receiver is moving, the change in frequency and consequently the Doppler shift will be different for each signal. However, this method still has a small probability of false detection and incorrectly identifying an authentic satellite signal as a spoofed signal, depending on the position of the satellites relative to the receiver.

A different approach to the subject is explored in the book "GNSS Interference Threats and Countermeasures" by [2], which delves into the concept of signal authentication through cryptographic encryption. Four of the most popular cryptographic techniques which can potentially be implemented in the GNSS technology, are reviewed and analysed in the chapter "Authentication Techniques". The four techniques are: Navigation Message Authentication (NMA), a method for satellite signal authentication with the use of digital signature; Spreading Code Authentication (SCA), a mechanism that introduces additional encrypted code segments into the PRN code and uses a secret authentication key; Navigation Message Encryption (NME), uses one secret keys to both encrypt and decrypt the message contained in the satellite signal using a symmetric encryption algorithm; Spreading Code Encryption (SCE), is used for signal authentication through encryption of the PRN code typically with a symmetric encryption and is already implemented for non-civilian signals in the GPS and in the Galileo for government use. These techniques are analysed and explained further in the book. Nevertheless, an issue with NME and SCE is that in order to be implement in GNSS devices, changes in the receiver's architecture need to be made.

It is seen through the literature review that there is not a perfect solution that protects from an attack with 100% effectiveness [10, 11]. It can also be observed that most literature focuses on the methods for detecting an attack while few turn the attention to finding methods for preventing an attack from occurring. The amount of literature that is focused on detection vastly outweighs the amount of literature which displays the current vulnerabilities of GNSS. In summary the GNSS security is still an open field of research where new security techniques are continuously being develop meanwhile, new attacking techniques are discovered for breach those security measures, as the attackers are trying hard to find new vulnerabilities in the system [16, 19, 20].

The literature review has revealed lack of sufficient research in regard to spoofing detection based on the Doppler effect, which occurs when a signal is transmitted by a moving transmitter and received by a moving receiver. During this project a method for detecting the presence of spoofed signals by monitoring the shift in frequency of different satellite signals at different directions relative to the receiver, will be examined with a set of experiments.

## 3 Research/Design/Process Methods

This chapter contains the necessary prior research made and development process of a detection method against spoofing attacks by taking a unique approach based on existing principles and fundamentals which are described analytically. The process is analysed along with the planning and designing of the experiment.

### *3.1 GNSS Messaging Protocols*

The GNSS receiver used for this investigation contains a u-blox 8/M8 chip using the National Marine Electronics Association (NMEA) ASCII and UBX messaging protocols. Message protocols are used as a method of one-way communication between the GNSS satellites and receivers. A message protocol is used to translate binary data which are transmitted in the analogue signal and converted to digital by the receiver. The NMEA 0183 messages contain a total of 19 sentences.

$GPGGA, $GPGLL and $GPRMC are the most frequently used sentences for most applications as they contain PVT related information, while the rest contain information used mainly by the receiver such as the precision of the calculated position, the number of connected satellites, noise to signal ration etc.

However useful the information contained within NMEA sentences might be for most applications, no information related to the signal's characteristics such as the frequency of received signal or Doppler shift is contained within any of the 19 NMEA sentences. For this reason, data were extracted using the UBX messaging protocol as it contains information related to the received satellite signals. UBX protocol contains broadcast navigation data and satellite signal measurements within the UBX-RXM-MEASX message [9]. This message reports the satellite reference time, the carrier to noise ratio, the measured code phase and most importantly the Doppler values for all connected satellite signals.

## 3.2 Doppler Effect and Special Relativity for Electromagnetic Waves

Electromagnetic waves are all transverse waves, oscillating perpendicular to the direction of energy transfer and in a vacuum all electromagnetic waves travel at the speed of light at about $3 \times 10^8$ ms$^{-1}$. Based on Maxwell's equations, a change in magnetic field generates a change in electric field and vice-versa. Electromagnetic waves can be generated by introducing alternating current to electrons or a discharging spark, changing their magnetic field, causing them to oscillate [17]. This is the basic principle of how RF transmitters produce signals.

The Doppler effect is a phenomenon that appears in wave propagation, causing an observed signal to experience shift in frequency from the originally transmitted frequency and it is caused by the motion of either the source or the receiver or both. An observer will experience either a frequency increase or decrease, depending on whether it's moving towards or away relative to the transmitting source, respectively. This phenomenon is also observed in GNSS signals as the satellites are in constant motion while transmitting electromagnetic signals. There are three different scenarios at which an increase in frequency can be observed:

1. The source moves towards the stationary receiver,
2. The receiver moves towards the stationary source,
3. Both the source and the receiver move towards one another.

In order for to observe a decrease in frequency there are also three different scenarios:

1. The source moves away from the stationary receiver,
2. The receiver moves away from the stationary source,
3. Both the source and the receiver move away from one another.

The hypothesis is that this phenomenon will be observed for every satellite signal when the GNSS receiver moves at any given direction, as the GNSS satellites are located at different directions around the receiver. For example, Fig. 1 shows an electromagnetic wave propagating from a stationary source "S" to a stationary observer "O" and "1" and "2" represent the first and second wave crests respectively. The distance between each consecutive wave crest is the resonant wavelength of the

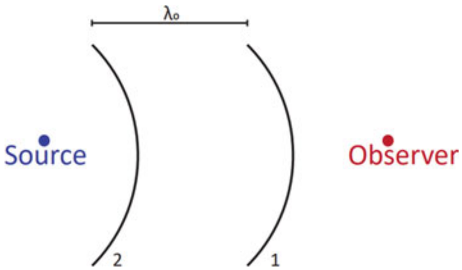


**Fig. 1** Stationary source and observer

signal and is denoted by "$\lambda_o$". Figure 2 shows a source moving towards an observer with velocity "v" relative to the observer while transmitting electromagnetic radiation. The source emitted the first crest of the wave at position "A"; by the time the source emitted the second wave crest, it has moved to the position "B". The distance the source has covered is equal to the velocity multiplied by the elapsed time (d = vT) while the observed wavelength will be equal to the distance the first wave crest has covered (which for electromagnetic waves is the speed of light multiplied by the time period since the wave was emitted) minus the distance covered by the source from A to B. The wavelength is given by the formula (1) below, where the wavelength is denoted by λ, the constant for the speed of light in a vacuum is c and the wave's period is T:

$$\lambda = cT - vT = (c - v)T \tag{1}$$

According to the special theory of relativity, the observer in Fig. 2 will experience time dilation, so the time of the wave's period as observed by the source will not be equal to the time observed by the observer [14]. To compensate for time dilation, Eq. (2) must be taken into account in order to find the time period between each wave crest as observed by the observer. Therefore, the period observed by the observer will be equal to the period as the wave is emitted by the source ($T_o$) divided by the square root of one minus the source's velocity (v) squared divided by the speed of light (c) squared.

$$T = \frac{T_o}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{2}$$

By applying the formulae $\lambda = \frac{c}{F}$, which is the standard formula of wavelength (c is the velocity of wave propagation and F is the frequency) and $T = \frac{T_o}{\sqrt{1-\frac{v^2}{c^2}}}$ in

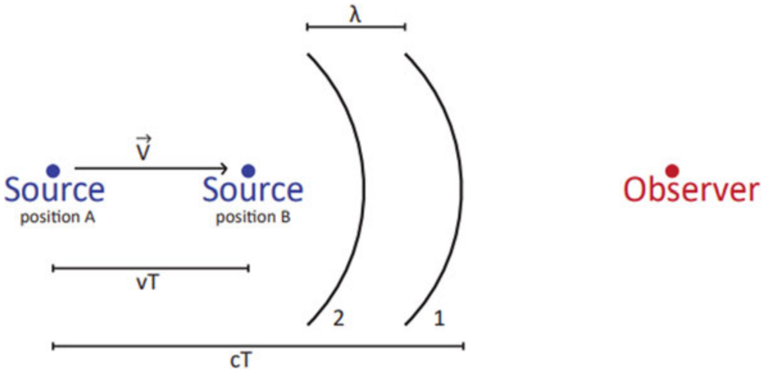


**Fig. 2** Source moving towards observer

Eq. (1), the frequency observed by the observer can be derived:

$$
\begin{aligned}
&\lambda = (c - v)T \\
&\frac{c}{F_{obs}} = (c - v) \cdot \frac{T_o}{\sqrt{1 - \frac{v^2}{c^2}}} \\
&F_{obs} = \frac{c}{c - v} \cdot \frac{1}{T_o} \cdot \sqrt{1 - \frac{v^2}{c^2}} \\
&F_{obs} = \frac{c}{(c - v)} \cdot f_o \cdot \sqrt{1 - \frac{v^2}{c^2}} \\
&F_{obs} = \sqrt{\frac{c^2}{(c - v)^2} \cdot \left(1 - \frac{v^2}{c^2}\right)} \cdot f_o \\
&F_{obs} = \sqrt{\frac{c + v}{c - v}} \cdot f_o
\end{aligned}
\tag{3}
$$

where $\mathrm{F_{obs}}$ is the frequency as seen by the observer, $f_o$ is the resonant frequency transmitted by the source, c is the speed of the wave which is equal to the speed of light at about $3 \times 10^8$ $\mathrm{ms}^{-1}$ and v is the relative velocity of the source with respect to the observer. Equation (3) can only be applied for when the source is moving towards the receiver as the relative velocity will be positive. In cases where the source is moving towards the observer, the signal is often referred as blue shifted as the wavelength shrinks, causing the visible light to shift to blue.

In cases where the source is moving away from the observer, the relative velocity will be negative and the frequency observed can be expressed by the Eq. (4):

$$
F_{obs} = \sqrt{\frac{c - v}{c + v}} \cdot f_o \tag{4}
$$

This case is known as redshift, as the wavelength expands, causing visible light to shift to red.

Doppler shift is the amount of frequency in Hertz, the received signal has shifted from the original transmission frequency and can be calculated by subtracting the original frequency (denoted by $\mathrm{f_o}$) from the frequency observed (denoted by $\mathrm{F_{obs}}$), as seen in Eq. (5) bellow:

$$
D_{shift} = F_{obs} - f_o \tag{5}
$$

where $D_{shift}$ is the amount of Doppler shift in Hertz. The sign of the resulting Doppler shift indicates whether the source and the observer are approaching (positive sign) or drifting apart (negative sign).

Notice that the velocity “v” in the Eqs. (3) and (4) are relative to the observer. The observer and the source are moving in a 3-dimensional space so the relative velocity will be for a 3-dimensional space as well. In order to calculate the relative velocity of a satellite with respect to the receiver, the velocities for the receiver and the satellite need to be expressed as vectors with the use of the following formula, where $V_{rec}$ denotes the receiver’s velocity, $V_{sat}$ denotes the satellite’s velocity and $V_{rel}$ denotes the relative velocity:

$$\overrightarrow{v_{rec}} = \begin{pmatrix} v_{rec-x} \\ v_{rec-y} \\ v_{rec-z} \end{pmatrix} \overrightarrow{v_{sat}} = \begin{pmatrix} v_{sat-x} \\ v_{sat-y} \\ v_{sat-z} \end{pmatrix} \overrightarrow{v_{rel}} = \begin{pmatrix} v_{sat-x} \\ v_{sat-y} \\ v_{sat-z} \end{pmatrix} - \begin{pmatrix} v_{rec-x} \\ v_{rec-y} \\ v_{rec-z} \end{pmatrix} \tag{6}$$

Apart from the movement of the receiver/source, Doppler shift can also be attributed to the movement of the medium, such as the movement of the ionosphere which refract the electromagnetic signals. In a moving medium, different frequencies travel with different velocities and the medium is called dispersive. A similar phenomenon with the same effect is observed in the dispersion of white light passing through a prism [17].

### 3.3 Experimental Setup

In order to identify whether the GNSS signals originated from actual satellites moving at different speeds relative to the observer, or if the signals were transmitted by a spoofer from a single source, the change in Doppler shift during a change in the velocity of the receiver, was measured and compared for each GNSS signal. SDR based spoofers can be programmed to transmit signals with variations in frequency to imitate the Doppler shift observed in authentic signals. The experiment accounts for that and uses the change in Doppler shift while the receiver changes its velocity as this parameter cannot be controlled or imitated by any spoofer with a single transmitter.

Based on Eq. (3), when a change in velocity is introduced, Doppler shift will also change. In an environment with only authentic satellite signals, the change in Doppler shift will vary for each satellite signal, as the satellites are at different positions, moving at different directions, with different relative velocities. However, if the signals originate from a spoofer transmitting from a single source, then the change in Doppler shift will be equal for every signal, as the change in relative velocity will remain the same for all supposed satellite signals as they all originate from the same source.

Based on the aforementioned principles, a test experiment was designed and carried out in order to measure and compare the Doppler shift from multiple satellite signals, originated from different positions in the sky, having different velocities relative to the receiver. To achieve this, the GNSS receiver was accelerated from a stationary position while recording measurements of Doppler shift. The experiment was repeated five times with different directions and different accelerations. The measurements were obtained and recorded with the use of u-center. The experiments were conducted in an open field space, with no buildings nor any high voltage power lines nearby so that the measurements would not be affected by interferences.

### *3.4 Detection Method*

The recorded measurements for Doppler shift were extracted from the UBX messages into excel spreadsheets to then be imported and analysed into MATLAB. To analyse the data, an alorithm was developed, seen in Fig. 3, to compare the change in Doppler shift for each satellite signal to one another. The algorithm checks and compares the value for the change in Doppler shift for every satellite signal at every instance a measurement was obtained. At the time of recording the experiments, the receiver was connected with five GPS satellites. If the algorithm detects at any time that the change in Doppler shift was the same for all five satellite signals, a message appears, aletrting the user for possible spoffing attack and also provides the exact time of detection. If on the other hand, the signals had different amounts of change in Doppler shift, then the algorithm considers the signals to be authentic.

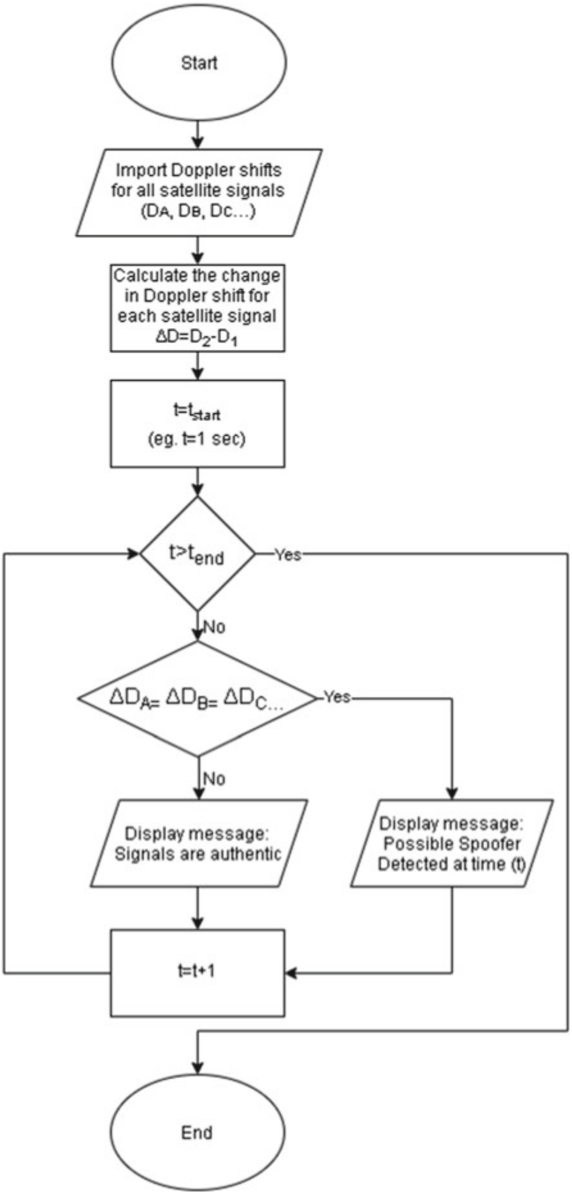
## 4 Results, Analysis and Discussion

In this chapter an experiment that was conducted, during which the receiver was accelerated and decelerated whilst it was connected to five GPS satellites, is reviewed and the results obtained are analysed. Measurements from the experiment are presented graphically and are then used by the algorithm developed to compare the signals and assess their authenticity. Then a simulation of a spoofer transmitting counterfeit GPS signals to an accelerating receiver, was performed, and the results obtained are reviewed and analysed. Measurements are also presented graphically and the algorithm is used to compare the simulated signals and assess their authenticity. Lastly, the results from the experiment and the simulation are compared and discussed.

### *4.1 Experiments and Results*

Figure 4 shows the orbital direction as well as the position of each GPS satellite connected with the receiver at the time of recording the experiments. The position of the receiver is indicated with the red arrows while the satellites that were used in the experiments are circled in red. For the first experiment, the receiver was accelerated from stationary until reaching a maximum velocity of 24.73 m/s after which it was decelerated to a stop. The receiver during the first experiment was moving towards

**Fig. 3** Algorithm's flowchart


Start
Import Doppler shifts for all satellite signals (DA, DB, DC...)
Calculate the change in Doppler shift for each satellite signal ΔD=D2-D1
t=tstart (eg. t=1 sec)
t>tend
Yes
No
ΔDA= ΔDB= ΔDC...
Yes
No
Display message: Signals are authentic
Display message: Possible Spoofer Detected at time (t)
t=t+1
End

the West. Satellites G02, G05 and G06 were located east relative to the receiver whereas the satellites G29 and G31 were located west relative to the receiver.

From the above figure, it is seen that the satellites G02 and G06 were moving away from the receiver while the satellites G05, G29 and G31 were moving towards the position of the receiver. Therefore, based on Eqs. (3) and (4), when the receiver is

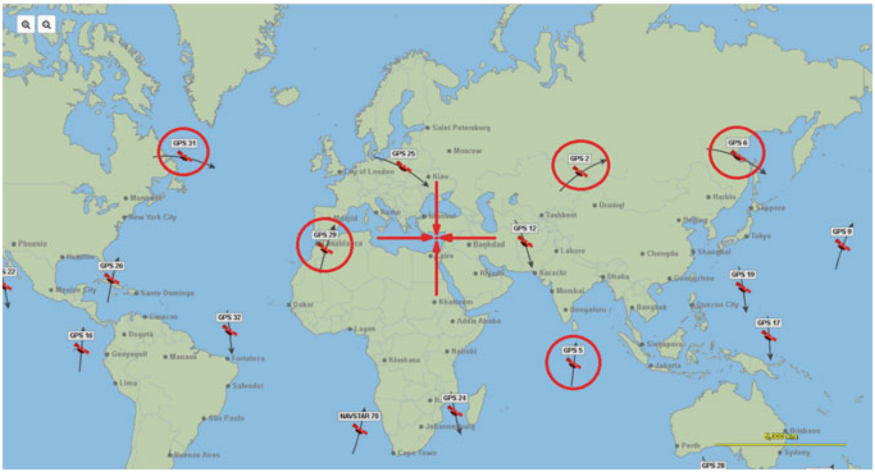


**Fig. 4** Map of satellite positions [13]

stationary, it is expected to observe a redshift (decrease in frequency) in the signals coming from satellites G02 and G06 and a blueshift (increase in frequency) in the signals coming from satellites G05, G29 and G31. When the receiver begins moving towards the West, it is also expected to observe a further increase in frequency in the signals coming from satellites G29 and G31 as the receiver will be moving towards them and at the same time a further decrease in frequency in the signals coming from satellites G02, G05 and G06 as the receiver will be moving away from them.

The following Figs. 5, 6, 7, 8 and 9 display the obtained measurements of Doppler shift against time, during the experiment for the five satellite signals. The receiver was initially at rest, then began accelerating at t = 1 s until t = 10 s at which point it began decelerating until it reached a stop at t = 19 s.

From the above figures it is seen that the satellite signals all had different values of Doppler shift; the satellites which were approaching the receiver had positive Doppler shift while the satellites which were moving away had negative Doppler

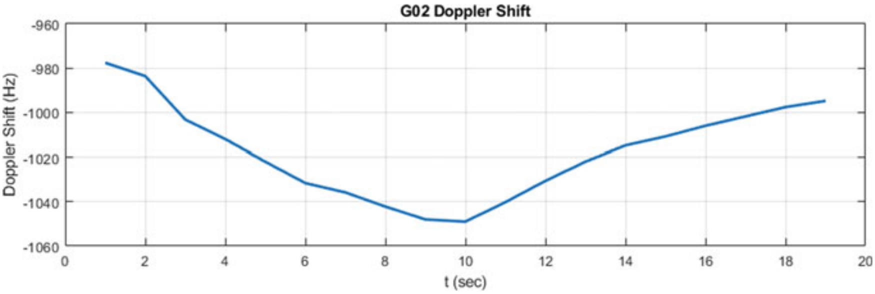


**Fig. 5** Doppler shift against time for G02

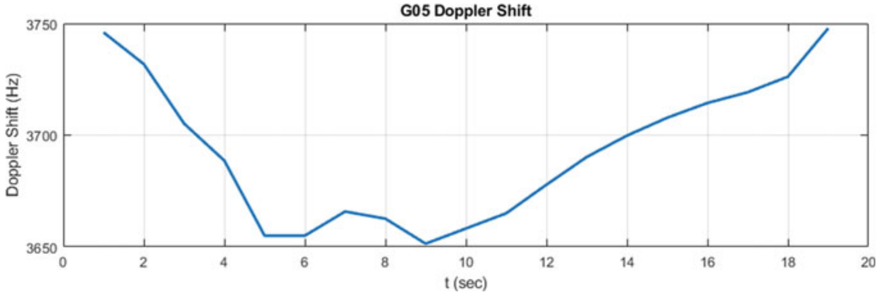


**Fig. 6** Doppler shift for G05

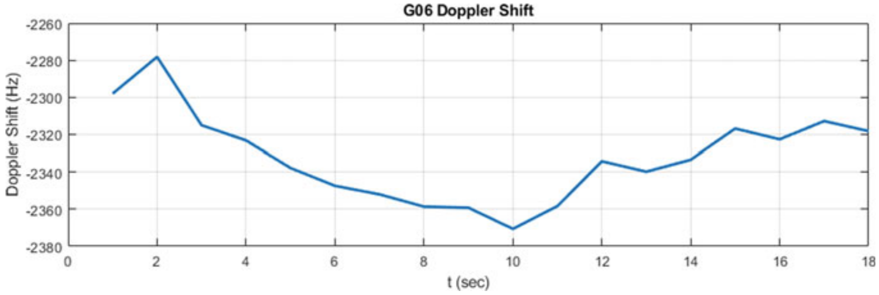


**Fig. 7** Doppler shift for G06

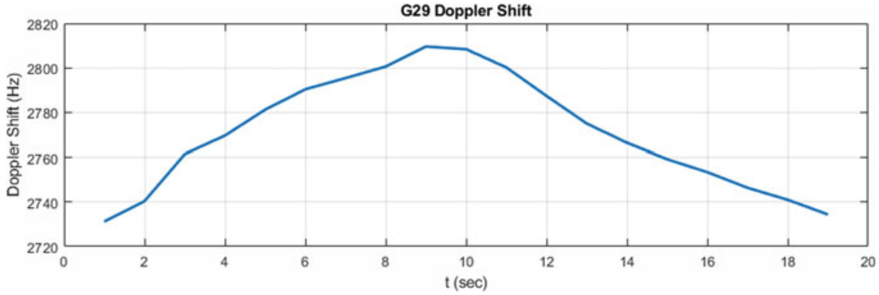


**Fig. 8** Doppler shift for G29

shift. It can be observed that as the receiver began accelerating, the Doppler shift in all satellite signals began to alter either increasing or decreasing with different rates and by different amounts for each satellite. Even though all the measurements were taken at exactly the same time during the same experiment with the same amount of acceleration, the results differ due to the fact that the relative velocity for each satellite with reference to the receiver was different.

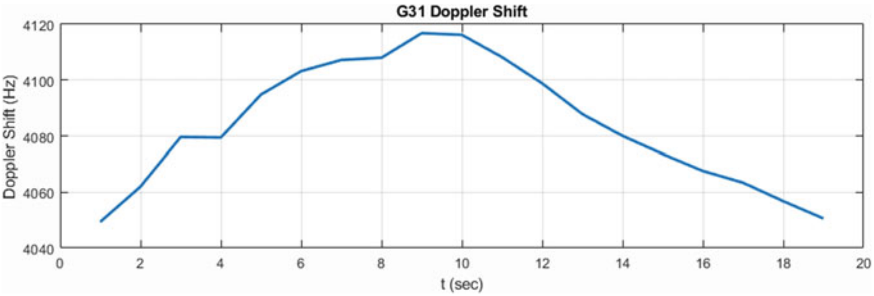


**Fig. 9** Doppler shift for G31

In Fig. 10, the change in Doppler shift over the course of the experiment for the five satellite signals is displayed. The change in Doppler shift was calculated using the formula:

$$\Delta Ds = Ds_2 - Ds_1 \tag{7}$$

where $\Delta Ds$ is the change in Doppler shift and Ds is the value of Doppler shift.

At no point does all of five satellite signals have the same change in Doppler shift. Although it is possible to measure the same change in Doppler shift in some satellite signals as seen with signals G29 and G31 at time t = 15 s, where both signals experienced a change in Doppler shift of −5.8 Hz, it is highly improbable for this to occur for the satellite signals, and the more the satellites the receiver connects to, the probability of them having the same change in Doppler shift is even lower

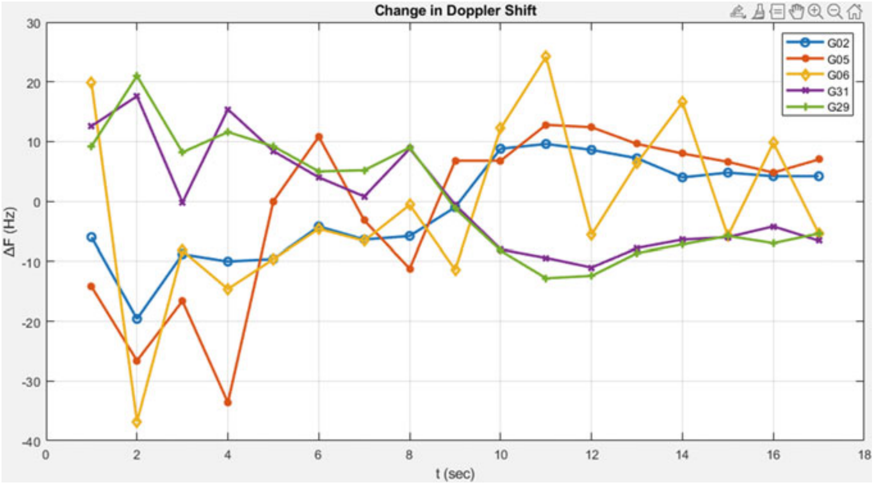
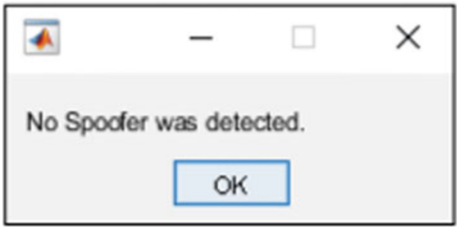


**Fig. 10** Change in doppler shift

**Fig. 11** No spoofer message



No Spoofer was detected.

especially during acceleration. The change in Doppler shift depends on the change in relative velocity, therefore for the satellites which are close to each other and move in a similar direction, the change in the relative velocity will be similar and consequently the change in Doppler shift in their signals will very likely also be similar. For example, signals from G29 and G31 overall have very similar trends, as seen in Fig. 10, as their changes in relative velocity are also similar.

As the change in Doppler shift was not the same at any instance for all the satellite signals, the algorithm did not detect the possibility of the signals being transmitted by a spoofer and thus a message was prompted informing the user that no spoofer was detected, as seen in Fig. 11.

## *4.2 Simulation*

After the experiments were done, a simulation of GNSS signals transmitted by a spoofer was performed. The experiment began with the receiver at rest and the Doppler shifts for each signal were recorded. Then the same initial values of Doppler shift were used as the initial values for the simulation. The subsequent values of Doppler shift were then calculated by rearranging the Eq. (7) to: $Ds_2 = Ds_1 + \Delta Ds$. The change in Doppler shift ($\Delta Ds$) was set the same for all signals, as it is assumed that the relative velocity between the source and the receiver, will be the equal for all the signals which are transmitted by the same source, the same way how a spoofer with a single transmitter operates. In the following figures, the simulated Doppler shifts against time for the spoofed signals are presented.

Through Figs. 12, 13, 14, 15 and 16 it is seen that, even though all the GNSS signals have different frequencies, their trends during acceleration and deceleration are identical. This is due to the fact that all GNSS signals are transmitted from the same source and thus, the receiver has the same relative velocity. If the GNSS signals were transmitted at the same frequency, they could have been picked-up and rejected by the receiver's existing anti-spoofing algorithm as reviewed previously in literature. However, when the spoofer transmits each GNSS signal at a different frequency, the signals can get past the detection algorithms without being detected.

Figure 17 shows the graph of the change in Doppler shift for the five spoofed GNSS signals obtained through the simulation. The graph further highlights that the

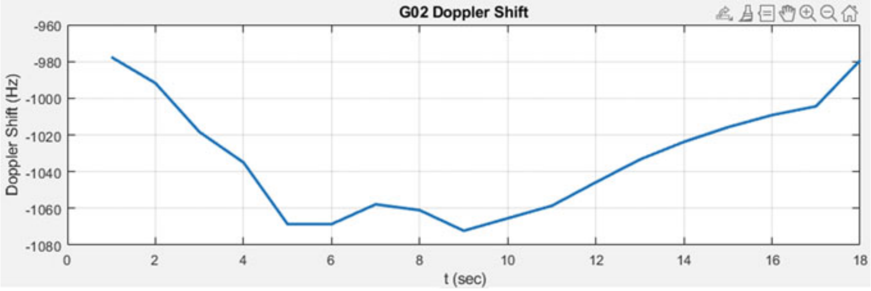


**Fig. 12** Simulated spoofed G02 doppler shift

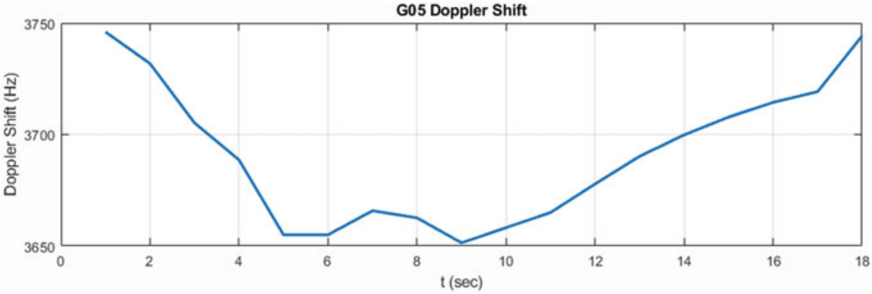


**Fig. 13** Simulated spoofed G05 doppler shift

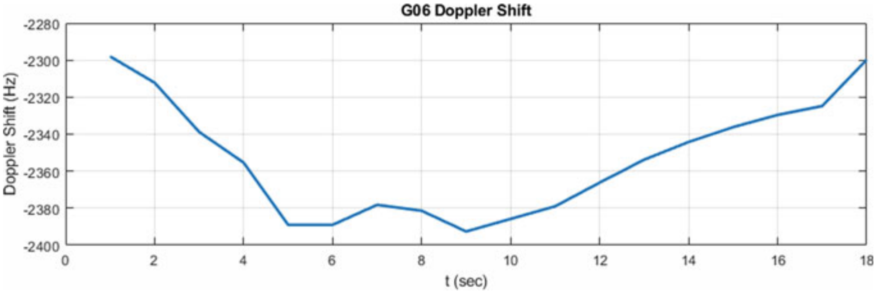


**Fig. 14** Simulated spoofed G06 doppler shift

signals originating from the same source can be linked together by analysing the trends of their change in Doppler shift for an accelerating receiver. As seen above, the graphs for the change in Doppler shift are perfectly aligned for every GNSS signal, unlike Fig. 10 where the signals originated from different sources.

Based on this association, the algorithm was able to detect and display a warning message for the possibility of the presence of a spoofing device. The warning will

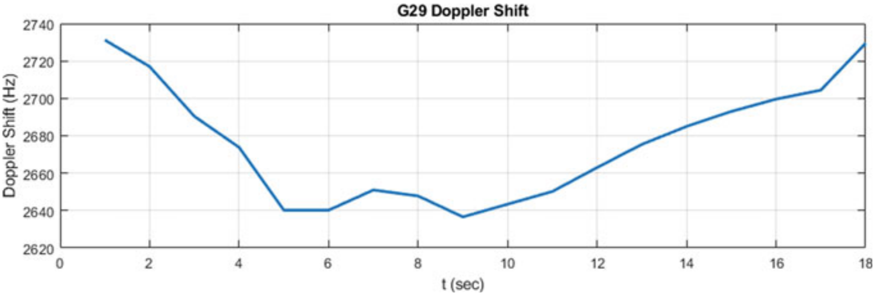


**Fig. 15** Simulated spoofed G29 doppler shift

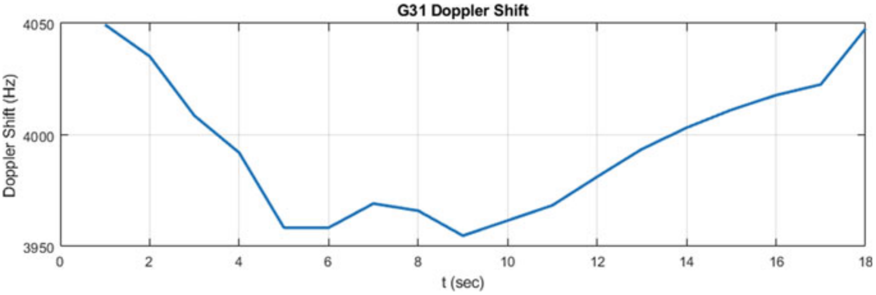


**Fig. 16** Simulated spoofed G31 doppler shift

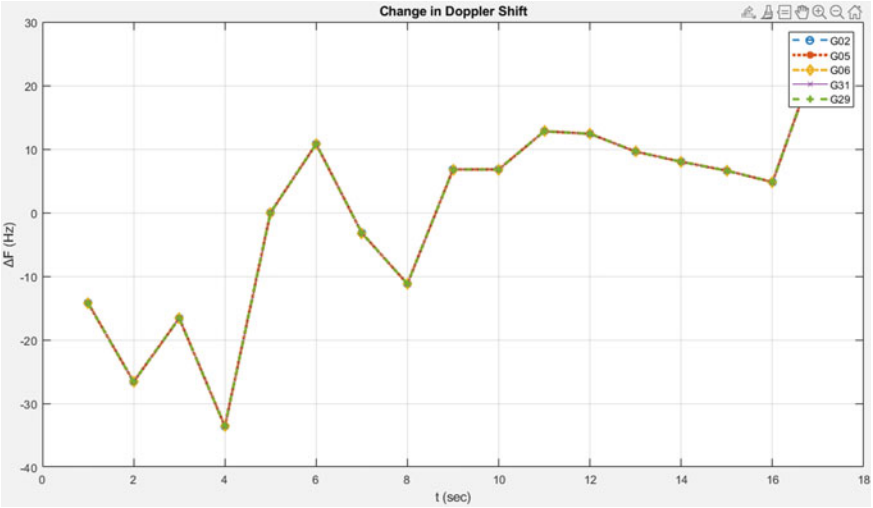


**Fig. 17** Simulated change in doppler shift of spoofed signals

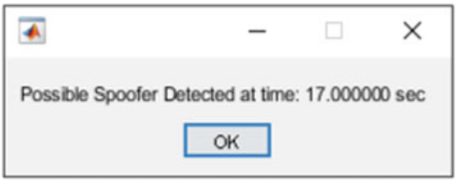

Possible Spoofer Detected at time: 17.000000 sec

**Fig. 18** Spoofer detection message

also show the exact time all the signals were linked together, as can be seen for example in Fig. 18.

### *4.3 Discussion*

Further experiments similar to the one viewed in Sect. 4.1 were conducted multiple times and the results obtained from all the experiments match and support the initial hypothesis that the receiver will observe different amounts of change in Doppler shift for signals originating from different sources over the course of an experiment where the velocity of the receiver will be constantly changing. The results from these experiments are included in the Appendix.

The results from the simulation in Sect. 4.2 have shown that signals transmitted from the same source can be linked together by comparing their change in Doppler shift. This method can be used to detect and mitigate a spoofing attack.

Comparing the results from the experiment to the results from the simulation, it is clearly observed that the signals can easily be separated into two categories; the signals with low spatial correlation, which will be the ones with dissimilar trends of their changes in Doppler shift, and the signals with high spatial correlation, which will be the signals with very similar or identical trends of their changes in Doppler shift. The ones separated into the category with high spatial correlation can be considered to be originating from a spoofer.

## 5 Conclusions

This research has revealed that it is possible to detect the presence of a GNSS spoofer by analysing the Doppler shift in GNSS signals and comparing how it changes for each signal while the receiver is accelerated. Spoofing attacks can be performed simply with the use of a readily available SDR, which can be programmed to transmit counterfeit GNSS signals, and paired with an RF jammer to jam weak GNSS signals, the attacker can avoid being detected by the currently used countermeasures in most cases. The need for implementing and improving countermeasures in GNSS is constantly becoming more essential.

## Bibliography

1. Broumandan A, Jafarnia-Jahromi A, Dehghanian V, Nielsen J, Lachapelle G (2012) GNSS spoofing detection in handheld receivers based on signal spatial correlation. In: Proceedings of the 2012 IEEE/ION position, location and navigation symposium
2. Dovis F (2015) GNSS interference, threats, and countermeasures. Artech House, Norwood
3. Jablonski D (2002) Radio navigation systems. Reference data for engineers, pp 37-1–37-9 (online). Available at: https://www.sciencedirect.com/science/article/pii/B978075067291750039X
4. Karagiannis D, Argyriou A (2018) Jamming attack detection in a pair of RF communicating vehicles using unsupervised machine learning. Veh Commun 13:56–63 (online). Available at: https://www.sciencedirect.com/science/article/pii/S221420961730222X
5. Qi W, Zhang Y, Liu X (2016) A GNSS anti-spoofing technology based on Doppler shift in vehicle networking. In: 2016 International wireless communications and mobile computing conference (IWCMC) (online). Available at: https://ieeexplore.ieee.org/document/7577146
6. Scott L (2021) Anti-Spoofing & authenticated signal architectures for civil navigation systems. In: Proceedings of the 16th international technical meeting of the satellite division of the institute of navigation
7. Sharifi-Tehrani O, Sabahi M, Danaee M (2021) GNSS jamming detection of UAV ground control station using random matrix theory. ICT Express 7(2):239–243 (online). Available at: https://www.sciencedirect.com/science/article/pii/S2405959520303040

## *References*

8. (2001) The effects of earth's upper atmosphere on radio signals. In: Radiojove.gsfc.nasa.gov. https://radiojove.gsfc.nasa.gov/education/materials/iono.htm
9. (2016) u-blox M8 FW SPG3.01 migration guide. In: U-blox.com. https://www.u-blox.com/sites/default/files/u-blox-M8-FWSPG301-MigrationGuide_AppNote_%28UBX-15028330%29.pdf
10. (2021) GPS—NMEA sentence information. In: Aprs.gids.nl. http://aprs.gids.nl/nmea/
11. Alfriend K, Vadali S, Gurfil P et al (2010) Introduction. Spacecraft formation flying, pp 1–11. https://doi.org/10.1016/b978-0-7506-8533-7.00206-2
12. Dietsche K, Reif K (2018) Automotive handbook. 10th ed, pp 1632–1635
13. Ford D (2021) Live world map of satellite positions. In: In-The-Sky.org. https://in-the-sky.org/satmap_worldmap.php
14. Kogut J (2001) Physics according to einstein. Introduction Relativ 7–22. https://doi.org/10.1016/b978-0-08-092408-3.50006-8
15. Leclère J, Landry R Jr, Botteron C (2018) Comparison of L1 and L5 bands GNSS signals acquisition. Sensors 18:2779. https://doi.org/10.3390/s18092779
16. Moore R (2003) Satellite RF communications and onboard processing. Encycl Phys Sci Technol 439–455. https://doi.org/10.1016/b0-12-227410-5/00884-x
17. Salous S (2013) Radio propagation measurement and channel modelling. 1–33
18. Warner J, Johnston R (2003) GPS spoofing countermeasures. http://lewisperdue.com/DieByWire/GPS-Vulnerability-LosAlamos.pdf
19. Xu G, Xu Y (2016) GPS: theory, algorithms and applications. 3rd ed, pp 1–13
20. Zhu Z, Li C (2016) Study on real-time identification of GNSS multipath errors and its application. Aerosp Sci Technol 52:215–223. https://doi.org/10.1016/j.ast.2016.02.032

# Using a Functional Voxel Model to Simulate Swarm Motion of a Multi-agent System in a Confined Space

**Kseniya Shutova and Anastasia Sycheva**

**Abstract** This paper proposes a way to implement potential attraction/repulsion fields. We consider the implementation of a multi-agent system that moves toward a common target in unbounded space, using swarming algorithms based on Reynolds rules. Three swarm modeling approaches are presented, jointly implementing a finite algorithm that ensures collision avoidance with all available barrier types. The algorithms are described in more detail in our previous paper [1]. We propose the joint application of the investigated swarming algorithm and a predator avoidance model based on reinforcement learning to provide collision avoidance with dynamically occurring barriers. Thus, it is possible to temporarily evade the target to ensure the safety of the agents' movement. The algorithm is based on Q-learning, the result of which is an action function. We consider the behavior of a multi-agent system modeled using the proposed approaches in a bounded space—a polygon or polygon. In this case, in addition to the described interactions, the movement of a group of agents is influenced by repulsive forces from walls. There is a problem of compensation of repulsive and attractive potentials, accompanied by braking of agent or ignoring of walls when moving to the target. This problem is proposed to be solved using function-voxel models. The principle of movement of agents according to the local geometric characteristics stored in the represented graphical M-images of the simulated polygon is described. In the paper, the solution of problems for hazard avoidance using the approach of potential fields, which are expressed by voxel surfaces, is obtained. The advantages of using these models and the need for an algorithm for predator avoidance are highlighted.

**Keywords** Flocking motion · Swarm modeling · Multiagent system · Machine learning · Reinforced learning · Functional voxel modelling (FVM) · Function of local zeroing out (FLOZ-function)

K. Shutova (✉) · A. Sycheva
Institute of Control Sciences of RAS, 65 Profsoyuznaya street, Moscow 117997, Russia
e-mail: shutova.k.u@yandex.ru

A. Sycheva
e-mail: a.a.sycheva@mail.ru

Y. A. Vershinin et al. (eds.), *Technologies for Smart Cities*,
https://doi.org/10.1007/978-3-031-05516-4_2

# 1 Introduction

The idea of problem solving by a group of simple systems has long been a focus in the field of artificial intelligence. Flock behavior is the basis of social system behavior. In today's world, mobile robots are designed to replace man-machine systems and single robots for tasks that are time-consuming, large, monotonous, or tedious, and for tasks that are hazardous to human health or life.

To perform the movement of a multi-agent system it is necessary to use methods of swarm behavior. This paper considers movement based on Reynolds rules: attraction to the target, speed alignment, attraction/repulsion from movement participants [2].

A group of dynamic agents and their relations $(\nu, \varepsilon)$ with equation of motion (1) is considered:

$$\begin{cases} \dot{q}_i = p_i \\ \dot{p}_i = u_i \end{cases} \tag{1}$$

where $q_i, p_i, u_i \in \mathbb{R}^{\mathrm{m}}$are coordinates, velocity, and accelerations of robots, respectively, $m = \overline{2, 3}$, $i \in \nu$. Agents can be cells and molecules, complex organisms such as insects, birds, animals, groups of people and crowds. We will call the members of their group α-agents.

The set $N_i = \{j \in \nu : \left\|q_j - q_i\right\| < r\}$ defines the neighbors of robot i. The range $d > 0$ is the desired distance between the agents, and $r > 0$ is the radius of the region in question. Let $z$ be the current distance between them. Let us introduce the scalar bump function $\rho_h(z)$, the action function $\phi_\alpha(z)$, represented by formulae (2), (3).

$$\rho_h(z) = \begin{cases} 1 & z \in [0, h); \\ \frac{1}{2}\left(1 + \cos\left(\pi \frac{z-h}{1-h}\right)\right), & z \in [h, 1); \\ 0 & \text{else.} \end{cases} \tag{2}$$

$$\phi_\alpha(\mathrm{z}) = \rho_h(z)\phi(z - d), \tag{3}$$

where $\phi(z) = \frac{1}{2}((a + b)\sigma(z + c) + (a - b))$ is the value ensuring the preservation of the distance between the agents. $\sigma(z) = \frac{z}{\sqrt{1+z^2}}$ is an unequal sigmoidal function. $0 < a \le b, c = \frac{|a-b|}{\sqrt{4ab}}$. At distance $z > d$ there is a force of attraction between the agents. It weakens as it approaches $z = r$. Here $h \in (0, 1)$ is responsible for the degree of steepness ("uphill gradient") of the function $\rho_h(z)$.

The formula (2) is needed to arrange the communication between the agents as a continuous change from maximum to minimum communication. Function (3) is used to study the system for stability.

Flocking motion is well simulated [3]. The following methods were investigated:

1. Reynolds rule algorithm [4]: $u_i = u_i^{\alpha} = f_i^g + f_i^d$—the sum of the gradient-based term ($f_i^g$) and the velocity consensus term ($f_i^d$) due to the change of connections between the agents:

$$f_i^g = \sum_{j \in N_i} \phi_{\alpha}\big(||q_j - q_i||\big)\vec{n}_{ij},$$

$$f_i^d = \sum_{j \in N_i} a_{ij}(q)\big(p_j - p_i\big), a_{ij} = \rho_h\left(\frac{||q_j - q_i||}{r}\right)$$

Here $\vec{n}_{ij} = \frac{q_j - q_i}{||q_j - q_i||}$ is the vector along the line connecting the i-th and j-th agents, $a_{ij}$ is the element in the relation matrix A responsible for the presence or absence of interaction between the objects. This method is key for all subsequent algorithms, as it forms the structure of the swarm.

Algorithm of the group having a certain target in the obstacle-free space: $u_i = u_i^{\alpha} + u_i^{\gamma}$, where $u_i^{\gamma} = -c_1(q_i - q_r) - c_2(p_i - p_r)$ is the navigation feedback used to move to the target. Here $c_1, c_2 > 0$. The pair $(q_r, p_r)$ is a static/dynamic target. The collective behavior of the group of robots applying the first algorithm differs from the behavior of the robots applying the second algorithm: no fragmentation occurs, the robots move cohesively to the target.

Algorithm of the group having a target in the space with obstacles with the possibility of avoiding multiple obstacles: $u_i = u_i^{\alpha} + u_i^{\beta} + u_i^{\gamma}$. Here $u_i^{\beta}$ is defined similarly to the first algorithm, except for the following: formula (3) is replaced by

$$\phi_{\beta}(z) = \rho_h(z)\big(\sigma\big(x - d_{\beta}\big) - 1\big),$$

which ensures constant repulsion. Here $d_{\beta}$ is the minimum distance from the agent to the object.

The robots avoid obstacles by separating from the group, and then reestablish their swarming behavior as they move toward the target.

The algorithms are described in more detail in the article [1].

It is worth noting that collision avoidance has the highest priority with respect to all available obstacles: robots, moving and stationary objects. Each algorithm is built based on the previous one. Each of them can be perceived as stages of construction of the final algorithm $u_i = u_i^{\alpha} + u_i^{\beta} + u_i^{\gamma}$.

## 2 Avoiding a Predator Attack

A basic requirement for most robotic systems is the ability to navigate safely in a common environment, that is, the ability to move toward certain targets while avoiding collisions with obstacles, teammates, and other groups. In some cases, agents need to change their destination to avoid a dangerous area.

The next algorithm under consideration is equipped with the ability to temporarily move away from the main target, if the group of agents is in danger: some object, called a predator, moves on them them [5]. The paper considers the application of the predator avoidance method in the case of a space that the robot cannot see at a given moment in its current position, or at the junction of two walls, behind which is a corridor. The predator is artificially placed in the passageway, or at the junction of walls, so that the agent at a greater distance makes a maneuver: slow down before the passageway to analyze the situation or with a large radius of curvature make a corner to have time to react to a possible obstacle.

Let's combine the advantages of coordinated movement, grouping, and reinforcement learning for a system with partial observability. This means that only agents close to the predator can see the direction of the predator's attack.

The algorithm is organized on Q-learning, which is based on the action value function $a_i$. This function gives the expected utility of doing action $a_i$in a given state $s_i$. The desired action of agents is to move in one of the eight directions to escape. This action is used to select a new reference point in the movement algorithm to escape from predators. The reward is an agent's motivation for the chosen action.

Independent (4) and cooperative (5) Q-learning are used for the cohesive pursuit of one goal.

$$Q_i^{k+1}(s_i, a_i) \leftarrow Q_i^k(s_i, a_i) + \alpha \left[ r_i^k + \gamma \max_{a_i^{'} \in A_i} Q_i^k\left(s_i^{'}, a_i^{'}\right) - Q_i^k(s_i, a_i) \right], \quad (4)$$

$$Q_i^{k+1}(s_i, a_i) \leftarrow \omega Q_i^k(s_i, a_i) + (1 - \omega) \frac{\sum_{j=1}^{|N_i|} Q_j^k\left(s_j, a_j\right)}{|N_i|}, \quad (5)$$

Here $\alpha$ is a learning rate, $\gamma$ is a discounting factor, and ω is the weight to determine trust, such that $0 \leq \omega \leq 1$. When $\omega = 1$ the agent trusts only itself, and when $\omega = 0$ the agent trusts only its neighbors. The implementation is demonstrated in Figs. 1 and 2.

If a predator approaches a group of agents, the robots coordinately move away from the interference to a safe distance by changing the target point. After making sure that the predator is out of sight, the agents return to the main target. This method is based on the algorithms of swarm behavior of agents described above. From the results of the simulation, we can conclude that using the procedure of escaping from a predator, agents can perform a greater number of maneuvers while moving. Using Q-learning allows agents to learn from their own and their neighbors' experiences in order to make more deliberate moves at each iteration.

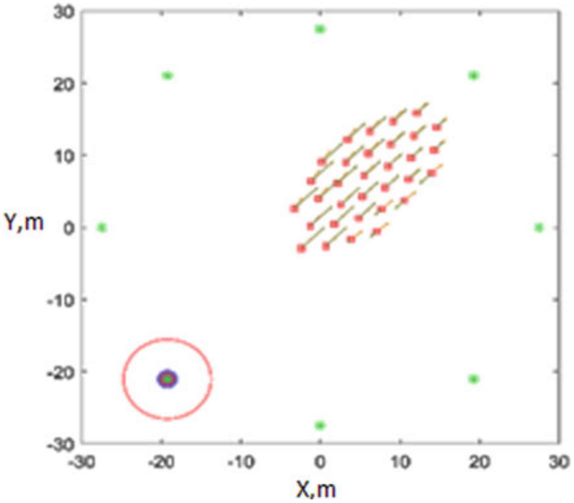


**Fig. 1** Escape from a predator with eight targets

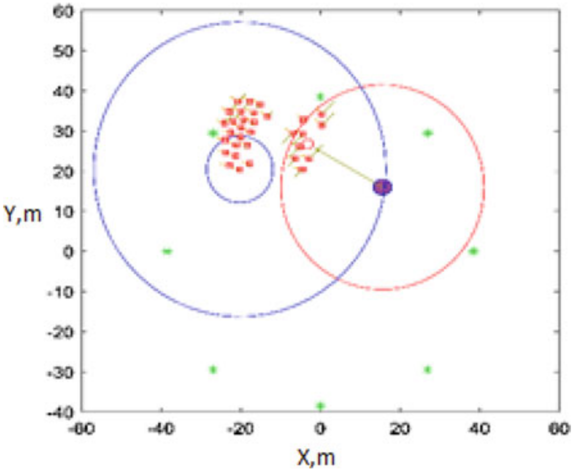


**Fig. 2** Escape from a moving predator in the presence of eight targets

## 3 Movement of an Agent in a Constrained Space

Above, the situations of agents moving in unbounded space were considered. Agents can, with situational awareness, move quite a distance away from the main target if the predator keeps moving in their direction. The question arises as to how agents will behave in conditions close to real: indoors.

When moving in an enclosed space (Fig. 3), agents perform deflection maneuvers away from obstacles, considering, in addition to the previously discussed interactions, also the repulsive force from the walls. The method of potential forces is the addition of all the forces of attraction to the target point and repulsion from obstacles, that

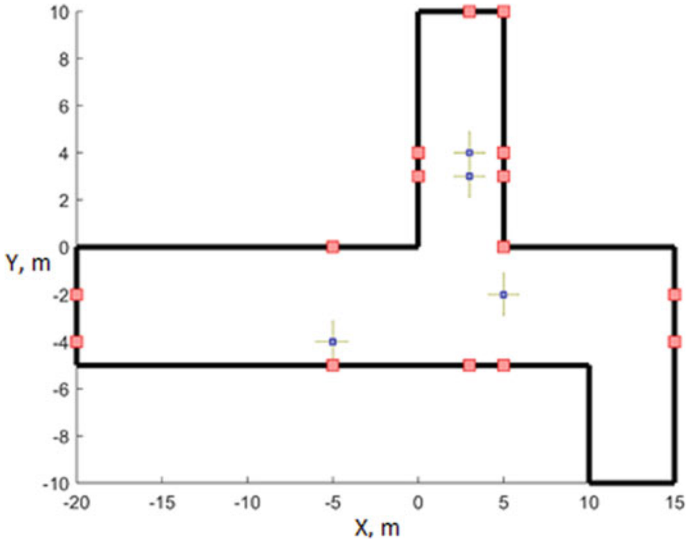


**Fig. 3** Simulation of a bounded geometric domain for four agents

is, it is based on the implementation of the motion of the mobile robot in the field of "information forces". The algorithms described above are based on this method of correlation of forces in the system. When an agent moves in a polygon (confined space), the minimum distance to the walls is calculated for the method of potential forces. In Fig. 3, this is the distance between the agent (blue dot)—and the transverse between the wall and the perpendicular drawn from the agent to the wall (red dot). The calculations are performed in the global coordinate system. Thus, the repulsion vector is composed of the repulsive forces of the agent from each visible wall.

It is possible for the forces of attraction and repulsion to cancel each other out. This leads to a situation in which the agent can either stand still until a motivating force emerges or start moving toward the goal without reacting to the walls, i.e., walking through them.

To solve this problem, it is advisable to consider other models of bounded region representation and corresponding motion algorithms.

The method of functional-voxel modeling [6] is a new approach to computer representation of geometrical information of a modeled function. Any algebraic function (in the two-dimensional case, the two-dimensional function of the form $z = f(x, y)$ is represented by a set of local geometrical characteristics $n_1, n_2, n_{3,} n_4$, calculated in the process of linear approximation of the function area and stored in graphic image-models (M-images) $C_1, C_2, C_3, C_4$. For example, 4 M-images, presented in Fig. 4, describe the geometry of the functional area of the expression:

$$z = \frac{\sin(\pi x)}{\sqrt{|x|} + 10} - y^3 \cos(x^2)$$

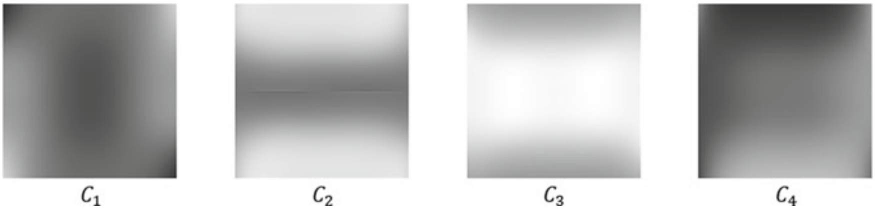


**Fig. 4** Functional-voxel model of function

The projection of its normals onto the XOY plane, expressed in two M-images shown in Fig. 5, will be a more obvious representation of this function.

Thus, the functional-voxel method makes it possible to represent local geometric characteristics of an algebraic function of any complexity in the form of graphical M-images, bringing it to a linear local function of the form:

$$z = -x\frac{n_1}{n_3} - y\frac{n_2}{n_3} + \frac{n_4}{n_3}$$

Moreover, this method provides quick access to the geometric information of each point in the function area, which allows to quickly make a decision about the agent's movement based on a functional-voxel model of the polygon.

It is possible to describe the contour of a bounded geometric area (polygon contour) algebraically by presenting each of the walls with a function for further construction of the function-voxel model However, flozation is a more convenient tool for solving this problem with the functional-voxel method. Flozation is the construction of the polygon contour by R-functional intersection [7, 8] of consecutive zero segments (two-dimensional functions of local zeroing (FLOZ-function)) on the positive area of the function.

Each of the walls of the polygon will be a separate FLOZ. Their consecutive intersection will create on the positive area the zero contour of the simulated polygon. The resulting functional-voxel model of the polygon will be for the agents a kind of map, calculated and modeled in predetermined obstacles. Thus, there is no need

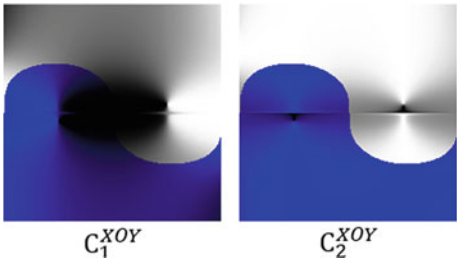


**Fig. 5** Function representation by projection of normals on the XOY

to constantly recalculate and control the situation—all local geometric information about the polygon will already be embedded in the functional-voxel model.

When constructing a functional-voxel model via flozation, each subsequent FLOZ affects the already formed geometrical configuration of the resulting surface, i.e. the geometrical properties of each of the individual segments are taken into account. Because of this, the problematic situation arising in the compensation of the repulsive and attractive potentials when considering the method of potentials is impossible.

The functional-voxel model of the presented polygon (Fig. 3) is a set of M-images shown in Fig. 6. The presented images $C_1$, $C_2$, $C_3$, $C_4$ characterize the local geometric characteristics $n_1$, $n_2$, $n_3$, $n_4$ of the simulated contour—the deviation of the normal at the function point from the OX, OY, OZ coordinate axes and the distance from the origin of coordinates.

The most informative in the context of this problem are the M-images of the projection of normal for the function of contour of polygon model on the XOY plane (Fig. 7).

These images show a clearer and more pronounced change in the angle of inclination of the normal to the surface at the point of the function. The first image contains the angle of slope of the normal from the OX axis, and the second from the OY axis.

It is possible to determine the proximity of the moving agent to the walls of the polygon by the color of the pixel on each of the two images. The movement of the robot will be carried out along the points with the maximum value of the local function. Such a trajectory will be close to the linear skeleton of the formed figure.

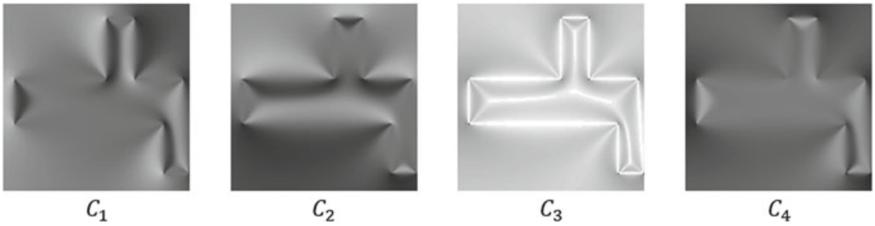


**Fig. 6** Functional-voxel model of polygon

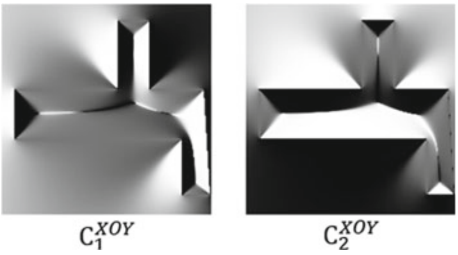
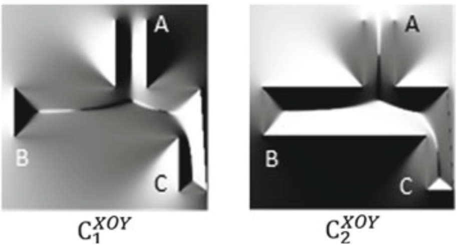


**Fig. 7** M-images of the normal projection of the polygon model on the XOY plane



$C_1^{XOY}$ $C_2^{XOY}$

**Fig. 8** M-images of the normal projection of the polygon model with the exit from the room on the XOY plane

At the joint of two walls, this trajectory envelopes the formed corner along the path on which the agents move, taking into account the inertial forces. The collision of the agents with the walls becomes impossible at any configuration of the walls, since the width of the corridors initially depends on the size of the agents.

Despite all the advantages of the functional-voxel method, the application of functional-voxel models in this formulation of the problem should be accompanied by the method of escaping from a predator. If the simulated polygon has an exit from the room, there is a probability that the agent will not have time to assess the situation when an obstacle appears in the passage, because the agent immediately before the exit increases its speed. Such a situation is presented in Fig. 8.

In the figure, there is no wall in area A, and dead ends in areas B and C. As you can see, triangular zones are formed in areas B and C, indicating a decrease in speed. A different situation is observed in area A: there is a smooth transition to a light tone, which means an increase in the speed of the robot.

## 4 Conclusion

Application of the functional-voxel method to the described motion problem in a multi-agent system will allow to solve the problem situation in the movement of agents on the described polygon, due to the consideration of mutual influence of zones of a functional area of obstacles, which form the polygon scene.

The linear approximation of any algebraic function underlying the functional-voxel model will simplify the calculations. The possibility to calculate in advance and save in graphical images the necessary geometric characteristics of the simulated polygon will accelerate the work of the algorithm of agent movement. The functional-voxel model allows to quickly access the required geometric characteristics of the polygon at any time, in contrast to the potential force methods that require constant recalculation if the agent position changes.

In the future, it is necessary to implement the proposed approach and study its work on the example of the motion of a set of agents in different configurations of the walls of a two-dimensional polygon.

In addition, thanks to the use of functional-voxel models, it is possible to further extend the developed algorithms of agent motion to the three-dimensional case, which will allow modeling the motion of not only ground agents, but also of aircrafts.

## References

1. Shutova KY, Modeling hazardous areas and obstacles for smart city technology based on the "predator avoidance" method of swarming autonomous agents, this compilation
2. Olfati-Saber R (2006) Flocking for multi-agent dynamic systems: algorithms and theory. IEEE Trans Autom Control 51(3):401–420
3. Helbing D, Farkas I, Vicsek T (2000) Simulating dynamical features of escape panic. Nature 407:487–490
4. Reynolds CW (1987) Flocks, herds, and schools: a distributed behavioral model. In: Proceedings of the 14th annual conference on computer graphics and interactive techniques, vol 21, pp 25–34
5. Zachary Y, Hung ML (2020) Consensus, cooperative learning, and flocking for multiagent predator avoidance. Int J Adv Robot Syst 1–19. https://doi.org/10.1177/1729881420960342
6. Tolok AV (2016) Functional voxel method in computer modeling. Fizmatlit, Moscow in Russian
7. Rvachev VL (1982) Theory of R-functions and some applications. Naukova Dumka, Kiev in Russian
8. Guo M, Wang W, Zhao G, Du X (2019) Bézier segmentation of T-spline solids in parametric domain. J Comput-Aided Des Appl 17(3):502–512

# Modeling of Unsafe Areas for Swarm Autonomous Agents

**K. Yu. Shutova**

**Abstract** In the motion tasks of an autonomous vehicle it is necessary to consider the position of individual obstacles and various unsafe zones. We consider a multi-agent system whose motion is carried out towards a common goal according to algorithms of swarm behavior based on Reynolds rules: speed matching, collision avoidance with neighbors and attraction to neighbors. Three approaches to modeling swarming behavior based on articles (Olfati-Saber, Flocking for multi-agent dynamic systems: algorithms and theory. IEEE Trans Autom Control, 51(3), 2016; Olfati-Saber, Murray, Flocking with obstacle avoidance: cooperation with limited communication in mobile networks. In: 42nd IEEE international conference on decision and control, vol 2, 2022–2028) are presented. The peculiarities of this work are that the approaches considered together help to realize such a model of motion, which ensures the avoidance of collision with all available types of obstacles. The method is intended for implementation in those spaces, where there will be many autonomous vehicles. The main problem is that when a dynamic obstacle is encountered, congestion can occur—the agents closest to the obstacle react quickly, while distant agents do so with a lag and create crush. The goals of this paper are to propose a method of indirectly transmitting danger information between swarm members without using communication channels. This means that those robots that do not see the danger can get information about it from other agents by observing their behavior. For this purpose, a method of escaping from a pack from a "predator" based on Q-learning is implemented.

**Keywords** Swarming behavior · Flocking · Multi-agent systems · Machine learning · Q-learning · Reinforcement learning

K. Yu. Shutova (✉)
Institute of Control Sciences of RAS, 65 Profsoyuznaya Street, Moscow 117997, Russia
e-mail: shutova.k.u@yandex.ru

Y. A. Vershinin et al. (eds.), *Technologies for Smart Cities*,
https://doi.org/10.1007/978-3-031-05516-4_3

# 1 Introduction

Pack behavior is the basis for the behavior of a system based on social interaction [3, 4]. A group of communicating entities is called a multiagent system.

Flocking is a form of collective behavior of many interacting agents with a common goal that the whole group has. Agents can be cells, complex organisms. For example, they can be birds, animals, groups of people and crowds. We will call members of our group α-agents.

Groups are examples of self-organizing networks of mobile agents. The first animation was created after Reynolds' introduction of three rules of agent behavior [6] based on swarm behavior:

1. Group cohesion. This means that agents should stay close to their neighbors;
2. Avoiding collisions. This means avoiding collisions with neighbors;
3. Speed matching. It means comparing the velocity with the nearest neighbors.

Three flocking algorithms are presented. Two of them are for free movement and one for constrained movement. The first algorithm embodies Reynolds rules [5]. The second is an algorithm for moving to the target in free three-dimensional space. The third algorithm makes it possible to bypass obstacles when moving to the target.

These algorithms combine consensus, cooperative learning, and crowding control to determine the direction of predator attack. They are also necessary for learning to run away from predators in a coordinated manner.

# 2 General Rules for Methods

The idea of solving problems by a group of simple systems has long been a focus in the field of artificial intelligence. In today's world, mobile robots in groups are designed to replace human-machine systems and single robots. For example, they are needed to perform labor-intensive, large-scale, monotonous, or tedious tasks, as well as tasks that are hazardous to human health or life.

Many concepts are based on graph theory [6]. Consider a graph $G$, representing a pair of $(\nu, \varepsilon)$ which consists of a set of vertices $\nu = \{1, 2, \ldots, n\}$ and edges$\varepsilon \subseteq \{(i, j) : i, j \in \nu, j \neq i\}$. The value $n$ is the number of nodes to be interpreted as agents of the system. The values $|\nu|$ and $|\varepsilon|$ are called the order and size of the graph, respectively. Let $q_i \in \mathbb{R}^m$ denote the position of the $i$-th node in the set $\nu$. The set of neighbors of node $i$ is defined as $N_i = \left\{j \in \nu : a_{ij} \neq 0\right\} = \{j \in \nu : (i, j) \in \varepsilon\}$, where $a_{ij}$ is the matrix of links between robots.

Consider a group of dynamic agents with equation of motion (1):

$$\begin{cases} \dot{q}_i = p_i \\ \dot{p}_i = u_i \end{cases} \tag{1}$$

where $q_i, p_i, u_i \in \mathbb{R}^m$ are the coordinates, velocity, and acceleration of the α-agents, respectively, $m = \overline{2, 3}$, $i \in \nu$.

We take as $r > 0$ the radius of the region under consideration, which is the range of interaction between agents. We consider $N$ agents in two-dimensional space. Let us take the range $d > 0$, which will be the desired distance between them.

Let us introduce a scalar interaction function:

$$\rho_h(z) = \begin{cases} 1, & z \in [0, h); \\ \frac{1}{2}\left(1 + \cos\left(\pi \frac{z-h}{1-h}\right)\right), & z \in [h, 1); \\ 0, & else. \end{cases}$$

Here $h \in (0, 1)$. The function $\rho_h(z)$ is needed to arrange the communication between the robots as a continuous change from maximum to minimum communication.

Let's set the action function $\phi_\alpha(z) = \rho_h\left(\frac{z}{r}\right)\phi(z - d)$, where $\phi(z) = \frac{1}{2}((a + b)\sigma(z + c) + (a - b))$. The value $\sigma(z) = \frac{z}{\sqrt{1+z^2}}$ is a non-uniform sigmoidal function with parameters that satisfy the conditions $0 < a \leq b$, $c = \frac{|a-b|}{\sqrt{4ab}}$. Such parameter values ensure that $\phi(0) = 0$. The value $a$ affects the repulsive force from the robots, and $b$ affects the attraction to them. This function ensures that the distance between the robots is maintained. The function $\phi_\alpha(z)$ is used to study the system for stability.

The swarm movement can be modeled well. The α-agent is a member of a group with dynamics $\ddot{q}_i = u_i$.

**Algorithm 1.** The simplest example of a group consisting only of α-agents in free space. $u_i = u_i^\alpha = f_i^g + f_i^d$ is the sum of the rate-based term ($f_i^g$) and the term based on the gradient ($f_i^d$) due to changes in the bonds between the agents. Values $f_i^g = \sum_{j \in N_i} \phi_\alpha\left(||q_j - -q_i||\right)\vec{n}_{ij}$; $f_i^d = \sum_{j \in N_i} a_{ij}(q)\left(p_j - p_i\right)$; $a_{ij}(q) = \rho_h\left(\frac{||q_j - q_i||}{r}\right)$.

Here $\vec{n}_{ij} = \frac{q_j - q_i}{||q_j - q_i||}$ is a vector along the line connecting i-th and j-th agents. $a_{ij}$ is an element in the relation matrix $A$, which is responsible for the presence or absence of interaction between the objects. This algorithm leads to a swarming motion with a limited set of initial states. When the number of α-agents is large, the algorithm leads to fragmentation of the group into separate subgroups. This method is key for all subsequent algorithms since it forms the structure of the swarm.

**Algorithm 2.** An example of a group when moving to the target in free space. $u_i = u_i^\alpha + u_i^\gamma$, where $u_i^\gamma = -c_1(q_i - q_r) - c_2(p_i - p_r)$—navigation feedback. Here $c_1, c_2 > 0$. The pair $(q_r, p_r)$ is the state of γ-agent, which represents the goal of the group. The goal of α-agent is to track γ-agent.

The collective behavior of a group of robots according to the first algorithm differs from the behavior of robots according to the second algorithm. That is, there is no fragmentation, the robots move cohesively towards the goal.

**Algorithm 3.** An example of a group when moving to a goal in a space with obstacles with the possibility of avoiding multiple obstacles.

u$_i = u_i^\alpha + u_i^\beta + u_i^\gamma$. Here $u_i^\beta$ is the distributed navigational feedback between the α-agent of the robot group and the obstacle: the β-agent. The expression is defined similarly to $u_i^\alpha$. The relationship between the robot in question and the set of obstacles in the field of vision is considered.

Here $u_i^\beta$ is defined similarly to the first algorithm, except for the following: $\phi_\alpha(z)$ is replaced by $\phi_\beta(z) = \rho_h(z)\big(\sigma\big(x - d_\beta\big) - 1\big)$, which ensures constant repulsion. Here $d_\beta$ is the minimum distance from the agent to the object.

## 3 Stability Analysis of Algorithms

The character of swarm movement of a group does not depend on the initial conditions: the location of agents, obstacles, and targets. However, it depends on the initial parameters of the model: the ratio of forces of attraction to each other and to the target, forces of repulsion from obstacles, the initial location of agents, the area of visibility, the reaction of agents to obstacles.

If the initial parameters are chosen correctly, the algorithm achieves the goal and does not depend on the dynamics of the situation, which confirms its stability.

## 4 Avoiding Attacks of Predators

The basic requirement for most robotic systems is the ability to navigate safely in general conditions. That is, the ability to move toward certain targets without encountering obstacles, teammates, or other groups. In some cases, agents need to change their destination to avoid a dangerous area.

The following algorithm is equipped with the ability to temporarily move away from the main target if the group of agents is threatened. For example, some object, which we will call a predator, moves on them [7, 8].

After constructing the flocking algorithms, let us combine the advantages of consensus, grouping, and reinforcement learning for a system with partial observability. This means that only agents that are close to the predator can see the direction of the predator's attack. Stages of execution of the predator avoidance algorithm:

- Consensus: informing the pack of the predator's direction of attack;
- Training with reinforcement: using the predator's direction of attack to inform the pack to move to a safe place (target);
- Moving each agent to the target in the group using coordinated movement.

The algorithm is organized on Q-learning, which is based on the action value function $a_i$. This function gives the expected utility of performing action $a_i$ in each

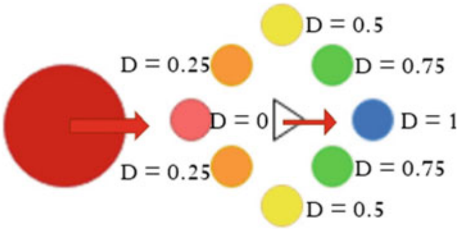


**Fig. 1** Directions of escape from a predator

state $s_i$. For robot i the state is defined as: number of predators $N_p$ in the detection range, direction of predator attack $d_p^k$, $k = 1, ..., n_p$, number of neighboring agents $\left|N_i^a\right|$ in range $r_\alpha$.

The desired action of agents is to move in one of the eight directions to escape. This action is used to select a new reference point in the movement algorithm to escape from predators.

Reward is a reward for the agent for the chosen action. The agent must maintain pack membership when escaping from a predator:

$$r_i = \begin{cases} \left|N_i^\alpha\right| D_r, & \left|N_i^\alpha\right| < 6 \\ 6D_r, & else. \end{cases}$$

The maximum reward an agent can get is 6 if it has all 6 neighbors. D is the scaling factor, which is chosen based on the direction of the predator Fig. 1.

The direction of escape from the predator with coefficient $D_r = 1$ has a vector that coincides with the movement vector of the predator. The worst direction is toward the predator with $D_r = 0$.

For cohesive pursuit of a single goal, the cooperative learning method is applied, for which each agent must first perform independent learning to obtain a separate table $Q_i$:

$$Q_i^{k+1}(s_i, a_i) \leftarrow Q_i^k(s_i, a_i) + \alpha\left[r_i^k + \gamma \max_{a_i^{'} \in A_i} Q_i^k\left(s_i^{'}, a_i^{'}\right) - Q_i^k(s_i, a_i)\right], \tag{2}$$

where α—learning rate, γ—coefficient of value strength.

After performing independent learning (2), each agent's Q-table is updated by interacting with its neighbors:

$$Q_i^{k+1}(s_i, a_i) \leftarrow \omega Q_i^k(s_i, a_i) + (1 - \omega)\frac{\sum_{j=1}^{|N_i|} Q_j^k\left(s_j, a_j\right)}{|N_i|}, \tag{3}$$

where $\omega$ is the weight for determining confidence, such that $0 \leq \omega \leq 1$. When $\omega = 1$ the agent trusts only himself, and when $\omega = 0$ the agent trusts only his neighbors.

Let us carry out the predator detection phase. If the agent is within reach of the predator, it performs a measurement of its relative direction. Set the direction from the measured angle $w_p$ between the agent and the predator:

$$info_i = \begin{cases} 1,\ 0 \leq w_p < 22.5, 337.5 \leq w_p \leq 360; \\ 2,\ 22.5 \leq w_p < 67.5; \\ 3,\ 67.5 \leq w_p < 112.5; \\ 4,\ 112.5 \leq w_p < 157.5; \\ 5,\ 157.5 \leq w_p < 202.5; \\ 6,\ 202.5 \leq w_p < 247.5; \\ 7,\ 247.5 \leq w_p < 292.5; \\ 8,\ 292.5 \leq w_p < 337.5, \end{cases}$$

The information vector of each agent is assigned a confidence factor $weight_{i,d}$. It corresponds to the number of agents by the number of directions and is determined by the agent's measurements from Eq. (4):

$$weigth_{i,d} = \begin{cases} \left(1 - \frac{||q_p - q_i||}{r_p}\right)\left(\frac{w_m - w_p + 45}{45}\right), & ||q_p - q_i|| < r_p, d = info_i \\ \left(1 - \frac{||q_p - q_i||}{r_p}\right)\left(\frac{w_m - w_p + 45}{45}\right), & ||q_p - q_i|| < r_p, d = info_i \\ \left(1 - \frac{||q_p - q_i||}{r_p}\right)\left(\frac{w_m - w_p + 45}{45}\right), & ||q_p - q_i|| < r_p, d = info_i \\ 0, & else, \end{cases}, \quad (4)$$

here $w_m$ is the average angle of the measured direction, $info_i + 1$ is the next direction counterclockwise, $info_i - 1$ is the next direction clockwise. In this formula, we get an inverse relationship between the distance to the predator by the first summand. The second summand divides the distance weight by the two directions of the weight vector. The idea is to assign a weight based on proximity to the predator and proximity to sector centers. Once the $weight_{i,d}$ is found, a consensus can be conducted based on the weighted vote.

Each agent updates its $info_i$ and $weight_{i,d}$ based on its neighbors. Goal: to reconcile the information by reaching consensus on the direction of the predator. The weights for the agents and its neighbors are summed into a weighted direction vector $weight_i$ such that: $\text{weight}_i = \text{weight}_i + \sum_{j=1}^{N_i} \text{weight}_j$. Then $info_i$ is set in the direction with the maximum weight $info_i = max_d(weight_{i.d})$. Thus, the weight and info are updated for all agents. Then the weight for each agent is updated to the maximum weight between it and its neighbors: $weight_i = \max\limits_{weight}\left(weight_{N_i} \cup weight_i\right)$. This allows all agents to converge quickly in the same direction. It is these calculations that allow those agents that are far away from the predator to get information about the state of the system from their neighbors that are closer to the obstacles. The resulting value of $weight_i$ is used to determine the state of $dir_p$ in the reinforcement learning component.

# 5 Simulation Results in Matlab

## *5.1 Algorithms of Swarming Behavior*

Algorithm 1: choose $a = b = 4.6$ to preserve symmetry of attraction and repulsion. When the number of agents is large, fragmentation occurs, robots gather into separate groups. The algorithm is stable when the number of agents is less than 10.

Algorithm 2: representation of Reynolds rules implemented in the first method, and function $f_i = -c_1(q_i - q_r) - c_2(p_i - p_r)$. Here $c_1, c_2$ are constants affecting the force of attraction to the target. In the example there is one static target with coordinates $(250, -25)$. The method preserves group cohesion when the number of agents is large.

Algorithm 3: the final algorithm for the movement of a group of robots. Obstacle avoidance is accompanied by an oscillatory process that occurs under the influence of repulsive forces from neighbors and obstacles and attraction forces to the goal. The robots bypass the obstacles by splitting the group, and then reestablish the swarm behavior as they move toward the goal Fig. 2. Constants $c_1 = c_2 = 0.2$. Collision avoidance has the highest priority with respect to all available obstacles: robots, stationary objects, moving fences. The units are presented in m.

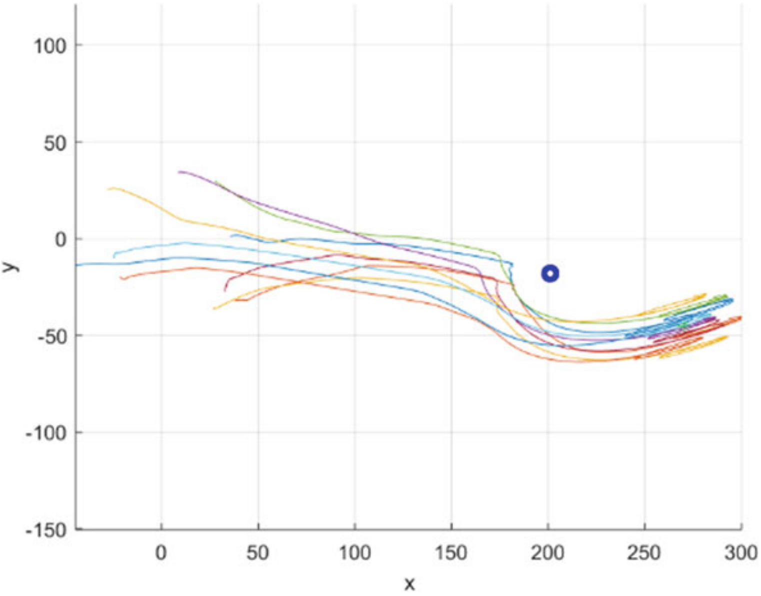


**Fig. 2** Simulation of Algorithm 3 for N = 8 robots

## 5.2 Predator Avoidance Algorithm Based on Swarm Behavior

The predator detection method is used to determine the direction of predator attack. The direction $dir_p$ obtained in Chap. 4 is used to determine the state of the system with respect to which Q-learning will be performed. Each learning episode consists of a certain period, which is sufficient for the agents to rush toward the target. The number of iterations is determined by the time step. In Fig. 4, the agents are represented by red rectangles and the predator is represented as a circle.

The robots' visibility area of the predators is larger than the predator's visibility area. This allows them to escape earlier.

Several variants of the agents' behavior were investigated:

- One target: the agents move only toward it, avoiding the predator only through repulsive forces Fig. 3;
- Two targets: the agents choose one of them, the better one, and move in its direction. The rest of the participants rush after the agents due to the large area of the pack's visibility and the force of attraction to it Fig. 5. When the area of visibility and the force of attraction are reduced, the swarm separates, but the movement to the same target is maintained;
- All targets are involved: in the first iterations, the agents split into separate swarms, choosing different directions, which leads to an oscillatory process. Also, oscillation occurs when all agents have chosen the same goal. This is since the agents coordinate speed and distance among themselves Fig. 4;
- All targets are engaged, and the predator moves along the same trajectory: the agents select the same target. If the predator approaches them, the agents move

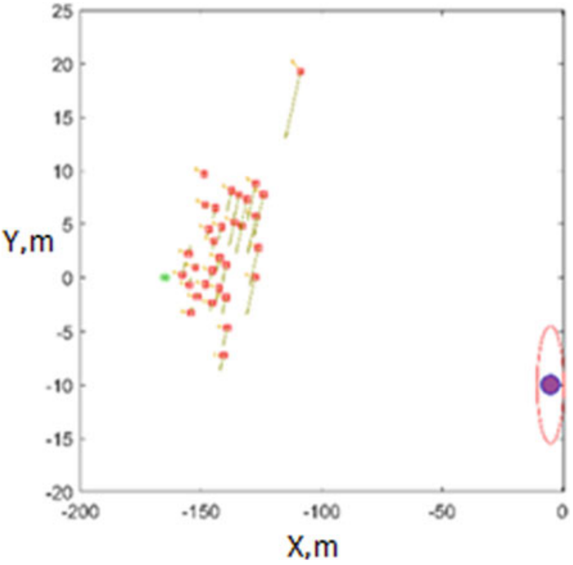


**Fig. 3** Running away from a predator with one target

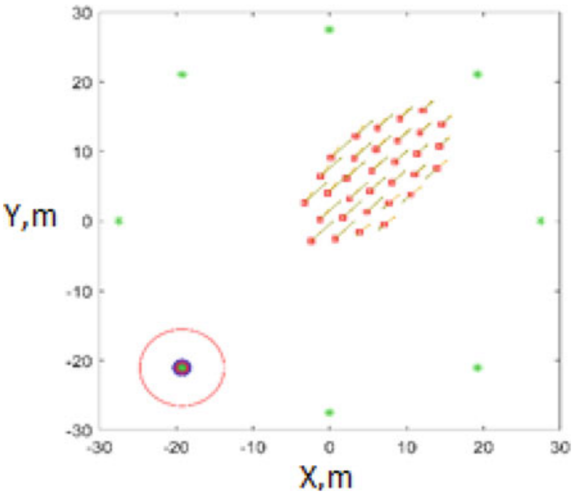


**Fig. 4** Escaping from a predator with eight targets

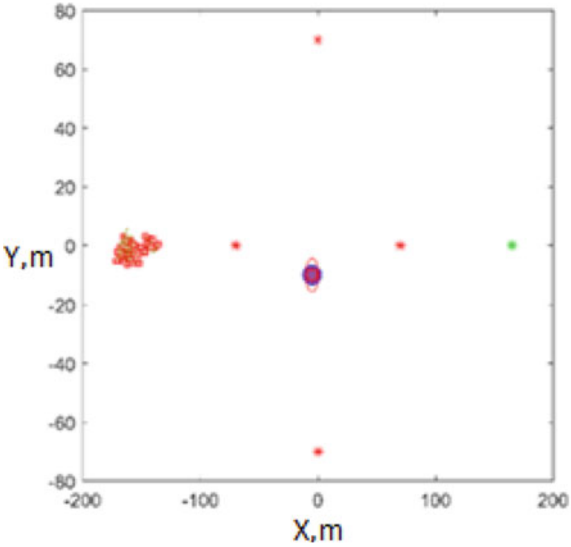


**Fig. 5** Escaping from a predator with two targets

away from it to a safe distance in a coordinated manner. It is worth noting that the agents only choose the direction of the temporary target. They do not have the task of reaching the temporary destination. If the predator is not in sight, the agents return to their main target, which was set before the encounter with the obstacle Fig. 6.

The units of measurement in Figs. 3, 4, 5 and 6 are in m. The horizontal axis is the X-axis, the vertical axis is the Y-axis.

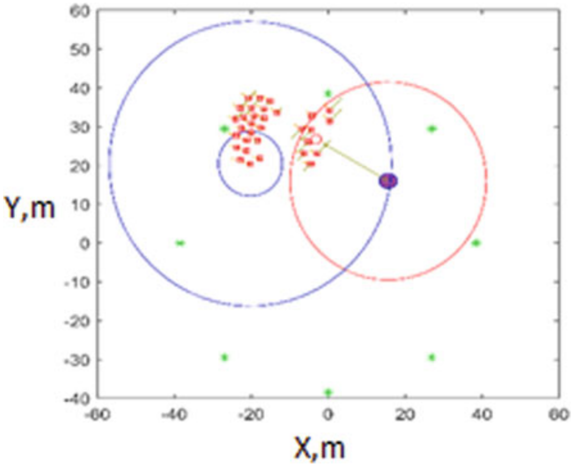


**Fig. 6** Running away from a moving predator along a trajectory with eight targets

The paper considers the application of the predator avoidance method in the case of such a space, which the robot cannot see now in its current position, or at the junction of two walls, behind which there is a corridor. The predator is artificially placed in the passage (blue circle in Fig. 7), or at the junction of walls in order to maneuver the agent at a greater distance: slow down before the passage to analyze the situation or with a large radius of curvature make a turn around the corner to have time to react to a possible interference.

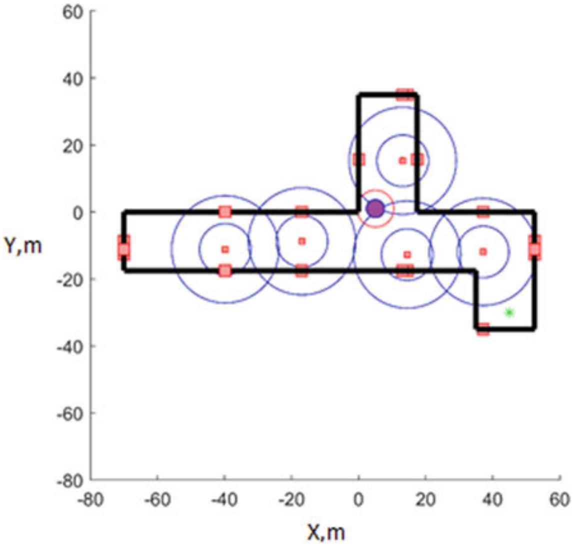


**Fig. 7** Movement of agents in the polygon in the presence of a predator

## 6 Movement of Agents in a Confined Space

Earlier we considered situations when agents move in unrestricted space. Agents can, oriented by the situation, go quite a long distance from the main target if the predator keeps moving in their direction.

When moving in an enclosed space, which is shown in Fig. 7, the agents perform deflection maneuvers away from obstacles. Thus, in addition to the previously considered interactions, the repulsive force from the walls is also considered. The method of potential forces is the summation of all forces of attraction to the target point and repulsion from the obstacles. It is based on realization of mobile robot movement in the field of "information forces". The algorithms described above are based precisely on this method of correlation of arising forces in the system. When an agent moves in the polygon, the minimum distance to the walls is calculated for the method of potential forces. In Fig. 7, this is the distance between the agent (blue dot) and the intersection between the wall and the perpendicular drawn from the agent to the wall (red square). The blue circles are responsible for the areas of visibility between the agents (small circle) and the radius of visibility of the predator (large circle). The red circle represents an overview of the predator. Calculations are performed in the global coordinate system. Thus, the repulsion vector is composed of the repulsive forces of the agent from each visible wall. The predator is represented as a purple circle. Located at the intersection of two walls. Targets are represented as "*" in green.

## 7 Conclusion

This paper considers a model of robot swarming in the presence of a predator, designed to move autonomous agents in confined spaces with insecure areas and obstacles.

Experiments were conducted under different initial conditions and scenes. Experiments have shown that when the number of agents is greater than 10, the swarm maintains its configuration and separates in the presence of an obstacle on the way to the target.

When considering the method of fleeing from a predator, the experiment showed that the agents change their main target to a temporary one and successfully get away from dynamic obstacles. There is still work to be done on the method, as in some experiments agents split into separate groups when fleeing from a predator.

When an agent on a simulated range has an exit from a room, there is a chance that he will not have time to assess the situation if an obstacle appears in the passage. This is because the agent increases speed immediately before exiting. The proposed method of predator avoidance will allow in such situations to put "special points" in dangerous places (for example, at the corner of the corridor bend or at the exit of the room) to reduce speed in advance and bypass this area at a greater distance to assess the space.

## References

1. Olfati-Saber R (2016) Flocking for multi-agent dynamic systems: algorithms and theory. IEEE Trans Autom Control 51(3)
2. Olfati-Saber R, Murray RM, Flocking with obstacle avoidance: cooperation with limited communication in mobile networks. In: 42nd IEEE international conference on decision and control, vol 2, pp 2022–2028
3. Helbing D, Farkas I, Vicsek T (2000) Simulating dynamical features of escape panic. Nature 407:487–490
4. Singh P, Tiwari R, Bhattacharya M (2010) Navigation in multi robot system using cooperative learning: a survey. In: 2016 international conference on computational techniques in information and communication technologies (ICCTICT). IEEE, New Delhi, India, 11–13 Mar 2010, pp 145–150
5. Reynolds CW (1987) Flocks, herds, and schools: a distributed behavioral model, in Comput. Graph. ACM SIGGRAPH'87 conference proceedings, vol 21, pp 25–34
6. Harary F (1969) Graph theory. Addition-Wesley
7. Sunehag P, Lever G, Liu S, et al. Reinforcement learning agents acquire flocking and symbiotic behaviour in simulated ecosystems. In: Artificial life conference proceedings, 2019, pp. 103–110.
8. Zachary Y, Hung ML (2020) Consensus, cooperative learning, and flocking for multiagent predator avoidance. Int J Adv Robot Syst 1–19. https://doi.org/10.1177/1729881420960342

# Guide to Governments for Successful Regional and Open MaaS Implementation

Scott Shepard

**Abstract** Mobility-as-a-Service (MaaS) is a digitalized platform that connects all forms of mobility (including public transportation) within a single app environment. The ability to find (discover), book (reserve) and pay within a single app, along with constructing your entire journey door to door in a seamless manner is a path that many providers and developers are currently undertaking. Successful MaaS environments are characterized by a fragmented, multiple player Mobility Service provider (MSP) ecosystem and strong public transit, as the lynchpin. There are four main MaaS environments, to be explored in this paper: Private MaaS (walled garden), Hybrid MaaS (middle path), Public MaaS (Business to Government B2G), and Open MaaS (Iomob). A global survey of unique MaaS approaches has been undertaken in this paper across the following cities: Denver, Colorado, USA—Private MaaS, Berlin, Germany—Public MaaS, Lisbon, Portugal—Hybrid MaaS, Singapore—Hybrid MaaS, and Skane Region (Malmo), Sweden—Open MaaS. Based on Iomob's MaaS journey experience, there are four main lessons to be learned: app stickiness, data quality and access, Integrated payments, importance of one-click door-to-door capability for travel experience, and modularized journeys. In conclusion, the main recommendations for MaaS success are: open platforms and data, flexible business models, and reducing friction. These are the key enablers for MaaS to help in the overall reduction of congestion and emissions, encouragement of public health, and improvement of the overall well being for cities and regions.

**Keywords** MaaS · Mobility · Digital · Payments · Transport · App · Technology

## 1 Introduction

Mobility as a Service (MaaS) has become a word synonymous with the sharing economy. In a world of massive disruption, based upon environmental, economic,

S. Shepard (✉)
Chief Commercial & Product Officer, Asistobe, Kanalveien 66 5068, Bergen, Vestland, Norway
e-mail: scott@asistobe.com

Y. A. Vershinin et al. (eds.), *Technologies for Smart Cities*,
https://doi.org/10.1007/978-3-031-05516-4_4

social, and cultural shifts, the trend towards sharing (whether it be housing, transportation, or other commodities) is increasing at a rapid pace. In light of this trend towards sharing, mobility is a key component in understanding our consumer habits, needs, and preferences. We are starting to see a shift within the mobility domain away from the desire for single car ownership. The explanations are many, but in a digitalized, shared, and on demand society, the requirement of owning your own automobile has diminished in the priorities of one's lifestyle.

Given this massive shift in how we view mobility and move across (and between) cities, there has been a surge in investment from Silicon Valley and China into the mobility ecosystem, from carsharing, to ridehailing, to micromobility. Shared mobility platforms currently reach customers and provide services though on demand access, ease of use, and a consistent digital offer. MaaS simply extends the on demand digital offer throughout the entire urban mobility ecosystem, allowing consumers to benefit from seamless intermodal journeys. As we move towards a shared economy in a highly urbanized context, the need for public transportation has never been greater, due to the change in ownership preferences and lifestyle changes. Given this recognition of the role that public transportation is to play now and into the future, it is important to understand how new technologies will enable movement in and across cities, as well as provide an enabler and "nudge" for people to [1] no longer choose single car ownership and [2] participate in this more sustainable paradigm shift towards multimodal mobility and active travel.

MaaS is a digitalized platform that connects all forms of mobility (including public transportation) within a single app environment. The ability to find (discover), book (reserve) and pay within a single app, along with constructing your entire journey door to door in a seamless manner is what providers and developers are currently bringing to market. Local governments and transportation policymakers are also recognizing the central role they play in this technical ecosystem, and are looking to position themselves in a manner that integrates all new forms of mobility in an urban context.

## 2 Purpose and Objective

The purpose of this paper is to analyze and provide best practices into how local and regional transit authorities are embracing MaaS on a global scale. The paper will primarily focus on the following areas:

(a) Formal assessment of the opportunities and barriers to implementing a multimodal MaaS journey
(b) Demonstrate lessons learned on a global and national scale, and
(c) Recommendations for fostering an environment for a successful long-term, sustainable multi-modal journey with payment system.

## 3 MaaS Typologies: Considerations for Local and Regional Governments

A key variable to be considered by local and regional governments is the "openness" of MaaS platforms and apps. The different typologies of openness relate to multiple factors, such as who builds the platform, or which government or public authority mandates a platform. As local governments look to explore specific technologies and business models, it is important to understand that there is a wide spectrum of MaaS, as seen in the matrix and sections below:

| Typology | Private MaaS | Hybrid MaaS | Public MaaS | Open MaaS |
|---|---|---|---|---|
| Business Model(s) | B2B, B2C | B2C, B2G | B2G | B2B, B2G |
| User Experience | Closed to | Disjointed | Open | Open |
| Deep Integrations | Only select MSPs | Partially available MSPs | All MSPs in a city/territory | All MSPs in a multi city region, nation |
| Digital Payments | Only curated MSPs (in ecosystem) | Ad hoc, MSPs deeply integrated | All MSPs and PTAs (if API available) | All MSPs and PTAs (if API available |

### Private MaaS

Private MaaS is by definition a platform that is a walled garden, meaning that only services that are within the closed ecosystem of the MaaS provider and are pre-approved are allowed to operate and become available to consumers in the app. For example, in this case only bikes or scooters that were part of the same corporate organizational structure (as the MaaS provider) would be allowed in the platform, blocking any competitor MSPs from being able to integrate and display on the app for consumer comparison. As can be derived from this perspective, this prevents consumers from making informed decisions related to their daily (and personal) mobility choices.

### Hybrid MaaS

Hybrid MaaS takes a balanced approach in the shared ecosystem. Notable examples include Business to Consumer (B2C) MaaS platforms developed commercially for consumers, yet bundled as white label solutions for PTAs and cities. The key difference between a Hybrid MaaS and a Public MaaS is that Hybrid MaaS platforms retain a direct relationship with the consumer, which is sort of a mix between a B2C and B2G business model. These are developed to be deeply integrated with available public transport and MSPs and offered directly to consumers in subscription plans. However, PTAs and cities lose the direct consumer relationship (as in Private MaaS) because mode split (e.g. who rides a taxi vs. a scooter vs. a train) is are bundled together in a fashion that can conflict with the primary policy goals of public authorities, who are tasked with delivering urban public transport services in the first place.

**Public MaaS**

Public MaaS, as the name implies is a carefully orchestrated Business to Government (B2G) platform that places cities and PTAs in the central orchestrating role of delivering shared mobility services to all urban inhabitants. This form of MaaS has been gaining quite a bit of attention in the past year, as there have been multiple deployments across Europe, most notably in Berlin (with the Jelbi app developed by Trafi). While Public MaaS places cities and PTAs in the center of the MaaS ecosystem, what is lacking is a ubiquitous, roaming functionality that enables consumers to benefit from a user experience that is frictionless and will thus increase app "stickiness" (users staying within one app, and not jumping to external apps).

**Open MaaS**

Open Maas is a vision that offers ubiquitous, decentralized, multi city urban mobility. What differentiates Open MaaS from Public MaaS is not a particular implementation of MaaS sponsored by an organization. The key attribute to Open MaaS is an open marketplace that allows for open mobility roaming. Instead, Open MaaS provides a base infrastructure for any player to provide mobility services, or to offer those services directly to consumers via an app. Open MaaS thus accommodates and enables all the previous approaches (private, hybrid, public) with subscription or pay-as-you-go models. The key attribute to Open MaaS is an open marketplace that allows for open mobility roaming.

## 4 MaaS Building Blocks and Considerations: Setting the Stage

Any successful MaaS service has key characteristics which help define its consumer offer, technical composition, and unique sales proposition. Below are a description of the characteristics that can be understood as part of most scenarios and use cases described throughout the paper.

(a) **MSP Integration (Barriers and Levels)**

MaaS Integration of MSPs is a time consuming and complex undertaking. It requires commercial agreements, coverage, and business models that benefit individual MSPs, MaaS providers, and local governments. Therefore, many MaaS providers have opted to either not integrate, or to "lightly integrate" MSPs. However, this provides a disjointed experience for the consumer. As there are multiple forms of integration, each with its API (discovery, registration, and booking/payments), local governments should consider developing or collaborating with MaaS providers that over the widest coverage and deepest integration for consumers of shared mobility services. MSPs are typically apprehensive to integrating on various (deeper) levels because of their [1] loss of direct consumer relationship and [2] commodification of service.

(b) **MaaS Business Models**

The three business models to be considered for MaaS include B2C, B2G, and B2B. B2C is the traditional MaaS business model, with a direct interface between the MaaS provider and consumer (e.g. consumers are the clients of the MaaS provider through bookings in app). B2G is a model for MaaS that seeks to develop MaaS platforms indirectly to consumers, through a government agency or enterprise (procurement, RFP, etc). B2B is a newer business model wherein private companies seek to offer shared mobility services to their consumers through a MaaS platform. There is not a one-size-fits all approach, but the general trend in the MaaS ecosystem is a shift away from B2C to B2B/B2G business models, as MaaS providers are struggling to monetize their offers strictly through transaction revenue.

(c) **MaaS Payment Integrations for Successful Multimodal Journeys**

Regarding payment integrations, as discussed in the previous section, depending on the level of integration possible options include: (i) Light integration, (ii) Light integration with deep linking and (iii) Deep integration.

i. Light Integration

A Light integration normally includes provision of API or GTFS data that could be either static or real time and enabling discovery of the services as well as multimodal journey planning inside the MaaS app with further redirection to the home page of the MSP app where users would be required to register or log in to perform further actions such as discovery, reservation of the vehicle and the payment for the services.

ii. Light Integration with Deep Linking

Further steps towards achieving more advance integration require from the MSP a deep link to a certain functionality from their existing app. This would allow the ability to locate a particular vehicle inside the MaaS provider app and be redirected to that vehicle inside the MSP app allowing reservation and payment for the services, skipping discovery part inside MSP app. It should be noted that deep linking is available to users once they have registered themselves in the MSP app and keep the session active, without being logged out during redirection from the MaaS pprovider app to the MSP provider app.

iii. Deep Integration

Deep integration is the most complex type of integration due to its versatility and the variety of features its offers. The scope of the integration typically includes a fully developed API which enables user to discover, build a multimodal journey, book, and pay for the service inside the MaaS provider app.

Furthermore, such integration allows users to benefit from features like seamless registration where user verified profiles are stored in the MaaS provider app. These profiles allow instant access to new services and mobility providers, and prevents the user from having to sign up to each MSP app individually.

## 5 MaaS Landscape: A Global Survey of Local and Regional Governments

Local and regional governments across the globe are starting to take an active role in implementing specific use cases and components of MaaS. Including functionalities such as stand alone journey planning apps, complex multi modal routing algorithms, deep mobility service provider (MSP) integration, and complete booking and payments, PTAs are exploring how MaaS can be leveraged to boost ticket sales, ridership, fulfill policy goals, and remain competitive in an ever-changing, disruptive mobility landscape.

Below is a survey of four PTAs across the globe, exploring their MaaS implementations and discovering how their experience can be relevant to other agencies looking to deploy similar systems in their service areas.

**Denver, Colorado, USA**

Denver Regional Transit District (RTD) is the regional agency operating public transit services in the Denver-Aurora-Boulder Combined Statistical Area in Colorado. It operates over a 2342 square mile area, serving 3.08 million people [1]. RTD has chosen to take an approach to MaaS that actively integrates ridehail and other on demand services as a key component of its services offering to its passengers. Given RTDs experience with the low usage of its previous white label branded journey planning and ticketing app, it chose to take a different approach. This approach could be considered a Private MaaS.

In this model RTD signed a non-exclusive agreement with Uber (in partnership with Masabi and Moovit) to provide the ability for consumers to book and pay for tickets on RTD bus and light rail lines directly in the Uber app. Uber has launched an aggressive marketing campaign in Denver (and across the United States) to promote its brand as providing the ability to book and pay for tickets within its app, and become a "partner" of public transit (Fig. 1).

However, this has raised many questions and concerns from other PTAs in the ecosystem, and now potentially pose as barriers to MaaS. The concerns center around privacy, control, and the role of PTAs. If a privatized MaaS app becomes the de facto white label for public transport, how does that affect the PTA in delivering services to its passengers? How is personal data protected in such an environment? And finally, how can PTAs retain the direct relationship with its passengers and not lose control, given its legislative mandate.

Many questions remain unanswered, but the Private MaaS model seen in Denver appears to be taking hold primarily in North America, through a similar non-exclusive arrangement with PTAs and municipalities.

**Fig. 1** Downtown Denver

**MaaS Characteristics**:

- Typology—Private
- Business Model—B2C
- Name of App—Uber (with RTD mobile transit ticketing)
- Year Implemented—2019.

### Berlin, Germany

Berlin has chosen to take a diametrically opposite approach to Denver when it comes to the implementation of MaaS in its urban region. The Berliner Verkehrsbetriebe (BVG) is the main public transport company of Berlin [2]. It manages the city's U-Bahn underground railway, tram, bus, and ferry networks. BVG recently undertook a complex implementation of B2G Public MaaS, with the unveiling in mid 2019 of the Jelbi app, developed by Trafi (Lithuania). The Jelbi MaaS app can be consider a Public MaaS, as it deeply integrates MSPs into the platform, along with BVG bus, rail, and tram for complete journey planning, discovery, registration, booking, and payments within the app (Fig. 2).

While the Jelbi app has demonstrated increased passenger usage and interest, there is always the question of complete availability. That is, it's a challenge for PTAs to recruit and deeply integrate all available MSPs in an urban mobility ecosystem, and that has been proven to be the case here. So even in a Public MaaS, which puts the PTAs at the center of the mobility orchestration, who is to form the commercial agreements with each MSP? The PTA? The MaaS provider? A third-party integrator?

**Fig. 2** Berlin—River Spree and Alexander Platz

While Public MaaS shows much promise in placing PTAs in the proverbial "driver's seat", it is yet to be seen if this a commercially sustainable business model for MaaS providers and PTAs, and how will platforms such as these scale across geographies and modes of mobility in the years to come.

**Berlin MaaS Characteristics**:

- Typology—Public
- Business Model—B2G2C
- Name of App—Jelbi
- Year Implemented—2019.

### Lisbon, Portugal

Lisbon with the advent of Web Summit, has become the startup hub of Europe. With this increased attention, the Camara Municipal de Lisboa - Lisbon City Hall (CML) has chosen to position the city as a test lab for all forms of mobility, ranging from carshare to micromobility. The city recently won the 2020 European Green Capital Award, and will be showcasing the 2020 ITS European Congress [3].

In light of this, CML is in the process of developing a "Mobility Data Catalog" which identifies the taxonomy, rules, infrastructure, and modes of mobility to operate within the city. In conjunction with this, CML has developed an open dialog with all MSPs that seek to test and operate in the city. As such, MaaS has become an increasingly important topic in the city, and CML has chosen to allow MaaS providers to plug into the ecosystem and develop/test their own apps, by connecting to open APIs provided by local PTAs such as Carris (bus) Lisbon Metro, and Fertagus (commuter rail).

In addition, with the introduction of the regional Viva Viagem smart card, users have seamless access to all major transport operators in the region.Therefore, through a mix of app-based and fare payment technologies, public transport ridership has increased and users are experimenting with all available mobility options that the city has to offer (Fig. 3).

This open approach allows the city to closely analyze business models, technological solutions, and platforms which could later be deployed on a regional basis. While this may ultimately morph into a Public MaaS to be deployed by the city (in partnership with PTAs), for now it is a pragmatic approach which seeks to first develop the framework, analyze solutions, and roll out a larger platform after enough data and performance is analyzed and assessed.

**Lisbon MaaS Characteristics**

- Typology—Hybrid
- Business Model—B2B2C
- Name of App—Multiple 3rd party providers
- Year Implemented—2019.

### Singapore

McKinsey's Urban Transportation Report ranks Singapore's transport system as the world's best overall based on five criteria—availability, affordability, efficiency, convenience, and sustainability [4].

**Fig. 3** Lisbon—Castelo Sao Jorge

Singapore also has one of the most cost-efficient public transport networks in the world, according to a study by London consulting firm Credo [4]. In light of an excellent backbone public transport network, Singapore was also named the "World's Smartest City" in 2019 by Juniper Research. These are no small achievements, and speak to the decades of careful urban planning, design, and development which have resulted in such a sustainable role model.

Singapore has chosen to leverage its excellent built environment and physical infrastructure to enable an open marketplace for multiple journey planning and MaaS apps, which have recently launched over the past two years. The world's smartest city enabled a city-wide Hybrid MaaS B2C app, Zipster. Zipster is the product of Singaporean start-up company MobilityX which was seed-funded by the country's leading rail operator SMRT [5] (Fig. 4).

Zipster provides a single point of access to all modes of transport across the main island from the train, metro and bus companies to privately-owned services such as taxis or car-, bike-. scooter- and even bus-sharing. Privately owned transport services such as the taxis, are deep-linked through the app so those wanting access to first/last mile options are diverted to the individual website to summon a cab or reserve a bicycle [5].

While transit-rich environments (such as Singapore) hold promise for MaaS platforms and applications, are they scalable? Certainly not every city is a Singapore, and each region has its unique circumstances in which to consider. As Singapore is a tightly integrated city-state with close cooperation between its public agencies,

**Fig. 4** Singapore Harbor

MaaS on the surface appears to be easy. But it's yet to be seen if an open, B2C Hybrid MaaS model such as this can solve urban mobility challenges across the globe.

**Singapore MaaS Characteristics**

- Typology—Hybrid
- Business Model—B2C
- Name of App—Zipster
- Year Implemented—2017/18.

### Region Skane (Malmo), Sweden

The MaaS in Skåne project proposed by Innovation Skane, Skanetrafiken, and Iomob was selected as a grant award recipient by the Swedish Energy Agency (SEA) to deliver a regional, open multimodal shared mobility platform for municipalities including Malmo, Lund, and Helsingborg. In 2019, Iomob was one of three winners of the 2019 Swedish Sustainable Mobility Challenge (SMC), which led to this opportunity with the Skåne region.

Iomob's project aims to develop a scalable MaaS solution for the Skåne Region in Southern Sweden. This will be accomplished by increased accessibility, reduced environmental impact and congestion, through seamlessly connecting buses, trains, taxi, bicycles, scooters, car sharing etc. in one platform (Fig. 5).

Iomob has developed a technology platform for mobility that enables seamless, multimodal travel over an open network with a large number of mobility service providers (MSPs). This open architecture empowers public transport authorities (PTAs) and 3rd party MaaS providers to deliver B2C consumer-facing mobility apps to their users.

**Fig. 5** Malmo Waterfront

The project partners and suppliers will use Iomob's MaaS platform with a unique ability to create open mobility markets which can interoperate across regions, and create a reliable and flexible service for consumers. The solution will create changed travel behaviors that will reduce travel by own car, promote $CO_2$-efficient modes of transport, be commercially viable, facilitate easy introduction of new mobility services, support Agenda 2030, reward travel on foot and bicycle and include public transport, rental bicycles, carsharing etc.

Iomob's core solution will not only be offered to Skanetrafiken for use within their own mobility app, but also the transit authority intends to allow Iomob to offer this same solution that includes Iomob's intermodal algorithms, integrations with private mobility services and Skanetrafiken's own transit services to private companies who wish to leverage Iomob's award-winning technology to offer their own B2C MaaS experience.

**Region Skane MaaS Characteristics**:

- Typology—Open
- Business Model—B2G2C
- Name of App—MaaS in Skane
- Year Implemented—2021.

## 6 MaaS Lessons Learned

There are unique lessons learned with each MaaS deployment and implementation across the globe. However, what we are seeing is a trend towards several challenges which can be seen across the entire ecosystem. Below is an overview of the lessons learned and benchmark experience which can be used in further identifying technical and functional requirements necessary to deliver MaaS platforms to consumers in a manner that is equitable, open, and accessible.

(a) **App Stickiness**

This is a term that is related to user growth such as app vitality, for instance, which is really just an indicator that a mobility platform is acquiring customers at a faster rate than it is losing them. High adoption means a shared mobility platform is "stickier" captures consumers and users in an unified experience. Customer Experience is the key to the adoption (and consumer retention) of new shared mobility services. By providing a seamless platform that integrates all primary functionalities, it has been proven that customers will more likely interact with and continually utilize (and return to) such platforms.

(b) **Data Quality and Access**

Many new MaaS digital platforms have been developed that "fetch" shared mobility data (typically in MDS and GTFS formats) and bundle into intuitive dashboards which municipalities and PTAs can utilize to monitor and enforce MSPs within MaaS environments. As such, MaaS Data quality and accuracy is imperative for such new opportunities to succeed in the long term. In order to enrich the environment that governments (and large corporates) require with regards to understanding the mobility patterns on a city scale, anonymized historical and real time data can empower regulators (and data scientists) with the insights required to understand the complete mobility picture.

(c) **Digital Payments**

Based upon user research conducted by Six Fingers for the Renfe as a Service RailMaaS pilot app in Spain, it has been identified through user research surveys that many consumers appreciate the ability to make a one-click purchase of their entire

door-to-door journey, simplifying the payment experience. The ability to enable a one-click door-to-door capability for travel experience is key to user adoption. Depending on the type of journey, consumers can opt for pay as you go, or subscription models. Fixed subscription models, while popular with MaaS providers, can be problematic for local and regional governments in trying to incentivize certain modes of mobility over others.

This is because a fixed subscription (e.g. $300/month - all modes) can result in consumers overutilizing specific modes, and thus reducing demand for public transport or higher capacity options.

(d) **Modularized Journeys**

MaaS consumers like to discover, book and pay for their mobility journey (especially intercity), in advance. Also, the ability to store their ticket and add connections at a later time is a critical feature.

MaaS solutions offering this combination of advance and just-in-time booking will increase consumer adoption of MaaS platforms.

## 7 MaaS Recommendations for Success

In order for MaaS to be a success on a local and regional scale, there are many domains to consider which will allow public transport agencies and operators to quickly scale and demonstrate success to consumers, passengers, and mobility stakeholders. These recommendations encompass technical and policy changes that regulators and agencies can look to consider for implementation.

(a) **Open Platforms and Data**

Constructing MaaS platforms that are consumer centric and consider the central, orchestration role of PTAs in delivering public transit are one of the most effective ways to demonstrate success. How this can be accomplished is by developing Public MaaS platforms that take into account openness. Openness regarding MSP access and data sharing is a key component. Data which is not real time and anonymized (shared by MSPs and other providers) in an open platform will be critical for cities and PTAs to make informed decisions on service planning and future capital investment in public infrastructure. As described earlier, by constructing MaaS platforms that actively recruit and integrate the highest number of MSPs, consumers will appreciate the ability to choose and compare their available options. PTAs and cities that encourage policies or issue public tenders to require open data from MSPs and full access to all providers boosts the relevance of MaaS usage in the mobility ecosystem.

(b) **Flexible Business Models**

Business models that encourage flexibility for both PTAs and MSPs guarantee the highest chance of commercially sustainable success. By structuring contracts and agreements between MaaS platforms, MSPs, and PTAs to collectively achieve long term mobility policy goals that cities seek to implement are key to success. Specifically, traditional B2C business models that solely focus on user acquisition prove to be a challenge for maaS platforms to effectively monetize. More flexible B2B / B2G SaaS business models that focus on long term subscription fees passed from MaaS platforms to PTAs/cities hold more promise. This means developing open and decentralized MaaS platforms to allow PTAs and MSPs to share in increased revenues due to increased usage and ridership, which promotes success for all stakeholders.

(c) **Reduce Friction**

Reducing friction is one of the most important recommendations for success in MaaS. By removing the frictions that consumers experience in using apps to navigate cities and switch modes of mobility, an increased adoption in MaaS will occur. These frictions include:

1. Juggling multiple apps
2. Fixed (non real time) journey planning and routing
3. Lack of notifications, difficult in app payment experiences, and light MSP integration.

Consumers will more quickly adopt MaaS platforms and perhaps experience a behavioral shift away from their personal vehicles in a reduction of the pain points as part of their daily journeys. PTAs and cities can reduce this friction in the MaaS ecosystem by requiring MSPs to openly share their data for location and payments, as part of the permit to operate in their cities. This approach ensures a level playing field, so PTAs can provide a landscape that encourages the best MaaS platforms to develop fully integrated apps (Hybrid or Public MaaS) for their consumers and passengers.

## 8 Regional MaaS Implementation Best Practices

Assessing the opportunities and barriers for the implementation of single application to provide seamless access to multimodal travel, with integrated payment and ticketing in regions is of utmost importance to ensure a successful and sustainable consumer offer. Below is an overview of best practices to consider for such an offer.

(a) **Mobility Service Providers and Public Transit Operators**

Most PTOs offer open API or GTFS data for each mode of transportation that facilitates basic integration enabling discovery and routing functionality. However, deeper integrations that enable seamless ticketing and payment solution typically require infrastructure upgrades and installation of modern hardware.

This includes QR codes and NFS readers on the access gates to the metro and train stations and buses that tend to be cost prohibitive.

Methods to overcome these technical integration barriers include the following:

- Provide a light integration at initial stage which will facilitate user access to ticketing and payments by way of an interactive step-by-step guidance inside an app during their journey.
- Integration with a simple iteration of mobile ticketing can be achieved by generating a copy of the actual paper/plastic ticket that can be displayed on the phone screen to the driver or gate security.

(b) **Key Recommendations for Regions to Offer Successful Multimodal Journeys**

In order to foster a MaaS-friendly environment in regions, the following recommendations are suggested:

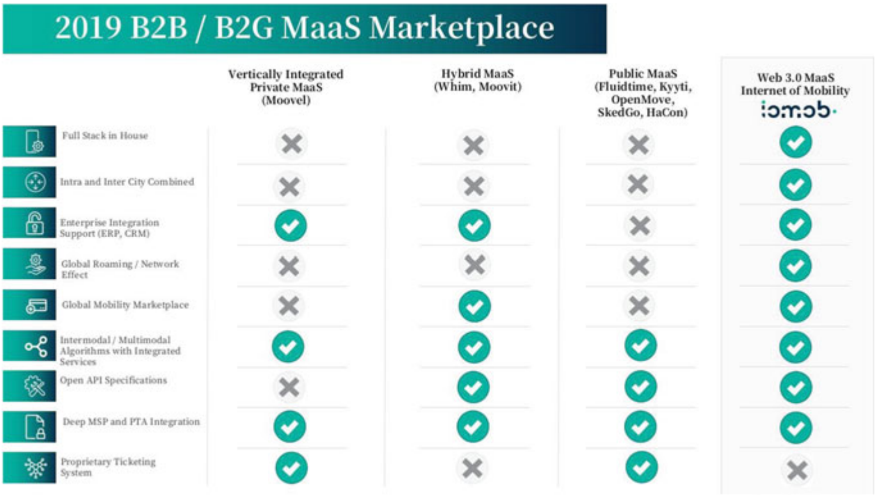
- Local policies and regulations should promote provision of mobility services by private companies like car sharing, bike sharing and micro mobility.
- Further development and improvement of the existing bike lane network will require modifications to provide uninterrupted lanes across the city, where possible
- Create an environment that's friendlier for more MSPs. This could include offering light integration MaaS platforms, encourage deeper integrations, and evaluate options between modes in a light-integration environment.
- Shared mobility and MaaS pilot programs should be encouraged to gather anonymized (and aggregated) data on urban shared mobility patterns, to assist in policy formulation and future regulations.
- Mobile parking technology providers could be integrated (and treated) as additional MSPs to further build the multi modal journey, and allow for a door to door experience for consumers
- Encourage localized, neighborhood-based MSPs to broaden the service offer, to build an ecosystem of providers for a successful multimodal MaaS platform
- Emphasize MaaS integration at intermodal transit stations and mobility hubs, to fill in the service gaps for public transport operators
- The ideal role of the PTA in establishing a successful MaaS is to serve as a technology orchestrator. This can result in outcomes-based solutions creating an open mobility marketplace that [1] deeply integrates MSPs and [2] allows [3] party MaaS providers to offer B2C apps for consumers

- Bundled payments / invoices is a potential option to explore, and can be leveraged to provide a superior customer experience related to thenmultiple MSPs that are booked in a typical multimodal MaaS journey
- Shared mobility data should be requested by MSPs to understand distribution, demand, and impact on urban street networks. However, careful consideration should be given to personal privacy and data should be anonymized and aggregated so as to not associate with PII (personally identifiable information, e.g. name, address, CC, etc.)

## 9 Conclusion

PTAs and cities hold the key to ensuring the success of MaaS in the future. By understanding what works and what doesn't, PTAs can leverage best practices deployed across the globe, structure sustainable business models, create open and competitive mobility marketplaces, and encourage MaaS platforms to develop apps and solutions that boost public transit ridership and promote mode shift away from personal vehicles.

## Appendix 1: MaaS Competitive Analysis



2019 B2B / B2G MaaS Marketplace

| | Vertically Integrated Private MaaS (Moovel) | Hybrid MaaS (Whim, Moovit) | Public MaaS (Fluidtime, Kyyti, OpenMove, SkedGo, HaCon) | Web 3.0 MaaS Internet of Mobility iomob |
|---|---|---|---|---|
| Full Stack in House | ✗ | ✗ | ✗ | ✓ |
| Intra and Inter City Combined | ✗ | ✗ | ✗ | ✓ |
| Enterprise Integration Support (ERP, CRM) | ✓ | ✓ | ✗ | ✓ |
| Global Roaming / Network Effect | ✗ | ✗ | ✗ | ✓ |
| Global Mobility Marketplace | ✗ | ✓ | ✗ | ✓ |
| Intermodal / Multimodal Algorithms with Integrated Services | ✓ | ✓ | ✓ | ✓ |
| Open API Specifications | ✗ | ✓ | ✓ | ✓ |
| Deep MSP and PTA Integration | ✓ | ✓ | ✓ | ✓ |
| Proprietary Ticketing System | ✓ | ✗ | ✓ | ✗ |

## References

1. Denver RTD Services. https://www.rtd-denver.com/services. Retrieved 19 Dec 2019
2. About Berlin BVG. https://www.bvg.de/en/Willkommen. Retrieved 19 Dec 2019
3. Lisbon European Green Capital Award Winner. https://ec.europa.eu/environment/europeangreencapital/lisbon-is-the-2020-european-green-capital-award-winner/. Retrieved 17 Dec 2019
4. Transport in Singapore. https://en.wikipedia.org/wiki/Transport_in_Singapore Retrieved 18 Dec 2019
5. Singapore Aims to Set MaaS Benchmark. https://www.itsinternational.com/sections/transmart/features/singapore-aims-to-set-maas-benchmark/. Retrieved Retrieved 17 Dec 2019

# Improving City Traffic Using Goal-Oriented Automotive Suspension Tuning with the Sub-Gramian Method

**Dmitry E. Kataev, Evgeniy Y. Kutyakov, and Yuri Vershinin**

## 1 Introduction

The simplest version of the sub-Gramian method is based on the spectral decomposition of a square $H_2$ norm of the transfer function. This case applies when a system can be described as an LTI dynamic one and therefore has a corresponding algebraic Lyapunov equation. In order to facilitate specific studies, e.g. small-signal stability analysis, even large-scale systems like power grids can be considered as LTI systems [1]. The finite sub-Gramian method [2] uses the spectral decomposition for the differential Lyapunov equation solution instead of algebraic one. Potentially it may allow its application to time-variant and certain types of non-linear systems.

A vehicle suspension appears to fit as an application for the sub-Gramian method. Its model has relatively low order and can be modified with time-variant, non-linear and controlled elements if needed. The system has multiple feedback loops. Modern literature contains detailed description of its dynamics as well. Vehicle suspension parameters tuning problem is still not fully formalized and its solution mostly exist as a set of empirically derived rules, even despite this problem exists for a long time and has a certain economical significance [3, 4]. This leads to another task of this study—to investigate the suitability of the sub-Gramian method as a base for the formal automotive suspension tuning problem statement.

One notable feature of sub-Gramian application to vehicle suspension is its ability to yield results that have physical meaning of energy or work. This includes work undone by the system itself, therefore providing some kind of formal approach to determine effort in addition to actual physical work. This allows potential goal-oriented ad hoc control targeting optimal passenger comfort, manoeuvre safety or energy efficiency while considering energy cost of such adjustments.

D. E. Kataev (✉) · E. Y. Kutyakov
Institute of Control Sciences of RAS, 65 Profsoyuznaya Street, Moscow 117997, Russia
e-mail: dekataev@gmail.com; dekataev@ipu.ru

Y. Vershinin
Coventry University, Priory Street, EC-Building, Coventry CVI 5FB, UK

Y. A. Vershinin et al. (eds.), *Technologies for Smart Cities*,
https://doi.org/10.1007/978-3-031-05516-4_5

## 2 Problem Statement

Let us consider a vehicle suspension model with four wheels, springs, dampers an tires, absolutely rigid body and no sway bars. The following forces affect front right (FR, fr) wheel and suspension:

$$\begin{aligned} F_{fr} &= K_{fr}(-L_f\theta_d + L_{fr}\theta_l + z - z_f rU) \\ &\quad + C_{fr}((-L_f\dot{\theta}_d + L_{fr}\dot{\theta}_l + \dot{z}) - \dot{z}_{frU}) \\ \ddot{z}_{frU} &= \frac{1}{M_{frU}}(F_{fr} - K_{frU}\dot{z}_{frU} - C_{frU}z_{frU}) \end{aligned} \tag{1}$$

where $F_{fr}$ is a force of FR suspension, $K_{fr}$—FR suspension stiffness, $L_f$—a distance from the center of gravity (COG) to the front axle, $\theta_d$—vehicle body pitch, $L_{fr}$—lateral distance from COG to FR suspension, $\theta_l$—vehicle body roll, $z$—COG vertical position, $z_{frU}$—FR unsprung mass position (m), $C_{fr}$—FR suspension damping, $M_{frU}$—FR unsprung mass, $K_{frU}$—FR tire stiffness (N/m), $C_{frU}$—FR tire damping. Note that vehicle body and unsprung masses positions are in different coordinate systems. Similar equations define other (FL, RR, RL) suspension and tire forces.

The vehicle body dynamics is the following:

$$\begin{aligned} M_b\ddot{z} &= F_{fr} + F_{fl} + Frr + Frl \\ I_y\ddot{\theta}_d &= L_f(-F_{fr} - F_{fl}) + L_r(F_{rr} + F_{rl}) \\ I_x\ddot{\theta}_l &= L_{fr}F_{fr} + L_{rr}F_{rr} - L_{fl}F_{fl} - L_{rl}F_{fl}, \end{aligned} \tag{2}$$

where $M_b$ is a body mass, $F_{fl}$—FL suspension force, $F_{rr}$—RR suspension force, $F_{rl}$—RL suspension force, $I_y$—body moment of inertia about y-axis, $L_r$—distance from COG to rear axle, $I_x$—body moment of inertia about x-axis (lateral), $L_{rr}$—lateral distance from COG to RR suspension, $L_{fl}$—lateral distance from COG to FL suspension, $L_{fl}$—lateral distance from COG to FL suspension. Figure 1 clarifies arms $L$ and angles $\theta$ in this model.

Such physical system being described by (1) for each wheel and (2) can be modeled as a space-state LTI system

$$\dot{x} = Ax + Bu, \quad y = Cx, \tag{3}$$

where state vector $x \in R^n$ represents body and suspension relative positions and velocities, as well as body pitch and roll, $u \in R^1$ is an external force caused by road bumps, vehicle acceleration and steering, $y \in R^1$ is an output signal like body position or roll. $A_{[n\times n]}$, $B_{[n\times 1]}$, $C_{[1\times n]}$ are real matrices defining the system dynamics. While further equations would be given for a SISO system, the method description and case study suggest implicit SIMO and MIMO systems analysis by presenting their transfer functions as sets of SISO ones with different matrices $C$.

The solution of a road vehicle suspension tuning problem is finding an optimal set of adjustable chassis parameters resulting in both minimal transient time and peak

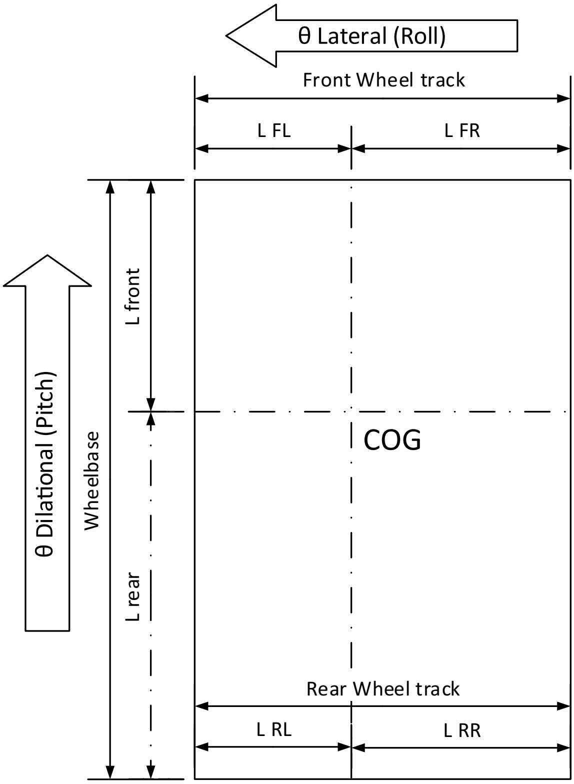


**Fig. 1** Vehicle body dimensions and angles

body accelerations. Both are important for passenger comfort and vehicle steering but require opposite technical solutions, e.g. soft shock absorbers are good for keeping lower body acceleration values but lead to longer transients.

Considering a suspension system as an LTI system with a body position as its output allows using its transfer function to build comfort and handling functionals. Such functionals can be sensitive both for high peak values and long transient durations. A transfer function square $H_2$-norm represent total output energy after a $\delta$-function input. The controllability Gramian can express it in the following way [1]:

$$||W||_2^2 = tr(C^T PC) \tag{4}$$

A controllability Gramian is defined as a solution of the following Lyapunov equation:

$$0 = AP + PA^T + BB^T \tag{5}$$

$$P = \int_0^\infty e^{A\tau} BB^T e^{A^T \tau} d\tau \tag{6}$$

A controllability Gramian on finite time interval is a solution of the following differential Lyapunov equation:

$$\frac{dP}{dt} = AP + PA^T + BB^T, \quad P(0) = 0. \tag{7}$$

While transfer function $||H_2||$-norm is defined only for an infinite time interval, by Parseval's theorem it coincide with its time-domain counterpart—impulse response function $||H_2||$-norm. This allows building finite time version of (4) and potentially adapt the proposed method for linear time-variant systems.

The sub-Gramian method is based on the decomposition of Lyapunov equation solution as a sum of parts corresponding to system eigenvalues and their pairs, therefore to certain suspension components and component sets. Elements of such decomposition are called sub-Gramians. This allows to decompose energy-based functionals proposed above and possibly make them more descriptive.

## 3 Sub-Gramians

One way to define the finite controllability sub-Gramian corresponding to a particular eigenvalue involves the solution of the Sylvester differential equation [5]

$$\begin{aligned} &\frac{d}{dt} P_k^c(t) = s_k I P_k^c(t) + P_k^c(t) A^T + A_{(k)} BB^T, \\ &P_k^c(0) = 0_n. \end{aligned} \tag{8}$$

Its general solution is

$$P_k^c(t) = \int\limits_0^t A_{(k)} BB^T e^{s_k \tau} e^{A^T \tau} \, d\tau, \tag{9}$$

$$A_{(k)} = Res\,(Is - A)^{-1}|_{s=s_k}. \tag{10}$$

The infinite sub-Gramian $P_k$ trace for a diagonalized SISO LTI system:

$$||W||_2^2 = \sum_k^n tr P_k = \sum_k^n tr(C^T P_k^c C), \tag{11}$$

$$tr P_k = tr(C^T P_k^c C) = \sum_l^n p_{k,l}, \tag{12}$$

$$p_{k,l} = -\sum_l^n \frac{1}{s_k + s_l} b_k b_l c_k c_l, \tag{13}$$

where $s_k$ is $k$th eigenvalue of $A$, $b_k$ is $k$th element of vector $B$ and $c_k$ is $k$th element of vector $C$. The finite sub-Gramian trace for a diagonalized SISO LTI system:

$$tr(CP^c(t_0,t)C^T) = \sum_{k=1}^{n}\sum_{\lambda=1}^{n} b_k b_\lambda c_k c_\lambda \frac{1}{s_k + s_\lambda}(e^{(s_k+s_\lambda)t} - e^{(s_k+s_\lambda)t_0}). \tag{14}$$

# 4 Case Study

## 4.1 Model Description and Methodology

Table 1 shows the default values of all model parameters. COG position is typical for compact front wheel drive cars with only a driver onboard.

The input is front right unsprung mass instantly moving 1 cm down. Such input signal represents an idealized case of front right wheel falling into a 1cm deep road dent with instant tire deformation. Outputs are body COG position $z$, pitch $\theta_d$ and roll $\theta_l$ angles. Studying these outputs allows analyzing basic suspension-dependent comfort and handling properties of a vehicle.

The following experiments involve linearizing the model with input and outputs defined above while continuously changing rear suspension damping in the interval from 0.5 to 2 of the default value (1000–4000 N/(m/s)). We repeat this procedure for three different front damping values: 1, 1.2 and 1.4 of the default value (2500, 4200 and 5880 N/(m/s)). Table 2 shows dominant connections between system modes and state variables acquired using matrix $A$ eigenvector analysis [6].

## 4.2 Numerical Results

At first we study square $H_2$-norms of the acquired transfer functions in order to clarify the energy-based approach. Figure 2 shows corresponding dependencies and allows to conclude that too low and too high damping values result in undesirably high energy estimations due to long-duration or highly accelerated transients respectively. It is notable, that optimal by the means of energy estimation rear damping values get proportionally higher as similar values for front dampers rise. The situation differ for transfer functions with body roll as output. As the input perturbation is localized in front, firmer rear dampers provide better roll resistance almost without local minimums, and vice versa for the front dampers.

Every mode M1–M10 has a significant impact on the resulting energy estimation as seen on Fig. 3 for a transfer function with body bounce as output. At the same time the ratios of their impact are very variable. It is expectable since COG doesn't

**Table 1** Default model parameters

| Parameter | Value | Unit |
|---|---|---|
| $L_f$ | 0.84 | m |
| $L_{fr}, L_{rr}$ | 0.88 | m |
| $L_r$ | 1.26 | m |
| $L_{fl}, L_{rl}$ | 0.72 | m |
| $K_{fr}, K_{fl}$ | 28,000 | N/m |
| $C_{fr}, C_{fl}$ | 2500 | N/(m/s) |
| $K_{rr}, K_{rl}$ | 21,000 | N/m |
| $C_{rr}, C_{rl}$ | 2000 | N/(m/s) |
| $M_{frU}, M_{flU}$ | 35 | kg |
| $K_{frU}, K_{flU}$ | $10K_{fr}$ | N/m |
| $C_{frU}, C_{flU}$ | $10C_{fr}$ | N/(m/s) |
| $M_{rrU}, M_{rlU}$ | 25 | kg |
| $K_{rrU}, K_{rlU}$ | $10K_{rr}$ | N/m |
| $C_{rrU}, C_{rlU}$ | $10C_{rr}$ | N/(m/s) |
| $M_b$ | 1200 | kg |
| $I_y$ | 2100 | kg m$^2$ |
| $I_x$ | 900 | kg m$^2$ |

**Table 2** System modes

| Mode | Dominant states | Mode | Dominant states |
|---|---|---|---|
| M1, M2 | $\theta_d, \dot{\theta}_d$ | M6 | $z_{flU}, z_{frU}$ |
| M3 | $z_{rlU}, z_{rrU}$ | M7, M8 | $\theta_l, \dot{\theta}_l, z, \dot{z}$ |
| M4 | $z_{rlU}, z_{rrU}$ | M9, M10 | $\theta_l, \dot{\theta}_l, z, \dot{z}$ |
| M5 | $z_{flU}, z_{frU}$ | M11–M14 | $\dot{z_{rlU}}, \dot{z_{rrU}}, \dot{z_{flU}}, \dot{z_{frU}}$ |

coincide with the geometrical center of the chassis, so any process in the suspension leads to COG movement.

Figure 4 shows dominant sub-Gramian traces for a transfer function with body roll as output. They correspond to modes M3 and M5 connected to rear and front unsprung masses relative positions respectively. In order to minimize body roll after hitting a road obstacle with one of forward wheels the methods advices for softer front dampers and firmer rear ones. Probably it will work the opposite way if only one rear wheel would hit a dent or bump on a road.

Figure 5 shows that the energy functional has a visible minimum in case of body pitch as output. Additionally, modes M4 and M9 start to take bigger part in forming the energy estimation as front damping values go higher. We may interpret this as more energy being not absorbed by too firm front suspension then going through the body to softer rear suspension.

**Fig. 2** Transfer functions square $H_2$-norms for different outputs

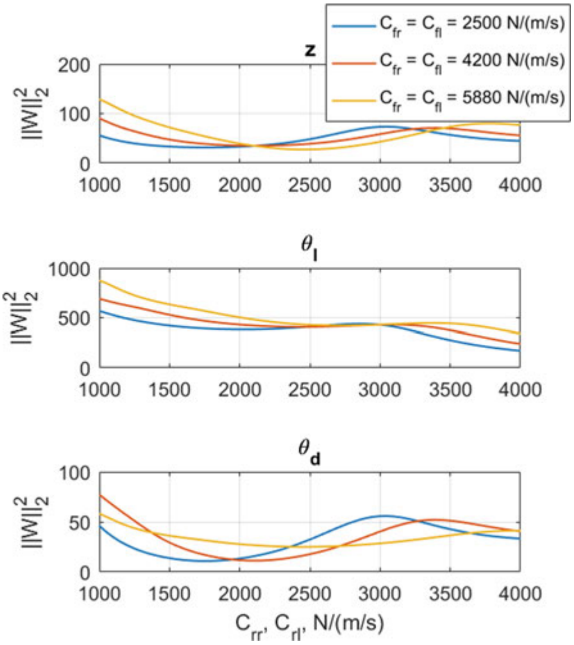
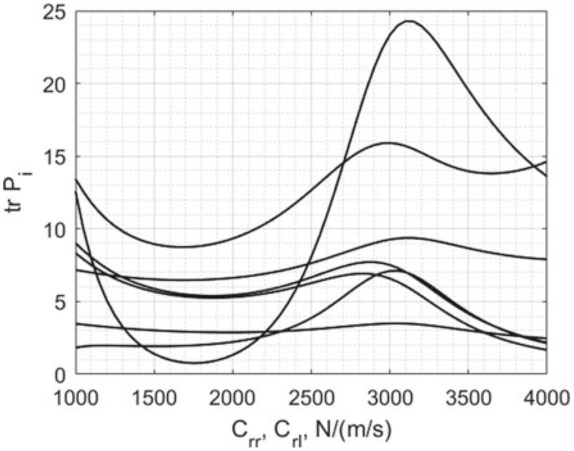


**Fig. 3** General view of the spectral decomposition of transfer functions square $H_2$ norms. Body bounce output. $c_f = 2500$



Let us study two cases from the transfer function set mentioned above using the finite sub-Gramian method. Let the case A correspond to the values $C_{fr}$, $C_{fl} = 4200$ N/(m/s) and $C_{rr}$, $C_{rl} = 1900$ N/(m/s), it is suboptimal in terms of proposed energy-based criteria. Let the case B correspond to the values $C_{fr}$, $C_{fl} = 5880$ N/(m/s) and $C_{rr}$, $C_{rl} = 1000$ N/(m/s), it is clearly unbalanced and far from optimal. Figure 6 presents finite sub-Gramian traces dynamics for modes M1 and M6 as well as infinite sub-Gramian trace for M1 as a reference. It shows that a finite

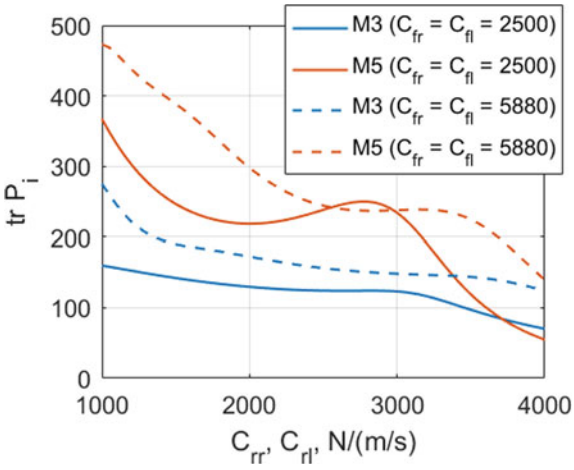


**Fig. 4** Dominant sub-Gramian traces. Body roll output. $c_f = 2500$

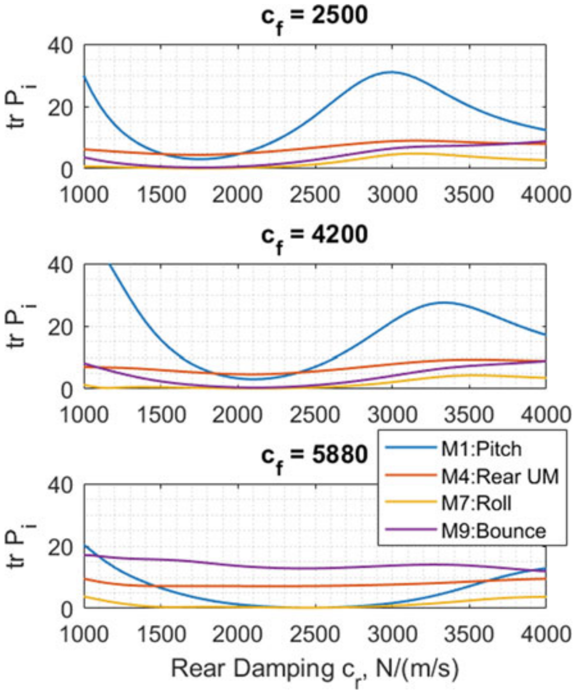


**Fig. 5** The most changing sub-Gramian traces. Body pitch output

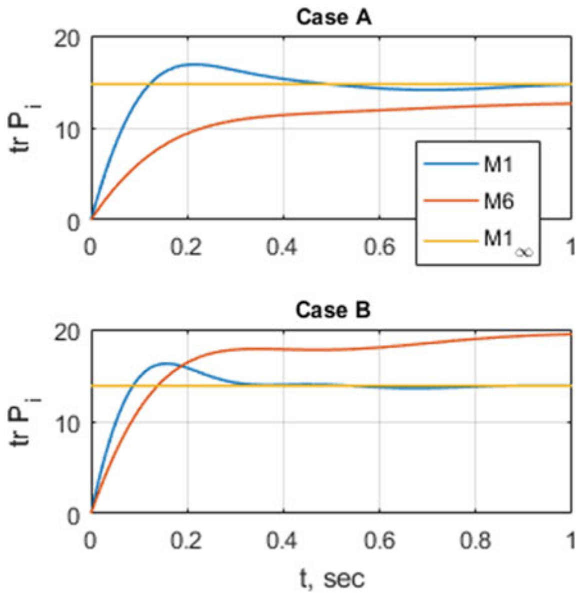


**Fig. 6** Finite M1 and M6 sub-Gramian traces, infinite M1 sub-Gramian trace

sub-Gramian trace value can temporarily exceed an infinite one. That does not strictly comply with earlier interpretation of finite sub-Gramians as total accumulated energy and likely means that the energy can pass from mode to mode, so it will be accounted in other modes on the infinite time interval.

The nature of physical processes indirectly proves such interpretation. After initial hit the suspension converses several types of kinetic and potential energy of the body, the springs and the unsprung masses into each other, which doesn't happen instantly. Figure 7 shows that in case B the damper generates greater force in shorter time during the first 400 ms. Its main goal is to slow down the spring, therefore the vertical movement of the body was reduced both by amplitude and speed during the first 600 ms, as seen on Fig. 8. However, after these 600 ms the delayed reaction from underdamped rear suspension takes place, which Fig. 6 also does reflect.

As the spring and damper exert their forces on both the body and unsprung mass, it is inevitable for certain modes finite-time energy estimations to temporarily exceed their final value. Therefore, the time to reach this value for the first time may be criterion for estimating the process speed. In the future work the system can be improved by the adding adaptive controller. And such additional process speed criterion may be of use for off-line adaptation algorithm tuning or short-term performance assessment. This may lead to further improvements in the adaptive control abilities to compensate the parameters change of the vehicle suspension system due to the change of the environmental conditions or the fault of the system's components [7].

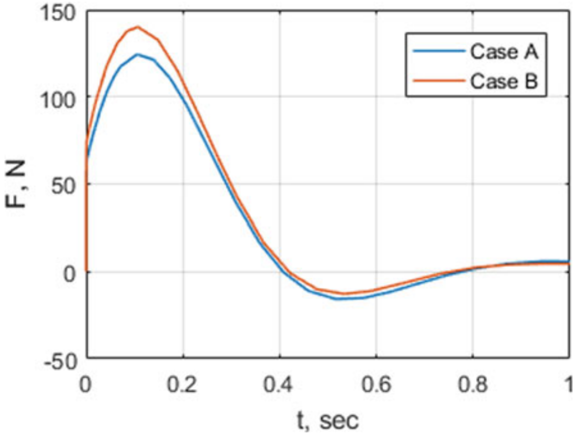


**Fig. 7** Force, generated by the FR damper

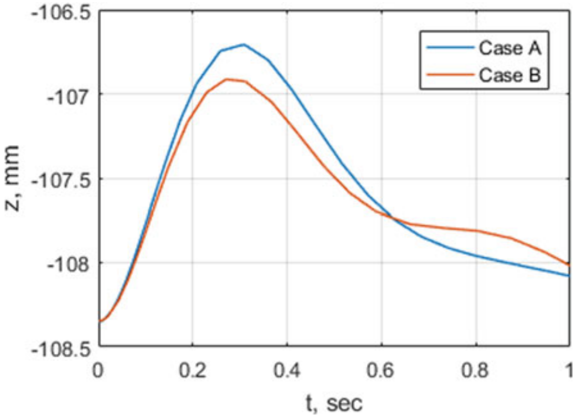


**Fig. 8** Vertical body movement

## 5 Conclusion

Both versions of the sub-Gramian method provide results that agree with engineering practice under given conditions. In particular, the method recommends correct parameters changes. The finite sub-Gramian analysis allows to locate undesirably fast or slow processes and to monitor energy conversion inside the system.

This will allow to connect finite sub-Gramian traces with physical energies in the system in further studies. More importantly, the results of this study provide a base for the formal suspension tuning problem statement using presented energy functionals.

## References

1. Yadykin IB, Kataev DE, Iskakov AB, Shipilov VA (2015) Characterization of power systems near their stability boundary using the sub-Gramian method. Control Eng Pract 53:173–183
2. Yadykin IB, Grobovoy AA, Iskakov AB, Kataev DE, Khmelik MS (2015) Stability analysis of electric power systems using finite Gramians. IFAC-PapersOnLine 48(30):548–553
3. Stone R, Ball JK (2004) Automotive engineering fundamentals. SAE, PA
4. Guiggiani M (2014) The science of vehicle dynamics. Springer, Netherlands
5. Kataev DE, Yadykin IB (2016) Solution of the Lyapunov matrix differential equations by the frequency method. J Comput Syst Sci Int 55(6):843–855
6. Vassiliev SN, Yadykin IB, Iskakov AB, Kataev DE, Grobovoy AA, Kiryanova NG (2017) Participation factors and sub-Gramians in the selective modal analysis of electric power systems. IFAC-PapersOnLine 50(1):14806–14811
7. Vershinin Y (2013) Adaptive control system for solution of fault tolerance problem for MIMO systems. Lecture Notes Eng Comput Sci 2202(1):117–120
8. Das RR, Elumalai VK, Subramanian RG, Kumar KVA (2018) Adaptive predator–prey optimization for tuning of infinite horizon LQR applied to vehicle suspension system. Appl Soft Comput (in press)
9. Seifi A, Hassannejad R, Hamed MA (2016) Optimum design for passive suspension system of a vehicle to prevent rollover and improve ride comfort under random road excitations. Proc Inst Mech Eng Part K J Multi-body Dyn 230(4):426–441
10. Kashem S, Nagarajah R, Ektesabi M (2018) Vehicle suspension systems and electromagnetic dampers. Springer, Singapore

# Mobility in 2050

**Jacques De Kegel, Sten Corfitsen, and Henk G. Hilders**

**Abstract** How our society and in particular our mobility will look like within 30 years from now is hard to predict. The authors take a bold position by looking at how mobility could look like in 2050, disregarding of any temporarily constraint and only then, define the way how to get there by analyzing the different steps, needed to pave the way. This methodology allows them to get rid of the famous tunnel-vision what most studies suffer from. It leads to very interesting and refreshing conclusions, in particular on congestion, traffic jams, car accidents and also on the evolution of the automotive industry in general. The key question, that gets an answer herewith, is "how many vehicles will we have on our roads in 2050?". The answer might be surprising.

**Keywords** Electric mobility · Battery swap · Automated driving · Innovation · Advanced technologies · Affordable mobility solutions

## Acronyms

| | |
|---|---|
| MaaS | Mobility as a Service |
| BaaS | Battery as a Service |
| IoT | Internet of Things |
| VR | Virtual Reality |
| AR | Augmented Reality |

J. De Kegel (✉)
Astridlaan 128, 9400 Ninove, Belgium
e-mail: jacques_dekegel@swap2drive.eu

S. Corfitsen
Villagatan 18, 104 51 Stockholm, Sweden
e-mail: sten.corfitsen@powerswap.se

H. G. Hilders
Arent Janszoon Ernststraat 215, 1083 JN Amsterdam, The Netherlands
e-mail: henk.hilders@cargo-box.com

Y. A. Vershinin et al. (eds.), *Technologies for Smart Cities*,
https://doi.org/10.1007/978-3-031-05516-4_6

| | |
|---|---|
| AI | Artificial Intelligence |
| FotA | Firmware over the Air |
| V2V | Vehicle to Vehicle communication |
| V2I | Vehicle to Infrastructure communication |
| ADEM | Autonomous Driving Electric Mobility |
| BEV | Battery Electric Vehicle |
| ICE | Internal Combustion Engine |
| PDC | Person-Driven Car |
| OEM | Original Equipment Manufacturer (here: Car Manufacturer) |
| BS | Battery Swap solution |
| B2C | Business to Consumer (here: Passenger Cars) |
| B2B | Business to Business |
| B2B-1 | Business to Business (here: Light Utility Vehicles ($\leq$3.5 t)) |
| B2B-2 | Business to Business (here: Heavy Utility Vehicles (>3.5 t)) |
| B2B-3 | Business to Business (here: Busses and Touring Cars) |
| CC | Conductive Charging (= charging at an energy pole with a cable) |
| BRS | Battery Recharging Station |

## 1 Lessons Learned from the Past

Following trends are identified

- Everything, related to environment, gets utmost priority by governments of all kind and gradually by Captains of Industry as well. Reduction of greenhouse gases ($CO_2$, $NO_x$, $CH_4$, etc.) and harmful particles (Black Carbon, $PM_{10}$, $PM_{2.5}$, etc.) are of utmost importance. The EU Green Deal and the Paris Treaty have to be respected and the objectives have to be realized. The health condition of Planet Earth is getting by far the highest priority. To illustrate this, we just refer at the recent judgement of the court at The Hague against Royal Dutch Shell and the presence of representatives of Engine No. 1 within the Board of Directors of ExxonMobil.
- The younger generation has different ideas about property and ownership of goods, be it an apartment or even a car. While for the baby boomers, ownership of a nice car was a major objective, for the generation x, y and z, this is no longer the case.
- Technology is becoming more and more important in our day-to-day lives. We see new types of technology popping up, like the Internet of Things (IoT), Artificial Intelligence (AI), Virtual and Augmented Reality (VR/AR), Blockchain, Cloud technology, Firmware over the Air (FotA), etc.
- Intermodal transport is not yet a reality, but a number of building blocks are already in place. We expect this to become a reality in 2050, as well for mobility of people as for transport of goods.
- New ways of mobility will be developed within the coming years, like hyperloop, flying taxis, autonomous driving vehicles, etc.

## 2 Mobility of People in 2050

Imagine following situation:

- In 2050, every single trip (person or goods) will be coordinated by an overall **Mobility Service Platform (MSP).** Interaction with this MSP might look like this:
  - Someone is preparing for a business trip to Brazil. His flight at Heathrow will leave at 14:00 h. The evening before, he contacts the MSP bot, saying: **“ *I have to catch a flight to São Paolo, Brazil, that leaves Heathrow Airport tomorrow at 14:00 h and I have only one bag of roughly 15 kg* ”.** Potentially, the MSP bot will ask a few more questions, in order to prepare this trip as good as possible. Finally, the MSP bot will return following answer **“ *At 09:30h, there will be an ADEM car (Autonomous Driving Electric Mobility) in front of your home, to bring you to the airport* ”**.
  - As he’s heading for a business trip, where he has to make an important presentation at a congress, the ADEM car, which is equipped with 360° video enabled windows, allows him to rehearse his presentation in front of a virtual interactive audience.
  - At his arrival at Heathrow airport, the on-board printer has already printed his boarding documents.

In 2050, the vast majority of vehicles will be autonomous driving vehicles. New wireless technologies, like 5G, 6G, etc., xG and short-range wireless communication like vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I) and many more will be available and will allow these vehicles to drive from A to B without any risk on car accidents anymore. Indeed, once human beings will stop driving themselves and entirely trust autonomous driving vehicles, the number of car accidents will drastically drop and eventually completely disappear.

This is a nice dream, but chances are real that it will become reality in 2050 or even some earlier.

Continental has produced a very nice clip, outlining their ideas on how the future of mobility may look like: https://youtu.be/mk52AaxofM8.

## 3 Intermodal Mobility of People

And what about intermodal transport systems?

In the example here above, that person has made use of only one single mode of mobility in order to get into the airport. But let’s take another passenger, who’s living in a small village in the neighborhood of Ghent, Belgium, and is heading for a congress in downtown Antwerp. A day ahead, while preparing her trip, she contacts the MSP bot, saying: **“ *Hello, tomorrow I have to be at the Mobility Congress in Antwerp at 9:00 h, but want to arrive at the Congress Hall at 8:30 h already.* ”** The

MSP bot might return following message: “***At 7:15h, ADEM car with id. 1701 will be in front of your front door. That one will bring you to the Hyperloop station of Ghent, where you will arrive at 7:46h. The hyperloop itself will leave at 7:55h and arrive in Antwerp at 8:05h. At the Antwerp hyperloop station, ADEM car with id. 2043 will wait for you and bring you to the auditorium of the University, where the congress will take place. In the meantime, I already have introduced all these data into your personal calendar on your smart phone. Your alarm clock will go off at 6:15 h. Is that OK for you?*** ”.

Did you notice that we never mentioned any payment so far? Payment handling will also be integrated and secured, in such a way that no one has to take care about. As all data, needed to handle payment transactions in a secure way, are available inside the MSP Central System (place of departure, place of destination, distance, mode of operation, etc.), the MSP Central System will match these with the payment schemes of the different operators, that provide services within the intermodal chain, and initiate the payment transaction. That whole transaction will be transparent for the user and secured via Blockchain transactions.

# 4 Intermodal Mobility of Goods

And what about transport of goods?

Like with persons, goods can and will be handled in an integrated way as well. Once the goods are stored at the producer plant into an intelligent box, the logistics operator can then get in touch with that same MSP bot and order a transport for that particular intelligent box towards the final destination location. Here, obviously, the operator will have to provide some more details, like size of the box, weight, upper and lower limits of temperature (e.g. in case of drugs, vaccines), … unless all these data are already prepared by the operator and collected into a standardized electronic logistics file, which then can be uploaded at once into the MSP Central System.

All requests will come together within the MSP Central System, where an algorithm will execute an optimization of logistics flows in order to avoid partially filled loads. Hence, the system will only return an answer to the operator, once the optimization routine has been executed. Then again, the answer will be that a large freight ADEM van will arrive at the production site to collect the sealed intelligent box and bring it to its final destination via different modes of mobility (ADEM, Train, Plane, Barge, Drone, etc.), where an authorized person, who’s entitled to unseal the intelligent box, will open this intelligent box and collect the goods.

The goods are secured at all time, since the intelligent box is sealed and protected against theft, damage and pilferage. Hence, we can speak about real door-to-door secured logistics trade lanes.

## 5 Is This Science Fiction or Reality, that Is Becoming True?

Now that we have a better view on how transport of people and goods might look like in 2050, let's elaborate these predictions a little more in detail. We will also take a look at the energy needs, since we may assume that, in case we all drive electric by 2050, the energy production plants and energy grids might be stressed to the limits. No? Let's make some calculations and discover the impact our new model of mobility will have on our day-to-day lives.

## 6 How Many Vehicles Will We Really Need in 2050?

That's a very important, if not the utmost important question, we have to ask ourselves. Nowadays, the majority of citizens own a private car or use a company car. This makes that 448 million inhabitants of the EU-27 own or use about 247 million passenger cars.[1] Adding to that 29 million Light Utility vehicles ($\leq$3.5 t), 6 million Heavy Utility Vehicles (>3.5 t) and 700,000 busses and touring cars, then we end up with a total figure of 283 million vehicles.[1]

Will we still need that huge number of vehicles in 2050 anymore? The answer is definitely: NO! Here's why:

Have you ever wondered what percentage of the time a car is in use and how much time it is simply parked? The equation is very simple. Let's take following assumptions:

- The average mileage of a vehicle is 15,000 km/year
- The average speed of a vehicle is 60 km/h.

Then we find out that this vehicle is in use during 250 h.

Since a year counts $365 \times 24 = 8760$ h, the vehicle is in use for 0.0285 or 2.85% of the time.

Hence following logical question arises: "Is it still justifiable to make an investment in equipment, which has 97.15% idle time?" Will we continue to own and use a vehicle that is only in use during 2.85%[2] of its time and is unused for the remaining 97.15% of the time?

Only the Happy Few will be able to afford such a "bad investment", while the vast majority of the population will be pushed towards more economically viable solutions. One of these solutions is the ADEM concept, since ADEM allows vehicles to be in use for almost 100% of the time. Obviously, an investment that can be depreciated over 100% of its time, is much more economic justifiable than one that is only in use for 2.85% of its time. This means that mobility will become much cheaper than it is today and that far less vehicles will be needed to satisfy the mobility needs of the population.

## 7 Autonomous Driving Electric Mobility (ADEM)

Let's concentrate for a while on the ADEM concept.

At first, we're looking at the **Autonomous Driving** vehicle, which means that fatigue or distraction of the driver is no longer an issue. This vehicle can be in use 24/7 and thus at 100% of the time. The only moment, when this vehicle is not in use, is when maintenance and cleaning is ongoing or at an unexpected breakdown of the equipment. Hence, this vehicle is unavailable only during planned and unplanned outages. But apart from that, the vehicle can be in use all time.

As ADEM handles about **Electric Mobility** and Battery Swap technology provides a full "recharge" in less than 2 min, queueing and long waiting times at energy charging poles will be a concept of the past. All business models, based on conductive charging of electric vehicles are doomed to fail, be it car sharing, taxi, ambulance, etc., since a vehicle, standing in front of a charging pole—even a fast charger—is economically speaking "dead" for a longer period of time. As long as there is no battery swap alternative, these business models will never fly. Hence, we assume that by 2050 the only way of charging will be by swapping batteries.

The driverless ADEM car, based on battery swap, is capable of being in use for almost 100% of the time. With this in mind, one can ask following question: "In that case, how many ADEM vehicles do we need in 2050 to offer a high quality MaaS (Mobility as a Service)?" The answer is "Less than you might imagine".

Indeed, as pointed out earlier, we, nowadays have 247 million passenger cars within EU-27, who are in use for only 2.85% of the time. Suppose that we are able to replace this fleet by vehicles, which are in use all time, then we only need about 7 million passenger vehicles to fulfill the mobility needs from the whole EU-27 population.

As this 7 million is theoretically speaking correct and is an optimization by a factor of 35, this number is not realistic, since there is cleaning and maintenance time needed, as well as very little time to perform the battery swaps as well. Moreover, there are hours of "low demand" for transportation of people. Between 1 a.m. and 5 a.m., there will be little need for mobility of people, even if price/km drops drastically within this timeslot. Hence, we assume an optimization of 25 as being realistic and yet of great improvement.

The biggest challenge within the years to come will be to guide the transition from the actual situation towards this projected situation in 2050. Let's take a closer look at it.

## 8 How to Move from ICE to BEV with Battery Swap and Ultimately to ADEM Between Now and 2050?

Today, we have about 247 million passenger cars registered in EU-27, of which only 0.2% or 0.5 million are BEV (Battery Electric Vehicles).[1] This means that 99.8%

are still equipped with an ICE (Internal Combustion Engine) and will have to be converted within the coming years towards Electric Mobility.

We may expect that, by 2050, the energy utilities will have made their turn-around as well and that at that very moment the electric energy will be provided for 100% from renewable energy sources (solar, wind, tide, bio, etc.).

Let's take a closer look at the usage mix of vehicles. The vast majority might be converted to ADEM, while some cannot be converted by 2050 from PDC (Person-Driven Car) to ADEM for obvious reasons:

- A medical doctor, who will have to visit patients at home, will always need an individually owned car, which will be available at any time.
- Home care givers will also need to have a dedicated car.
- Technicians (plumbers, HVAC-technicians, etc.) as well will have the need for an individual vehicle, although most of them will make use of a Light Utility vehicle (≤3.5 t), of which there are about 29 million enrolled today within EU-27.

Let's assume this figure being around 10% or 25 million units within EU-27. These vehicles will evolve over time from ICE towards Battery Swap BEV (BEV-BS). Anyhow, we assume that even these vehicles might evolve, within the time frame between 2050 and 2060 towards an ADEM equivalent. In the event this comes thru by 2060 we'll have only automated driving vehicles on the roads, thus removing almost every single risk for a car accident.

The remaining 222 million passenger vehicles are candidate for an "investment optimization" through the ADEM MaaS offering. This will not happen overnight, but will take some time. We assume that the first roll-out of ADEM will only start in 2030, with 50% conversion in 2040 and finally a 100% conversion in 2050. As we anticipate a 25 times improvement in efficiency, we can assume that the number of vehicles needed will be about 25 times smaller. Overcrowded routes will then be a concept of the past and some roads will be redesigned to host more pedestrians, bikes, steps, etc.

In the intermediate timeframe between now and 2030, a number of ICE cars will be converted towards the intermediate Conductive Charging technology of Battery Electric Vehicle (BEV-CC), thus bringing the BEV penetration ratio from a 0.2% in 2020 towards about 5% in 2030 before being halved in 2040 and completely disappear and being replaced by the more effective and efficient Battery Swap solution. At that moment, the millions of charging poles will also disappear from our streets, freeing space for trees.

On top of these 247 million passenger cars, we also have nearly 36 million Utility vehicles on the European streets, including busses and touring cars. Over time, they will all have to become green as well and thus migrate to a battery swap model too in order to be economic viable.

Likewise, this evolution will not happen overnight, but will need "some time". We foresee a start of this migration in 2030, with a 50% uptake in 2040 and a 100% completion in 2050. Mission accomplished!

But before we'll reach this point, quite a number of hurdles will have to be taken, as well technical as political or societal. Not everyone is yet ready to abandon private ownership of a car.

The key trigger will be an economic one, since owning a car is expensive, as we have illustrated, and will become many times more expensive in the future. Or, said in a different way, standard individual mobility will become much cheaper once we can make use of a vehicle for nearly 100% of its time, instead of only 2.85% as of today. Will it become 35 times cheaper? No, certainly not, but in case we are able to offer a 25 times cheaper mobility, the vast majority of the population will be interested in moving that way. Don't you think so?

As you can see in Table 1, the number of vehicles will drop considerably by 2050. There will be nearly 80 million vehicles on the European roads, compared to the 283

**Table 1** Overview of the actual vehicle mix in Europe (*source* ACEA—The European Automotive Manufacturers Association) and forecast for an evolution towards a full MaaS and ADEM based mobility model by 2050

| | | | Actual | | Forecast | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 2020 | | 2030 | 2040 | 2050 | 2060 |
| B2C | B2C-ICE | (1) | 222,000,000 | 78.39% | 210,900,000 | 105,400,000 | 0 | |
| | | (2) | 25,000,000 | 8.83% | 23,750,000 | 11,875,000 | 0 | |
| | BEV-CC | (3) | 500,000 | 0.18% | 12,850,000 | 6,425,000 | 0 | |
| | BEV-BS | (4) | | | 0 | 18,300,000 | 36,600,000 | * |
| | ADEM | (5) | | | 0 | 4,218,000 | 8,436,000 | |
| B2B-1 | B2B-1 | (6) | 29,000,000 | 10.24% | 29,000,000 | 14,500,000 | 0 | |
| | | (7) | | | | 14,500,000 | 29,000,000 | * |
| B2B-2 | B2B-2 | (8) | 6,000,000 | 2.12% | 6,000,000 | 3,000,000 | 0 | |
| | | (9) | | | | 3,000,000 | 6,000,000 | * |
| B2B-3 | B2B-3 | (10) | 700,000 | 0.25% | 700,000 | 350,000 | 0 | |
| | | (11) | | | | 350,000 | 700,000 | * |
| Total | | | 283,200,000 | | 283,200,000 | 181,968,000 | 80,736,000 | |
| PDC | | | 283,200,000 | | 283,200,000 | 177,750,000 | 72,300,000 | |

* Over time, these vehicles can also become autonomous driving, thus also becoming ADEM, in which case PDC will be reduced to 0% and ADEM becoming 100% by 2060

(1) ICE cars (convertible from PDC towards ADEM)
(2) ICE cars (convertible from ICE towards battery swap)
(3) BEV cars (from conductive charging towards battery swap)
(4) BEV cars (battery swap)
(5) ADEM (autonomous driving electric mobility)
(6) Light utility vehicles ($\leq$3.5 t)
(7) LUV conversion from ICE towards battery swap
(8) Heavy utility vehicles (>3.5 t)
(9) HUV conversion from ICE towards battery swap
(10) Busses and touring cars
(11) B&TC conversion from ICE towards battery swap

million nowadays, of which 8.5 million will be Automated Driving (ADEM) and the remaining 72 million will still be Person Driven. Hence, traffic jams will definitely be a concept of the past at that moment in time. Road works can be executed without any problem during daytime. Car accidents will also disappear to a large extend, since Automated Driving vehicles do not suffer from distraction, fatigue, hubris, … as humans do. Searching for a free parking spot will become a concept of the past as well since ADEM cars do not park for a longer time at the road side. They will just collect or drop passengers and immediately continue driving towards the next mission. And, last but not least, moving from A to B will no longer be a waste of time, since, while being inside the ADEM vehicle, one can do many useful things, from attending a video conference, rehearsing a presentation up to enjoying the landscape or listening to music, podcasts, etc. And the tiring job of conducting a car will finally be delegated to the onboard computer.

In a plotted format, Fig. 1a and b looks like this.

## 9 How Much Energy Will Be Needed to Power All These BEVs?

Next question that will pop-up is: "How many GWh do we need to power this Electric Mobility in Europe?". The answer is not obvious and needs some more in-depth reflection, but here is the answer.

First, we have to agree on some assumptions[2]:

- An electric PDC, as we know it actually (BEV-CC), will need to have a battery capacity onboard of at least 90 kWh, in order to remove range anxiety (e.g. Tesla, Audi e-tron, etc.).
- A car, equipped with a Battery Swap solution (BEV-BS), will only need an onboard capacity of about 30 kWh, since for these passenger vehicles, range anxiety won't exist anymore.
- As each of them will drive on average 15,000 km/year and consume about 0.2 kWh/km, they will each consume 3000 kWh or 3 MWh per year.
- For an ADEM vehicle, we consider likewise a 30 kWh-battery capacity, although models with higher capacity (e.g. 2 × 30 kWh … up to 6 × 30 kWh) might be designed as well, thus considerably reducing the number of swap-stops.
- An ADEM vehicle, which is operating 24/7, will cross 375,000 km/year and will thus consume 75,000 kWh or 75 MWh.
- We will include the Utility Vehicles as well, in order to be complete:
    - A B2B-1 Light Utility Vehicle ($\leq$3.5 t) has a battery capacity of 2 × 30 kWh = 60 kWh and has an average energy consumption of 0.5 kWh/km.
    - A B2B-2 Heavy Utility Vehicle (>3.5 t) has a battery capacity of 8 × 30 kWh = 240 kWh and an average energy consumption of 1 kWh/km.

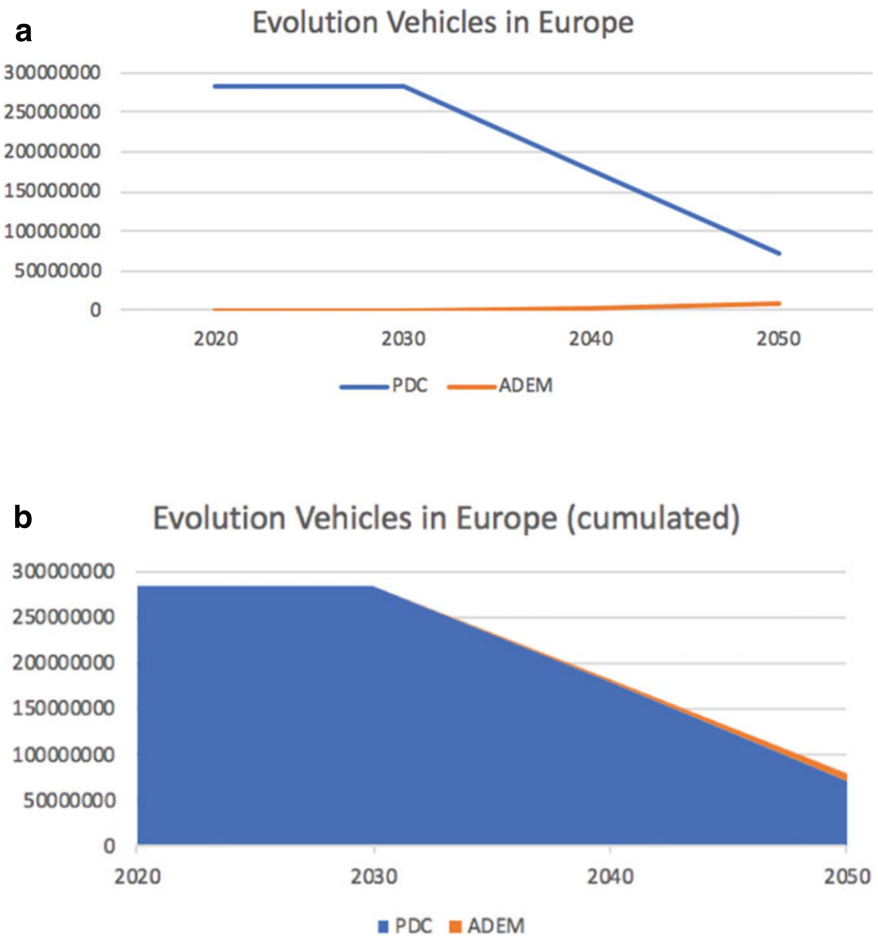


**Fig. 1** **a** Forecast on the evolution of the vehicles in Europe. **b** Forecast of the evolution of the vehicles in Europe (cumulated view)

– A B2B-3 Bus and Touring Car needs a much larger battery capacity. Here we anticipate a total battery capacity of 1000 kWh and an average energy consumption of 1 kWh/km.

With all these figures combined, we can forecast the potential energy need as follows (Table 2 and Fig. 2).

**Table 2** Forecast of the energy requirements for mobility in Europe, in the event of a total shift towards electric mobility by 2050

| | Battery capacity | Distance (km/year) | Consump. (kWh/km) | Annual El. (kWh) | Energy need (GWh) | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | 2020 | 2030 | 2040 | 2050 |
| BEV-CC | 90 | 15,000 | 0.2 | 3000 | 1500 | 38,550 | 19,275 | 0 |
| BEV-BS | 30 | 15,000 | 0.2 | 3000 | 0 | 0 | 54,900 | 109,800 |
| ADEM | 30 | 375,000 | 0.2 | 75,000 | 0 | 0 | 316,350 | 632,700 |
| B2B-1 | 60 | 30,000 | 0.5 | 15,000 | 0 | 0 | 217,500 | 435,000 |
| B2B-2 | 240 | 10,000 | 1.0 | 10,000 | 0 | 0 | 30,000 | 60,000 |
| B2B-3 | 1000 | 50,000 | 1.0 | 50,000 | 0 | 0 | 17,500 | 35,000 |
| Totals (GWh) | | | | | 1500 | 38,550 | 655,525 | 1,272,500 |

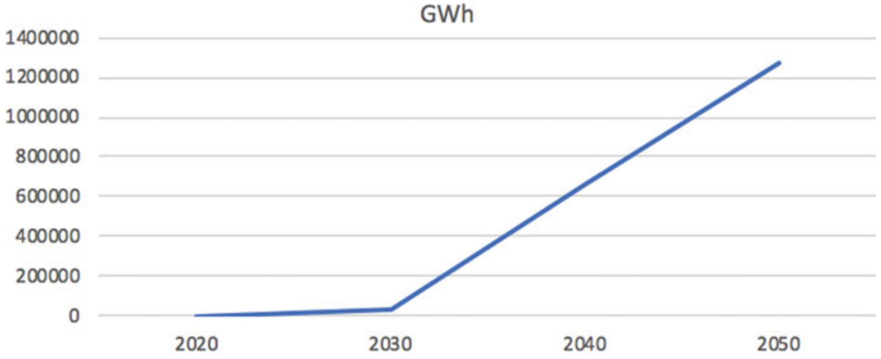


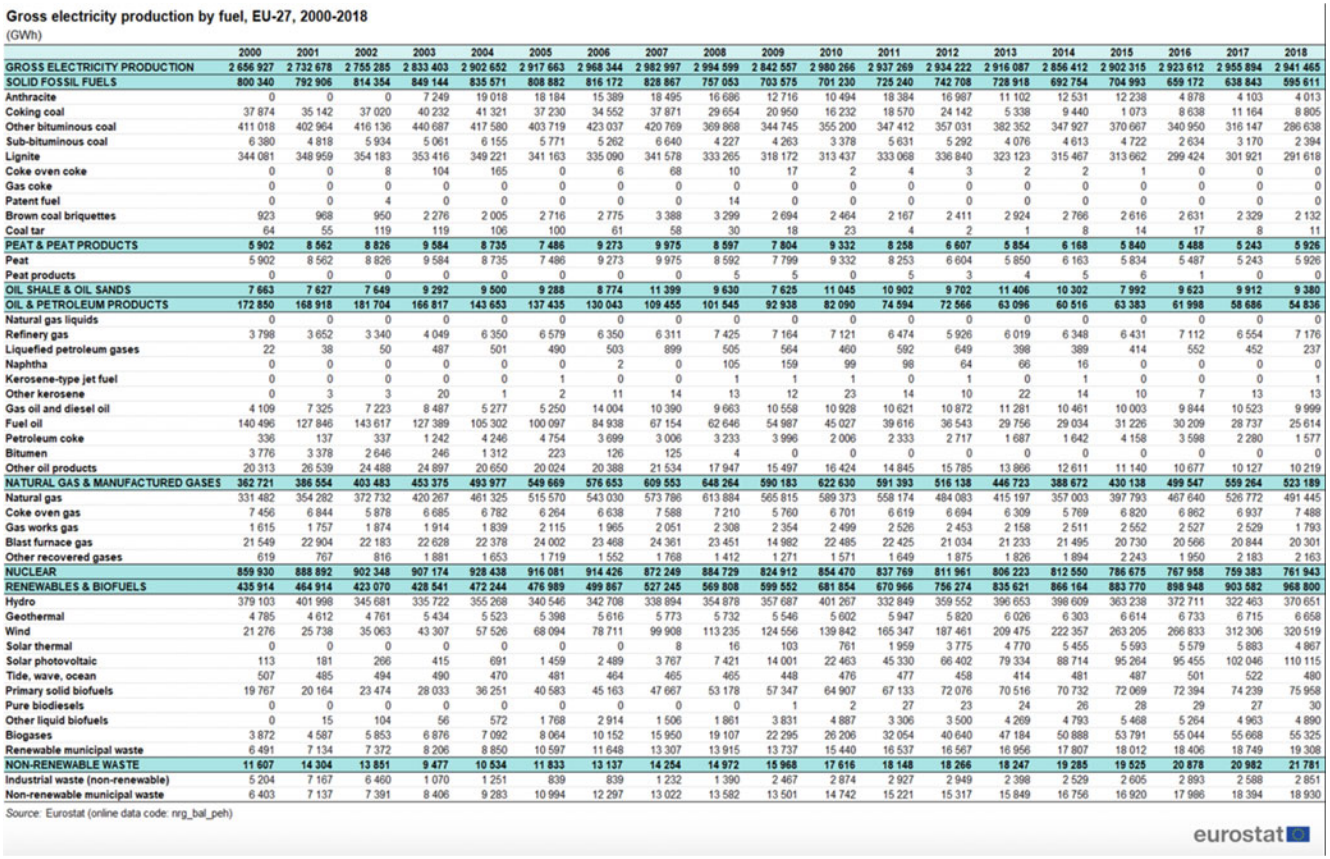
**Fig. 2** Graphical representation of the energy requirements for mobility in Europe EU-27, in the event of a total shift towards electric mobility by 2050

## *9.1 More Than 1272 TWh by 2050, You Say?*

To make it a little more tangible, we will now take a look at the European energy production. According to Eurostat, the Gross Electricity Production for EU-27 in 2018 was 2941.47 TWh.

This means that, in case our energy production capacity remains unchanged until 2050, at that moment, almost 43% of our generated energy have to be allocated to mobility. Since it would be naïve to suppose that energy needs for all other purposes except mobility will decrease in that same period by 43%, we have to grow our production capacity to catch up with the transition from ICE towards BEV (Fig. 3).

Observe, while looking up this table, that Energy Utilities across Europe still have a long way to go in order to reach the 2050 target of 100% renewable energy. Hence, we may conclude that Energy Utilities worldwide have multiple major challenges ahead within the coming 30 years.



Gross electricity production by fuel, EU-27, 2000-2018

(GWh)

| | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GROSS ELECTRICITY PRODUCTION | 2 656 927 | 2 732 678 | 2 755 285 | 2 833 403 | 2 902 652 | 2 917 663 | 2 968 344 | 2 982 997 | 2 994 599 | 2 842 557 | 2 980 266 | 2 937 269 | 2 934 222 | 2 916 087 | 2 856 412 | 2 902 315 | 2 923 612 | 2 955 894 | 2 941 465 |
| SOLID FOSSIL FUELS | 800 340 | 792 906 | 814 354 | 849 144 | 835 571 | 808 882 | 816 172 | 828 867 | 757 053 | 703 575 | 701 230 | 725 240 | 742 708 | 728 918 | 692 754 | 704 993 | 659 172 | 638 843 | 595 611 |
| Anthracite | 0 | 0 | 0 | 7 249 | 19 018 | 18 184 | 15 389 | 18 495 | 16 686 | 12 716 | 10 494 | 18 384 | 16 987 | 11 102 | 12 531 | 12 238 | 4 878 | 4 103 | 4 013 |
| Coking coal | 37 874 | 35 142 | 37 020 | 40 232 | 41 321 | 37 230 | 34 552 | 37 871 | 29 654 | 20 950 | 16 232 | 18 570 | 24 142 | 5 338 | 9 440 | 1 073 | 8 638 | 11 164 | 8 805 |
| Other bituminous coal | 411 018 | 402 964 | 416 136 | 440 687 | 417 580 | 403 719 | 423 037 | 420 769 | 369 868 | 344 745 | 355 200 | 347 412 | 357 031 | 382 352 | 347 927 | 370 667 | 340 950 | 316 147 | 286 638 |
| Sub-bituminous coal | 6 380 | 4 818 | 5 934 | 5 061 | 6 155 | 5 771 | 5 262 | 6 640 | 4 227 | 4 263 | 3 378 | 5 631 | 5 292 | 4 076 | 4 613 | 4 722 | 2 634 | 3 170 | 2 394 |
| Lignite | 344 081 | 348 959 | 354 183 | 353 416 | 349 221 | 341 163 | 335 090 | 341 578 | 333 265 | 318 172 | 313 437 | 333 068 | 336 840 | 323 123 | 315 467 | 313 662 | 299 424 | 301 921 | 291 618 |
| Coke oven coke | 0 | 0 | 8 | 104 | 165 | 0 | 6 | 68 | 10 | 17 | 2 | 4 | 3 | 2 | 2 | 1 | 0 | 0 | 0 |
| Gas coke | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Patent fuel | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Brown coal briquettes | 923 | 968 | 950 | 2 276 | 2 005 | 2 716 | 2 775 | 3 388 | 3 299 | 2 694 | 2 464 | 2 167 | 2 411 | 2 924 | 2 766 | 2 616 | 2 631 | 2 329 | 2 132 |
| Coal tar | 64 | 55 | 119 | 119 | 106 | 100 | 61 | 58 | 30 | 18 | 23 | 4 | 2 | 1 | 8 | 14 | 17 | 8 | 11 |
| PEAT & PEAT PRODUCTS | 5 902 | 8 562 | 8 826 | 9 584 | 8 735 | 7 486 | 9 273 | 9 975 | 8 597 | 7 804 | 9 332 | 8 258 | 6 607 | 5 854 | 6 168 | 5 840 | 5 488 | 5 243 | 5 926 |
| Peat | 5 902 | 8 562 | 8 826 | 9 584 | 8 735 | 7 486 | 9 273 | 9 975 | 8 592 | 7 799 | 9 332 | 8 253 | 6 604 | 5 850 | 6 163 | 5 834 | 5 487 | 5 243 | 5 926 |
| Peat products | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 5 | 0 | 5 | 3 | 4 | 5 | 6 | 1 | 0 | 0 |
| OIL SHALE & OIL SANDS | 7 663 | 7 627 | 7 649 | 9 292 | 9 500 | 9 288 | 8 774 | 11 399 | 9 630 | 7 625 | 11 045 | 10 902 | 9 702 | 11 406 | 10 302 | 7 992 | 9 623 | 9 912 | 9 380 |
| OIL & PETROLEUM PRODUCTS | 172 850 | 168 918 | 181 704 | 166 817 | 143 653 | 137 435 | 130 043 | 109 455 | 101 545 | 92 938 | 82 090 | 74 594 | 72 566 | 63 096 | 60 516 | 63 383 | 61 998 | 58 686 | 54 836 |
| Natural gas liquids | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Refinery gas | 3 798 | 3 652 | 3 340 | 4 049 | 6 350 | 6 579 | 6 350 | 6 311 | 7 425 | 7 164 | 7 121 | 6 474 | 5 926 | 6 019 | 6 348 | 6 431 | 7 112 | 6 554 | 7 176 |
| Liquefied petroleum gases | 22 | 38 | 50 | 487 | 501 | 490 | 503 | 899 | 505 | 564 | 460 | 592 | 649 | 398 | 389 | 414 | 552 | 452 | 237 |
| Naphtha | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 105 | 159 | 99 | 98 | 64 | 66 | 16 | 0 | 0 | 0 | 0 |
| Kerosene-type jet fuel | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| Other kerosene | 0 | 3 | 3 | 20 | 1 | 2 | 11 | 14 | 13 | 12 | 23 | 14 | 10 | 22 | 14 | 10 | 7 | 13 | 13 |
| Gas oil and diesel oil | 4 109 | 7 325 | 7 223 | 8 487 | 5 277 | 5 250 | 14 004 | 10 390 | 9 663 | 10 558 | 10 928 | 10 621 | 10 872 | 11 281 | 10 461 | 10 003 | 9 844 | 10 523 | 9 999 |
| Fuel oil | 140 496 | 127 846 | 143 617 | 127 389 | 105 302 | 100 097 | 84 938 | 67 154 | 62 646 | 54 987 | 45 027 | 39 616 | 36 543 | 29 756 | 29 034 | 31 226 | 30 209 | 28 737 | 25 614 |
| Petroleum coke | 336 | 137 | 337 | 1 242 | 4 246 | 4 754 | 3 699 | 3 006 | 3 233 | 3 996 | 2 006 | 2 333 | 2 717 | 1 687 | 1 642 | 4 158 | 3 598 | 2 280 | 1 577 |
| Bitumen | 3 776 | 3 378 | 2 646 | 246 | 1 312 | 223 | 126 | 125 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Other oil products | 20 313 | 26 539 | 24 488 | 24 897 | 20 650 | 20 024 | 20 388 | 21 534 | 17 947 | 15 497 | 16 424 | 14 845 | 15 785 | 13 866 | 12 611 | 11 140 | 10 677 | 10 127 | 10 219 |
| NATURAL GAS & MANUFACTURED GASES | 362 721 | 386 554 | 403 483 | 453 375 | 493 977 | 549 669 | 576 653 | 609 553 | 648 264 | 590 183 | 622 630 | 591 393 | 516 138 | 446 723 | 388 672 | 430 138 | 499 547 | 559 264 | 523 189 |
| Natural gas | 331 482 | 354 282 | 372 732 | 420 267 | 461 325 | 515 570 | 543 030 | 573 786 | 613 884 | 565 815 | 589 373 | 558 174 | 484 083 | 415 197 | 357 003 | 397 793 | 467 640 | 526 772 | 491 445 |
| Coke oven gas | 7 456 | 6 844 | 5 878 | 6 685 | 6 782 | 6 264 | 6 638 | 7 588 | 7 210 | 5 760 | 6 701 | 6 619 | 6 694 | 6 309 | 5 769 | 6 820 | 6 862 | 6 937 | 7 488 |
| Gas works gas | 1 615 | 1 757 | 1 874 | 1 914 | 1 839 | 2 115 | 1 965 | 2 051 | 2 308 | 2 354 | 2 499 | 2 526 | 2 453 | 2 158 | 2 511 | 2 552 | 2 527 | 2 529 | 1 793 |
| Blast furnace gas | 21 549 | 22 904 | 22 183 | 22 628 | 22 378 | 24 002 | 23 468 | 24 361 | 23 451 | 14 982 | 22 485 | 22 425 | 21 034 | 21 233 | 21 495 | 20 730 | 20 566 | 20 844 | 20 301 |
| Other recovered gases | 619 | 767 | 816 | 1 881 | 1 653 | 1 719 | 1 552 | 1 768 | 1 412 | 1 271 | 1 571 | 1 649 | 1 875 | 1 826 | 1 894 | 2 243 | 1 950 | 2 183 | 2 163 |
| NUCLEAR | 859 930 | 888 892 | 902 348 | 907 174 | 928 438 | 916 081 | 914 426 | 872 249 | 884 729 | 824 912 | 854 470 | 837 769 | 811 961 | 806 223 | 812 550 | 786 675 | 767 958 | 759 383 | 761 943 |
| RENEWABLES & BIOFUELS | 435 914 | 464 914 | 423 070 | 428 541 | 472 244 | 476 989 | 499 867 | 527 245 | 569 808 | 599 552 | 681 854 | 670 966 | 756 274 | 835 621 | 866 164 | 883 770 | 898 948 | 903 582 | 968 800 |
| Hydro | 379 103 | 401 998 | 345 681 | 335 722 | 355 268 | 340 546 | 342 708 | 338 894 | 354 878 | 357 687 | 401 267 | 332 849 | 359 552 | 396 653 | 398 609 | 363 238 | 372 711 | 322 463 | 370 651 |
| Geothermal | 4 785 | 4 612 | 4 761 | 5 434 | 5 523 | 5 398 | 5 616 | 5 773 | 5 732 | 5 546 | 5 602 | 5 947 | 5 820 | 6 026 | 6 303 | 6 614 | 6 733 | 6 715 | 6 658 |
| Wind | 21 276 | 25 738 | 35 063 | 43 307 | 57 526 | 68 094 | 78 711 | 99 908 | 113 235 | 124 556 | 139 842 | 165 347 | 187 461 | 209 475 | 222 357 | 263 205 | 266 833 | 312 306 | 320 519 |
| Solar thermal | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 16 | 103 | 761 | 1 959 | 3 775 | 4 770 | 5 455 | 5 593 | 5 579 | 5 883 | 4 867 |
| Solar photovoltaic | 113 | 181 | 266 | 415 | 691 | 1 459 | 2 489 | 3 767 | 7 421 | 14 001 | 22 463 | 45 330 | 66 402 | 79 334 | 88 714 | 95 264 | 95 455 | 102 046 | 110 115 |
| Tide, wave, ocean | 507 | 485 | 494 | 490 | 470 | 481 | 464 | 465 | 465 | 448 | 476 | 477 | 458 | 414 | 481 | 487 | 501 | 522 | 480 |
| Primary solid biofuels | 19 767 | 20 164 | 23 474 | 28 033 | 36 251 | 40 583 | 45 163 | 47 667 | 53 178 | 57 347 | 64 907 | 67 133 | 72 076 | 70 516 | 70 732 | 72 069 | 72 394 | 74 239 | 75 958 |
| Pure biodiesels | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 27 | 23 | 24 | 26 | 28 | 29 | 27 | 30 |
| Other liquid biofuels | 0 | 15 | 104 | 56 | 572 | 1 768 | 2 914 | 1 506 | 1 861 | 3 831 | 4 887 | 3 306 | 3 500 | 4 269 | 4 793 | 5 468 | 5 264 | 4 963 | 4 890 |
| Biogases | 3 872 | 4 587 | 5 853 | 6 876 | 7 092 | 8 064 | 10 152 | 15 950 | 19 107 | 22 295 | 26 206 | 32 054 | 40 640 | 47 184 | 50 888 | 53 791 | 55 044 | 55 668 | 55 325 |
| Renewable municipal waste | 6 491 | 7 134 | 7 372 | 8 206 | 8 850 | 10 597 | 11 648 | 13 307 | 13 915 | 13 737 | 15 440 | 16 537 | 16 567 | 16 956 | 17 807 | 18 012 | 18 406 | 18 749 | 19 308 |
| NON-RENEWABLE WASTE | 11 607 | 14 304 | 13 851 | 9 477 | 10 534 | 11 833 | 13 137 | 14 254 | 14 972 | 15 968 | 17 616 | 18 148 | 18 266 | 18 247 | 19 285 | 19 525 | 20 878 | 20 982 | 21 781 |
| Industrial waste (non-renewable) | 5 204 | 7 167 | 6 460 | 1 070 | 1 251 | 839 | 839 | 1 232 | 1 390 | 2 467 | 2 874 | 2 927 | 2 949 | 2 398 | 2 529 | 2 605 | 2 893 | 2 588 | 2 851 |
| Non-renewable municipal waste | 6 403 | 7 137 | 7 391 | 8 406 | 9 283 | 10 994 | 12 297 | 13 022 | 13 582 | 13 501 | 14 742 | 15 221 | 15 317 | 15 849 | 16 756 | 16 920 | 17 986 | 18 394 | 18 930 |

*Source:* Eurostat (online data code: nrg_bal_peh)

**Fig. 3** Gross electricity production by fuel, EU-27, 2000–2018. *Source* Eurostat

Another assumption, we can make is following: suppose we intend to generate our energy for mobility solely from wind energy and we assume that a 3 MW wind turbine on land generates on average 6.5 GWh, then to produce the required 1272 TWh, Europe needs to have **195,692** 3 MW wind turbines, dedicated to mobility. Given the European surface of 4,272,000 km$^2$, this means on average 1 turbine per 22 km$^2$. Since a number of area's are not suited to install turbines, the density in other areas will be much higher. It's clear that the needed energy will have to come from other sources as well.

## 10 Next Steps

### *10.1 Quick Look Backwards*

Now that we've a better view on where we are going at, it's good to ask ourselves: "where are we today?". Quite remarkably, one can find very diverging figures, depending on the source and the interpretation of these figures. In order to create a common ground, the figures we use herewith are coming from the ACEA report[1] "Vehicles in Use in Europe—January 2021". We can consider these figures as recent, objective and undisputable.

When figures are different from these, it is very often due to the fact that PHEV (Plug-in Hybrid Electric Vehicles) are considered as well as "Electric Vehicles", although they aren't. In most cases, drivers are using them the same way they are driving a traditional ICE-vehicle, thus generating 140 g $CO_2$/km or more, instead of the figures, mentioned within the marketing brochures. Hence, these vehicles are NOT a contributor to the Clean Air efforts, we all have to make within the coming decades, … on the contrary! PHEVs are jeopardizing most efforts! We expect them to disappear very rapidly.

But, let's look backwards for a short while and have a look at where it all started. In an article on the website of the US Energy Department,[3] we read: "By 1900, electric cars were at their heyday, accounting for around a third of all vehicles on the road. During the next 10 years, they continued to show strong sales." So, back in 1900, the market share of the BEV was about 33%. Today, in Europe, the actual market share is 0.2%.[1] This is 150 times smaller! Hence, we can only conclude that so far, it went completely wrong.

### *10.2 The Ambition Is to Evolve Towards 100% BEV by 2050!*

When we hear politicians of all color declare that we will have energy neutrality by 2050 and that we need to have an emission free mobility by that moment in time, it's clear that the way ahead will be steep and long. Is the European Green Deal a dream?

Is the Paris Climate agreement an unrealistic ambition? Maybe. In case we continue to tackle the problems the same way, we did during past decades, for sure it will! Hence, we have to change our approach, make a 180° U-turn and make BEV "sexy" enough for citizens to move from ICE towards BEV. That's the real challenge. A challenge that—surprisingly enough—the OEMs didn't discover so far.

## *10.3 How to Close the Gap?*

How to close the gap between 0.2% and 100% … and do this between now and 2050? Let's take a closer look on how a smooth transition between our actual mobility model and what it is supposed to be in 2050 and beyond may look like. In this chapter, we will formulate an answer to the overall sounding question: "How to get there?".

At first, we will come back on Table 1 and elaborate this one a little more in detail. At that table, it was mentioned, and indicated with a (*), that the very last PDCs (Person Driven Cars) might also become Autonomous Driving, hence ending up with a 100% automated fleet on our roads. Although Automation is not the key subject of this chapter, we'll include it for a while and extend our forecast onto 2060, when the last PDC might be converted into its ADEM equivalent and road mobility will be completely automated.

## *10.4 The Evolution of the Fleet on Our Roads Within the Coming Decades*

The different categories of vehicles will be split up a little more in detail in order to have a better view on the individual evolutions (Table 3 and Fig. 4).

*Remark*: We deliberately do not make any projection on growth in population or on increase or decrease in mobility, which both might influence our mobility as well. That's completely out of scope and might being influenced by factors, which are hard to predict. The Corona pandemic and consequent home office labor has made mobility drop significantly in 2020, compared to 2019. In this analysis, total distance/year per person has remained unchanged, as well as the number of vehicles in the B2B-categories. It's only within the B2C-category, that we've projected an increase in efficiency from 2.85% per car towards nearly 100% of efficiency.

**Table 3** Overview of the vehicle mix in in Europe (*source* ACEA—The European Automotive Manufacturers Association) and forecast for an evolution towards a full MaaS and 100% ADEM based mobility model by 2060

| | | | Actual | | Forecast | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 2020 | | 2030 | 2040 | 2050 | 2060 |
| B2C | B2C-ICE | (1) | 222,000,000 | 78.39% | 210,900,000 | 105,400,000 | 0 | 0 |
| | | (2) | 25,000,000 | 8.83% | 23,750,000 | 11,875,000 | 0 | 0 |
| | BEV-CC | (3) | 500,000 | 0.18% | 12,850,000 | 6,425,000 | 0 | 0 |
| | BEV-BS | (4) | | | 0 | 18,300,000 | 36,600,000 | 0 |
| | ADEM | (5) | | | 0 | 4,218,000 | 8,436,000 | 45,036,000 |
| B2B-1 | B2B-1 | (6) | 29,000,000 | 10.24% | 29,000,000 | 14,500,000 | 0 | 0 |
| | | (7) | | | | 14,500,000 | 29,000,000 | 0 |
| | | (12) | | | | | | 29,000,000 |
| B2B-2 | B2B-2 | (8) | 6,000,000 | 2.12% | 6,000,000 | 3,000,000 | 0 | 0 |
| | | (9) | | | | 3,000,000 | 6,000,000 | 0 |
| | | (13) | | | | | | 6,000,000 |
| B2B-3 | B2B-3 | (10) | 700,000 | 0.25% | 700,000 | 350,000 | 0 | 0 |
| | | (11) | | | | 350,000 | 700,000 | 0 |
| | | (14) | | | | | | 700,000 |
| Total | | | 283,200,000 | | 283,200,000 | 181,968,000 | 80,736,000 | 80,736,000 |

(continued)

**Table 3** (continued)

| | Actual | | Forecast | | | |
|---|---|---|---|---|---|---|
| | 2020 | | 2030 | 2040 | 2050 | 2060 |
| PDC | 283,200,000 | | 283,200,000 | 177,750,000 | 72,300,000 | 0 |

(1) ICE cars (convertible from PDC towards ADEM)
(2) ICE cars (convertible from ICE towards battery swap)
(3) BEV cars (from conductive charging towards battery swap)
(4) BEV cars (battery swap)
(5) ADEM (autonomous driving electric mobility)
(6) Light utility vehicles (≤3.5 t) with ICE
(7) LUV conversion from ICE towards PDC with battery swap
(8) Heavy utility vehicles (>3.5 t) with ICE
(9) HUV conversion from ICE towards PDC with battery swap
(10) Busses and touring cars with ICE
(11) B&TC conversion from ICE towards PDC with battery swap
(12) LUV conversion from PDC with battery swap towards ADEM
(13) HUV conversion from PDC with battery swap towards ADEM
(14) B&TC conversion from PDC with battery swap towards ADEM

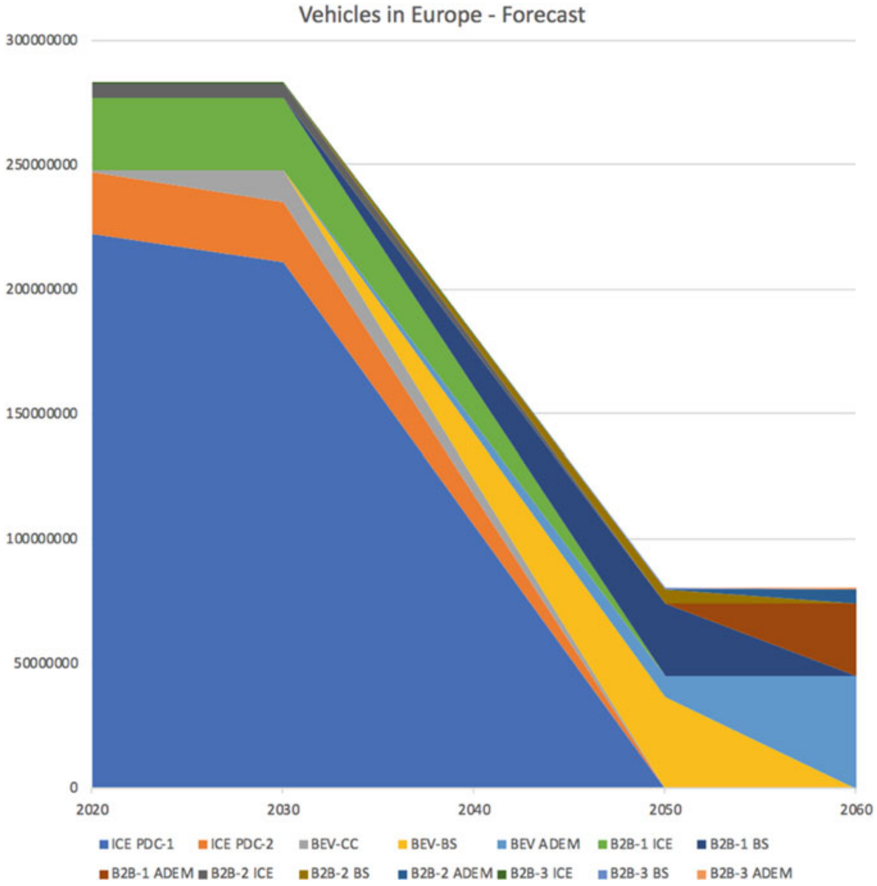


**Fig. 4** Graphical representation of the vehicle mix in in Europe (*source* ACEA—The European Automotive Manufacturers Association) and forecast for an evolution towards a full MaaS and 100% ADEM based mobility model by 2060

## *10.5 B2C Vehicles, Also Called Passenger Cars*

At first, we'll take a closer look at the passenger cars. This includes all type of passenger cars (Table 4):

- Individually owned
- Collectively owned fleets (e.g. company cars, etc.)
- Professional used cars
- Car sharing
- Taxi
- etc.

**Table 4** Overview of the B2C vehicle mix (passenger cars) in in Europe (*source* ACEA—The European Automotive Manufacturers Association) and forecast for an evolution towards a full MaaS and 100% ADEM based mobility model by 2060

| | | | Actual | | Forecast | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 2020 | | 2030 | 2040 | 2050 | 2060 |
| B2C | B2C-ICE | (1) | 222,000,000 | 89.70% | 210,900,000 | 105,400,000 | 0 | 0 |
| | | (2) | 25,000,000 | 10.10% | 23,750,000 | 11,875,000 | 0 | 0 |
| | BEV-CC | (3) | 500,000 | 0.20% | 12,850,000 | 6,425,000 | 0 | 0 |
| | BEV-BS | (4) | | | 0 | 18,300,000 | 36,600,000 | 0 |
| | ADEM | (5) | | | 0 | 4,218,000 | 8,436,000 | 45,036,000 |
| Total | | | 247,500,000 | | 247,500,000 | 146,268,000 | 45,036,000 | 45,036,000 |
| PDC | | | 247,500,000 | | 247,500,000 | 146,268,000 | 36,600,000 | 0 |

(1) ICE PDC-1 cars (convertible from PDC with ICE towards ADEM)
(2) ICE PDC-2 cars (convertible from ICE towards battery swap)
(3) BEV cars (from conductive charging towards battery swap)
(4) BEV cars (battery swap)
(5) ADEM (autonomous driving electric mobility)

We make distinction between cars, used by professionals, whose job is exclusively possible thanks to the use of that car and cars, used by users, be it private or even professional ones, for whom the use of that car is fantastic, but an instantaneous availability is not really necessary. This last type of users might make use of a so-called ***On Demand*** vehicle and is thus candidate to make use of an ADEM car as soon as this one becomes available.

We indicate them as **ICE PDC-1** and indicate them as "***convertible from PDC with ICE towards ADEM***". We estimate their number at roughly 90% of the European B2C-fleet or 222 million. As the ADEM cars will only be available between 2030 and 2040, we have to consider an intermediate step, where a number of them will first migrate from ICE towards BEV-CC (Conductive Charging) in 2030, before being converted into Battery Swap (BEV-BS) from 2030 onwards. That's why we see an increase of BEV-CC in 2030, up to 5% of the global B2C fleet, before being halved by 2040 and finally completely disappear by 2050.

The other group of drivers, indicated as **ICE PDC-2** and estimated at roughly 10% or 25 million cars, will continue to drive an ICE until 2030, before starting to migrate towards BEV. Since we expect at that moment Battery Swap to be available, we proclaim them as "***convertible from ICE towards Battery Swap***". We expect them to be converted for 50% by 2040 from ICE towards Battery Swap, before the ICE finally will completely fade away by 2050.

The 3rd group, indicated as **BEV-CC**, is the group of BEV as we know them actually, equipped with Conductive Charging capability. Despite the limited success of this group so far, we expect them anyhow to grow still within the coming 10 years, before fading away, being halved by 2040 and disappear completely by 2050. At that moment, the charging poles can be removed as well, thus freeing our streets from these hindering installations.

So far, the actually existing categories B2C cars: the ICE vehicles and the BEV-CC.

What about the newcomers: **BEV-BS** and **BEV-ADEM**?

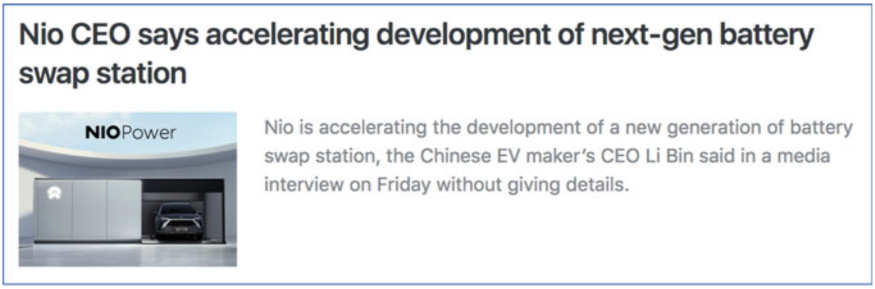
First, we have to clarify that ADEM is equipped with Battery Swap as well, but on top of that is also equipped with Automated Driving, hence the denomination Automated Driving Electric Mobility or ADEM.

Actually, both categories are not yet present on the European market, while Battery Swap is gaining momentum in China with protagonists like NIO, Beijing EV (subsidiary of the BAIC Group), Geely and others.[4] This whole development is pushed forward by China's Ministry of Industry and Information Technology (MIIT) through research at the China Automotive Technology Research Center. It is clear that the Chines EV industry is cementing a leadership position onto the future of the global EV market. Moreover, the Chinese Government recently approved the first official swappable EV battery standard and safety guidelines, which are set to go into effect on Nov. 1, 2021.

Nio even introduced an own Battery as a Service (BaaS) model, comparable to the European Swap2drivE model (Fig. 5).

More recently, Geely has announced the "1-min Swap station" (Fig. 6).

In the event that Europe very soon does realize it has to change strategy, in order to avoid leadership from China and thus embrace Battery Swap and ADEM, we might see a quick take off with roughly 105.5 million ICE passenger cars in Europe



Nio CEO says accelerating development of next-gen battery swap station

Nio is accelerating the development of a new generation of battery swap station, the Chinese EV maker's CEO Li Bin said in a media interview on Friday without giving details.

**Fig. 5** Screen shot from an article on Nio's efforts to accelerate development of the next-gen battery swap station

**Fig. 6** Screen shot from Geely's commercial on their 1-min Battery Swap solution

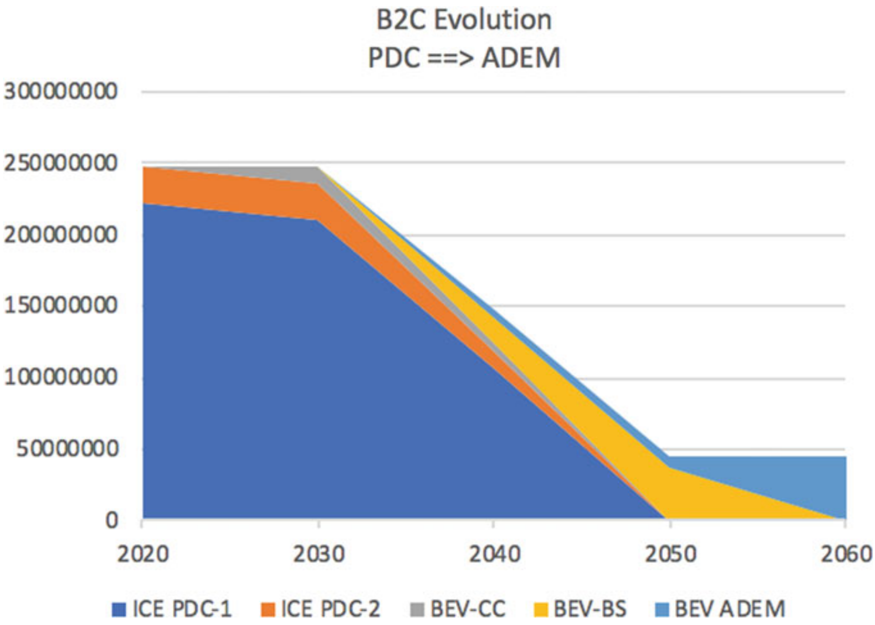


**Fig. 7** Graphical representation of the B2C vehicle mix in in Europe (*source* ACEA—The European Automotive Manufacturers Association) and forecast for an evolution towards a full MaaS and 100% ADEM based mobility model by 2060

converted towards ADEM by 2040 and the remaining 105.5 million finally by 2050. This will result in about 8.5 million ADEM vehicles, since they are available at nearly 100% of the time and thus about 25 times more efficient, as we already mentioned here above.

The ADEM itself, which strongly depends on the progress on Level 5 Automated Driving, may be expected to become certified and start to appear on our roads by 2030 and take-off within the decade that follows. By 2040, we can expect to have already roughly 4 million units on our roads and this will be doubled between 2040 and 2050 (Fig. 7).

Hence, by 2050, we will thus be able to enjoy a 100% clean fleet in our streets, which will be much smaller than today's fleet, due to vehicle usage optimization.

To some, this may look strange, but clearly, this is the only way to get there!

## 10.6 B2B-1 Vehicles, Also Called Light Utility Vehicles (≤3.5 t)

We will now take a closer look at the Light Utility Vehicles, of which there are about 29 million units in Europe.[1] We expect these to evolve from a full ICE fleet towards

**Table 5** Overview of the B2B-1 vehicle mix [light utility vehicles (≤3.5 t)] in in Europe (*source* ACEA—The European Automotive Manufacturers Association) in 2020 and forecast for an evolution towards a fully electrified fleet by 2050 and 100% ADEM by 2060

| | | | Actual | | Forecast | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 2020 | | 2030 | 2040 | 2050 | 2060 |
| B2B-1 | B2B-1 | (6) | 29,000,000 | 100% | 29,000,000 | 14,500,000 | 0 | 0 |
| | | (7) | | | | 14,500,000 | 29,000,000 | 0 |
| | | (12) | | | | | | 29,000,000 |
| Total | | | 29,000,000 | | 29,000,000 | 29,000,000 | 29,000,000 | 29,000,000 |
| PDC | | | 29,000,000 | | 29,000,000 | 29,000,000 | 29,000,000 | 0 |

(6) Light utility vehicles (≤ 3.5 t) with ICE
(7) LUV conversion from ICE towards PDC with battery swap
(12) LUV conversion from PDC with battery swap towards ADEM

a full BEV fleet by 2050, with an intermediate step of roughly 50% penetration in 2040. Ultimately, this fleet might also become Automated Driving by 2060, but that's not the subject of this document, as outlined earlier on.

In or around 2030, some of them might make an intermediate stop at CC, prior to move to Battery Swap, although chances are very little since these are commercially used vehicles and a commercial vehicle, standing still at a charging pole, in order to get energized, is considered as "dead capital" and no single entrepreneur is interested in dead capital. That's why we deliberately ignore this option as the likelihood is really negligible (Table 5 and Fig. 8).

## 10.7 B2B-2 Vehicles, Also Called Heavy Utility Vehicles (>3.5 t)

When looking at the Heavy Utility Vehicles, of which there are about 6 million units in Europe,[1] the situation is even more outspoken. Likewise, they will evolve from a full ICE fleet towards a full Electric fleet by 2050, with an intermediate step of roughly 50% penetration in 2040. Ultimately, this fleet might also become Automated Driving by 2060, once all other vehicles are becoming Automated Driving, but that's not the subject of this document, as outlined earlier on.

In this category, the likeliness of an intermediate stop at CC, prior to move to Battery Swap, is for sure not going to happen, since these are 100% commercially used vehicles and a commercial vehicle, standing still at a charging pole, in order to get energized, is "dead capital". No single entrepreneur will invest in dead capital. That's why this option is completely ignored, since the likelihood is really nonexistent (Table 6 and Fig. 9).

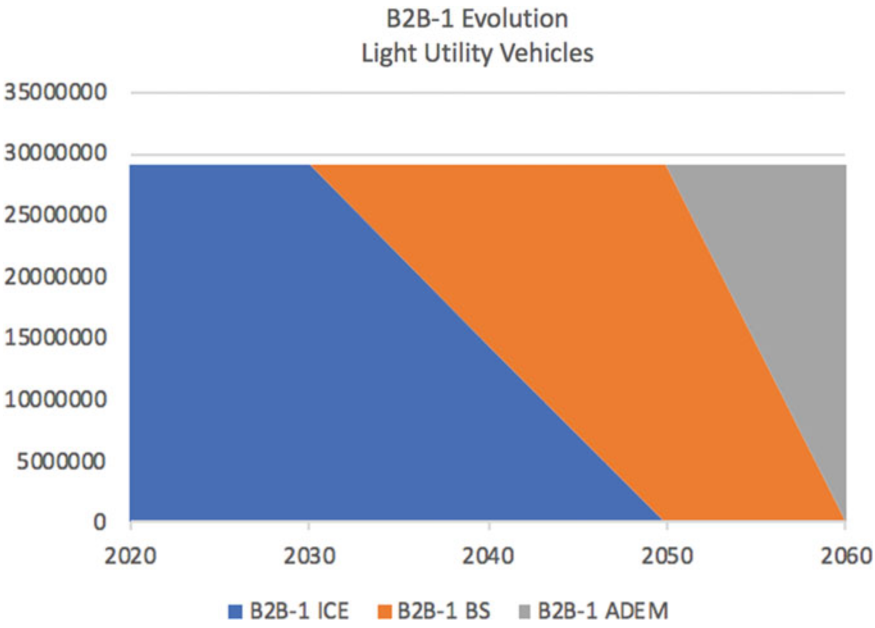


**Fig. 8** Graphical representation of B2B-1 vehicle mix in in Europe (*source* ACEA—The European Automotive Manufacturers Association) and forecast for an evolution towards a full electric fleet by 2050 and 100% ADEM based mobility model by 2060

**Table 6** Overview of the B2B-2 vehicle mix [heavy utility vehicles (>3.5 t)] in Europe (*source* ACEA—The European Automotive Manufacturers Association) in 2020 and forecast for an evolution towards a fully electrified fleet by 2050 and 100% ADEM by 2060

| | | | Actual | | Forecast | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 2020 | | 2030 | 2040 | 2050 | 2060 |
| B2B-2 | B2B-2 | (8) | 6,000,000 | 100% | 6,000,000 | 3,000,000 | 0 | 0 |
| | | (9) | | | | 3,000,000 | 6,000,000 | 0 |
| | | (13) | | | | | | 6,000,000 |
| Total | | | 6,000,000 | | 6,000,000 | 6,000,000 | 6,000,000 | 6,000,000 |
| PDC | | | 6,000,000 | | 6,000,000 | 6,000,000 | 6,000,000 | 0 |

(8) Heavy utility vehicles (>3.5 t) with ICE
(9) HUV conversion from ICE towards PDC with battery swap
(13) HUV conversion from PDC with battery swap towards ADEM

A Heavy Utility Vehicle will contain multiple 30 kWh batteries, thus giving it an autonomy of several hundreds of km, which is sufficient for a full working day (Fig. 10).[2]

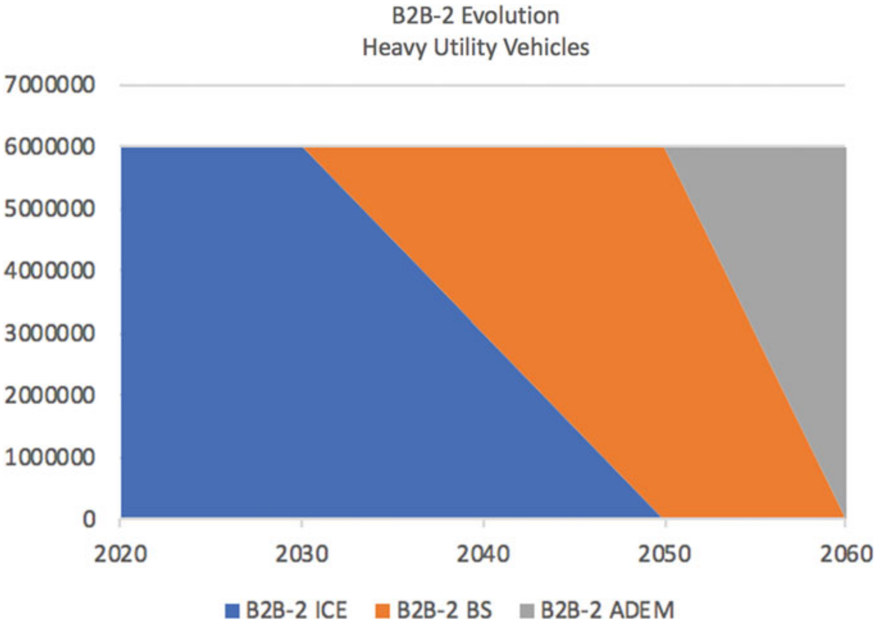


**Fig. 9** Graphical representation of B2B-2 vehicle mix () in Europe (*source* ACEA—The European Automotive Manufacturers Association) and forecast for an evolution towards a full electric fleet by 2050 and 100% ADEM based mobility model by 2060

### *10.8 B2B-3 Vehicles, Also Called Busses and Touring Cars*

While Busses have completely different needs than Touring Cars, they mostly are accumulated within the same category. The total number for Europe is about 700,000 units.[1]

A Bus might need an autonomy of a few hundred km/day, while a Touring Car needs a lot more. In some cases, when it's driven by a team of drivers, this can go up to 2000 km within 24 h. Those Touring Cars, once they evolve towards electric propulsion, not only need "swappable drivers" but also swappable batteries.

For both categories, multiple technical implementations are possible.[2] Here, we'll only look at the numbers (Table 7 and Fig. 11).

## 11 Which Are the Success Factors?

We've asked ourselves: "Where are we today?" and the hard but honest answer is "Almost nowhere!". It's hard to attribute to a market penetration of only 0.2% a more appealing predictive than this one. One can ask at the very same moment "How could it come so far?".

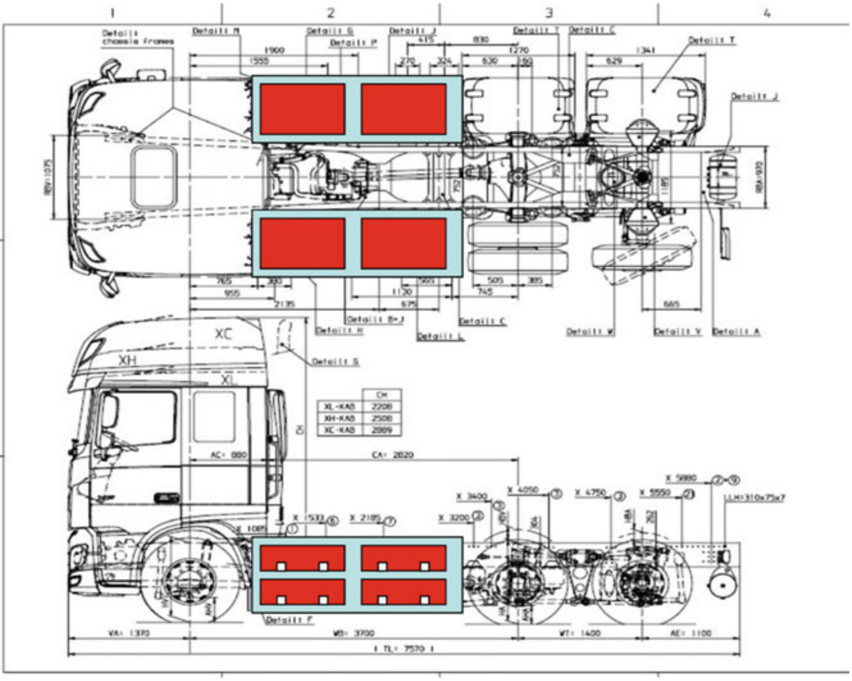


**Fig. 10** A heavy utility vehicle, transformed to contain up to 8 × 30 kWh swappable batteries, giving it an autonomy of about 240 km after a full swap. *Courtesy* DAF Trucks

**Table 7** Overview of the B2B-3 vehicle mix (busses and touring cars) in Europe (*source* ACEA—The European Automotive Manufacturers Association) in 2020 and forecast for an evolution towards a fully electrified fleet by 2050 and 100% ADEM by 2060

| | | | Actual | | Forecast | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 2020 | | 2030 | 2040 | 2050 | 2060 |
| B2B-3 | B2B-3 | −10 | 700,000 | 100% | 700,000 | 350,000 | 0 | 0 |
| | | −11 | | | | 350,000 | 700,000 | 0 |
| | | −14 | | | | | | 700,000 |
| Total | | | 700,000 | | 700,000 | 700,000 | 700,000 | 700,000 |
| PDC | | | 700,000 | | 700,000 | 700,000 | 700,000 | 0 |

(10) Busses & touring cars with ICE

(11) B&TC conversion from ICE towards PDC with battery swap

(14) B&TC conversion from PDC with battery swap towards ADEM

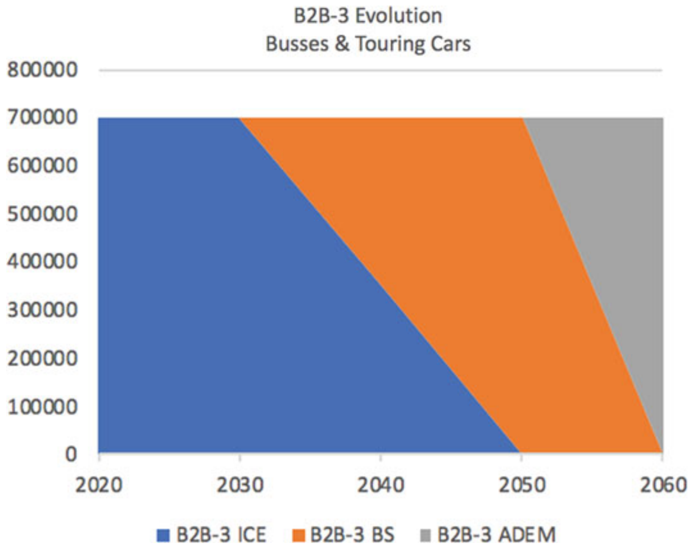


**Fig. 11** Graphical representation of B2B-3 vehicle mix in in Europe (*source* ACEA—The European Automotive Manufacturers Association) and forecast for an evolution towards a full electric fleet by 2050 and 100% ADEM based mobility model by 2060

But then pops up next question: "What can we do to improve this situation?" and "How do we intend to reach the objectives of the EU Green Deal and the Paris Climate agreement?".

Here are the answers:

On the 1st set of questions, as said earlier, the OEMs didn't discover so far how to get their offering appealing and sexy. Indeed, up to now, the guys and girls in the executive boardrooms didn't go out and talk with their prospects and clients. They only looked up their spreadsheets and decided to add another year "of the same stuff" on top of it. Pushed by the governments, they only found out some software tricks, now indicated as Dieselgate or published some misleading documents, indicated as Astongate or buy "regulatory credits" at Tesla. By the way, Tesla generated in 2020 a net profit of 438 million $, coming from these regulatory credits, that they sold to other OEMs.

When you hear OEMs saying: "Citizens have **fear** to move towards EV", one should correct this phrase and tell them, they have to say: "Citizens have **fear** to move towards **our** EV". They should stop blaming the citizen, but instead start talking to him or her, start listening and once they've clearly understood the requirements, return to the drawing tables and design an appealing and sexy BEV.

Here's why! Up to now, all BEVs have the same inconveniences:

1. Their **range** is limited. In order to overcome this inconvenience, they bring more battery capacity on board, thus increasing the weight and by consequence the energy consumption/km, as well as the price. On top of that, heavy vehicles

produce more Black Carbon and Fine Dust ($PM_{10}$ and $PM_{2.5}$) compared to lighter vehicles.

2. The **charging time** still is a huge inconvenience, since "Time = Money". Even for unemployed or retired people, time still = money. The OEMs, jointly with the "charging pole lobby", are increasing the charging capacity per charging pole up to fast chargers, who charge at 150 kW and more. This is killing the lifetime of the battery, which in the end is by far the most expensive component of the vehicle. On top of that, it's killing the energy grids as well. Moreover, the hassle of handling a cable in all weather conditions is also a burden.
3. The **acquisition cost** of such a BEV-CC is very high. The more battery capacity they install, in order to cope with the range issue, the more expensive the vehicle becomes. Hence a BEV-CC is much more expensive than an ICE equivalent car. Subsidies might help for a while, but in the end, these are coming from taxes and excises. Hence, this is a Shifting of Funds transaction.

On the 2nd set of questions, the answer is quite obvious as well: by migrating towards a Battery Swap model. In that case, all the inconveniences, mentioned here above, are disappearing at once and new business models are becoming available, like Battery as a Service (BaaS). Moreover, the Battery Recharging Stations (BRS) will act as energy storage hubs within the Smart Energy Grids, once they're fully powered by Renewable Energy (Solar, Wind, etc.). At the same time, these hubs can be used by the Energy Utilities to perform Peak Shaving, at moments when Demand is higher than Supply. It can even be used to offer Frequency Regulation as well, thus allowing energy utilities to keep full control on the grid.

## 12 Would You Buy an EV from Sony[5]?

This question may surprise, but it's the title of an article, published recently on Autoweek.com.[5] The barrier between electronics giants and automakers might soon fade away, as the industry turns to EVs. Indeed, after newcomers like Tesla and Nio and after also Apple announced having some interest in becoming a Car Producer and Seller, other companies, like Sony, are knocking on the door and announcing their interest in this new and very promising BEV opportunity. A car nowadays is becoming more and more an "ICT-solution on wheels", rather than a vehicle and since those companies have learned to listen to their clients and seduce them with equipment that responds to their needs, they might have a good chance to beat the traditional OEMs onto their own battleground. Very surprising, but inspiring at the time.

## 13 Conclusions

Mobility will be quite different in 2050, compared to the one we know and use actually. Changes will take place within the coming years in order to evolve towards a 100% green mobility. How this will look like, is hard to predict. Will it be the one, that we have described herewith? Nobody knows. But this is for sure: in case we move towards a 100% electric mobility, in order to make it viable and economically feasible, we have to anticipate a number of evolutions:

- We have to anticipate huge additional energy production capacity. The actual energy production in Europe of 2941 TWh will get an additional load of nearly 43%, in order to satisfy the upcoming mobility needs.
- Automation is gaining momentum everywhere. Mobility will follow as well:
  - Automated Driving will become standard
  - Automated battery swapping will also become mainstream
- Once this automation realized, a multimodal MaaS concept can be put in place, in which moving from A to B will be an enjoyment and idle time of equipment will be reduced as much as possible. Hence, mobility will become much cheaper and thus more affordable for everyone.

We're living a pivotal moment in history at which the future of mobility is taking shape. But prior to that, industry and government jointly have to take a number of bold decisions.

### Notes

1. Figures, coming from ACEA Report "Vehicles in Use in Europe"—January 2021. https://www.acea.be/uploads/publications/report-vehicles-in-use-europe-january-2021.pdf.

   **Remark**: note that figures from European Countries Bulgaria, Cyprus and Malta are missing within the ACEA report. We have calculated them, based on the average number of vehicles per capita within the remaining 24 countries of the EU, thus giving us a quite correct number of the total amount of vehicles, available within the EU-27.
2. EV 2.0 Electrical Vehicle: Friend or Foe of Smart Grids?—ir. Jacques De Kegel. https://www.linkedin.com/in/jacques-de-kegel.
3. Website of the US Energy Department about the History of the Electric Car. https://www.energy.gov/articles/history-electric-car.
4. China's Battery Swap Trend is Way Ahead. https://guidehouseinsights.com/news-and-views/chinas-battery-swap-trend-is-way-ahead.
5. Would you buy an EV from Sony? https://www.autoweek.com/news/green-cars/a35226640/would-you-buy-an-ev-from-sony/.

## References

1. De Kegel J (2020) Rechargeable battery for an electric vehicle, vehicle comprising such a battery, method and system for replacing a rechargeable battery of an electric vehicle; and system for transferring energy to the battery. European Patent 2,543,093, 12 Feb 2020
2. De Kegel J (2019) EV 2.0—electrical vehicle: friend or foe of smart grids? v2019-06-03. https://www.linkedin.com/in/jacques-de-kegel/
3. De Kegel J (2020) Mobility in 2050—Part I. https://www.linkedin.com/pulse/mobility-2050-jacques-de-kegel. Accessed 7 Dec 2020
4. De Kegel J (2021) Mobility in 2050—Part II. https://www.linkedin.com/pulse/mobility-2050-part-2-jacques-de-kegel. Accessed 20 Jan 2021
5. De Kegel J (2021) Mobility in 2050—Part III. https://www.linkedin.com/pulse/mobility-2050-part-iii-jacques-de-kegel. Accessed 5 Feb 2021
6. De Kegel J (2021) Imagine you can swap … https://www.linkedin.com/pulse/imagine-you-can-swap-jacques-de-kegel. Accessed 2 Mar 2021
7. De Kegel J (2021) Mobility in 2050—EU-27. https://www.linkedin.com/pulse/mobility-2050-eu-27-jacques-de-kegel. Accessed 18 Jun 2021
8. De Kegel J (2021) How to become "Fit for 55"? … https://www.linkedin.com/pulse/how-become-fit-55-jacques-de-kegel. Accessed 4 Dec 2021
9. De Kegel J (2021) Swap2drivE—a systemic approach-V2.1. https://www.youtube.com/watch?v=1NJxJee9T1o. Accessed 1 Dec 2021
10. Corfitsen S (2017) System and method for performing payments from a vehicle. US Patent 2,017,076,290 A1
11. Corfitsen S (2016) Method and device for automatic refuelling of vehicles. US Patent 2,016,251,213 A1
12. Corfitsen S (2016) Method and device for replacement of a battery in a vehicle. US Patent 2,016,221,543 A1
13. Corfitsen S (2015) Device and method for replacement of batteries in battery driven vehicles. US Patent 2,015,129,337 A1
14. Corfitsen S (2013) Method for exchanging batteries in battery-powered vehicles. US Patent 2,013,104,361 A1
15. Corfitsen S (2009) Method for payment in connection with automatic fuelling of vehicles. WO Patent 2,009,072,973 A1
16. Corfitsen S (1995) Method and arrangement for automatically refueling automotive vehicles. US Patent 5,393,195 A
17. Corfitsen S (1993) Apparatus for the automatic fuelling of automotive vehicle. US Patent 5,393,195 A
18. Corfitsen S (1997) Apparatus for automatic refuelling of vehicles. US Patent 5,671,786 A
19. Corfitsen S (1997) Apparatus for automatic refuelling of vehicles. US Patent 5,638,875 A
20. Corfitsen S (2000) Arrangement at airplane seats. WO Patent 0,038,986 A1
21. Corfitsen S (1999) Device for automatic refuelling of vehicles. WO Patent 9,959,912 A1
22. Corfitsen S (1998) Device for positioning at automatic fuelling of vehicles. WO Patent 9,854,084 A1
23. Corfitsen S (1998) Device for automatic fuelling of vehicles. WO Patent 9,854,083 A1
24. Corfitsen S (1996) Arrangement for docking at automatic fuelling of vehicles. WO Patent 9,605,136 A1
25. Corfitsen S (1996) Adapter for automatic fuelling of vehicles. WO Patent 9,605,135 A1
26. Hilders HG. Collapsible lightweight air cargo container. US Patent 7,681,752 B2
27. Hilders HG (1991) The Advanced Sandwich Panel. Paper presented at the University of Ferrara, at the celebration of its 600th anniversary. Refers to EU Eureka Grant 272
28. Bont WS, Blumendal H, Hilders H et al (1965) Results of experiments to isolate membrane bound ribosomes from rat liver. Biochim Biophys Acta

29. Hilders HG (1988) The Advanced Sandwich Panel Concept. The application of multi-axial warp knitted fabrics in aerospace components. Eureka 272, presented at the annual conference of Ministers of Economy in Copenhagen
30. Hilders HG (2009) Conversations for a smarter planet. Paper: Innovative Air Cargo Logistics. IBM, Brussels
31. Koo V, Hilders HG (2019) The intelligent air cargo container. Presented at the sixth China International Logistics Development Conference in Shijiazhuang, PRC
32. Hilders HG (2019) The intelligent air cargo container. Presented at TUSPARK Innovation Research Institute, Nanjing, PRC. Publicly available preprint
33. Hilders HG (2014) The intelligent air cargo container. CITRIMACC, the approved proposal for the EU Commission under the Horizon 2020 program. Online Document
34. Hilders HG (2016) Cargobox, the Intelligent Multi-modal Air Cargo System, Interbox International BV website: http://cargo-box.com
35. Hilders HG (2019) Coolboxx, an integrated solution for temperature-controlled air cargo shipment of temperature-sensitive products, such as medicines, pharmaceuticals, and perishables. Interbox International BV website: http://cargo-box.com
36. Hilders HG (2020) Controlled shipping of COVID-19 vaccines. Summary, disclosed to relevant parties in the Logistics Supply Chain

## *Hashtags*

#futureofmobility #automotive #innovation #sustainability #technology #mobility #elektrischrijden #electromobility

**Jacques De Kegel** is a Belgian citizen, born in 1954 in Aalst. He graduated as Master in Engineering Sciences at the KU Leuven University in Leuven (Belgium). He worked for IBM for 33 years, deepened his knowledge with additional studies at London Business School and Boston University and finally became a mobility expert within the Emerging Business Opportunities and Smarter Cities Solutions teams. He is patent holder on the universal Swap2drivE Battery Swap solution. After his retirement, he runs his own engineering company DK Engineering & Services.

**Sten Corfitsen** is a Swedish citizen born 1952 in Stockholm with a MSc degree from Royal Institute of Technology. With background from a inventor family with creations as the pace maker, the inkjet print technology, Sten has innovations in his blood. After having built his first flying gyrocopter followed by a number of employments at IBM and Philips, Sten became the pioneer with developing automatic fueling of ICE cars (Fuelmatics). Lateron he spun-off Powerswap AB, a company dedicated to launch battery swapping as the solution to speed up the transition to electric vehicles. He has extensive experiences with automation, the innovation process and to introduce new technology in conservative industries.

**Henk G. Hilders** (1940) studied Biochemistry and Medicine and worked for almost 10 years in a Coronary Care Unit of an Amsterdam NL hospital. He made a career switch in 1986 into advanced composites with focus on aerospace applications. He has several patents on his name, the first one, based on multi-axial warp knitted fabrics received an EU grant in 1988.

He invented the Cargobox in the early years of the 21st Century and is still working on improvements and technical additions.

# Smart Transport as an Enhancement of the Urban Infrastructure

Ekaterina Zakharova, Inna Minashina, and Fedor Pashchenko

**Abstract** The concept of a Smart City is to use the existing resources in an optimal way to provide the greatest convenience to its residents. This requires close integration of all components, for example, street video surveillance, public services, intelligent transport systems and others, on the scale of a megalopolis. Every year, the world's megacities are becoming more comfortable for residents due to the introduction of newest technologies. First, such technologies include intelligent control systems in the transportation field. The main goals of Smart Transportation are the efficient and coordinated movement of people, monitoring the location of objects, fast and reliable interaction of vehicles with each other, as well as guaranteeing road safety. This paper represents examples of artificial intelligence technologies and optimization methods applications to create such smart systems.

**Keywords** Smart city · Smart transport · Artificial intelligence · Vehicle rescheduling problem · Forecasting · Railway traffic

## 1 Introduction

A smart city is the integration of information and communication technologies for the management of urban infrastructure: transport, education, healthcare, networks serving residential and security. The goal of a smart city is effective governance and a high standard of living for the population using innovative technologies. The smart city infrastructure consists of the following elements:

E. Zakharova (✉) · I. Minashina
Moscow Institute of Physics and Technology, Moscow, Russia
e-mail: zakharova.em@mipt.ru

I. Minashina
e-mail: minashina.ik@mipt.ru

F. Pashchenko
Institute of Control Sciences, Moscow, Russia
e-mail: feodor@ipu.rssi.ru

Y. A. Vershinin et al. (eds.), *Technologies for Smart Cities*,
https://doi.org/10.1007/978-3-031-05516-4_7

- Water supply and energy supply management—Smart housing and communal services
- Waste Management—Smart Garbage;
- Ensuring the mobility of citizens within the city—Smart transport;
- Digitalization and provision of reliable communication;
- Citizens' participation in city management—e-government;
- Environmental protection—control of pollution and noise levels, creation of "green" neighborhoods;
- Safe City—ensuring the safety of citizens;
- Affordable e-education and healthcare—smart healthcare, telemedicine, distance learning.

Smart transport is a common name for all types of vehicles using modern communication technologies for efficient movement of people, location monitoring, the interaction between vehicles and other traffic elements, improving the environmental friendliness of transport and the safety of road use in general [1]. Because of the increasing number of vehicles on the roads, there are problems of congestion and irrational use of road resources, which entail an increase in travel time, the amount of fuel consumed, and emissions polluting the environment.

The main trend of creating smart transport systems is the creation of intelligent automated control systems, which are able not only to find optimal solutions for emerging traffic situations but also to conduct analysis to identify "bottlenecks" of technological processes on the roads.

### 1.1 Artificial Intelligence for Smart Transport

Trends in the introduction of modern technologies in the field of train traffic control are dictated by changes in the principles of implementing responsible technological processes aimed not just at improving all components of the railway infrastructure and rolling stock, but also at obtaining maximum effect for all participants in the transportation process [2, 3]. The most relevant areas of smart transport development are the active digitalization of archived traffic data, as well as the use of big data and artificial intelligence. More and more intelligent systems are being developed that provide automation and decision support at all stages of the organization of transportation.

In the works [4, 5], the most actual tasks of Smart Transportation are highlighted, such as automation of the main control functions; the using analysis and Big Data to find solutions; the transition to intellectual planning of maintenance; the development of expert systems capable of finding solutions in emergency situations. Most of the technological tasks solved in railway transport are quite complex, and their full formalization is almost impossible. It is almost impossible to automate such processes using only standard analytical methods and linear algorithms. Therefore, there was a need to develop and apply new methods capable of solving these problems. Such approaches, for example, are AI methods and expert systems [6, 7].

## 2 Practical Methods of Problem Solving in Smart Transport Systems

Improvement of the economic efficiency of any transport system is impossible without solving the problem of automatic train schedule planning and managing the maintenance of traction facilities. In order to ensure the correct operation of the transportation process, it is necessary to create appropriate conditions for the maintenance of the operational fleet, guaranteeing its supply with the required amount of traction resources in the prescribed time. In turn, when planning the maintenance of traction resources, an important condition is the behavior of traffic flows affecting the simulated system. These three tasks are interrelated and underlie the creation of modern intelligent smart transport systems.

### *2.1 Forecasting Problem of the Traffic Flow*

Many studies have been conducted on modelling the traffic flow. Regression models, probabilistic methods and others were used for this purpose. However, the existing methods are no longer so effective due to the lack of accurate models in this area. As an alternative, methods from the field of artificial intelligence can be used to model and predict such flows. They can be considered as more flexible, reliable and suitable for solving problems in conditions of great uncertainty, in comparison with conventional deterministic statistical approaches. In addition, the combination of various artificial intelligence methods and optimization methods can lead to behavior that is more efficient and give greater flexibility when working with real and large-scale tasks.

If the transport flow is described as an incompressible fluid, then the main predicted macro parameters should be:

- Flow is the total number of people on the train per time unit.
- Density is the number of people on each unit of the train.

Based on this, the density of the passenger flow will be the output parameter for our model.

The problem under study as a whole can be formulated as follows: the density of the flow of passengers for the previous period for this train is given. It is required to give a forecast for the density of the flow of passengers in the current time period. The flow density $N(t)$ is the number of passengers on the train at the current time t per unit area. Based on these results, the amount of traction resources required in the current time interval is optimized.

Most control systems of modern technological processes are built using mathematical modelling and forecasting. Many technological processes and systems have important features: complexity, nonlinearity, poor knowledge of the connections within the system, high inertia, the presence of lag and interference, unsteadiness, etc. These uncertainty conditions lead to the fact that the standard deterministic approach to modelling may become unacceptable. One of the approaches to solving this problem is to attract high-quality information [8]. In this regard, fuzzy logic and neural networks will be used along with analytical algorithms to construct the rules of conclusions of the decision-making system. In this paper, a fuzzy difference model TSK is used to model transportation processes-a model built based on the statistical fuzzy model Takagi (Takagi) and Sugeno (Sugeno), called the TS model:

$$R^{\theta} : if\ v_1(t)\ is\ V_1^{\theta}(t),\ \ldots,\ v_{m+l+1}(t)\ is\ V_{l+m+1}^{\theta}(t),$$

$$then\ N^{\theta}(t) = w_0^{\theta} + \sum_{j=1}^{m+l+1} w_j^{\theta} v_j(t),\ \theta = \underline{1,\ n}$$

where $V_k^{\theta}$—fuzzy sets characterized by membership functions $V_k^{\theta}\left(v_k, d_k^{\theta}\right)$.

The advantages of the TS model are as follows: parametric identification can be used for a fuzzy dynamic model; it describes nonlinear dynamic processes with high accuracy; the averaging properties of the output mechanism and the specific type of membership functions provide high noise immunity of the model [9].

In the fuzzification block, the values of variables $v_1(t), v_2(t), \ldots, v_{M+L+1}(t)$ are converted into a matrix

$$V = \begin{bmatrix} V_1^1(v_1(t)) & \cdots & V_{m+l+1}^1(v_{m+l+1}(t)) \\ \vdots & \ddots & \vdots \\ V_1^n(v_1(t)) & \cdots & V_{m+l+1}^n(v_{m+l+1}(t)) \end{bmatrix}$$

In the fuzzy inference block, the value of the truth of the θ-th rule is calculated by the formula

$$\beta^{\theta}(t) = V_1^{\theta}(v_1(t)) \oplus V_2^{\theta}(v_2(t)) \oplus .. \oplus V_{l+m+1}^{\theta}(v_{l+m+1}(t))$$

from where the fuzzy function is determined

$$u^{\theta} = \frac{\beta^{\theta}(t)}{(\beta^1(t) + \beta^2(t) \ldots + \beta^n(t))} +, \theta = 1, \ldots n$$

In the defuzzification block, the specific value of the $\mathrm{N_t}$ output is determined by the formula (1).

### 2.1.1 Predicting Traffic Flow Approach

Hybrid approach of model identification is proposed to use structural and parametric identification [10, 11]. Structural identification is the determination of the number of rules and the order of a fuzzy model. Parametric identification—determination of coefficients of linear difference equations, as well as parameters of membership functions. The stochastic approximation method is used as a parametric identification algorithm [12].

Initial data for the algorithm: the number of rules $n$, the order $r$, $s$ of the fuzzy model, the value of the coefficient vectors c and the parameters of the membership functions. The criterion for stopping the algorithm is the average modular error:

$$J(c) = \frac{1}{T}\sum_{t=1}^{T}\left(\left|N(t) - \hat{N}(t)\right|/N(t)\right)$$

The operation of the entire hybrid algorithm ends when J < 0.03.

### 2.1.2 Passenger Traffic Simulation

Training data is needed to start the hybrid algorithm. Input data for the hybrid algorithm: time for prediction t, the flow density at the moment t − 1 and t − 2, maximum number of standing and sitting places. Membership functions and a set of fuzzy rules are calculated for each of these parameters.

Forecasting the density of passenger traffic over the entire time period of the selected train determines the optimal planning of transportation, as well as the necessary amount of trains and wagons to ensure the transportation process. The train and wagon numbers may vary depending on the magnitude of the flow density.

This model can be modified in the following ways. It is possible to train the system to predict the density of passenger traffic not only by the time periods of the day but also by the days of the week, as well as by various seasons. This is applicable when the flow changes depending on the day of the week and the period of the year. For example, during the summer season, the number of people travelling from the city to the region for the weekend increases dramatically. Therefore, it is worth running longer trains on Friday evening from the city and on Sunday evening to the city.

Thus, the proposed model makes it possible to describe the behavior of traffic flows with sufficient accuracy using intelligent systems. This advantage can be used to solve the traction resources control problem on the transport network. Especially in the case when the transport situation is changing dramatically and it is necessary to find the optimal solution according to some parameters.

### 2.1.3 Using a Passenger Flow Simulation Algorithm to Regulate the Interval of Movement and the Number of Subways Trains

Planning the schedule for the transportation process is the basis for correct work of all transport systems, including the subway. At the same time, this process always requires compliance with a large number of technological factors, such as ensuring passenger transportation demand, traffic safety, efficient use of the capacity of sections and stations, rational use of traction resources. The main global trends in the development of rail transport systems suggest an increase in the automation of the transportation process. One of the profitable improvements in this area can be calculating the required number of trains and the time of their departure from the depot for efficient transportation of all passenger traffic.

The discussed above method of passenger traffic forecasting allows one to solve the problem of smart predicting the number of trains and their departure time from the depot for optimal covering the entire passenger traffic, taking into account its intensity and unevenness. The implemented in such a way smart transport management complex can automatically coordinate, by means of a schedule, the processes of movement and departure at stations, the delivery, and dispatch of rolling stock from the depot, taking into account the specified pair and passenger traffic on the line.

Thus, the proposed model provides an opportunity to solve the specified problem of regulating the interval of movement and the number of trains, from which we can conclude about the potential of using the built systems in this area.

### 2.1.4 Providing Railway Traffic with Transport Resources Problem

Planning the Locomotives Fleet Quantity for a Specified Period

To solve this problem, it is necessary to automate the calculation of the forecast locomotive number and build a plan of operations to ensure this at a specified time interval. All calculations are carried out separately for each directorate (except the case of moving locomotives) [13]. Locomotives differ in the type of traction and the type of movement. The following information is received at the input of this module:

- The number of the required locomotive for the period by directorate and type;
- Locomotives with their attributes and current location;
- The railway graph with its parameters.

The main steps to bring the number of the locomotive fleet to the required are:

**Stage 1**. The number of the required locomotive is calculated as following (given the percentage of defective locomotives for the period)

$$N_{plan} = N(k + 1),$$

where N is the number of the required locomotives, k is the factor accounting for defective locomotives (for the past year).

**Stage 2**. The forecasted locomotives number is calculated and their state (for example, for which of them will be decommissioned):

$$N_{cur} = N - N_{cons} - N_{dec} + N_{bt}$$

where $N$ is the total number of all locomotives of this directorate at the moment; $N_{cons}$ is the number in unexploited fleet; $N_{dec}$ is the number for decommissioning or transferring into the unexploited fleet; $N_{bt}$ is the number for purchasing.

Therefore, the directorate state is calculated at the period start time, taking into account factors affecting the locomotive number. Further, if there is a deficit in any directorate, a certain list of operations is being taken.

### Multi-agent Approach for the Locomotive Relocation Task

To solve this problem, the following algorithm was compiled based on technological experts' experience. The algorithm describes all available activities, sorting and search parameters. The architecture of this system based on multi-agent technologies, in which the following types of agents were defined:

- The planner is a central agent that synchronizes the work of all other agents, manages the sequence of events, and processes input and output information;
- The directorate is regional directorates' agents, which contains information about the locomotive fleet and their technical state, and the behavior of the agent when receiving various commands from the main one.

Since the system has multiagency, each agent in it strives to achieve its specific goal, using the strategies laid down in it to achieve him or her, but at the same time taking into account the main limitations of the entire system. Almost all operations for the shortage or getting rid of the excess are carried out within the directorate. However, there is a locomotive relocation from. This task was reduced to a transport problem and its solution using an auction algorithm.

### The Solution to the Transportation Problem for the Locomotives Relocation

The task is formulated as follows: it is necessary to make an optimal plan for the relocation of locomotives from one directorate to another.

To solve it, the algorithm of asynchronous parallel auctions was used under the condition of uniformity of objects, described in [14]. This method determines flows that maximize the overall utility of moving similar objects at a given cost of transporting one unit from supplier $i$ to consumer j, varying flows. Each supplier and consumer is assigned an intelligent agent responsible for storing internal information for him and exchanging information with other agents and a managing agent to coordinate the work of other agents.

Directorates with a deficit are consumers in terms of the auction method, and directorates with excess are suppliers. Let's denote the variables s—flow, u—value of utility function. Consumers contain information about a set of already distributed pairs $\langle s, u \rangle$, upon receipt of which suppliers select locomotive available for relocation. However, there are often cases of "competition" of directorates with a deficit for the same locomotives. In this case, it is necessary to determine where it is more profitable to send this transport unit. This decision is made based on the calculation of the new utility value according to the following formula:

$$u_i^{new} = u_i^{old} + b_{ij} + \varepsilon,$$

where $u_i^{new}$ and $u_i^{old}$ are the new and old values of the utility function, respectively, $\varepsilon$—is an arbitrarily small value, and bij is the difference between the best choice and the choice that is second after the best. This value is constantly growing, and then eventually this process stops. Huang's method is used to estimate the completion phase of the algorithm.

The algorithm main steps are the following:

Step 1. Initialization. Each participant contains information about the pairs < s, u > . At the initial moment of time, there is only one such pair, the utility of which is 0. With changes in the course of the algorithm, the utility and power change, and at the same time this participant sends the rest an updated set of pairs < s, u > indicating its own unique identifier.

Step 2. Formation of applications. If there are unallocated repairs at some point in time, then the process of creating and sending applications to all repair companies takes place. After that, the process of calculating the utility function is carried out, as a result of which flows are formed, sent to repair companies with the selected values of the utility function and repair ids.

Step 3. Processing of applications. Each repair company, upon receipt of such a flow, selects repairs for the distribution of its repair capacity. After that, messages are sent about updating information about the flows and values of the utility function of the assignment for them until either all repairs are placed or all capacities are filled.

## Algorithm Testing Results

The result of solving this transport problem is the optimal distribution of locomotives from the directorates with an excess to the directorates with a disadvantage, indicating specific locomotives that allow you to bring the current number of locomotives to the required in the forecast period. Table 1 shows an example of the result of the algorithm for a year.

**Table 1** Results of the system of construction and distribution of repairs

| Input data | | | |
|---|---|---|---|
| Region | Required quantity | Forecasted quantity | Deficit/excess |
| Reg_1 | 1061 | 1214 | 153 |
| Reg_2 | 1290 | 1118 | -172 |
| Reg_3 | 1601 | 1674 | 73 |
| Reg_4 | 1664 | 940 | -724 |
| *Output data* | | | |
| Reg_1 | 1061 | 1214 | 0 |
| Reg_2 | 1290 | 1118 | 0 |
| Reg_3 | 1601 | 1674 | 0 |
| Reg_4 | 1664 | 940 | 0 |

**Table 2** Results of the locomotive fleet planning for a specified period

| Region | From an unused fleet | Relocation | | ESLR | Decommission |
|---|---|---|---|---|---|
| | | Receiving | Sending | | |
| Reg_1 | 6 | 0 | 293 | 130 | 0 |
| Reg_2 | 2 | 164 | 0 | 6 | 0 |
| Reg_3 | 6 | 0 | 86 | 7 | 0 |
| Reg_4 | 8 | 215 | 0 | 0 | 0 |

The required quantity and the forecasted fleet quantity are estimated in locomotives. The deficit and excess of the fleet is estimated by taking into account the purchase of new locomotives, the decommissioning of old ones and providing an extended service life repair (ESLR). In addition to this, the following information is displayed for each operation performed (Table 2).

In addition, for each event and each directorate, there is an opportunity to view a list of locomotives that need to hold an event, indicating the numbers of locomotives and their parameters. At the same time, the user has the opportunity to change the decision and start recalculation.

### 2.1.5 Providing the Required Resource Quantity for Transport Traffic Nodes

Due to the large volume of freight transportation by rail, transport companies are tasked with providing such a number of locomotives that are able to realize the transportation plan. The volume change is carried out by transferring units between the operated and non-operated fleet. In addition, an analysis of the condition of each locomotive.

The main goal of solving this task is increasing the economic efficiency of fleet use and transportation planning by automating locomotive fleet planning and management by improving the quality, reliability and systematization of information for management decision-making. The main technological requirements are the following: the selection of locomotives should be carried out according to certain criteria and must take into account their technical condition and current location. In addition, it is necessary to take into account the possibility of manually adjusting the list of locomotives in connection with the necessary repairs. Therefore, the task is divided into three stages.

Stage 1. An initial solution is created in which there is no decommissioning (this locomotive has a special status for manual correction by the user. In addition, an ESLR schedule is planned.

Stage 2. The user adjusts the schedule and manually sets locomotives for decommission or an ESLR.

Stage 3. The schedule is submitted taking into account the immutability of user edits. After that, the graph is checked at the time of the technical feasibility of carrying out an ESLR of the specified locomotives and a decision is issued on the possibility (or impossibility) of carrying it out, after which the final schedule of repairs is displayed.

### Locomotive Repairs Schedule Making

When solving the problem of ensuring the necessary volume of traffic, it is not enough just to bring the maintenance of the locomotive fleet to the required number, it is also necessary to maintain each element of this fleet in working condition. To do this, each section of the locomotive must periodically undergo maintenance in accordance with technological standards, and therefore a schedule of repairs must be calculated. This task formulates as follows: it is necessary for each locomotive in the operated fleet to make a schedule for each type of repair, taking into account the adjustments of repairs according to the priority of repairs. All repairs are divided by seniority, each of which has its own frequency; each type of locomotive has its own mileage rate for them.

The capabilities of repair facilities are limited and to ensure the optimal distribution of repairs to the relevant facilities in order to maximize the workload. This task is reduced to a transport task, which in turn can be formulated as follows: it is necessary to distribute the repair schedule to suitable repair facilities, taking into account the type of repair, the locomotive parameters, as well as the maximize distribution efficiency [15, 16]. Each enterprise has the ability to repair a limited set of series for specified repair types and a capacity. In addition, for the specified period, the relocation norm also serves as an additional restriction.

This approach provides automation of the following technological processes:

- Creation senior factory repairs schedule and additional maintenance, which should be carried out independently;

- Creation of an ESLR schedule for locomotives;
- Creation repairs scheduled for various time periods with the indication of detailed information about every repair.

### Solving the Task of Planning Locomotives Repairs

To solve this problem an object-oriented software architectural model was implemented with the following objects classes:

- Classes for the sequence of execution of the general algorithm for creating a repairs schedule and drawings up individual repair plans for each locomotive;
- Classes for storing information about locomotives and its current parameters;
- Classes for storing information for all types of repairs for each locomotive;
- Classes for recording and storing all calculated repairs by type.

### The Task of Repairs Distribution by Repair Facilities Solving

The task of distributing repairs to repair companies can be reduced to solving a multi-threaded transport problem, which can be formulated as follows. There is a capacitive network consisting of nodes and branches, to which objects are transferred along with branches. The main task is to determine the maximum capacity for each arc. The transport distribution problem with minimal costs is positioned as a linear programming model. A modified auction method is used to solve this problem [17]. The auction method is a parallel relaxation method for solving the classical assignment problem. After calculating all the permissible utility functions, the repair company with the maximum utility function is selected. To reassign a repair its utility function must be higher than the previous one. This leads to the convergence of the algorithm to obtain the final solution.

The algorithm can be represented as follows.

There are N repair facilities and N repairs to distribute the task. For each factory, there is a non-empty subset of $A(i)$ objects that can be assigned to it. Solution S is a set of <repair company-repair> (i, j) pairs such that:

- $j \in A(i)$ for all $(i, j) \in S$;
- each repair company i has at most one pair $(i, j) \in S$;
- each object j has at most one pair $(i, j) \in S$.

The full purpose is a set of N pairs <repair company, repair> . In the context of this assignment S, a repair facility i is assigned if there is such a repair j that $(i, j) \in S$. There is some given integer value $a_{ij}$ in which repair company i is associated with repair j ∈ A(i). It is necessary to find a complete assignment that: maximizes the following function for all completed assignments S:

$$\sum_{(i,j)\in S} a_{ij}$$

**Table 3** Results of repair schedule making for the year planning period

| Input data | |
|---|---|
| Sections | 5017 |
| Repair facilities | 570 |
| Repairs type systems | Old and new |
| Repairs entered by the user | 0 |
| *Output data* | |
| Required repairs quantity | 18,614 |
| Distributed repairs quantity | 17,205 |
| Undistributed repairs quantity | 287 |

**Table 4** Results of repair schedule making for the 10 days

| Input data | |
|---|---|
| Sections | 5056 |
| Repair facilities | 570 |
| Repairs type systems | Old and new |
| Repairs entered by the user | 10 |
| Output data | |
| Required repairs quantity | 486 |
| Distributed repairs quantity | 392 |
| Undistributed repairs quantity | 72 |

The auction algorithm is used here, which solves the dual-use problem: minimize the function under the condition. A detailed description of the algorithm of auctions and dual price is given in the articles [16].

### Results of Planning Repairs Schedule

The results of the work of this module are the schedule of repairs, indicating the start and end dates of repairs, the repair facilities, the type of repair, the ID of the locomotive. Consider as an example the solution obtained based on the following data for annual planning (Table 3).

Table 4 shows the solution for the ten days period.

At the same time, in addition to obtaining total indicators for repair needs, as well as for occupied and unoccupied capacities, the user can view the schedule of repairs for each locomotive on the form.

## 2.2 Freight Scheduling Problem

While managing the transportation process the task of finding an effective timetable and providing trains with traction resources is decisive. Figure 1 illustrates the main idea of the freight scheduling problem.

Having received the generated train routes, locos and locomotive crews' parameters, the system should build a list of trains that need traction and a list of available locomotives and crews. For trains from the first list, it is necessary to find the optimal attachment of locomotives and crews from the other two lists. In other words, an assignment that provides the best implementation of the railway transportation management problem and complies with specified restrictions and technical regulations in relation to traction resources.

Freight scheduling problem includes three sub problems: train scheduling problem, locos assignment problem and locomotive crew assignment problem. Here, for brevity, we will dwell in more detail on only one of them—the locos assignment problem.

Figure 2 illustrates the problem of assigning locomotives to trains.

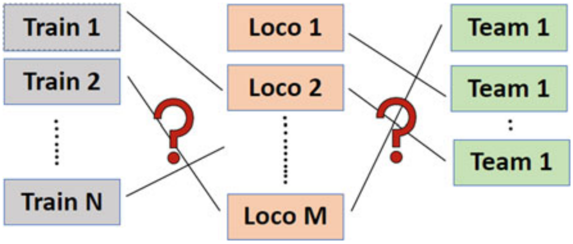
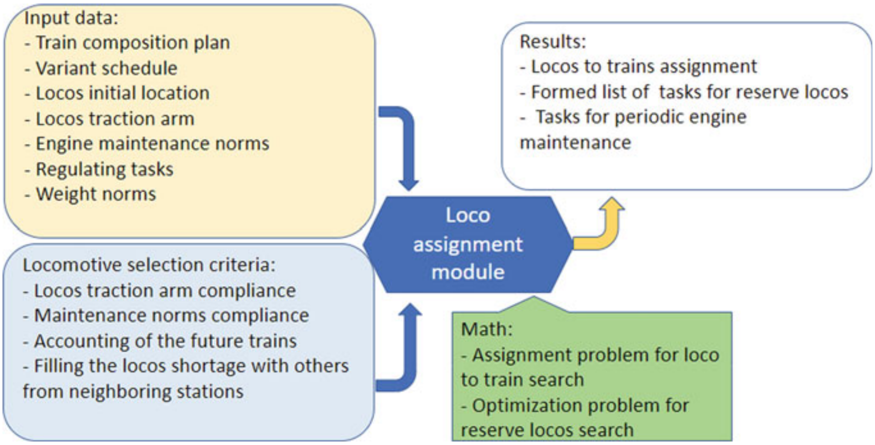


**Fig. 1** Main idea of freight scheduling problem



Input data:
- Train composition plan
- Variant schedule
- Locos initial location
- Locos traction arm
- Engine maintenance norms
- Regulating tasks
- Weight norms

Locomotive selection criteria:
- Locos traction arm compliance
- Maintenance norms compliance
- Accounting of the future trains
- Filling the locos shortage with others from neighboring stations

Results:
- Locos to trains assignment
- Formed list of tasks for reserve locos
- Tasks for periodic engine maintenance

Math:
- Assignment problem for loco to train search
- Optimization problem for reserve locos search

**Fig. 2** Problem of assigning locomotives to trains

The specified data is supplied to the module input. The task is to assign locomotives to trains in the best possible way according to the specified criteria and constraints.

### 2.2.1 Formalization of the Locomotives Assignment Task

The technological process of transportation entails many restrictions that are imposed on the train and traction resources schedule generation rules. To take into account these limitations and features of the railway area, a task formalization method was developed, which formed the basis for creation of a mathematical model of the described control system.

Formalization of the task includes building a model of the freight management system, forming a utility function and formulation of an optimization problem for an effective train and traction resources scheduling.

#### Multi-agent Transportation Management System

Due to a large number of technological limitations in cargo transportation planning, it was necessary to develop a system that would take into account and comply with all requirements and constraints of the described task. For this purpose, a multi-agent control system (or "planner") was proposed. It consists of the following components:

- The execution environment, which interacts with external systems, receives all the input data from them and sends the received data to the main agent.
- Agents:
  - Station agents, which contains their parameters, data about neighboring stations, the number of available locomotives and locomotives teams for each hour of the planning period. Restrictions for them:

    The trains number at the station should not exceed the number of station tracks at any given time  
    The locomotive's number and locomotives teams at the station should not be negative at any hour
  - Line sections agents. Two agents are created for each line (in the forward and reverse direction). The line agent knows the transfer capacity for each hour on the planning horizon. Restrictions:

    The number of trains sent to the line every hour should not exceed the capacity value for this hour.
  - Train agents. Agent's goals:

    Get departure and arrival times on each included in the route line.  
    Minimize the dwell time of the train on intermediate stations of the route.
  - Loco agents. Restrictions and goals:

The railway traffic norms are should not exceed
Minimize the loco's journey time without a train.
Minimize locomotive downtime at stations.
Maximize the effective work of the locomotive

- Team agents. Its similar restrictions to locomotives.

To solve the entire planning period of the railway modeling problem, the system was decomposed into 4 subsystems:

- The Volumetric Train Planner operates with train, loco, team and stations. This part of the system splits the initial number of trains into hourly intervals, taking into account the restrictions of all agents.
- The Volumetric Loco Planner includes locomotive and station agents that eliminate conflicts related to the lack of locomotives.
- The Volumetric Team Planner is similar to the previous one.
- Object-by-Object Planner contains station, train, locomotive and team agents. It binds trains to the train schedule paths, assigns locomotives, and crews to their abstract transposition plans (which are the results of the work of the described above planners). Optimization of the locomotives (teams) assignment is the main goal of this work. It is a subtask of the general freight scheduling problem that the optimization algorithms are aimed at.

### The Utility Function Definition

The key question for taking into account technological constraints and formulating train scheduling tasks as an optimization problem is the utility function formation. The implementation of the railway control system is formed on its basis. The main criteria of the utility functions are given in Table 5. A normalized numerical value *uk* is assigned to each criterion. After that the value for the utility function of the pair <locomotive, locomotive slot (an abstract plan for moving a locomotive between stations)> is calculated as $U = \sum_k c_k u_k$, where $c_k$ is the weight of the corresponding k-th criterion.

Thus, the evaluation rules for the described above subtask solution efficiency were formed. After the Object-by-Object Planner has compiled a list of available slots and locomotives, one can proceed to the heart of planning—solving the assignment of traction resources to the appropriate slots task. This problem should be formulated as a general mathematical optimization problem with constraints.

### Reducing the Task of Attaching Traction Resources to Trains to the Optimization Problem

Above, we have considered the type of utility functions for pairs <locomotive slot, assigned to it locomotive> . Considering the optimization of the trains to traction resources attachment in the entire planning period, cumulative utility function was

**Table 5** Information on the calculation of utility function criteria for the locomotives to a trains linking task

| № | Criteria name | Description | Sign |
|---|---|---|---|
| 1 | Working time | The difference between the arrival time at the terminal station of the slot and the departure time from the initial station | + |
| 2 | Locomotive waiting time | The difference between the departure time of the slot and the arrival time of the locomotive at the departure station. If this difference is negative, then take the value of the criterion = 0 | – |
| 3 | Slot waiting time | The difference between the arrival time of the locomotive at the departure station and the departure time of the slot to the first section. If this difference is negative, then take the value of the criterion = 0 | – |
| 4 | Time till service station | The time of the engine relocation from the end station of the train route to the nearest station providing the required service type | – |
| 5 | Time until the next periodic engine maintenance | The operating time remained until the next periodic engine maintenance at the moment of arrival at the terminal station of the slot. If the value is <0, then the value of the criterion is set to $\boxed{\text{OBJ}}$ | – |
| 6 | Weight restrictions | Difference between maximum permissible weight of the train that this locomotive can carry on this route and the weight of the train corresponding to the locomotive slot. If the value is <0, then the criterion value is set to -infinitive. Further calculations on this branch are not required | + |
| 7 | Time to transfer the locomotive by reserve | The travel time between the station where the locomotive is currently located and the station from which the locomotive slot begins | – |
| 8 | Future plans | Consideration of possible “future” slots from the terminal station | + |

used, which is equal to the sum of all objective functions of constituents locomotive slots in the whole schedule. Thus, the total objective function is a multidimensional function of the vector argument, where vector—any solution of the locomotives to slots assignment problem, presented in the form, where—the number of the locomotive (crew) assigned to the i-th slot,—number of slots in the current planning period.

Then, in order to solve the optimal railway scheduling problem, it is necessary to solve the general optimization problem of finding the extremum of the above-mentioned function with given constraints:

$$F(X) = \sum_{i} U_i(x_i) \rightarrow extr(max);\ X^* = \arg extr F;\ r_j\big(X^*\big) \le b_j,\ j = 1, \ldots, l$$

where $X^*$—the desired solution; $r_j(X^*)$—given set of constraints; $l$—number of constraints.

The search is performed on a set of constraints related to technological features in the railway complex. Then it is necessary to find such a kind of iterative operator for finding the extremum of the function F(X) to make the iterative process converge to the best solution or, in other words, to the extremum of the total utility function.

### 2.2.2 Solving the Problem of Building an Effective Schedule

The assignment problem is a special case of a transport problem. It, in turn, is a special case of a linear programming problem. Such problems are solved using classical methods, but using a specialized method gives faster results.

#### Hungarian Algorithm

The most common way to solve the assignment problem is the "Hungarian algorithm", which is a combinatorial optimization algorithm.

It was proposed in 1955 and its computational complexity is, where n—is the number of tasks and employees (the number of jobs must be equal to the number of employees). The algorithm can be modified up to the level of complexity. The Hungarian method was implemented to solve the problem of assigning traction resources to trains, but its speed turned out to be unacceptable. The results of this experiment are discussed in more detail below. Therefore, another algorithm was found, that is, solving the assignment problem using auctions.

#### Auctions Algorithm for Assignment Problem Solution

D. Bertsekas proposed the solution of the assignment problem using auctions. The main ideas of the auction algorithm are described in [14]. In relation to our problem, the steps of the algorithm can be described as follows [15, 16]:

**Step 1**. The best slot j is selected for the detached locomotive $i$:

$j_i = \arg \max\limits_{j \in A(i)} \big(U_{ij} - p_j\big)$,where $p_j$—-th slot price, $A(i)$—set of slots allowed for the i-th locomotive.

**Step 2**. The second-highest value of the "benefit" from the assignment of the locomotive to the $j_i$-th slot is calculated:

$$\omega_i = \max_{j \in A(i), j \neq j_i} U_{ij} - p_j$$

If except $j_i$, there are no other possible locomotive assignments, then $\omega_i = -\infty$.

**Step 3**. Next, the new price for the slot $j$ is calculated:

$$p_j = \max\{\lambda; U_{ij} - \omega_i + \varepsilon\}$$

where λ—the threshold value (constant) below which it is forbidden to set the price, $\varepsilon$—infinitesimal number of the order 1/N.

**Step 4**. If $\lambda \leq u_{ij_1} - \omega_i + \varepsilon$ then locomotive $i$ is assigned to the slot $j$, and the previous assignment to slot j is reset.

Steps 1–4 are repeated until all locomotives are assigned to locomotive slots.

A rigorous proof of the convergence of this algorithm, as well as a justification for the choice of values and ε is given in the article [13].

This method allows you to quickly get a solution, but it is not able to meet a number of planning criteria, as well as to solve the optimization problem in compliance with the maintenance percentages of crews and reserve locomotives. The mentioned criteria include:

- Obtainment of several solutions with specified accuracy for selection of the best according to additional technological criteria;
- Solving a multidimensional assignment problem for the simultaneous solution of all three subtasks of the freight scheduling problem;
- Adjustment of the resulting solution accuracy.

### Simulated Annealing Method

Due to the inability of the auction algorithm to satisfy the listed above criteria, there is a need to develop another method of multidimensional optimization that solves the described situations. As a trade-off, the simulated annealing method was chosen. In general, the algorithm for assigning traction resources to trains with the method of simulated annealing can be represented as follows:

- The initial temperature and minimum temperature are initialized together with the temperature change function.
- The first locomotives assignment is chosen randomly and its overall utility is calculated.
- As long as the temperature is greater than the minimum value, the new state and its overall utility are calculated. If the utility of the new state is greater than the utility of the current one, then it switches to the new one. If the new state has a lower utility, the transition is carried out with a certain probability, depending on the temperature. For example: $P = exp(-\Delta t), where \Delta = F_1 - F_0$—is the difference between the total utility of the new and current states, and t is the current temperature.
- Transition to a new temperature according to the temperature change function.

To obtain a new state from the current one, the following algorithm was proposed: the coordinate of the utility matrix is randomly selected and any two assignments at this coordinate are swapped.

The peculiarity of the considered optimization task is the diversity of the utility function, which appears from launch to launch at various planning steps. The described method of simulated annealing, based on a random approach, copes well with cases of non-unimodal utility function, and allows you to meet the additional planning criteria described above. On other hand, its main disadvantage is a long operating time, which can be reduced by adjusting the appropriate parameters. In this regard, it is convenient to use a hybrid of the two methods described in this paragraph, using each of them for the purposes corresponding to their capabilities.

### Genetic Optimization Algorithm

The optimization problem is reformulated as the problem of finding the maximum of some function $f(x_1, x_2, ..., x_n)$, called the fitness function. It is necessary that on a bounded domain of definition $f(x_1, x_2, ..., x_n) \geq 0$, while continuity and differentiability are not required. A string of bits encodes each parameter of the fitness function. An individual will be called a string that is a concatenation of strings of an ordered set of parameters $|x_1|x_2|...|x_n|$. The universality of GA lies in the fact that only the fitness function and coding of solutions depend on a specific task. The remaining steps for all tasks are performed the same way.

With the help of the fitness function, among all individuals of the population, the most (crossing capable) and the least (removed from the generation) adapted individuals are distinguished. Thus, the fitness of the new generation is on average higher than the previous one. The final solution to the problem is the fittest individual of the last generation with the greatest fitness function.

The algorithm step consists of the following stages:

Stage 1. Setting initial states and parameters. Before starting the GA, the program receives a list of stations, locomotives, and a set of volumetric plans for the purpose of locomotives. In addition, to start the GA, the following parameters are set: the maximum population size P, the number of populations M, the number of the worst individuals D.

Stage 2. Creation of the initial population. At the start of the generation of the initial parent population of each individual, all agents of stations and locomotives are created based on the starting information indicating the number of the current population $Pn_{Id} = 0$.

To create one individual of the parent population, the following mechanism is used. The system receives an input volume plan for the transfer of locomotives, train routes formed with the threads of the variant schedule. The route lists all stations in the order of route. When linking trains to threads, there are the following features:

- The train can change the threads during the movement.
- Different trains can travel on the same line in different sections.

For each volume plan, the algorithm replaces the planned departure and arrival times with real ones taken from the detailed route of the train. Then there is an object-by-object assignment of real locomotives. For each time at least one locomotive is planned to depart, a departure stations list is compiled and the evaluation function of all existing locomotive-plan combinations at this station is calculated.

The evaluation function has the following form:

$$u(Plan, Loco) = C_1\frac{T}{N_1} + C_2\frac{D}{N_2} + C_3\frac{S}{N_3}$$

where $Plan$ is the identifier of the binding plan, $Loco$ is the locomotive identifier assigned to this $Plan$, $C_1$, $C_2$, $C_3$ are configurable coefficients, $N_1$, $N_2$, $N_3$ are normalization coefficients, T is the time remaining until the next repair, given the time of the relocation to the repair facility, D is the distance remaining until the next repair, given the distance of the relocation to the repair facility, S is the coefficient of correspondence between the train route and the traction arm of the locomotive. The selected locomotives are assigned to the plans, their location and parameters are updated according to the forwarding plan for further work.

If there is no one locomotive suitable for the train according to the restrictions, a return to the previous time step occurs and the assigned task is solved again.

For the found solution the following assignments matrix is created:

$$M_{Pn_{id}} = (m_1(Plan_1, Loco_1, PnId, A_1)\ldots, m_i(Plan_i, Loco_i, PnId, A_i))$$

where $Plan$ is the locomotive forwarding plan, $Loco$ is the locomotive assigned to the $Plan$ (if no suitable locomotive is found Loco $= -100$), $PnId$ is the population ordinal number, A is the identifier of assign (if the locomotive is assigned it is equal to 1, otherwise—0).

Thus, one of the initial individuals of the population is created. For each individual, the following total utility function is calculated:

$$F_i = \sum_{i,j} u_{i,j}$$

Stage 3. Crossover. Two parents are selected from the starting population—$P_1$ and $P_2$. The child $C_1$ is generated in the following way:

$$\begin{aligned} {m_i}^{C_1} = \ \{{m_i}^{P_1}, if\ An_{P_1} = 1\ \ {m_i}^{P_2}, if\ A_{P_2} = 1\ \text{и}\ Loco_{Plan}^{P_1} \\ \neq -100\ {m_i}^{P_1}, if\ A_{P_2} = 0\ and\ A_{P_1} = A_{P_2} = 0 \end{aligned}$$

The usefulness of each individuals is calculated as follows:

**Table 6** Comparison of the Hungarian method and the auction method

| Quality indicators | Hungarian algorithm | Auction method |
|---|---|---|
| Run time (simulation of the assignment task on AgentSpeak) | 3–4 min | 20 s |
| Complexity | O(n3) | Nlog N |

$$\begin{aligned} U_{C_1}(Loco, Plan) = \{U_{P_1}(Loco, Plan), if\ {m_i}^{C_1} = \ {m_i}^{P_1} \text{ и } A_{P_1} \\ = 1\ \ U_{P_2}(Loco, Plan), if\ {m_i}^{C_1} = \ {m_i}^{P_2}\ and\ A_2 \\ = 1\ 0, if\ A_{P_1} = A_2 = 0 \end{aligned}$$

that is, if $m_i^{child\_1}$ was copied from the genotype of any parent, the usefulness of a pair $(Loco_i, Plan_i)$, selected based on the genotype of a parent was used for this.

Stage 4. Mutation. With some probability, a mutation may occur in the population. The parameter TPlanId is time, which is randomly determined for each individual. The TPlanId redesigns assignments for plans.

Stage 5. Selection and removal of the worst individuals. After all the individuals for a given population have been generated, the planner selects a given number of the worst individuals in terms of a utility function from all the individuals and then removes them from the offspring.

Stage 6. Stop criterion. The generation of new populations stops if one of the following stopping criteria is reached: the final population size is reached or the improvement of the utility function with a new individual does not exceed the specified.

Stage 7. Final selection of the population. The best individual with the maximum utility function is selected.

## Results

### *Comparison of the Hungarian Algorithm and the Auction Method for Solving the Assignment Problem*

Comparing the proposed approach to solving the assignment problem by the auction method with the application of the Hungarian algorithm, it can be noted that the auction algorithm shows a significant increase in the convergence rate compared to the Hungarian algorithm: 20 s instead of 3–4 min (these time indicators should not be evaluated in absolute terms, since this study was conducted in the AgentSpeak language, which itself is a rather "slow" language) [17].

This time estimate shows that the auction algorithm converges in $Nlog(N)$ time, which is a significant improvement over the $O\left(n^4\right)$ Hungarian method (Table 6).

Despite the noticeable advantage in speed, the auction method in some cases is powerless to fulfill the mentioned above additional planning requirements. Let's

**Table 7** Comparative analysis of the auction method and the simulated annealing method

| Criteria | Auction method | Simulated annealing |
|---|---|---|
| Capability to return several solutions | No | Yes |
| Working with "elongated" matrices | Good | Time costs increasing |
| Working with sparse matrices | Manual verification is required for some special cases | Time costs increasing |
| Run time | 36 s | 3 min |
| Scalability of the dimension of the problem | No | Yes, but also time costs increasing |
| Accuracy adjustment | No | Yes |

consider the results of another optimization algorithm that would take into account these requirements.

*Comparison of the Auction Method and the Simulated Annealing Method*

To evaluate the characteristics of these algorithms, a series of experiments was carried out, in which the following criteria of their work were considered: run time, capability to return several solutions, working with sparse matrices, adjusting accuracy, the possibility of increasing the dimension of the problem, working with "elongated" matrices [18]. A comparative analysis of these characteristics is shown in Table 7.

The test data used for the analysis are close to the real traffic and distribution of transportation resources on the Eastern Range of the Russian Railways network. The traction resources to trains assignment was carried out on the planning horizon in 24 h.

On this data set, the freight transportation control module was executed twice: in the first run, the auction algorithm performed the assignment of traction resources, in the second by the annealing simulation algorithm. The main advantage of the auction algorithm is its operating time. As can be seen from the table, according to the carried out tests, planning ends almost four times faster with auctions than with the simulated annealing method. The auction also shows better performance on datasets, where the number of locomotives significantly exceeds the number of trains. On the other hand, the simulated annealing method allows high flexibility in adjusting the accuracy of the final solution. If the set of acceptable solutions includes a number of solutions with very similar total utility, the simulated annealing method can be run with adjusted parameters in order to reduce the number of iterations, which leads to a significant reduction of the execution time without loss of optimality of the solution. This is possible, for example, when there are several assigned teams that come to the depot of the registry at about the same time.

The main advantage of the simulated annealing method is that it can be easily extended for a multidimensional task. For example, all three subtasks of the freight

**Table 8** Simulation data. Simulation data and work results (P is the maximum size of one population, M is the maximum number of populations, D is the number of deaths per generation)

| | Without GA | Genetic algorithm | | | |
|---|---|---|---|---|---|
| Parameter | | P = 15, M = 10, D = 4 with F | P = 15, M = 20, D = 4 with F | P = 20, M = 10, D = 4 with $N_p$ | P = 15, M = 20, D = 4 with $N_p$ |
| Summary F | 139,440 | 335,760 | 414,960 | −109,920 | −109,920 |
| $N_p$ | 27 | 14 | 13 | 47 | 47 |
| Working time | 1 m, 13 s | 15 m, 24 s | 10 m, 31 s | 19 m, 13 s | 9 m, 29 s |

scheduling problem can be performed simultaneously using this method (four-dimensional assignment). The disadvantage of the simulated annealing method is that the scheduler's operating time increases significantly in the case of an dimension increase (rough estimate:O(nm), where n is the maximum dimension for one of the coordinates, and m is the number of coordinates). For this reason, one has to introduce additional heuristics to achieve an acceptable operating time.

Taking into account the described characteristics of the proposed algorithms, it is convenient to use a hybrid solution of the assignment problem using auction and simulated annealing methods, applying each of them in accordance with the current set of requirements.

*Results of Attaching Traction Resources Problem Solving Using GA*

As a simulation, a railway section was selected, including 566 stations, 772 volumetric locomotive plans, 469 train plans, 4723 train route lines, 783 real locomotives with their parameters, 44 real trains.

The standard planning mechanism has the following features: all utility functions for pairs $(Plan, Loco)$ are calculated several steps ahead, then the total utility function is calculated relative to this depth and the option with the highest value of this function is selected. Table 8 shows the results of the algorithm with the optimization criterion by F—by the utility function or by Np—the total number of assignments.

From the simulation results, it seems that the locomotive binding plan changes depending on the choice of the parameter for which optimization was carried out.

## 3 Conclusion

A smart city is a combination of smart management systems and technologies for managing various areas of activity of an urban agglomeration. The main task of creating such a smart city is to improve the quality of life of the population using the latest technologies and methods. Smart transport is one of the main directions of a smart city responsible for the movement of people, optimization of the interaction between vehicles and efficient use and change of transport infrastructure. The tasks that arise in this area are characterized by the complexity of their formalization and solution, often associated with increased security requirements. To solve them, the main trend today is the development of intelligent automated control systems. capable of not only finding optimal solutions for emerging emergency situations but also providing analysis for the premature identification of "bottlenecks" in the technological processes of management and on the roads.

## References

1. Valdez AM, Cook M, Potter S (2018) Roadmaps to utopia: tales of the smart city. Urban Studies 55(15):3385–3403
2. Nikitas A, Michalakopoulou K, Njoya ET, Karampatzakis D (2020) Artificial intelligence, transport and the smart city: definitions and dimensions of a new mobility era. Sustainability 12(7):2789
3. Rjab AB, Mellouli S (2018) Smart cities in the era of artificial intelligence and internet of things: literature review from 1990 to 2017. In: Proceedings of the 19th annual international conference on digital government research: governance in the data age, pp 1–10
4. Thaduri A, Galar D, Kumar U (2015) Railway assets: A potential domain for big data analytics. Procedia Comput Sci 53:457–467
5. Reinhold S, Laesser C, Bazzi D (2015) The intellectual structure of transportation management research: A review of the literature
6. Pashchenko FF, An BT, Hieu TD, Pashchenko AF, Van Trong N (2020) Intelligent technologies in decision-making support systems. In 2020 international conference engineering and telecommunication (En&T), pp 1–4
7. Pashchenko AF, Pashchenko FF (2013) Smart technologies for solving the transport problems of the city. In: Proceedings of the seventh international conference management of the development of large-scale systems (MLSD'2013), vol 2, pp 141–144
8. Van Gerven M, Bohte S (2017) Artificial neural networks as models of neural information processing. Front Comput Neurosci 11:114
9. Pashchenko FF (2007) Introduction to consistent methods of system modeling. Finance and Statistics, Moscow
10. Pashchenko FF, Minashina IK, Zakharova EM (2013) Neuro-fuzzy modeling of passenger flows, management of large-scale systems development (MLSD'2013). V.A. Trapeznikov Institute of Management Problems, 2:144, ISBN 978-5-91450-138-6
11. Pashchenko FF, Kuznetsov NA, Minashina IK, Zakharova EM (2013) Using relaxation algorithms for estimating parameters of neuro-fuzzy models. In: 4th international conference ICDQM-2013 proceedings. Belgrade, Serbia, 94–100
12. Wasan MT (2004) Stochastic approximation. Cambridge University Press, 58
13. Zakharova EM (2018) Development of planning and management algorithms in timetable tasks in railway transport. MIPT, Dolgoprudny
14. Bertsekas DP, Castanon DA (1989) The auction algorithm for the transportation problem. Annal Oper Res 67–96
15. Pashchenko FF, Pashchenko AF, Kuznetsov NA, Minashina IK, Zakharova EM (2017) Analysis of the adaptive algorithms behaviour applied to the railway optimization problems. In: The 8th international conference on ambient systems, networks and technologies, vol 10. ANT 2017, Madeira, pp 560–567
16. Pashchenko FF, Takmazian AK, Kuznetsov NA, Minashina IK, Zakharova EM (2017) Intelligent control systems for the rolling equipment maintenance of rail transport. In: AICT2017

11th international conference on application of information and communication technologies, Moscow

17. Kuznetsov NA, Pashchenko FF, Ryabykh NG, Minashina IK, Zakharova EM, Tsvetkova OA (2015) Implementation of train scheduling system in rail transport using assignment problem solution. Procedia Comput Sci 63:154–158
18. Kuznetsov NA, Minashina IK, Ryabykh NG, Zakharova EM, Pashchenko FF (2016) Design and comparison of freight scheduling algorithms for intelligent control systems. Procedia Comput Sci 98:56–63

# Experimental Approximation of a Vehicle's Fuel Consumption Using Smartphone Data

**Stavros-Richard G. Christopoulos, Stratis Kanarachos, and Konstantina A. Papadopoulou**

**Abstract** An algorithm is developed in order to record a vehicle's fuel consumption using data from a smartphone's sensors. Six field tests were conducted: (1) Ford Fiesta car with automatic transmission driven around Coventry, UK, with "passive" and "restless" driving behaviors, (2) Ford Fiesta car with manual transmission under heavy traffic driven around Athens, Greece, (3) Ford Fiesta car with manual transmission driven around Athens, Greece, in heavy traffic, in a highway, with "passive" and "restless" driving behaviors and high variation in altitude during the trip, (4) Ford Fiesta car with manual transmission driven around Athens, Greece, with "passive" and "restless" driving behaviors and even higher variation in altitude during the trip, (5) Suzuki Swift car with manual transmission, a route including highway, streets and alleys, in the west Attica in the surrounding area of the capital of Greece, and (6) Suzuki Swift car with manual transmission, with "passive", "normal" and "restless" driving behaviors in West Attica, in a place with a small hill with a very high slope that we "climbed" three times in a row. The results show that the proposed algorithm improves the smartphone-recorded GPS data so that they show high accuracy when compared to the GPS data extracted from each vehicle's on-board diagnostic system.

**Keywords** Fuel consumption · Smartphone · GPS data · Exhaust emissions

## 1 Introduction

Modeling of a vehicle's fuel consumption and exhaust emissions could be an effective tool to help develop vehicle technologies towards a greener future with low carbon emissions. Exhaust emissions affect not only sustainability of an urban environment, but also have a negative effect on human health.

During the early stages of studying fuel emissions, researchers used data from a large spatiotemporal scale [1, 2]. Those data, however, did not provide information

S.-R. G. Christopoulos (✉) · S. Kanarachos · K. A. Papadopoulou
Faculty of Engineering, Environment and Computing, Coventry University, Priory Street, Coventry CV1 5FB, UK
e-mail: ac0966@coventry.ac.uk

Y. A. Vershinin et al. (eds.), *Technologies for Smart Cities*,
https://doi.org/10.1007/978-3-031-05516-4_8

about the vehicle and its moving parameters. To that end, new models have been developed that take under consideration vehicle technology and moving [1].

For the calculation of a vehicle's exhaust emissions, the first step is to approximate the fuel consumption during a trip. However there are numerous factors to affect the estimation of fuel consumption, like, for example, the weather, the type and characteristics of the vehicle, the type of the fuel, the driver and the traffic conditions.

Smartphones equipped with various sensors such as GPS, compass, accelerometer etc., can provide a means with which one can collect driving data [3–5]. As the smartphone industry is continuously expanding, there has been a series of applications developed to support driving, in general, and provide data about fuel consumption and exhaust emission in particular [6].

Approximating a vehicle's fuel consumption using smartphone-recorded data has a few fundamental difficulties. One of them is the low-accuracy of a smartphone's GPS data [7]. Furthermore, the data from the smartphone's sensors are extremely noisy and there are no other additional vehicle data that would help in our analysis, like angular velocity of the engine, horsepower, torque and gear ratio.

In the present study, we aim to develop an algorithm to be used for estimating fuel consumption using only a smartphone's sensors, while exhibiting the same accuracy in our results as when using the vehicle's on-board diagnostics, regardless of the aforementioned factors that affect fuel consumption.

In order to estimate the emissions using a smartphone's GPS data after we have calculated the fuel consumption, we follow the following calculations:

One liter of petrol corresponds to around 0.737 kg/l density [8]. The chemical reaction of the production of the emissions using petrol is given by [9]:

$$2C_8H_{18} + 25O_2 \rightarrow 16CO_2 + 18H_2O$$

That means that the specific $CO_2$ emission is 3.30 (kg of $CO_2$)/(kg of petrol) = 2.43 kg of $CO_2$ per liter of petrol [8].

For diesel engines, the emission formula is given by:

$$C_{12}H_{24} + 18O_2 \rightarrow 12CO_2 + 12H_2O$$

The density of diesel is 0.848 kg/l, which means that the specific $CO_2$ emission is 3.15 (kg of $CO_2$)/(kg of diesel) = 2.67 kg of $CO_2$ per liter of petrol [8].

Thus, using the fuel consumption algorithm we developed and the emission formulas for petrol and diesel engines we can estimate the $CO_2$ emissions using the smartphone data.

## 2 Data Collection

Data collection for the present study is based on the use of a combination of the Androsensor [10] and Torque Pro [11] android applications, alongside the ELM327 device [12]. Androsensor is designed to record data from all the available sensors in a smartphone in.csv format with a sampling rate up to 10 Hz. Torque Pro, on the other hand, uses Bluetooth to extract the real on-board diagnostics data (OBD) [13] from the vehicle with the help of ELM327, a programmed microcontroller produced by ELM Electronics for translating the OBD interface found in most modern cars.

### 2.1 Data Accuracy and Fuel Consumption Algorithm

In order to approximate the fuel consumption of the vehicle using only smartphone data we designed an algorithm that can account for several parameters, namely speed, kinetic energy, dynamic energy, aerodynamic and the road friction, rolling resistance, idling fuel consumption, fuel consumption when the vehicle stops, transmission-engine load, mass of the vehicle, drag coefficient, tire pressure, frontal area of the vehicle, density of the air and one free factor, $Mf$, that depends on the vehicle's model. It is necessary to point out that the poor performance of the GPS is a known problem, especially, in urban areas [14–16] where narrow streets and high buildings are present, hence the GPS speed estimation is poor.

The algorithm accepts as inputs both the altitude and the speed from the smartphone's GPS data, as well as the vehicle's characteristics (i.e. model, mass, drag coefficient, frontal area, tire pressure) and air density (see Fig. 1).The outcomes of the algorithm are the fuel consumption approximation and a better estimation of the vehicle's speed after improving the GPS data. The schematic representation of the algorithm's operation is represented in Fig. 1.

More specifically, in order to achieve a better estimation of the GPS speed (from here on denoted as iGPS speed, where 'i' stands for 'improved') we calculate for every single point in a GPS-speed diagram the least square line [17, 18] that interpolates $n$ points before and $n$ points after this single point (in total $2n + 1$ points). Subsequently, we attribute to the central point the corresponding value of the least square line. Of course, with this approximation we lose the first $n$ and the last $n$ points of our measurements, however, the $n$ value has been selected to correspond to a few number of points, e.g. $n = 5$, which amounts to 0.5 s.

For the purpose of the current analysis, the $n$ value is selected equal to 5 ($n = 5$). The reason for this value is the sampling rate of 10 Hz for both of the applications that we are using. In addition, the closest number of points in order to cover a time frame of almost (and simultaneously greater than) one second is 11. Thus, filtering the GPS speed signal through this process, we increase the accuracy compared with the "real" speed data that were obtained from the OBD interface via the ELM327 device.

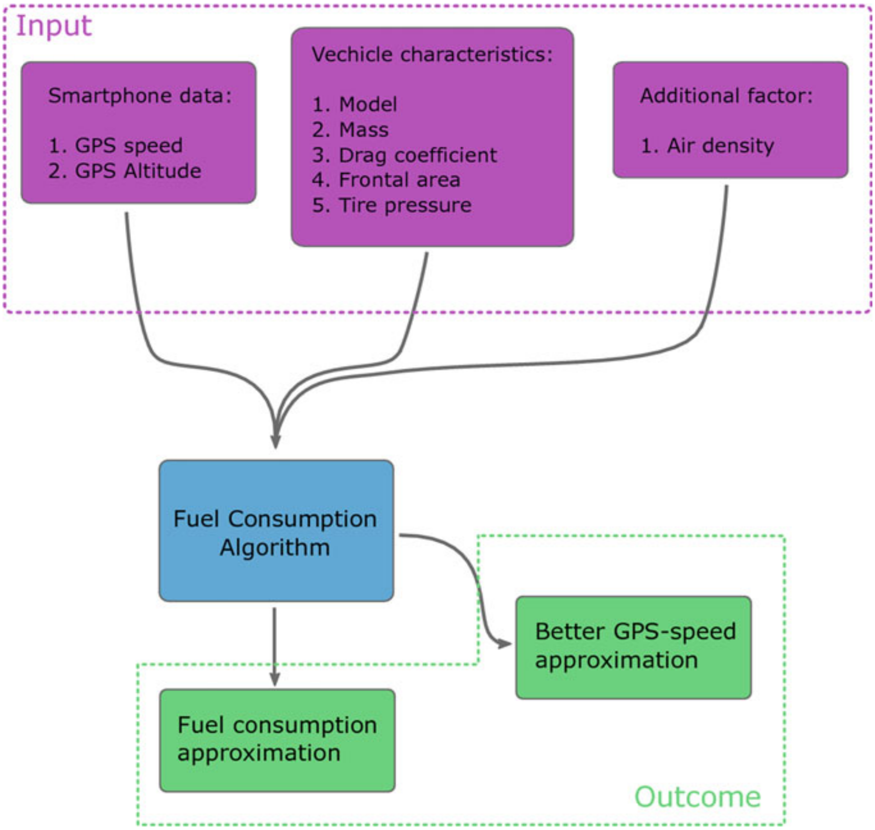


**Fig. 1** Schematic representation of the operation of the fuel consumption algorithm

## 3 Fuel Consumption Analysis

### *3.1 Field Tests*

For the purpose of this study, we designed an experiment of several routes with a wide range of different parameters (e.g. car, driver, cities, geography, weather, time), as well as the effect of traffic conditions and the driving behavior to the fuel consumption rate. In particular, we followed six different testing routes with three different drivers (Driver1, Driver2 and Driver3 hereafter) that were not aware of the exact reason of the measurements (thus avoiding purposeful driving), in two different cities in different European countries (i.e. Coventry, UK and Athens, Greece, see maps in Fig. 2) and three different vehicles, a three year old Ford Fiesta 1600 cc petrol with automatic transmission (Vehicle1, hereafter), a five year old Ford Fiesta 1600 cc petrol with manual transmission (Vehicle2, hereafter), and an eleven year old Suzuki Swift 1300 cc 92HP petrol with manual transmission (Vehicle3, hereafter).

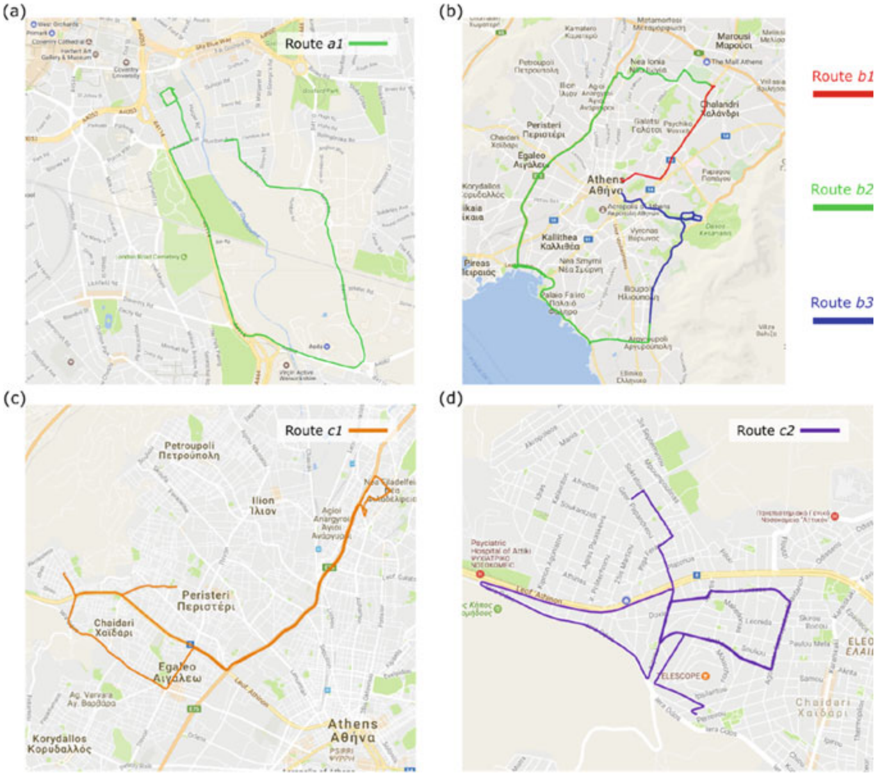


**Fig. 2** Maps of the routes followed during the field trials: **a** Route *a1* in Coventry city, UK, with the first driver and the first Vehicle (Automatic Ford Fiesta), **b** Routes *b1*, *b2*, *b3* with the red, green, and blue solid lines respectively, in the centre and the surrounding area of Athens, Greece, with the second driver and the second vehicle (Manual Ford Fiesta), **c** Route *c1* in the west suburbs of Athens covering the regions (Municipalities) of Peristeri, Nea Philadelphia, Egaleo and Chaidari. And **d** Route *c2* in the region of Chaidari in Athens, Greece

For the better understanding of the results, the discussion is more extensive for the first route than the others. However, we follow exactly the same analysis for all cases.

## 3.2 Results

In the first field trial, Driver1 drove Vehicle1 in Coventry City, UK. Figure 2a shows the route *a1* that has been followed for the first test. We asked the driver to drive in a range of driving styles from "passive" driving behavior to "restless". Thus, the average speed of the trial was 31.8 km/h, the standard deviation of the speed was 22.5 and the max speed was 85.0 km/h. The altitude variation in this route was 29 m.

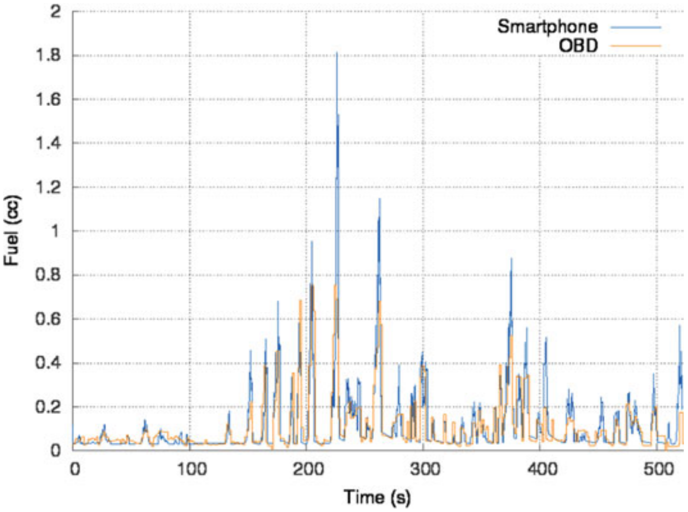


**Fig. 3** A typical excerpt of Route *a1*, of the "real" fuel consumption extracted from the OBD with the solid orange line, together with the smartphone approximation algorithm outcome with the blue solid line

Figure 3 shows the comparison between the "real" (OBD) fuel and the smartphone approximation algorithm's outcome for every single point. One can observe the high-quality matching of the two time-series.

Nevertheless, the right measure to evaluate our method would be the total estimation of the fuel consumed. The calculation of the Cumulative Distribution Function (CDF) [19] gives a different perspective on the whole behavior of the fuel consumption data during the time of the test as well as the total fuel consumed.

Figure 4a presents the CDF for both the "real" fuel consumption and the approximation using the smartphone data. As we can see, the algorithm's outcome follows exactly the behavior of the "real" data extracted from the OBD and the total estimation error is only 1.17%. We can also notice that there is a slight difference in the beginning of the trial that could be interpreted by the fact that the engine was cold and, for this reason, the fuel consumption rate was higher than normal. It is worth to mention that this difference is observed any time that we are starting the vehicle with a cold engine.

A second example of the behavior of the algorithm's operation is shown in Fig. 5a, where we see the total fuel consumption of the smartphone approximation and OBD fuel measurements for a sliding window of 901 points (close to 1.5 min). Each point in the diagram, therefore, represents the total fuel consumption for the time window starting 450 points earlier to this point and ending 450 points after this point. We chose this length since is close to 1000 points, while it is worth to note that we have the same accuracy for different window lengths.

For the second field trial, we used Vehicle2 and different Driver2. The field test took place in Athens, Greece (see route *b1* in the map of Fig. 2b), in a heavy traffic

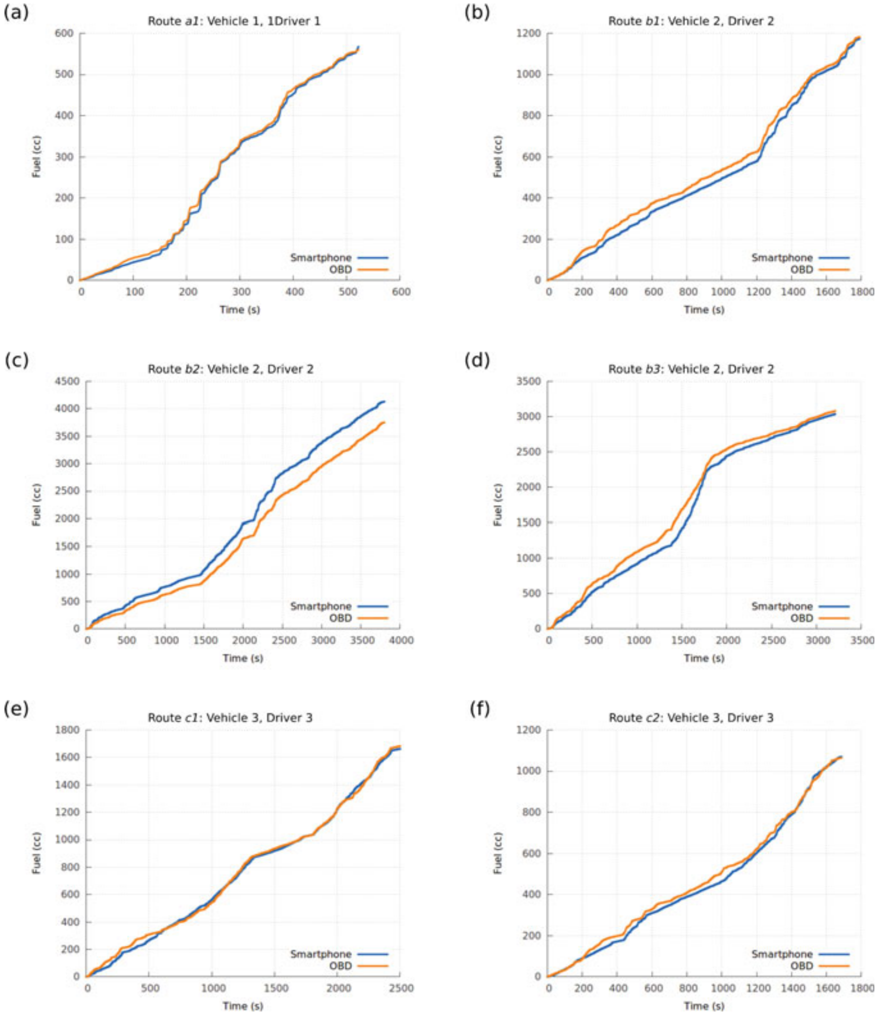


**Fig. 4** The cumulative distribution functions of the "real" fuel consumption as extracted from the OBD with the solid orange lines, together with the Cumulative distribution functions of the smartphone approximation algorithm's outcome with the blue solid lines for the Routes **a** *a1*, **b** *b1*, **c** *b2*, **d** *b3*, **e** *c1* and **f** *c2*

route, especially during the first part of the route. The average speed of this trial was 17.5 km/h, the standard deviation of the speed was 19.5, the max speed was 76.0 km/h and the altitude variation in this route was 80 m. We kept exactly the same values for the algorithm's operational parameters as in the previous test, since the vehicle is a similar model of Ford Fiesta, but with a manual transmission.

The CDF of the fuel consumption approximation when we use the smartphone data together with the real fuel consumption data is presented in Fig. 4b and the total

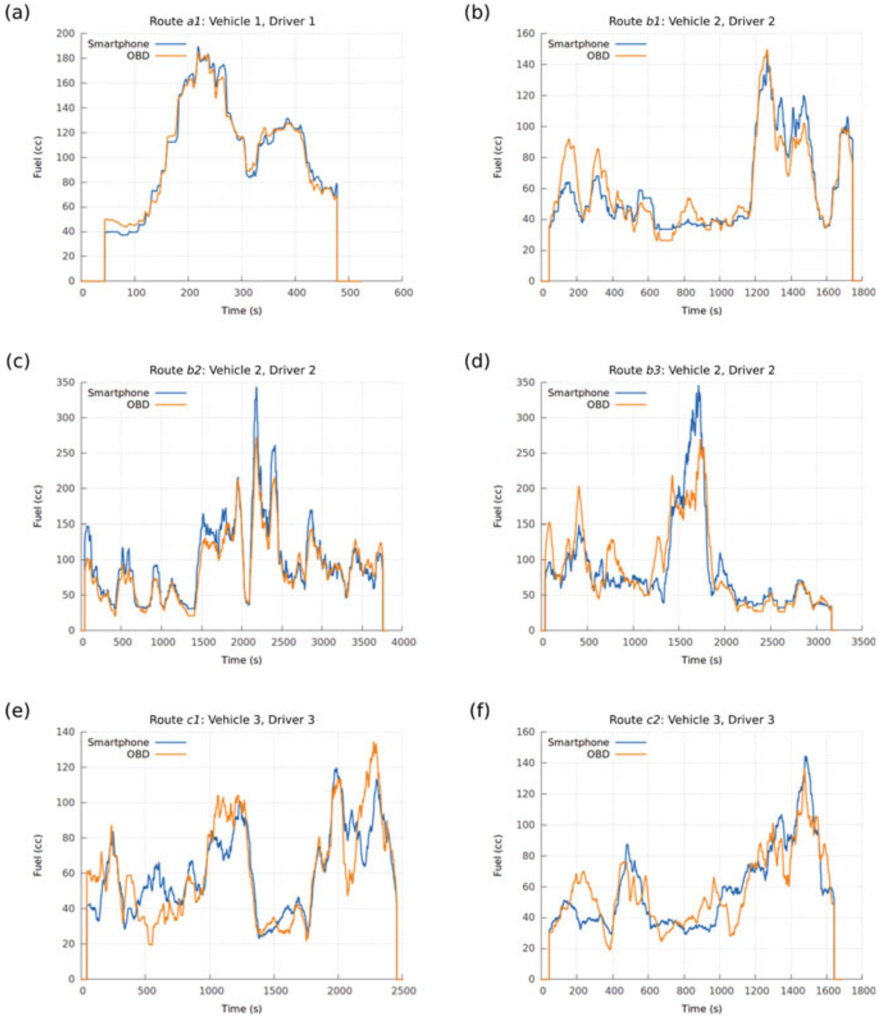


**Fig. 5** The total fuel consumption for the smartphone approximation (blue solid line) and OBD fuel measurements (orange solid line) for a sliding window of 901 points (close to 1.5 min) is presented for the Routes **a** *a1*, **b** *b1*, **c** *b2*, **d** *b3*, **e** *c1* and **f** *c2*. Each point in the diagram represents the total fuel consumption for the time window starting 450 points earlier to this point and ending 450 points after this point

fuel consumption's comparison of the sliding window of 901 points is presented in Fig. 5b. The results show that the final fuel consumption approximation error in this case is 0.7%. As in the previous case, the "smartphone" CDF curve follows precisely the behavior of the "real" data. We can also see the small difference when the engine is cold (the air temperature was close to 0 °C) in the beginning of the route. The fact that the approximation is independent of the car's gearbox type for a similar car model in different driving conditions shows the exceptional operation of the algorithm.

The third field trial was conducted with the same driver (Driver2) and car (Vehicle2) as in the second field test in Athens, for the route *b2* in the map of Fig. 2b. This test includes different types of road, i.e. Highway, Avenue, Street, Alley, normal driving, periods of testing with different driving styles (restless and passive) and a high variation in altitude. The average speed of this trial was 37.4 km/h, the standard deviation of the speed was 29.7, the max speed was 119.9 km/h and the altitude variation in this route was 172 m.

In this case, the CDF (see Fig. 4c) of the smartphone's data follows once again exactly the same behavior as the "real" data. The estimation error in this trial is the greatest of all cases, reaching 10.1%. However, the excellent behavior of the improved (via our algorithm) data is observed in Fig. 5c, which shows the total fuel consumption's comparison, as in the previous cases, of the sliding window of 901 points between the "real" and the "smartphone" data.

The fourth field trial was held by the same driver (Driver2) and vehicle (Vehicle2), again in Athens, Greece (see route *b3* of Fig. 2(b)). The challenges to this route were the high variation in altitude and the fact that one part of the route was made with different driving styles (restless / passive). The average speed of the trial was 26.5 km/h, the standard deviation of the speed was 22.0, the max speed was 96.5 km/h and the altitude variation in this route was 156 m. In this case, the final fuel consumption approximation error is only 1.4%. The CDF of the fuel consumption approximation when using the smartphone data together with the "real" fuel consumption data is presented in Fig. 4d. The "smartphone" curves, as the previous times, follow the "real" data (see Figs. 4d and 5d).

Throughout the last three field tests, we covered a large area, including the centre and the surrounding region of the city of Athens from the East to the West and from the North to the South (see the map of Fig. 2b). The following two field trials, however, took place with a different driver (Driver3) and car (Vehicle3) again in Athens, but in completely different routes (see the maps of Fig. 2c and d). In these tests we inserted into the algorithm the specific values for the mass, the drag coefficient, the frontal area of the Suzuki Swift with the different factor *Mf* for this vehicle model, and then followed the same procedure as in the previous tests.

The fifth field trial (see route *c1* of Fig. 2c) is a normal route, including highway, streets and alleys, in the west Attica in the surrounding area of the capital of Greece. In this route the average speed of the trial was 37.5 km/h, the standard deviation of the speed was 25.8, the max speed was 97.1 km/h and the altitude variation in this route was 111 m. The total approximation error here was only 1.2%. One can observe again the remarkable efficiency of our algorithm (see the corresponding diagrams in Figs. 4e and 5e). As we observed in the previous cases the only difference in this test between the OBD and the smartphone data is again at the beginning of the route when the vehicle's engine is cold. The air temperature during the route was about 1 °C. It is worth to mention here that, although the vehicle was produced eleven years ago, the curves of the CDFs of "smartphone" and "real" fuel consumption data are incredibly close.

The last field trial (see route *c2* of Fig. 2d) was carried out with three different driving styles (passive, normal and restless) in a place with a small hill with a very high slope that we “climbed” three times in a row. The average speed of this trial was 24.7 km/h, the standard deviation of the speed was 14.9, the max speed was 66.7 km/h and the altitude variation in this route was 53 m. Once again, the operation (see Figs. 4f and 5f) of the algorithm was excellent since the total fuel consumption’s estimation error is only 0.5%.

## 4 Concluding Remarks

In the present study, we performed six field tests in order to determine a vehicle’s fuel consumption. We developed an algorithm in order to improve the accuracy of GPS data recorded via a smartphone with respect to the real vehicle data extracted from OBD. We found that the approximation error between the two data sets (smartphone and OBD) varies between 0.3 and 3.7%, except one case where we observed 10.1% error and in which the fuel consumption rate of the real data is changing rapidly.

In general, the results show that the smartphone’s data could be used for the identification of a vehicle’s fuel consumption during a trip. Using this approach, a driver could monitor their driving behavior in order to reduce their fuel consumption through specially designed applications. This method can become a valuable and inexpensive tool for reducing carbon emissions and developing macro-traffic simulation models. For this to happen, however, implementation of smartphone applications at a large scale is required.

## References

1. Kan Z, Tang L, Kwan MP, Zhang X (2018) Estimating vehicle fuel consumption and emissions using GPS big data. Int J Environ Res Public Health 15(4):566
2. Cai H, Xie SD (2007) Estimation of vehicular emission inventories in China from 1980 to 2005. Atmos Environ 41:8963–8979
3. Onyekpe U, Palade V, Kanarachos S, Szkolnik A (2021) IO-VNBD: Inertial and Odometry benchmark dataset for ground vehicle positioning. Data Brief 35:106885
4. Weber Y, Kanarachos S (2020) CUPAC–The Coventry University public road dataset for automated cars. Data Brief 28:104950
5. Christopoulos SRG, Kanarachos S, Chroneos A (2018) Learning driver braking behavior using smartphones, neural networks and the sliding correlation coefficient: road anomaly case study. IEEE Trans Intell Transp Syst 20(1):65–74
6. Kanarachos S, Mathew J, Fitzpatrick ME (2019) Instantaneous vehicle fuel consumption estimation using smartphones and recurrent neural networks. Expert Syst Appl 120:436–447
7. Kanarachos S, Christopoulos SRG, Chroneos A (2018) Smartphones as an integrated platform for monitoring driver behaviour: the role of sensor fusion and connectivity. Transp Res part C Emerg Technol 95:867–882
8. Combustion of Fuels—Carbon Dioxide Emission (n.d.) available from <https://www.engineeringtoolbox.com/co2-emission-fuels-d_1085.html> [2 Feb 2018]

9. Pinto G, Oliver-Hoyo MT (2008) Using the relationship between vehicle fuel consumption and $co_2$ emissions to illustrate chemical principles. J Chem Educ 85(2):218
10. Asim F (2015) 'AndroSensor'. https://play.google.com/store/apps/details?id=com.fivasim
11. Husnjak S, Forenbacher I, Bucak T (2015) Evaluation of eco-driving using smart mobile devices. PROMET-Traffic Transp 27(4):335–344
12. Campolo C et al (202) SMaRTCaR: an integrated smartphone-based platform to support traffic management applications. In: First international workshop on vehicular traffic management for smart cities (VTM)
13. Moniaga JV et al (2018) Diagnostics vehicle's condition using obd-ii and raspberry pi technology: study literature. J Phys Conf Ser 978(1). IOP Publishing
14. Groves PD (2011) Shadow matching: a new GNSS positioning technique for urban canyons the. J Navig 64:417–430
15. Wang L, Groves PD, Ziebart MK (2012) Multi-constellation GNSS performance evaluation for urban canyons using large virtual reality city models. J Navig 65(3):459–476
16. Wang L, Groves PD, Ziebart KZ (2013) Urban Positioning on a Smartphone: Real-Time Shadow Matching Using GNSS and 3D City Models. In: Proceedings of the ION GNSS, 'ION GNSS'. held 2013 at Nashville, Tennessee
17. Abramowitz M, Stegun I (1970) Handbook of mathematical functions. Dover, New York
18. Bronshtein IN, Semendyayev KA, Musiol G, Mühlig H (2015) Handbook of mathematics [online], 6th edn. Springer-Verlag Berlin Heidelberg. Available from http://www.springer.com/gb/book/9783662462201
19. Walck C (2007) Hand-book on statistical distributions for experimentalists. University of Stockholm, 10

# Passive and Active Suspension Systems Analysis and Design

Yuri A. Vershinin

**Abstract** Active suspension systems allow one to improve the performance of a vehicle, to reduce the fuel (energy) consumption and exhaust emissions. This in turn allows one to improve the transport traffic and well-being in cities. The analysis and design of several types of semi-active and active suspension systems is provided in this paper.

**Keywords** Automotive suspensions systems · Active suspension systems · Modelling and simulation · Vehicle performance

## 1 Introduction

Increased competition on the automotive market has forced companies to research into alternative strategies to classical passive suspension systems [1, 2]. To improve handling and comfort performance, instead of a conventional static spring and damper system, semi-active and active systems are being developed [5, 6]. A semi-active suspension system involves the use of a dampers or spring with variable gain [3, 4]. Such systems can only operate on three fixed positions: soft, medium and hard damping or stiffness. Additionally, a semi-active system can only absorb the energy from the motion of a car body.

Alternatively, an active suspension system possesses the ability to reduce acceleration of sprung mass continuously as well as to minimise suspension deflection, which results in improvement of tyre grip with the road surface, thus, brake, traction control and vehicle maneuverability can be considerably improved.

Y. A. Vershinin (✉)
Coventry University, Coventry, UK
e-mail: o292j@gmx.com

Y. A. Vershinin et al. (eds.), *Technologies for Smart Cities*,
https://doi.org/10.1007/978-3-031-05516-4_9

## 2 Passive Suspension System

A one-degree-of-freedom (1-DOF) of a passive suspension system is given in Fig. 1. (We do not take into account the mass and stiffness of a wheel in this example).

A mathematical model of a passive suspension system can be obtained from the Newton's second law according to the free-body diagram, given in Fig. 2, as

$$m\ddot{x}_1(t) = -k(x_1(t) - x_0(t)) - b(\dot{x}_1(t) - \dot{x}_0(t)) \tag{1}$$

where

m is the ¼ car body mass,
k is the suspension spring coefficient,
b is the suspension damping coefficient,
$x_0$ is the road vertical disturbance (input signal),
$x_1$ is the vertical displacement of the sprung mass (output signal).

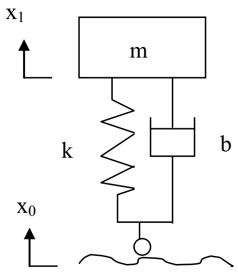
Re-arrange Eq. (1) as follows:

$$m\ddot{x}_1(t) + b(\dot{x}_1(t) - \dot{x}_0(t)) + k(x_1(t) - x_0(t)) = 0 \tag{2}$$

Equation (2) can be represented in the standard form as:

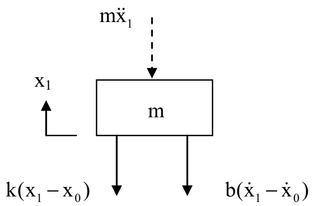
$$\ddot{x}_1(t) + \frac{b}{m}(\dot{x}_1(t) - \dot{x}_0(t)) + \frac{k}{m}(x_1(t) - x_0(t)) = 0 \tag{3}$$

**Fig. 1** A 1-DOF passive suspension system



Road surface

**Fig. 2** The free-body diagram



Represent Eq. (3) in the Laplace form

$$\mathrm{s}^2\mathrm{X}_1(\mathrm{s}) + \mathrm{s}\frac{\mathrm{b}}{\mathrm{m}}(\mathrm{X}_1(\mathrm{s}) - \mathrm{X}_0(\mathrm{s})) + \frac{\mathrm{k}}{\mathrm{m}}(\mathrm{X}_1(\mathrm{s}) - \mathrm{X}_0(\mathrm{s})) = 0 \tag{4}$$

or

$$\mathrm{s}^2\mathrm{X}_1(\mathrm{s}) + \mathrm{s}\frac{\mathrm{b}}{\mathrm{m}}\mathrm{X}_1(\mathrm{s}) + \frac{\mathrm{k}}{\mathrm{m}}\mathrm{X}_1(\mathrm{s}) = \mathrm{s}\frac{\mathrm{b}}{\mathrm{m}}\mathrm{X}_0(\mathrm{s}) + \frac{\mathrm{k}}{\mathrm{m}}\mathrm{X}_0(\mathrm{s}) \tag{5}$$

Equation (5) can be represented as

$$\mathrm{X}_1(\mathrm{s})\left[\mathrm{s}^2 + \frac{\mathrm{b}}{\mathrm{m}}\mathrm{s} + \frac{\mathrm{k}}{\mathrm{m}}\right] = \mathrm{X}_0(\mathrm{s})\left[\frac{\mathrm{b}}{\mathrm{m}}\mathrm{s} + \frac{\mathrm{k}}{\mathrm{m}}\right] \tag{6}$$

The transfer function between the vertical displacement of the sprung mass and the road vertical disturbance can be obtained from (6) as

$$\frac{\mathrm{X}_1(\mathrm{s})}{\mathrm{X}_0(\mathrm{s})} = \frac{\frac{\mathrm{b}}{\mathrm{m}}\mathrm{s} + \frac{\mathrm{k}}{\mathrm{m}}}{\mathrm{s}^2 + \frac{\mathrm{b}}{\mathrm{m}}\mathrm{s} + \frac{\mathrm{k}}{\mathrm{m}}} \tag{7}$$

The transfer function (7) can be represented in the form of polynomials

$$\mathrm{G}(\mathrm{s}) = \frac{\mathrm{X}_1(\mathrm{s})}{\mathrm{X}_0(\mathrm{s})} = \frac{\mathrm{b}_1\mathrm{s} + \mathrm{b}_0}{\mathrm{s}^2 + \mathrm{a}_1\mathrm{s} + \mathrm{a}_0} \tag{8}$$

where

$$\mathrm{b}_0 = \frac{\mathrm{k}}{\mathrm{m}}, \quad \mathrm{b}_1 = \frac{\mathrm{b}}{\mathrm{m}}, \quad \mathrm{a}_0 = \frac{\mathrm{k}}{\mathrm{m}}, \quad \mathrm{a}_1 = \frac{\mathrm{b}}{\mathrm{m}}$$

The standard form of a second order system in the transfer function representation is given as

$$\mathrm{G}(\mathrm{s}) = \frac{\mathrm{k}_{\mathrm{ss}}\omega_n^2}{\mathrm{s}^2 + 2\varsigma\omega_n\mathrm{s} + \omega_n^2} \tag{9}$$

where

$k_{ss}$ is the steady-state gain,
$\varsigma$ is the damping ratio,
$\omega_n$ is the natural frequency.

Compare Eqs. (7) and (9) we can obtain the damping ratio $\varsigma$ and natural frequency $\omega_n$ of the passive suspension system as

$$\omega_n = \sqrt{\frac{k}{m}}, \quad \varsigma = \frac{\frac{b}{m}}{2\omega_n}. \tag{10}$$

The simulation of the passive suspension system has been performed using the MATLAB package.

The following parameters are used:

$$\begin{aligned}
&\text{1/4 body mass,}\ \ m\ = 530.6\ \text{kg}\\
&\text{suspension spring coefficient, } k = 22{,}750\ \text{N/m,}\\
&\text{suspension damping coefficient,}\ \ b = 700\ \text{Ns/m.}
\end{aligned} \tag{11}$$

The transfer function of the passive suspension system (7) is obtained in the form:

$$G_p(s) = \frac{1.3193\ s + 42.876}{s^2 + 1.3193\ s + 42.876} \tag{12}$$

The frequency response characteristics of the mass vertical displacement $x_1(t)$ versus the road vertical disturbance $x_0(t)$ is given in Fig. 3:

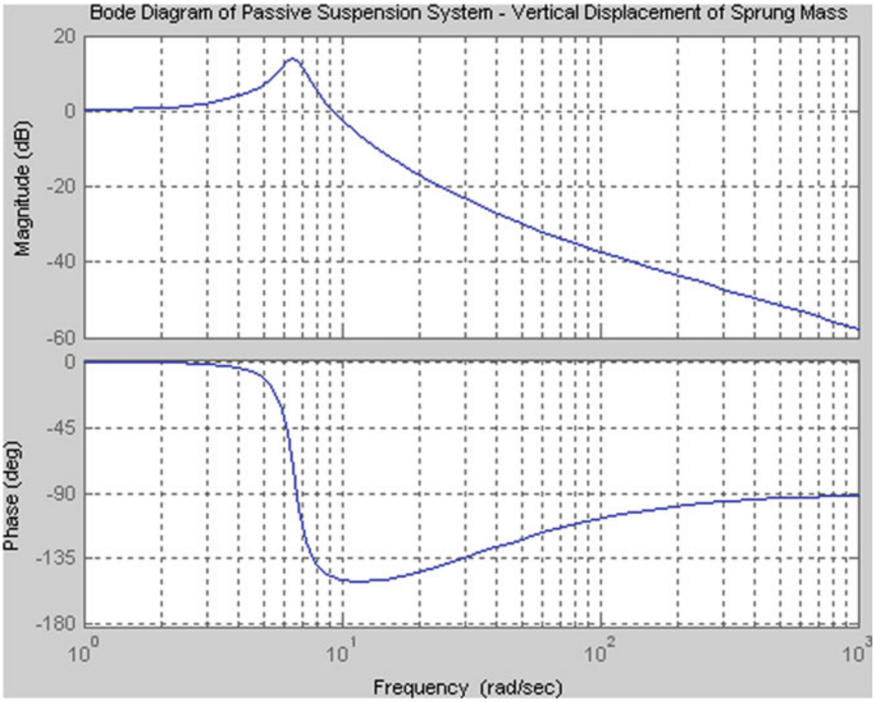


**Fig. 3** Frequency response characteristics of the mass vertical displacement versus the road vertical disturbance of the passive suspension system

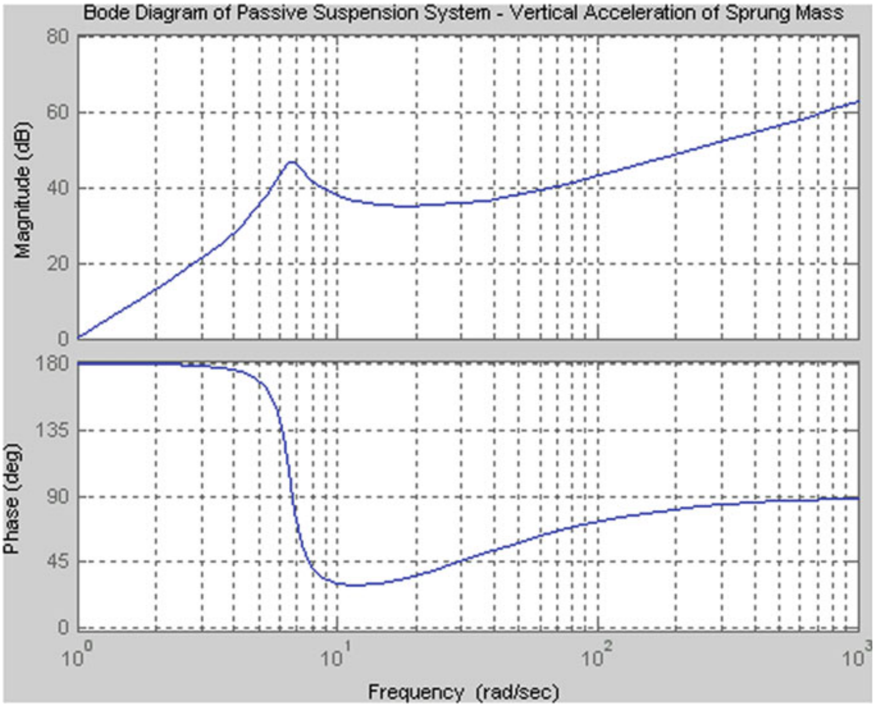


**Fig. 4** Frequency response characteristics of the mass vertical acceleration versus the road vertical disturbance of the passive suspension system

The frequency response characteristics of the mass vertical acceleration $\ddot{x}_1(t)$ versus the road vertical disturbance $x_0(t)$ is given in Fig. 4.

Represent the system (8) in the block-diagram form in order to obtain the behaviour of the passive suspension in the time domain. The input-output relationship can be obtained from (8) as

$$X_1(s) = \frac{b_1 s + b_0}{s^2 + a_1 s + a_0} X_0(s) \tag{13}$$

If we define:

$$Q(s) = \left[\frac{1}{s^2 + a_1 s + a_0}\right] X_0(s) \tag{14}$$

then, we can re-write (13) as

$$X_1(s) = (b_1 s + b_0) Q(s) \tag{15}$$

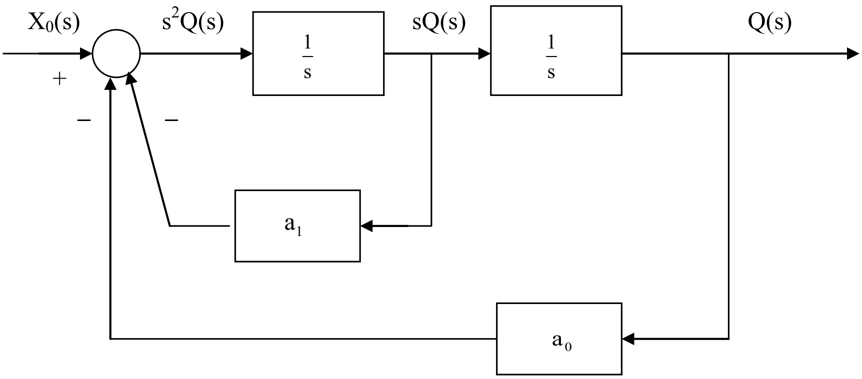


**Fig. 5** Representation of the system (8) in the block-diagram form

Therefore,

$$X_1(s) = b_1 sQ(s) + b_0 Q(s) \tag{16}$$

From (14) we can obtain the following:

$$s^2Q(s) + a_1 sQ(s) + a_0 Q(s) = X_0(s) \tag{17}$$

Re-arrange Eq. (17) in the following form:

$$s^2Q(s) = -a_1 sQ(s) - a_0 Q(s) + X_0(s) \tag{18}$$

Represent (18) in the block-diagram form (Fig. 5):

Now, we can represent the complete system using Eq. (16) in the following block diagram form (Fig. 6):

Denote:

$$\begin{aligned} Q_1(s) &= Q(s) \\ Q_2(s) &= sQ_1(s) \end{aligned} \tag{19}$$

Then, Eq. (18) can be represented in the following form:

$$sQ_2(s) = -a_1 Q_2(s) - a_0 Q_1(s) + X_0(s) \tag{20}$$

The output can be obtained from Eq. (16) as

$$X_1(s) = b_1 Q_2(s) + b_0 Q_1(s) \tag{21}$$

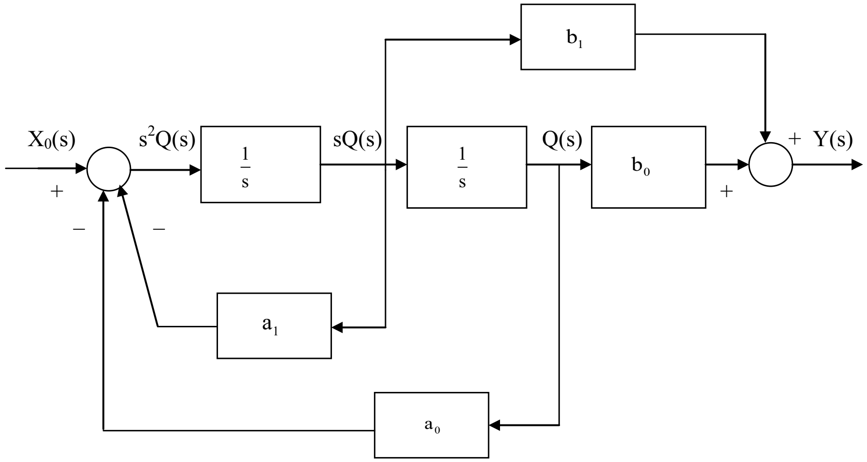


**Fig. 6** Representation of the equation (18) in the block-diagram form

Assuming that the initial conditions are zero, the state-variable model in the time-domain can be represented in the following form:

$$\begin{aligned} \dot{q}_1(t) &= q_2(t) \\ \dot{q}_2(t) &= -a_1 q_2(t) - a_0 q_1(t) + x_0(t) \\ x_1(t) &= b_1 q_2(t) + b_0 q_1(t) \end{aligned} \tag{22}$$

The system is designed using standard blocks from the Simulink library. This is given in Fig. 7.

The impulse response of the passive suspension system is given in Fig. 8.

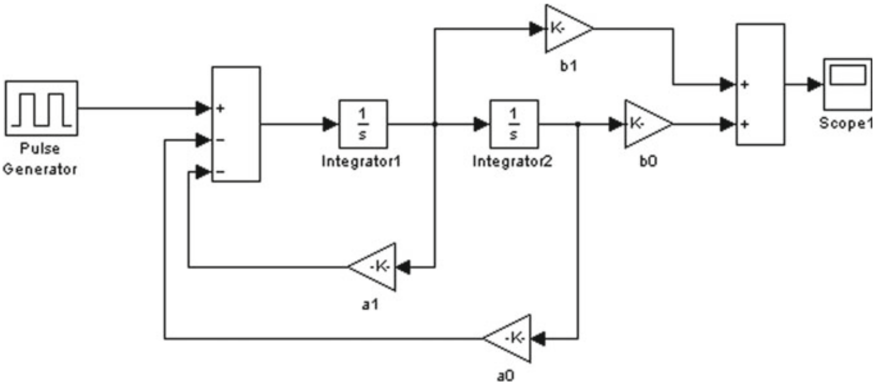


**Fig. 7** Representation of the complete system in the Simulink form

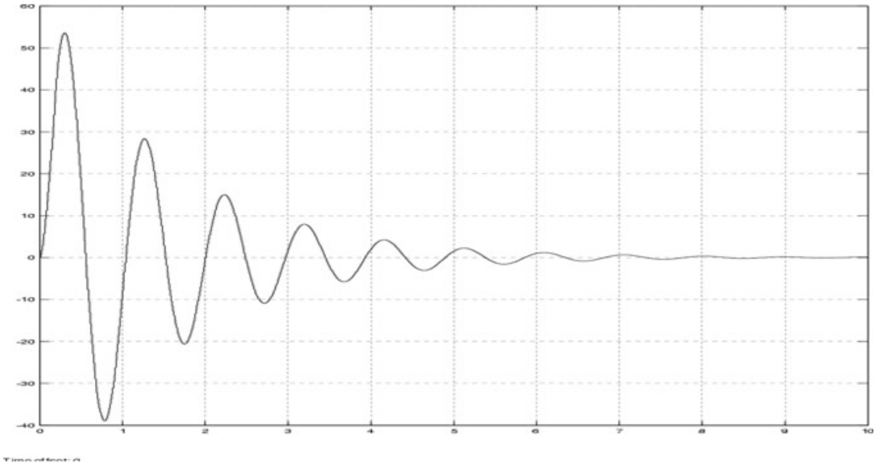


**Fig. 8** The impulse response of the passive suspension system.

## 3 Active Suspension System

***Sky-hook damper control system*** (Fig. 9)

It is well known that sky-hook control is effective in suppression of sprung mass vibrations. Thus, implementation of this system can dramatically improve comfort of driving. A theory of operation of a sky-hook damper system is given in Fig. 10.

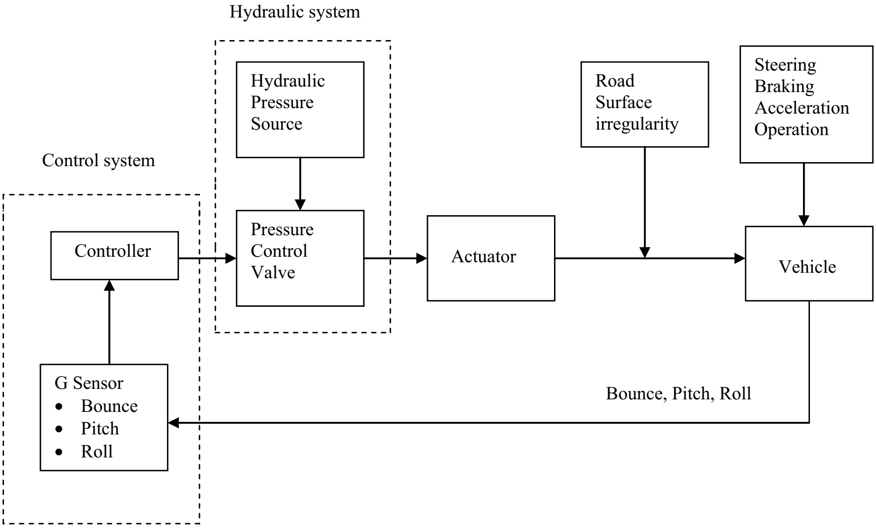


**Fig. 9** Hydraulic and control systems for an active suspension

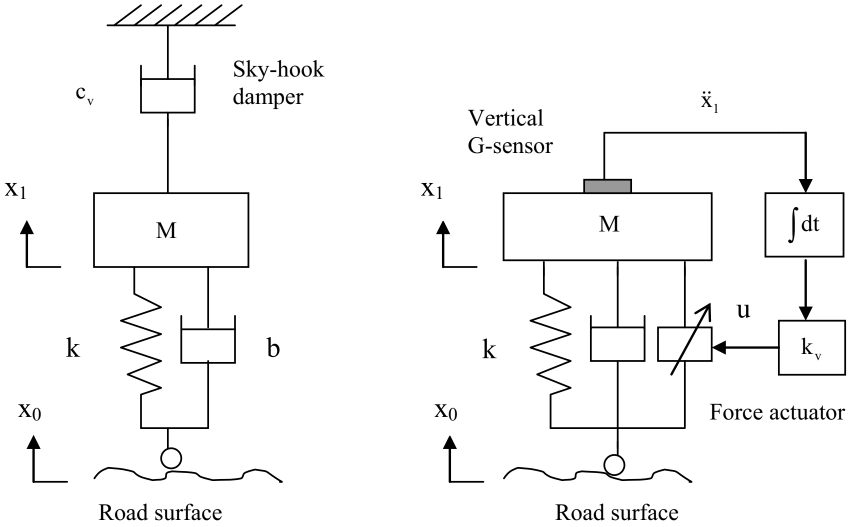


**Fig. 10** The idea of sky-hook damper control for a 1-DOF active suspension system

A sky-hook damper is a virtual damper and control signal $\boldsymbol{u(t)}$ is calculated using absolute velocity of a car body as follows:

$$\mathrm{u(t)} = -\mathrm{c_v}\dot{\mathrm{x}}_1(t) \tag{23}$$

where

$\dot{x}_1(t)$ is the absolute velocity of vertical sprung mass motion,
$k_v$ is the coefficient of the sky-hook damper.

A mathematical model of an active suspension system can be represented in the following form:

$$m\ddot{x}_1(t) + b(\dot{x}_1(t) - \dot{x}_0(t)) + k(x_1(t) - x_0(t)) = u(t) \tag{24}$$

Substitute (23) into (24):

$$m\ddot{x}_1(t) + b(\dot{x}_1(t) - \dot{x}_0(t)) + k(x_1(t) - x_0(t)) + k_v\dot{x}_1 = 0 \tag{25}$$

Represent the Eq. (25) in the Laplace form as

$$ms^2X_1 + bsX_1 - bsX_0 + kX_1 - kX_0 + k_vsX_1 = 0 \tag{26}$$

Re-write Eq. (26) in the following form

$$X_1(ms^2 + bs + k + k_vs) = X_0(bs + k) \tag{27}$$

The transfer function between the mass vertical displacement and the road disturbance can be obtained from Eq. (27) as:

$$\frac{X_1(s)}{X_0(s)} = \frac{\frac{b}{m}s + \frac{k}{m}}{s^2 + \frac{b+k_v}{m}s + \frac{k}{m}} \tag{28}$$

The simulation of the active suspension system (28) has been performed on MATLAB with the parameters given in (11) and $k_v = 650$ Ns/m.

The transfer function of the active suspension system (28) is obtained in the form:

$$G_a(s) = \frac{1.3193\ s + 42.876}{s^2 + 2.827\ s + 42.876} \tag{29}$$

The frequency response characteristic of the mass vertical displacement $x_1(t)$ versus the road disturbance $x_0(t)$ is given in Fig. 11.

The frequency response characteristic of the mass vertical acceleration $\ddot{x}_1(t)$ versus the road disturbance $x_0(t)$ is given in Fig. 12.

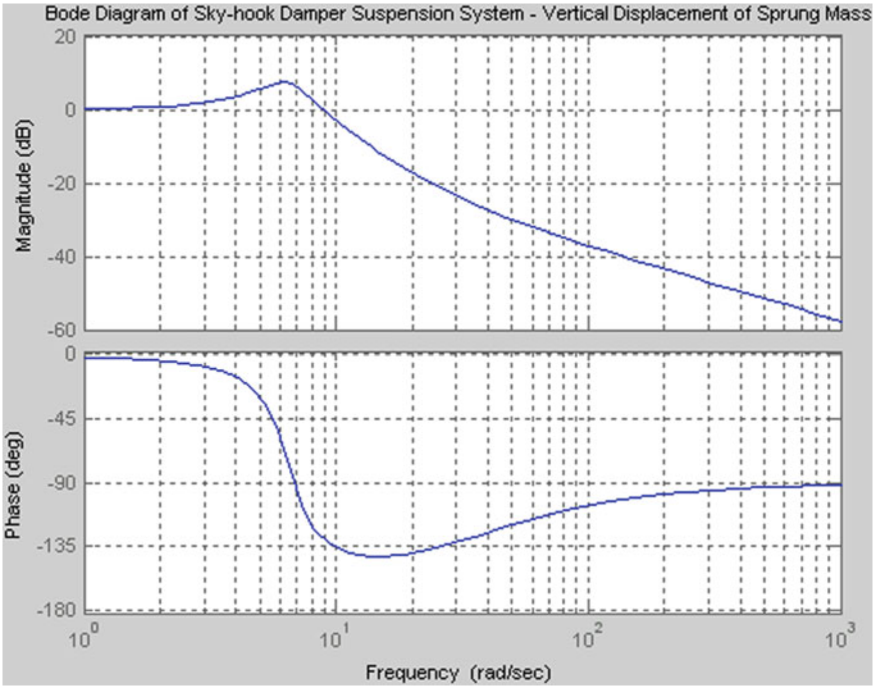


**Fig. 11** Frequency response characteristics of the mass vertical displacement versus the road vertical disturbance of the active suspension system

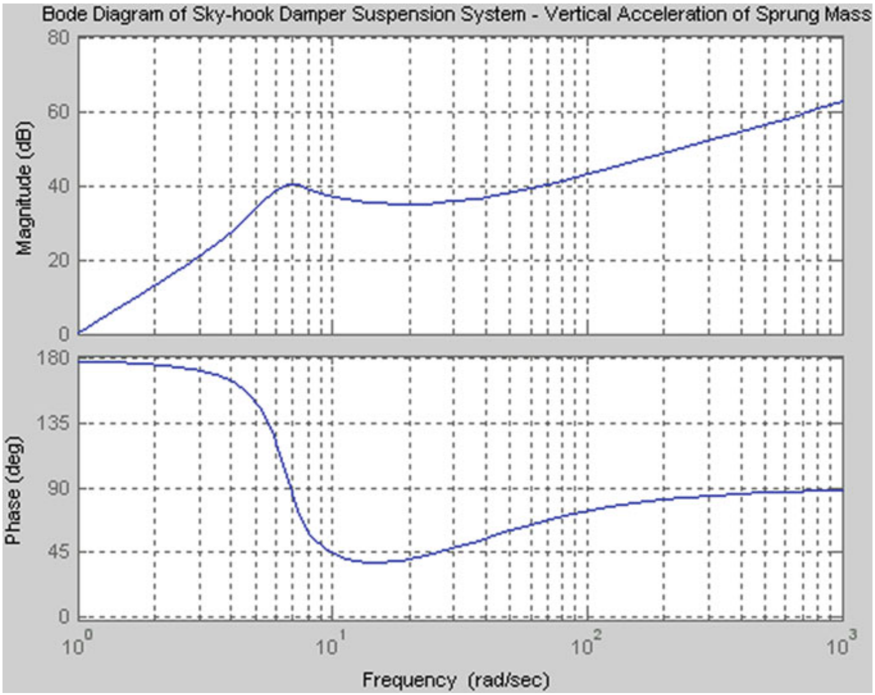


**Fig. 12** Frequency response characteristics of the mass vertical acceleration versus the road vertical disturbance of the active suspension system

It can be seen from Figs. 3 and 11 that the magnitude of the peak (on the natural frequency) of mass vertical displacement of the active suspension is lower than that of the passive suspension. It follows from Figs. 4 and 12 that the peak of mass vertical acceleration has also been suppressed on the active suspension.

The impulse response of the active suspension system is given in Fig. 13.

It can be seen from Figs. 8 and 13 that the transition response has been reduced from 9 s (on the passive suspension) to 3 s (on the active suspension).

Thus, the given above MATLAB/Simulink results prove the advantages of the active suspension system when compare with the passive suspension.

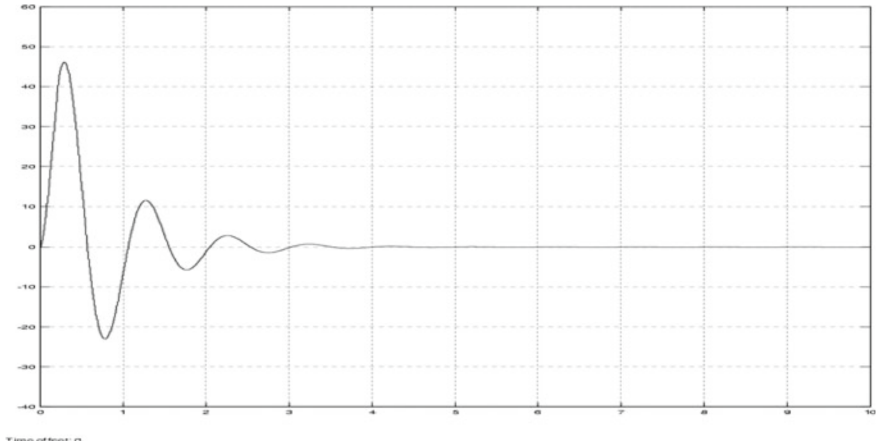


**Fig. 13** The impulse response of the active suspension system

# 4 Conclusions

The analysis of several types of semi-active and active suspension systems is given in this paper. It has been demonstrated that active suspension systems provide the better performance in the frequency domain. The time response of an active suspension system on the applied input is faster as well. The signal from an accelerometer is used for the control system. Three-axes accelerometers are usually available on vehicles. Such accelerometers are specifically designed for the applications on vehicles. The accelerometers are not expensive and robust devices. They work in the acceptable range of driving and environmental conditions. Therefore, accelerometers are the preferable alternative to liner potentiometers and velocity sensors to measure the vertical displacement and velocity of the displacement of a car-body. The numerical integration of the signal from an accelerometer can be implemented on the automotive ECU (Electronic Control Unit).

# References

1. Alexander D (2005) Handling the ride. Autom Eng Int 44–50
2. Cao D, Rakheja S, Su C-Y (2010) Roll- and pitch-plane coupled hydro-pneumatic suspension. Part 1: feasibility analysis and suspension properties. Veh Syst Dyn 48:361–386
3. Ivers DE, Miller LR (1991) ‘Semiactive suspension technology: an evolutionary view. In: ASME advanced automotive technologies, DE-40, Book No. H00719-1991, pp 327–346
4. Song X, Ahmadian M (2010) Characterization of semiactive adaptive control algorithms with application of magneto-rheological dampers. J Vibr Control

5. Williams RA (1997) Automotive active suspensions. Part 1: basic principles. J Automob Eng 211:415–426
6. Williams RA (1997) Automotive active suspensions. Part 2: practical considerations. J Automob Eng 211:427–444

# Architecture and Its Vulnerabilities in Smart-Lighting Systems

**Florian Hofer and Barbara Russo**

**Abstract** Industry 4.0 embodies one of the significant technological changes of this decade. Cyber-physical systems and the Internet Of Things are two central technologies in this change that embed or connect with sensors and actuators and interact with the physical environment. However, such systems-of-systems undergo additional restrictions in an endeavor to maintain reliability and security when building and interconnecting components to a heterogeneous, multi-domain *Smart-** systems architecture. This paper presents an application-specific, layer-based approach to an offline security analysis inspired by design science that merges preceding expertise from relevant domains. With the example of a Smart-lighting system, we create a dedicated unified taxonomy for the use case and analyze its distributed Smart-* architecture by multiple layer-based models. We derive potential attacks from the system specifications in an iterative and incremental process and discuss resulting threats and vulnerabilities. Finally, we suggest immediate countermeasures for the latter potential multiple-domain security concerns.

**Keywords** Industry · 4.0 · CPS · Security · Smart-city · Smart-lighting

## 1 Introduction

Progressive computerization brings technology into every corner and improves the automation and performance of environmental and manufacturing processes. The German Government envisions the fourth industrial revolution as an inevitable prospect for future development. This revolution endows modern systems with "Smart" attributes to increase operational efficiency, share information, and improve

F. Hofer (✉)
Post-Doctoral Researcher, SwSE, Free University of Bolzano, Piazza Domenicani 3, Bolzano, Italy
e-mail: Florian.Hofer@unibz.it

B. Russo
Full Professor, SwSE, Free University of Bolzano, Piazza Domenicani 3, Bolzano, Italy
e-mail: brusso@unibz.it

Y. A. Vershinin et al. (eds.), *Technologies for Smart Cities*,
https://doi.org/10.1007/978-3-031-05516-4_10

their services' quality [20]. These endowments allow the creation of fully flexible production systems. They bring in new business models, services, and products via *Smart-** systems such as *Smart-Home* or *Smart-City* [23].

Smart technologies rely on Cyber-physical systems (CPS) and the Internet of Things (IoT) to achieve such goals. *Smart-** systems operate via an autonomous, decentralized decision-making process that allows for local and faster reaction and thus enables higher efficiency and production quality [21]. They run on a network of intelligent devices of diverse make and function, requiring standardized interfacing and communication. This heterogeneity could lead to inconsistencies making a system vulnerable. System attacks can exploit those vulnerabilities to eavesdrop or harm an asset's value, causing virtual and physical loss. This distress is particularly the case of *Smart-Lighting* systems, where publicly installed devices may be subject to physical and cyber-attacks [3].

Consequently, modern architectural designs needs must include security (*security by design*). However, this is a hard-to-achieve task in a multi-domain environment where definitions and analysis models defer between the application-relevant functional models. Furthermore, a recent systematic mapping study identifies a lack of research on security for Industry 4.0 architectures [19]. As a result, security often experienced neglect, and existing architectural models frequently suffer from simplifications and assumptions from the "offline"-secure world. Thus, there is a need for guidelines to build architectures and their models that incorporate security concerns. Such guidelines would also help assess systems' multi-domain vulnerabilities and propose strategic and preventive countermeasures [21] or determine corrective mitigation measures that could reduce or eliminate a vulnerability [25].

This study presents a layer-based analysis and classification technique of architectural vulnerabilities for multi-domain systems. While specific new weaknesses may appear when interconnecting such heterogeneous systems, this chapter focuses on a technique that extends our knowledge of a system through existing results from relevant domains. We explore security concerns within several reference models designed from connected CPSs or IoT sub-domains to extend our assessment beyond a single-domain analysis systematically. As a practical example of such a process, we examine the decentralized *Smart-Lighting* architecture of an in-field case study. Overall, the contributions of this work are:

- A technique to associate and classify architectural layers of existing reference models for CPS and IoT environments;
- A technique to consolidate the taxonomy of threats and attacks for vulnerability analysis of multi-domain, multi-layer Smart-* architectures;
- An architectural analysis and multi-domain taxonomy on vulnerabilities and attacks of a Smart-Lighting system.

We organized the rest of the chapter as follows. Sections 2 and 3 present related work and our methodology and evaluation strategy. In Sect. 4, we examine the case study and create its multi-domain architecture reference and security taxonomy tables. In Sect. 5, we apply our iterative classification technique using these tables continued by discussing threats, vulnerabilities, and possible countermeasures. Finally, we conclude in Sect. 6.

# 2 Related Work

We identified three major security topics: assessment through architecture layers, (traditional) offline analysis tools, and architecture design and patterns. In addition, we select cornerstone studies and describe their relevance in Industry 4.0 in the following.

## *2.1 Security and Layers*

Lezzi et al. [21] analyze how research deals with the current cybersecurity issues in Industry 4.0 contexts, laying down the state of development regarding Smart-* architectures. The authors argue that an ideal design and development strategy considers cybersecurity from the start. The study identifies norms and guidelines for architecture security. It proposes structured solution approaches along with the taxonomy of standard cybersecurity terms. Within their list of threat identification methods, they mention a three-layer-based attack assessment technique. While they do not discuss the method's efficiency further, their concluding remarks highlight the lack of an all-layer cybersecurity analysis.

Although little research exists on vulnerability classifications in these new Smart contexts, we can adopt some published results on CPS architectures. Ashibani and Mahmoud [3] redacted a generic security analysis comparing CPS technologies to traditional IT security. The article is among the first to discuss the analysis and detection of multi-layer security requirements, possible attacks, and issues for information security on three architectural layers. However, their theoretical considerations appear limited to their feasibility, and many discussed terms had non-traceable sources.

Varga et al. [32] created an analogous, IoT focused overview. The study targets the automation domain with a fourth architecture layer for data processing, highlighting security threats and mitigation. The paper displays how similar analyses can impact results from their biased viewpoint. While IoT and CPS security present similarities, the article disregards typical distributed control and treats issues as binary problems making the analysis incomplete. However, the strong data-centric viewpoint helps in the assessment of data processing systems.

Han et al. [16] submit in their layer analysis a different aspect to vulnerabilities by classifying them as internal or external. They propose a four-plus-one layer architecture and a framework for an intrusion detection system (IDS). Due to the lack of a unique definition of CPS, the authors suggest an iterative application of appropriate mitigation strategies. Unfortunately, they apply this iterative notion to IDS design only. Furthermore, even though they deliver a control-centered selection of attacks for each layer, the article also admits definition issues.

## *2.2 (Traditional) Offline Analysis Tools for Security and Safety*

Safety and security relied on design time offline analysis tools for many years, a tradition that did not change much for cybersecurity. Bolbot et al. [6] describe the relationship between the two as a conditional dependence. Their article focuses on design-time safety assurance methods, their modifications, and their integration. They identify sources of CPSs' complexity and test offline assessment techniques against them. Within the remarks of this investigation, we find the need for a systematic method for issue identification. They highlight the importance of mixing and adapting existing techniques to deal with CPS's complexities to tackle cybersecurity issues.

Subramanian and Zalewski propose in [29, 30] an alternative assessment approach for non-functional requirements to connect security and safety in the CPS domain. The non-functional domain's well-defined ontology allows for an inter-dependency graph, which then propagates information as needed. The method shows how the dependencies of a single requirement can change an issue's weight. Majed et al. [24] suggests a framework for evaluating security exposure by weight on a connected graph. Via the shortest path, we can then identify the most accessible vulnerability. Although both are exciting approaches, the distribution of weight and path for each node remains unclear.

## *2.3 Architecture Design and Patterns*

Alguliyev et al. [2] analyze and classify in a recent literature review existing research on CPS security using the CIARR model, a variant of the CIAA security requirements. This variant separates availability into resilience and reliability, suggesting that CPS's non-functional requirements vary from traditional IT. The analysis discusses approaches of architectural design to improve system security. It draws up the context and risks, offers a generalized attack tree, proposes mitigation strategies, and informs about found countermeasures and dominant future research areas.

Ryoo et al. [26] try to break assessment conventions by proposing a generic new three-stage approach. The three phases collect information based on tactics, patterns, and vulnerabilities. The process guides an analyst through three security analysis phases with an improved weakness (CWE-1000) and entirely new architecture pattern databases. However, the method is still subject to refinement and tuning.

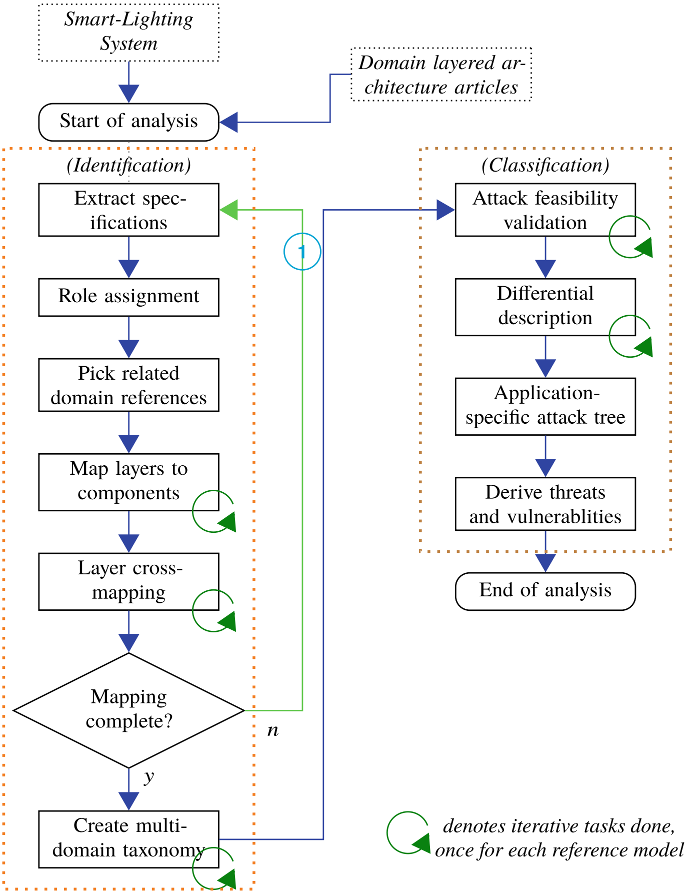

Smart-Lighting System
Domain layered architecture articles
Start of analysis
(Identification)
Extract specifications
1
Role assignment
Pick related domain references
Map layers to components
Layer cross-mapping
Mapping complete?
n
y
Create multi-domain taxonomy
(Classification)
Attack feasibility validation
Differential description
Application-specific attack tree
Derive threats and vulnerablities
End of analysis
denotes iterative tasks done, once for each reference model

**Fig. 1** Design cycle steps carried out for a Smart-* system vulnerability identification

## 3 Method

For our layer-based technique, we processed the target *Smart-Lighting* architecture (SLA) in all its components, Fig. 1. Inspired by design science [18], our technique combines results and experiences from a knowledge base of related domains. An iterative design process (design cycle) integrates them to create artifacts assessed in-field (relevance cycle). Finally, the present technique and resulting artifacts will grow the knowledge base (rigor cycle). The focus of this chapter's design process is described in the following consists of two steps: identification and classification.

*Identification.* As with any modern system-of-systems, an SLA contains multiple heterogeneous systems, each with its domain-specific constraints. Hence, we first analyze the system's composition by gathering specifications from technical data

**Table 1** Example LoRaWAN end-node layer role and definition differences for reference models

| Role | $RM_A$ [3] | $RM_H$ [16] | $RM_L$ [22] |
|---|---|---|---|
| Control | Application: […] process the received information from the data transmission level and issue commands to be executed by the physical units, sensors and actuators | Supervisory control: by aggregating the measurement data from multiple points in the network, the supervisory sub-control level creates system-level feedback control loops, which make system-level control decisions | Application: […] receives the data transmitted from network layer and uses the data to provide required services or operations. For instance, the application layer can provide the storage service […] or provide the analysis service to […] predicting the future state of physical devices |
| Communication | Transmission: is responsible for interchanging and processing data between the perception and the application. […] are achieved using local area networks, communication networks, the Internet or other existing networks […] | Network: […] takes charge of networking sensors and actuators as well as bridging the sensor/actuator layer and the higher control layer with a variety of communication devices and protocols | Network: […] used to receive the processed information provided by perception layer and determine the routes to transmit the data and information to the IoT hub, devices, and applications via integrated networks |

sheets and reconstructing its architecture diagram. In particular, each component gets assigned one or more architectural roles. For example, we will see that the LoRaWAN end-node controllers take up two roles, (B) in Fig. 2. They act as a communication gateway (networking) and perform decentralized supervision of the connected lighting-bus devices (control).

Domain-specific security aspects further characterize a component. For example, physical tampering characterizes a light device. Based on previous work [19], we select domain-specific layered architectural reference models ($RM_i$) from representative research papers. Each model holds a different architectural focus (e.g., control flow) or domain (e.g., IoT) and carries related information on possible layer-level attacks and vulnerabilities.

Based on component roles, we draw a map $M_i$ between the components and the layers of a model $\Lambda(RM_i)$. For example, the sensor and actuator layer contains a light device. Starting bottom-up, we iterate through the architecture components and survey each $RM$ for matching role descriptions. We ensure that: (a) every component fits into at least one layer of a reference model, and (b) for each layer, there exists a component that maps into it. The former ensures that each component can be described in each reference model and enriched with the information of a layer's attacks and vulnerabilities. The latter secures that all attacks described in each reference model find a target in our system. As an example, Table 1 shows the roles of a LoRaWAN lighting controller and their assigned layer mapping among three reference models. For $RM_A$'s generic CPS and IoT-oriented and $RM_L$'s service-oriented model, the "Control" role maps best to their Application layer. In contrast, in the control-oriented $RM_H$ model, this supervising component best fits the Supervisory Control sub-layer. While layer descriptions are similar, the focus diverges slightly between models. Also, the attack definitions for the mapped layers reflect such proximity. For Example, the definitions for *Malicious Code* $(RM_H)$ and *Malicious virus/worm* $(RM_L)$ refer to the same type of attack. However, they diverge due to focus, i.e., performance versus data-centric, emphasizing the importance of creating a unified taxonomy.

Each component now equips a role and projected attacks for each mapped reference model layer. Consequently, we can link layers from the different reference models through their common component mapping. This consolidation phase constructs thus cross-mappings $CM_{ij}$ among the reference model layers $\Lambda(RM_i)$ and $\Lambda(RM_j)$ based on component allocation. However, as one model's layer definition may encircle only a subset of another model's layer, cross-mappings are unidirectional and may not hold in reverse. It is typically the case for more abstract models that cross-map to layer-rich models. Therefore, we suggest starting with the former as it allows for easier and better associations.

In the final identification step, we use the cross-mapping $CM_{ij}$ to create a differential attack and threat table. This table contains relevant definitions to verify a Smart-Lighting system for attacks, threats, and vulnerabilities. First, we create an initial base table that records layer relationships and enriches each entry with its original research's attack taxonomy. We further append each attack's origin and original meaning for a more accurate understanding by following citations. Next, starting again from the least detailed model, e.g., $RM_A$, we review attack definitions of layers connected by the cross-mapping, $CM_{ij}$. We remove duplicates, integrate definitions, or highlight differences or ambiguities. If an undefined term appears, we define it according to its context with the help of other domain-related and reference sources. Once complete, we group the remaining entries by layers and threads to build the differential, application-specific taxonomy table. Consequently, the consolidation outputs are a layer mapping to the architecture and between models and a taxonomy table that includes the analyzed multi-domain perspective.

*Classification.* With the preceding table, we perform a differential weakness discovery for the SLA. We evaluate each attack's definition and assess if, in the Smart-Lighting domain, the proposed attacks remain possible or sensible. Starting bottom-up in the architecture, we pick a network or its next component and verify each attack in the differential taxonomy table for every assigned layer in the model. The process repeats until it analyzes all reference model layers and SLA components. We

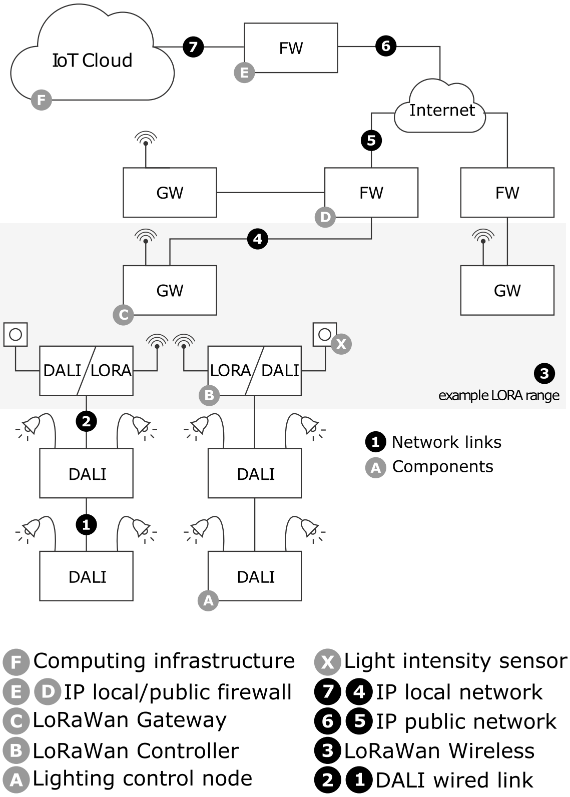


**Fig. 2** Smart-Lighting case study architecture

summarize and discuss the results of the attack analysis in a differential description that presents newly found attacks while honoring a previous model's results.

To highlight differences and commonalities to CPS of a Smart-Lighting system, we create a domain-specific attack tree based on the generic tree created by Alguliyev et al. [2]. Then, using their attacks-threats functional CPS model, we reuse or define further attacks and threats in the taxonomy derived from our study. The reviewing of vulnerability definitions of the reference articles and the resulting attack tree will then serve as input for a final assessment of the possible vulnerabilities in a Smart-Lighting system.

# 4 Identification

## 4.1 *Smart-Lighting Architecture Under Study*

To explain our method, we use the case study architecture of a Smart-Lighting installation. The *Smart-City* pilot project runs in the city of Merano, Italy. It covers an area of 26 km$^2$ and more than 6.700 distributed light posts. Figure 2 illustrates the result of the identification step. The figure shows a simplified version for the demonstrative purpose of the installed system's architecture containing all the elements needed to create smart, remotely controlled lighting infrastructure. Three different networking technologies convey the control and status information between the end-nodes and the computing cloud: light devices (1–2), wireless network (3), and a traditional IP-based network (4–7). Each of the three networks varies in function and timing constraints, acting thus as a leveled control model [15].

*Dali end-nodes*. Digital Access Light Interface (DALI), a master-slave two-wire message-based bus for lighting and illumination systems, interconnects the light devices [5]. Its self-clocked differential encoding runs the data on a low data rate of 1200 baud in half-duplex, and when externally powered, for multiple hundred meters and resilient to interference [5, 17]. The DALI end device controllers (A), also called ballast controllers, execute simple application-specific programs and require only small micro-controllers [10]. In 2017, the Digital Illumination Interface Alliance (DiiA) released a revised version of the standard. DALI 2 standardizes timing requirements and signal slopes, increasing interoperability [12]. It also adds multi-master operation or multiple logical units per bus device while maintaining backward compatibility with DALI 1.

*The LoRaWAN network*. The wireless star-of-stars network is designed on a LoRaWAN (Long Range Wide Area Network) master-slave protocol that runs on top of a Semtech LoRa wireless transmitter [8]. The transmitter operates in Industrial-Scientific-Medical (ISM) band with either 250 half-duplex channels of 5.5 kbps and one at 11 kbps in chirp spread spectrum (CSS), or one channel of 50 kbps in frequency shift key (FSK) modulation (Europe channels). Its transmission robustness outperforms traditional systems, enabling servicing thousands of devices and reducing the need for a mesh network [4]. The LoRaWAN network associates nodes (B) through gateways (C) to network and application servers (F). A LoRaWAN end-node (B) can have different modes: event-driven sensors, beacon scheduled actuators (usually both battery-powered), or always online. It stores two AES128 keys, securing the communication to the *network* and *application* server. The installed gateways (C) serve as bi-directional relays and mount multichannel-multi-modem units for simultaneous reception on different frequencies and data rates without any end-node association or handover. A network server manages the distribution of data flow between an application and nodes. It reconfigures a gateway's multi-modem channels and data rate according to needs and environmental conditions. Such an adaptation targets the shortest air time (Adaptive data rate) and the best channel diversity (Channel maps) while increasing overall transmission efficiency and total throughput [8]. As the entry and exit-point of data flow are not binding, LoRaWAN supports redundancy by default. However, the network setup makes direct end-node communication impossible [4]. The used LoRaWAN end-device mounts a LoRaWAN/DALI master controller for routing and the timed control of connected DALI devices, and Bluetooth LE hardware for the initial configuration setup [31]. It offers over-the-air (OTA) firmware update and OTA device activation and features digital and analog inputs and outputs to attach optional sensor-actuator hardware. The hardware of the used LoRaWAN gateway mounts an ARM Cortex-A™processor running a Linux kernel. It allows user program deployment and features a backup uplink over 3G.

*The IP infrastructure*. The IP-based infrastructure is configured as in a traditional IT system. Local networks use IPv4 or IPv6 connectivity through Internet (5–6) and unite firewalls with gateways (4) and computation cloud (7). Within and between networks, standard protocols (IPSec/HTTPS) secure connections. The firewalls (D–E) perform routing and protection tasks, providing traditional intrusion detection algorithms. The cloud environment (F) stores and analyzes data. The data from the on-site gateways enters the cloud through a software firewall, which forwards it to the "Loriot IoT" network server running as an IaaS instance. The latter forwards the message payload to a PaaS application server operating an "Azure IoT" service running a set of custom-developed "Azure Functions" and micro-services. These gather and store the acquired information in a "Kosmos DB" No-SQL database and take control measures accordingly. The virtual LAN and firewall (E) configuration allow internal data flow governance and additional fine-grained protection mechanisms.

## 4.2 Layer Mapping

To cover all aspects of our SLA, we use four reference models, $RM_i$, originating from varying domains: one model to cover generic aspects of CPS's information security, one to highlight and stress the importance of information and control flow in CPS, and two to extend aspects peculiar to IoT, Big Data, and service orientation. The SLA maps then to the reference models. Reference models $RM_i$, their mappings $M_i$ and cross-mappings $CM_{ij}$ are further detailed in the rest of this section.

$RM_H$ [16]: the architecture of systems arranges in a 4 plus 1 layer model, Fig. 3 center. Its layer stack contains the Physical, Sensor and Actuator, Network, and Control layers. The latter divides into three control-oriented sub-layers: Local-distributed control action layer, Supervisory sub-control level, and Supervisory higher control level. This division highlights hierarchical separation and enables distributed independent control. The plus one (Information) layer interfaces transversely, acting on all four stacked layers. It represents the information flow sinking to the top and sourcing from among all layers in the architecture or vice versa, supporting the notion of shared information for distributed control.

$RM_A$ [3]: uses a three-layer approach defined as the Perception, Transmission, and Application layer. Its architectural distribution is similar to $RM_H$ in that both propose a centrally layered stack with similar features. While the authors acknowledge that three layers are not enough to abstract all CPS functionality, the model suffices to capture the functional core. Without a Physical and an Information layer, $RM_A$ proposes a more generalized view that allocates all control and computation on the top layer.

**Table 2** Identified architecture roles and description

| Role | Description |
|---|---|
| Store and process | Elaboration and analysis of information |
| Control | Feedback management informed via sensing and actuating |
| Communication | Data exchange and bridging between locations, systems and technologies |
| Sensing | Capturing and quantifying of the physical world |
| Actuating | (Inter)action on the physical world |

$RM_V$ [32]: focus on IoT and distributed data acquisition. It draws on the previous three-layer model but adds a Data processing layer to take care of the vast data mole entering the IoT hub. This addition suggests a strong focus on data processing and process automation analytics.

$RM_L$ [22]: details the aspect of service orientation in an IoT-based layered architecture. Similar to $RM_V$, they extend the three-layer model with an additional Service-oriented layer between Network and Application. The layer orchestrates and manages the available services to translate, process, and store incoming and outgoing data. As this role is passive, it suffers from adjacent layers' vulnerabilities, making it transparent. As we will see during mapping, $RM_L$ ends up virtually equivalent to $RM_A$.

Next, we compose layer mapping ($M_i$) for each model and component in Fig. 2 based on the identified roles in Table 2. An integrated streetlight (A) senses the lamp current and actuates lamp illumination levels. A generic sensor, such as a light intensity sensor (X), captures light intensity. We map both thus to $RM_H$'s Sensor and Actuator layer. $RM_V$'s Sensors and Actuators layer matches the same description, while the Perception layer offers the best match for $RM_A$ and $RM_L$. The light device further uses lamp data and control to monitor and govern light intensity and system health. In $RM_H$, the Local(-distributed) Control sub-layer manages the given sensory information locally, acting as a local control entity. All other $RM$ refer to a single Application layer for this purpose. Figure 2 shows these devices connected to peer nodes and a master controller (B) through wired couplings (1)–(2). The latter firstly functions as a network bridge between DALI and LoRaWAN. It forwards the information over wireless connections (3) and IP networks (4-5-6-7) through firewalls (D–E) and gateways (C) to the IoT Cloud (F). The Network layer of $RM_H$ and $RM_L$ best describes these devices' and links' connectivity roles. It is responsible for distributing and interconnecting devices, sensors, actuators, and services the Control layer. Similar descriptions for the Transmission layer in $RM_A$ and the Networking

**Table 3** Device layer allocations for the SLA of Fig. 2

| Device | | $RM_A$ | | $RM_H$ | $RM_V$ | $RM_L$ |
|---|---|---|---|---|---|---|
| IoT cloud | (F) | Application | | Higher supervisory Control | Application | Application |
| | | – | | – | Data processing | – |
| | | – | | – | – | Service-oriented |
| Firewall | (D–E) | Transmission | Information | Network | Networking | Network |
| Gateway | (C) | | | | | |
| Controller | (B) | Application | | Supervisory Control | Application | Application |
| Lighting node | (A) | | | Local control | | |
| Light Sensor | (X) | Perception | | Sensor and Actuator | Sensors and Actuators | Perception |
| Physical world | | – | | Physical | – | – |

layer in $RM_V$ fall into place. Secondly, LoRaWAN end-node controllers (B) perform minor decentralized supervision, switching, and timing operations of the connected light devices. This description maps to the Supervisory Control sub-layer of $RM_H$ to which all local controllers subside. The nodes report back to a higher instance, a business process located in the IoT cloud. This process is in charge of management and control of the system's overall operation, i.e., the city, and relates to the Supervisory Higher Control of $RM_H$. All other $RM$ refer to both mentioned control sub-layers to the single Application layer. $RM_V$ and $RM_L$, however, have different role associations for the IoT cloud. $M_V$ includes an assignment of the Data processing layer in charge of information pre-processing. At the same time, $M_L$ foresees a Service-oriented architecture to manage service interaction and processes. All the above components handle or contain information of some sort. $RM_H$'s Information layer applies thus to all components and roles. Further, the physical world is specified only in $RM_H$ mapped, thus, to the physical layer. Table 3 summarizes the device-layer assignments of this paragraph.

$CM_{AH}$: The perception layer maps to the Sensor and Actuator layer from Han et al., the Transmission to the Network layer, and the Control to the Application layer. However, $RM_A$ has no reference neither for the physical nor for the information layer. While the latter might blend into the existing three layers of $RM_A$, no notion of physical components other than sensors or actuators is present in $RM_A$, making this a partial extending cross-mapping.

$CM_{AV}$, $CM_{AL}$: The two IoT models show only minor mapping differences to the three-layer model of $RM_A$. Both map almost directly with minor differences in naming for Transmission/Networking and Perception/Sensor and Actuator. The fourth layer in both proposals shares some functionality with their lower layer. However, it communicates to the upper layer, partially parallel to $RM_A$'s application layer. Their function distribution on the example architecture emerges thus almost identical and

**Table 4** Excerpt from the perception layer of the multi-domain taxonomy table, with notes of $RM$ origin

| Threat | Attack | Note | Also seen |
|---|---|---|---|
| Resource exhaustion | Path based DoS (A) | Sending messages along the routing path, flooding | |
| | Flooding (A) | Flooding a service or resource with requests, ex join | Flooding (V) |
| | (D)DoS attacks (A) | Attempt to make network or service unavailable | |
| | Replay attack (A) | Replay authentication to deplete authentication authority resources; also delayed, to external entities or to other node to gain access or trust (H) | Replay (H), replay (V), replay (L) |

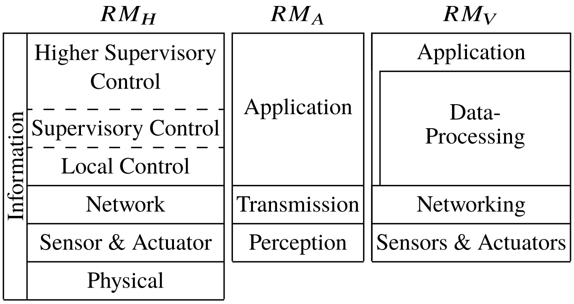


**Fig. 3** Comparison of layer allocations for the SLA of Fig. 2 in relation to $RM_A$, the least detailed model

transparent. These models can thus extend $RM_A$'s notions of attack for application layers with details based on data processing.

Finally, we build a common taxonomy starting with the most abstract model $RM_A$. First, we compare definitions with the more detailing classification of $RM_H$ and the definition extensions for service and data-centric architectures of $RM_V$ and $RM_L$. Then, we align definitions with $RM_A$'s layers in a separate spreadsheet, mark inconsistencies, additions, and duplicates in color, and finally, filter and merge them. It is worth noting that we encountered and marked multiple definitions during this step with untraceable origins. The grouped remaining attacks by layer and threat, where we use $RM_A$ as layer base, result like the excerpt in Table 4. The table lists threats for each layer, filtered attacks, duplicates, and the definitions, possibly integrated with the duplicate's details. For example, the last row lists three attack duplicates in other $RM$s. It integrates the $RM_A$ definition of "Replay attack" with detail from $RM_H$.

Figure 3 displays the cross-comparison result of the consolidation phase. All tables containing attack and threat descriptions of the different stages of this work are available as download.[1]

# 5 Classification

In this section, we iterate through the SLA in Fig. 2 and verify if the attacks remain possible or sensible. As the capabilities of the three network technologies differ significantly, we analyze each network separately by component. Then, through their component's layer allocations (Table 3), we verify the feasibility of a layer's attack using the unified differential taxonomy. Finally, we determine threats and connected vulnerabilities, concluding with suggestions for countermeasures.

## *5.1 Attack Validation*

### 5.1.1 DALI Network

The DALI network consists of streetlights (A) interacting with the physical world, connected through a two-wire bus (1–2) to a LoRaWAN end-node (B) that acts as network master, Fig. 2.

$RM_A$ [3]: At the Perception layer, we mainly see two types of node attacks: attacks that *physically* act on the node and attacks that *virtually* interact with the node. The former requires some form of physical activity on the node where an attacker can get, alter, or make information inaccessible through node capture, tampering, or destruction. The latter type interferes with the node's function by intervening in sensor measurement or corrupting data integrity. Physical attacks may cause information disclosure by, e.g., replacing the node with a duplicate, stealing its data, replicating its functions, and attacking information integrity through false information. Such attacks may cause system malfunction, e.g., darkening specific city areas, as lampposts are publicly accessible. For virtual attacks on systems using the new DALI 2 standard, a captured or inserted false node could act as master and take over other nodes (i.e., spoofing). It may actively poll or even change another node's data and configuration, directly controlling, corrupting, and desynchronizing a network. DoS attacks, such as flooding, can take out a node and make its services unavailable. Finally, electromagnetic interference attacks influence sensory measurements and actuation control, e.g., through action on the system's resonance frequency, corrupting measured values, or feedback loops.

Misleading attacks and buffer overflow occur at the application layer in (A) and (B). The former attacks attempt to make status or value readings unreachable (Denial

[1] Industry 4.0—Smart-Lighting Taxonomy table https://bit.ly/3nHaxjN.

of Service—DoS), forge commands, or intercept and manipulate loops through altered information (Man in the Middle—MitM). The latter injects malicious code. Such attacks to be successful require specialized knowledge of the attacked microcontrollers [10, 31]. On this layer, all attacks interact virtually with the node. Thus, an attacker with network access can systematically trial all reachable nodes.

Virtual attacks gain even more visibility on the Transmission layer of a controller (B). Namely, physical access to the two-wire DALI bus allows an attacker to perform DoS or selective collision attacks (including the mere cutting of wires), muting targeted nodes, and disrupting or desynchronizing control loops. Flooding, or attacker-initiated off-the-schedule polls, can quickly exhaust the network's limited relay capacity. While the simple bus intrinsically avoids routing-based vulnerabilities such as MitM or selective forwarding, its standard lacks authentication and encryption. That enables eavesdropping or, on the new DALI 2 standard, data tampering and forging control messages. Finally, it is worth mentioning that an attacker can orchestrate most of the listed attacks remotely through, e.g., a captured LoRaWAN gateway node (B).

$RM_H$ [16]: The model redefines the desynchronization attack (called Control Forgery in $RM_A$) for the Control layer. The new definition calls it specifically designed to damage a system, e.g., delayed instrument readings that dis-align physical and cyber worlds. In $RM_A$, it only causes generic system misbehavior. For the Network layer, $RM_H$ suggests the spoofing attack may also aid in transmitting false error messages. These messages suggest fictitious lamp failures to the supervisory control, disabling the lamp. On the Physical layer, attacks on external system components are considered. An attacker may intervene on DALI infrastructure and hinder its operation, e.g., cover a lamp. Finally, the Information layer highlights privacy issues that might arise through the information extracted from the transmitted data. For example, the presence/passage of persons in motion-activated areas hints at vacancy. It may cause burglary of adjacent housing units.

$RM_V$ [32]: The Application layer of this model adds configuration tampering attacks for both nodes. Due to resource constraints or programming errors, the embedded code on ballasts and LoRaWAN end-nodes might not verify control parameters for limits and constraints. Such attacks set invalid operating values, e.g., default illumination values to zero, disabling illumination, and threatening safety.

$RM_L$ [22]: This model considers unauthorized users' implications on the network level. Like configuration attacks, unprotected DALI allows an attacker to alter device settings with comparable results (Perception layer). Furthermore, the model identifies malicious code injection attacks as a source of access for multiple system levels. Thus, the node would act as a vehicle for diffusion on all levels. However, as for the attack in $RM_A$, resource constraints and specificity make it hard to predict their success.

### 5.1.2 LoRaWAN Network

A typical SLA combines multiple gateways (C), LoRaWAN end-nodes (B), and at least one Network server (F) through LoRa (3) to create a city-wide LoRaWAN network, Fig. 2. In addition, these LoRaWAN end-nodes may feature sensors and actuators (X) which interact with the physical world for further illumination control or monitoring.

$RM_A$ [3]: On the Transmission layer, the LoRaWAN network exposes multiple availability-related attacks as adversaries can directly access the ether. For example, despite the robust multi-channel multi-modem gateway configuration, typical DoS attacks are feasible through multi-channel frequency jamming, intentional collision, or flooding. Notably, a random message flooding attack targeting those gateways might disrupt a network section. However, the latter solely react to preambles and, by default, cannot handle more than ten packets at a time [28]. Furthermore, a targeted DoS attack will cause collisions and force re-transmissions, ultimately exhausting a battery-powered node's energy.

Suppose these messages are further "replays" of join requests (replay attack). In that case, forwards to Join or Network servers add computation burden and eventually exhaust available resources. Collision attacks have a similar overpowering effect. Unverified transmission practice on the medium and its protocol similarities to ALOHA impact severely on successful message reception, i.e., channel exhaustion at 60% load and only 18% of total capacity [4, 13]. A jamming attack is harder to perform and requires at least three parallel transmissions on the default LoRaWAN frequencies close to a device [8]. Namely, an adequately configured device will see jamming as radio interference and re-transmit on a different channel. Related attacks, such as a resonance attack, will also identify as interference and cause the same response [4]. However, if unconfirmed messaging is used, the data loss may go unnoticed and not trigger any change. Listen-in and analyzing this high number of re-transmitted packages enables side-channel and time analysis attacks to deduce session-key composition. However, the used two-layered encryption limits attack effectiveness. Other typical attacks for the Transmission layer, such as MitM, Sybil, and eavesdropping, remain ineffective until successful capture of the key. Despite missing keys, traffic analysis helps conclude origin, network configuration, and message function. Alternatively, an attacker can attempt node capture and tamper with its memory. The node hosts the necessary keys to send manipulated messages, opting to disrupt the network using valid credentials.

A node (B) is again subject to misleading and buffer overflow attacks at the Application layer. However, while session-keys-protected channels harden a manipulation by resorting to MitM, command forgery, and interception, attempts to make status or value readings unreachable (DoS) stay valid. Thus, even though transmission requires master capabilities and session keys, the same risks for code injection as for the DALI network apply. Again, an attacker with network access can systematically trial all reachable nodes. The only attack at the Perception layer refers to sensors connected to the LoRaWAN controller. Electromagnetic interference attacks can influence sensory measurements and actuation control.

$RM_H$ [16]: We encounter privacy and policy-related issues at the Information layer, desynchronization problems at the Supervisory Control sublayer, and issues with direct intervention at the Physical layer. The Information layer is mainly protected by encryption; however, this does not stop attackers from traffic analysis;

gathering event-based information such as pedestrian or vehicle passing results are helpful, e.g., to assess citizens' behavioral patterns in their neighborhood. Furthermore, multiple join attempts caused through interference signals may help an attacker de-crypt keys used for network and application sessions through excuse attacks. In addition, an adversary can tamper with sensory devices on the physical layer to manipulate measurements and influence lamp control, e.g., artificially raise sky illumination levels, tricking the system into believing that a shallow street illumination level suffices. Finally, a control issue that might emerge is a side effect of scalability. Similar to the situation described in Sect. 5.1.1, the size of the network influences the throughput capabilities. Even though the control loop involving LoRa is less tight, an extended period of reduced or interrupted communication with a gateway or network server could lead to unpredictable behavior.

$RM_V$ [32]: Similar to DALI, configuration tampering attacks at the Application layer may befall LoRaWAN end-nodes with similar side effects. Finally, at the Network layer, the model adds fairness mechanism attacks and extends the definition of DoS flooding. The former attack tampers with the open-source WAN algorithm to elude medium sharing mechanisms and exhaust transmission resources. Flooding's extended definition reveals a similar purpose: malformed packets flood a targeted network or application to overload and corrupt resource availability.

### 5.1.3 IP-Based Infrastructure

The most traditional network in our SLA, the IP infrastructure, connects multiple IP-based devices. It transports data between the on-site LoRaWAN gateways (C) and the IoT computing cloud (F) through dedicated firewalls (D–E), Fig. 2. In addition, the network is in charge of a higher level of connectivity, servicing LoRaWAN and DALI for the applications supervising the city's lighting.

$RM_A$ [3]: On the Transmission layer, we find typical communication-related attacks that target resource availability or intercept or manipulate messages. Most parts of the network apply double-encryption, making attacks such as MitM and eavesdropping onerous. Routing-based attacks are most effective on routed LAN packets, available at the Cloud internal LAN (7). Selective forwarding, routing, sinkhole, wormhole, replay, spoofing, or compromised key attacks could occur here. They help an attacker to weaken and delay network traffic or reroute data for traffic and side-channel analysis. If integrated with traffic analysis, such attacks get more efficient and difficult to detect.

Besides, despite tunneling and encryption, most of the DoS attacks keep their effectiveness. A typical DDoS attack could target VPN end-points, e.g., FW (E), which makes up a single point of failure for the two sub-nets, and a bottleneck on high traffic. Similarly, all network components are susceptible to exhaustion attacks. Finally, tampering and node capture, e.g., the external firewall, could help acquire stored secrets and, e.g., reroute VPN tunnels for general data capture. The primary function of the Application layer is storing and elaboration of information. Primary attacks to this layer identify thus as Database attacks, including data alteration and

User Privacy leakage through data mining on the sensed data. An attacker may gain access to a system via malicious code on shared instances or buffer overflow and consequent code injection.

Furthermore, along with continuously more service-oriented systems, service discovery spoofing helps integrate malicious services into the system, gathering data access. Replayed messages on this service plane may help an attacker to get the trust of the system. Message interception and alteration (MitM) and eavesdropping can cause data leaks or corruption. A malicious service can flood other services until exhaustion, making them unavailable. Such attacks' effectiveness depends on the architecture and implementation of the data processing cloud, not specified by any examined standard.

$RM_H$ [16]: The Network layer presents re-definitions of Sybil and spoofing. For example, injected routing error messages at the inter-VM LAN connection (7) make the grid seem partially offline. At the same time, Sybil attacks target fake network size. Finally, on the Control layer, the system keeps being the target of desynchronization attacks. An attacker can, e.g., tamper with time-servers to misalign lamp control from status.

$RM_V$ [32]: $RM_V$'s application layer considers user interaction with system and data separately from its computation. Thus, if a user connects remotely to the system, a new path opens, allowing network-based threats similar to $RM_A$, including eavesdropping, MitM, routing, or system exhaustion attacks. Configuration tampering attacks may further affect the new terminal, attempting to influence the lighting system's function remotely. On the Data processing layer, we identify Malware attacks again to gain system-level access. $RM_V$ further highlights the interactions and attacks that might occur inter-VM and based on shared resources' contention. The former include instant-on gap attacks, where immediate demand requirements allow initial unrestrained executions due to performance concerns. The latter rely on the exhaustion of shared resources. As a result, the attacked service is depleted and unable to perform the requested services. Another mentioned attack, exhaustion flooding, achieves a similar result. The flooding with requests requires additional resources, slows down the system, and finally exhausts all resources. Side-channel attacks could extract information from non-sanitized shared memory or CPU caches among the VMs. The model does not include additional attacks for the Networking layer.

$RM_L$ [22]: This model sees user-focused attacks on the application layer during client interaction. They try to leak data and capture user credentials through infected emails, phishing websites, and malicious scripts. The model then adds two more definitions on the Network layer: the sinkhole attack, as a maneuver to get more input data routed through for traffic analysis and device tampering and to secure a device's configuration data and secrets, and consequently, gain unauthorized access to devices and networks.

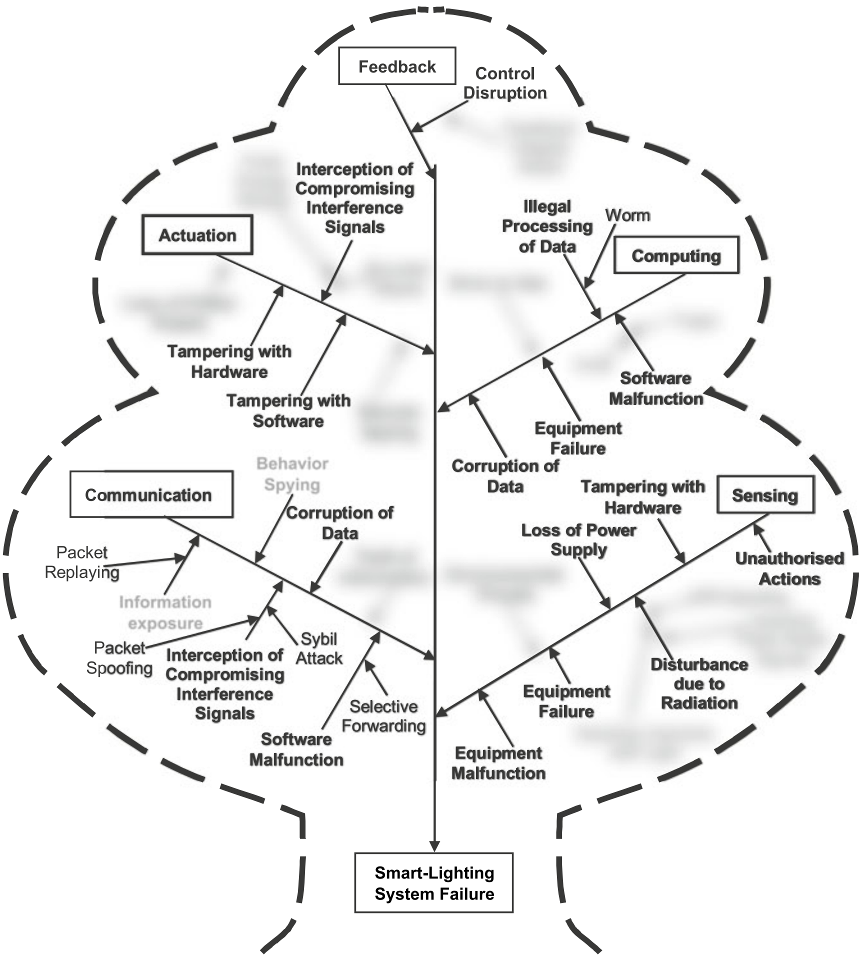


**Fig. 4** Attack tree for a Smart-Lighting architecture, modified (gray), blurred removed from [2]

## 5.2 *Attack Tree and Vulnerabilities*

### 5.2.1 Attack Tree for Smart-Lighting Architecture

Based upon the attack and threat tree developed by Alguliyev et al. [2], we created a version dedicated to SLA attacks. Figure 4 illustrates the resulting attacks-threats CPS functional model where the gray highlighting marks alterations w.r.t. the original, i.e., renamed or relocated branches. In the following, we detail threats connected to the attacks.

*Attacks on actuation*. Our SLA of Fig. 2 contains two actuators: DALI ballasts that control the lamps and LoRaWAN timed controllers to manage these ballasts. Both are installed mostly on or near a light pole. A threat of *Tampering with Hardware* results when physical interaction with the node can occlude actuation. An attacker can manipulate a LoRaWAN end-node or DALI ballast to take control, disable or extract secrets with device tampering or node destruction attacks. *Tampering with Software* occurs when changes on it make actuation non-functional. For instance, Integrity attacks on a lamp-driving LoRaWAN-node can cause incorrect lamp switching times, impeding proper lighting. Finally, *Interception of compromising interference signals* refers to actuation instability caused by external intervention on the actuator signals in closed-loop systems. For example, an attacker can destabilize lamp control and manipulating switching behavior through command-control forgery attacks on DALI ballast.

*Attacks on Communication*: The communication infrastructure of our SLA is represented by DALI, LoRaWAN, and IP-network infrastructures. These include bridges between DALI, LoRaWAN, and IP-based components, i.e., LoRaWAN end-nodes and Gateways, the two firewalls, and the Internet. In addition, all connections, except DALI, are encrypted at least once; AES128-CBC for LoRaWAN, IPSEC, and HTTPS for IP-based connectivity. *Information exposure* refers to the threat that allows data gathering on a non-protected communication channel. An attacker can listen in and obtain information on system and encryption passively via an eavesdropping attack on DALI networks or actively through polling via replay attacks on LoRaWAN channels. *Behavior spying* results when an attacker can gather long-term information on the system's operation, people, and activity remotely. Via traffic analysis attacks, e.g., an adversary, can inspect a LoRaWAN end-node's event-based transmissions that report pedestrian movement. As stated in Sect. 5.1, such circumstantial information can help determine citizens' whereabouts for planned burglary. *Software malfunction* results from circumstances that cause incoherent, incomplete, or timely inadequate data transmission that inhibit the system's correct operation. Typical attacks that might cause such behavior are selective forwarding or flooding attacks, applicable to every IP network link. Another example is collision attacks that delay the successful reception of event-based packets from the LoRaWAN end-node until a successful re-transmission attempt. The threat of *Corruption of data* occurs when an attacker can manipulate information and thus void data integrity. On DALI networks, e.g., the adversary could easily tamper with the transit data as the protocol has no encryption or access control. *Interception of compromising interference signals*, again, refers to communication instability caused by external intervention on the data transmission. Such instability can be caused by jamming attacks on the LoRaWAN network or spoofing attacks on the IP connectivity and flooding attack with consequent loss or alteration of packets or connectivity.

*Attacks on feedback*: Feedback refers to the control function that Cyber-physical systems perform when acting through actuators on sensory input or computational status changes. These include, thus, control algorithms and systems for their implementation. *Control disruption* occurs when the system cannot react to sensory input or status changes, thus destabilizing a system. Via a control-command forgery attack, an attacker could, e.g., manipulate the status of a DALI ballast, desynchronizing feedback control and influencing correct actuation.

*Attacks on Computing*: Computing refers to the equipment used for data storage and elaboration. Cloud services and infrastructure (F) serve data mining, user interaction, and process performance improvement. The threat of *Corruption of data* refers to manipulating information, stored and computed values, e.g., programmed light switching times, to secretly damage the system. A data tampering or integrity attack can alter stored control information. The *Equipment failure* occurs when the computing infrastructure is unable to fulfill the requested computation task. These failures can happen due to physical wear-out and resource exhaustion, an attack that depletes computing resources. *Software malfunction*, yet, results when the computation does execute as requested, but not correctly. These malfunctions are often caused by bugs but can also be due to malicious code installed in the cloud servers, e.g., viruses and trojans, which tamper with software functionality. Finally, *Illegal data processing* happens when an unauthorized agent or a user accesses more than the allowed amount of resources and data and discloses user privacy. For example, such exposure can be a consequence of installed malware (Worms) or an attacker that performs side-channel attacks. In addition, a maligned virtual machine on a multi-tenant cloud could tap shared memory and manipulate the computing instance.

*Attacks on Sensing*: Sensing in our SLA is performed on two locations: DALI ballasts that inform about the lamps' real-time data, and LoRaWAN controllers, sometimes battery-powered, that sense the environment, e.g., luminosity or movement sensors. *Loss of Power Supply* is relevant for devices with reduced energy resources that may suffer from exhaustion and fail service. Battery-powered LoRaWAN end-node may experience an outage due to forced repeated transmissions through LoRa jamming attacks that sleep-deprived the node. *Equipment failure*, yet, refers to the inability of nodes to perform the required task. A node outage attack can put a LoRaWAN or DALI node out of order via physical destruction. *Tampering with hardware* on sensing identifies issues that might arise when hardware modifications impede correct measurement. Direct physical intervention attacks can cover a lighting sensor, making it inoperable. *Unauthorized actions* recall the possibility of prohibited intervention on sensors that access or alter data, misuse the node, or impede its function. The sensing configuration data on the unprotected DALI nodes can be manipulated through data tampering attacks during writes on the bus link, altering measured results. The same attack can also be the source of other threats. *Equipment malfunction* is the result of incorrect sensing due to technical hindrance. Tampering with a sensor's configuration would cause sensing to fail its function. Finally, devices are subject to the *Disturbance due to radiation* when an attacker interferes with the normal sensory function by manipulating the measured physical unit. The LoRaWAN node. e.g., can be fooled through a physical direct intervention attack, irradiating the luminosity sensor with a torch.

### 5.2.2 Vulnerabilities for Smart-Lighting Architecture

After the evaluation of attacks and threats for this SLA, we now identify the causing vulnerabilities. Tracing vulnerability records from the related papers [3, 16], we examine threats and attacks to detect possible vulnerabilities.

At the perception and transmission layers of $RM_A$, most of the attacks identified have two common causes: the low resource constraint the devices withhold and their physical size and exposure. Resource limitation is mostly the enabler of attacks that hinder proper communication, protection, and access control. Unprotected DALI allows an attacker to eavesdrop or inject any command or data. Targeted LoRa or DALI network attacks can deplete available communication or energy resources, disabling parts of the network and feedback control. Similarly, the limited ether availability constraints the transmission capacity of LoRaWAN and eases the attacker's channel interference.

Furthermore, the large scale of an SLA contributes to resource scarcity. It increases channel contention and utilization and co-existence problems [14], finally forcing airtime management or transmission power throttling to reduce range and interference rate. The Wide distribution of a lighting system conduces to the vulnerability of physical exposure. Unattended areas ease network integrity attacks through device tampering, targeted interference, and device destruction. It makes nodes accessible and allows for physical interaction, altering measurements and feedback. Similarly, on the transmission layer, the SLA's wide distribution and large scale cause LoRa's ether resources to incur bottlenecks if an incorrect device configuration neglects available channels. The same holds for gateway setup, where incorrect settings can ease preamble-based resource availability attacks.

Software bugs and inconsistent protocols may enable unauthorized access to infrastructure and information on the transmission and application layer. Human-made error or incorrect device configuration may allow attackers to access systems due to incorrect or mixed permissions schemes or cause system failure. Further vulnerabilities present in the IP and Cloud infrastructure are mostly the typical issues encountered in modern systems. We find missing specification details for the software components running the Smart-* architecture's back-end in addition to service attacks and information leakage issues. Indeed, the two non-standard components, an IDS (D–E) for CPS and the network server (F), have not been defined thoroughly in their specification and architecture [7, 16]. While we can secure the rest of the IP system by applying traditional architectural patterns and techniques, these two components suffer from inconsistent or incomplete specifications.

## *5.3 Reflection on Countermeasures*

This section offers some countermeasures specific to the Smart-Lighting system's weaknesses under study that we leverage from the analysis in the previous sections, the existing literature, and specifications of the technologies. The list should not be seen as exhaustive.

At the lamp end-posts, we have to deal primarily with the physical exposure of the DALI bus, the LoRaWAN controllers, and the sensors. Using cabinets and locks that require a specific tool or key and mounting controllers at height might impede immediate access to wires and devices. Wires should further be carried through shielded conducts, diminishing the risk of interference. Unfortunately, resource constraints and the limiting standards do not permit protection measures against eavesdropping or MitM attacks for DALI communication; a replacement with more powerful hardware could significantly impact unit installation cost and solution attractiveness.

One main point that helps mitigate the attacks on the limited ether availability is a balanced configuration of the LoRaWAN network. The LoRaWAN standard provides the network server with the ability to reconfigure channels and optimize ether usage for gateways and end-nodes. However, there is no binding requirement for such capability. To our knowledge, no network server product includes neither an automatic channel distribution on gateways and nodes nor the ability to select channels beyond the eight frequencies in the Semtech default profile.

Proper distribution of bandwidth and frequencies can drastically increase the resilience of the infrastructure. The sixteen-plus available settings-slots prescribed by the LoRaWAN standard allow a complementary configuration. Furthermore, each node should reach at least two gateways using the lowest spread factor. This approach increases the communication robustness due to LoRa's high channel selectivity and switchable, more robust, higher symbol rates. The latter makes it easy to increase the allowed SNR range and decode messages despite interferences [8, 27].

Likewise, adjacent nodes' downlink and uplink settings should be distributed equally among reachable gateways and channels. End-nodes typically communicate on two channels: one randomly selected among the enabled channels for bidirectional transfer and a second shared RX window from the gateway to all nodes. This second link adds resilience to the network. As long as one downlink is available, the network server can reconfigure a node to use new uplink frequencies [8]. If possible, node channel configurations should contain a disabled configuration of all gateway channels in reach. Disabled channels are automatically enabled after several unsuccessful transmissions, empowering a node in distress to reach all available gateways.

An algorithm running on the network server may manage such additional channel configurations to exploit maximum robustness. It might use geo-information and empirical measurement results to compute channel distribution appropriately and send updates over the secondary RX window. An algorithm for this purpose has been developed by Demetri et al. [11]. Satellite imaging and experimental measurements approximate signal coverage and considers the environment, locations, buildings, and city structure. To avoid the issue of limited throughput and co-existence interference, the number of nodes per channel and gateway should also be equally distributed [4]. A tool called LoRaSim by the university of Lancaster[2] helps this purpose. Although the tool does not consider the environmental situation, it can verify if a configuration

[2] https://www.lancaster.ac.uk/scc/sites/lora/lorasim.html.

is viable. It selects optimal frequencies, captures situations of hidden terminals and exposed nodes, and determines the best-case range and coverage for a given network configuration. Lower spread factor and less interference reduce required air-time and repetition. A service incorporating such an algorithm could improve overall resilience, optimize hardware use and increase end-node battery lifetime.

Unfortunately, LoRa (physical layer) and specification-dependent vulnerabilities cannot directly be dealt with. The specification of protocols is an alliance product (DiiA and LoRa Alliance) and might be open to improvement proposals [8]. At the moment, different proposals exist for both vulnerabilities [1]. The alliance also recently proposed an intra-channel hopping technique (FHSS) to mitigate collisions and contention [9]. The higher robustness comes at a price of a very low throughput of only a few hundred bps. It promises to be an elegant solution for high-density and coexisting networks. However, such changes need time for validation and processing and can therefore be considered only in the long run.

Simple stateful packet inspection is not enough for a CPS's IP-based network. Han et al. [16] identifies security challenges not uniquely at the border to the external networks, but everywhere in this complex interconnected and heterogeneous system. Therefore, intrusion detection must be entwined in the whole CPS system according to each node's limits. Furthermore, each node could be tampered with by generating invalid data. Thus, the solution extends from brute physical force and consequent failures to uncertain information degraded and influencing a system's control. Finally, border firewalls are often the responsible routing point for point-to-point networks. Ideally, multiple connections between IP-based networks can avoid bottlenecks and targeting attacks by routing traffic as needed.

The final set of discussed vulnerabilities connects solely to the application layer. Most of the software modules of the control units and in the application cloud work with parameters. To avoid that those settings are invalid, ideally, the final device or application that uses the information must verify correctness. Han et al. see this also as a possible application for an IDS as an adversary could inject invalid values to cause a control deviation or misbehavior. However, the limited resources make a distributed IDS difficult on some devices. Therefore, a parameter check, a distributed IDS, or both should be installed based on resource availability.

More problems arise if the specifications for these software modules have errors or are incomplete. Unfortunately, in this case, the specifications should also follow standards and might suffer from this dependency. Nevertheless, many details can be derived and adapted following best practices and generalizations from experience with similar installations and architectures. Multi-tenant microservice-based systems are popular in cloud-based computation, making them an excellent architectural template source. Therefore, a computing cloud infrastructure could be derived from microservice-based architectures for data elaboration, integrated with the knowledge gained from running experiments and prototypes. These should finally help achieve the highest security standards without impacting the overall performance. Lastly, most software is following new technology trends, subject to multiple changes in a short time, and suffering from high defect probability. Therefore, agile practice and testing tool-chains are the only suggestions to be given from the development standpoint.

## 6 Discussion and Conclusions

This security analysis presented a technique for the offline analysis of a Smart-* multi-domain system-of-systems. We proposed a design science approach that relied on the connected domains' experiences and performed a layer-based cross-analysis on a Smart-Lighting use case. Using four distinct layered architecture modeling approaches, we identified architectural roles, assigned model layers. Then, we created a unified taxonomy that reflects and extends each involved domain's attack definitions, threats, and vulnerabilities. In an iterative process, we determined possible attacks, valid threats and discussed vulnerabilities for a merged-domain Smart-Lighting architecture. Finally, we discussed some first possible countermeasures.

After the execution of our analysis, we can assess three significant discoveries for Industry 4.0. Firstly, the domain overlapping configuration of such a system-of-systems makes it infeasible to cover all threats and attacks based on a single domain's viewpoint. The integrative approach we presented detected more issues than a single model would. Interestingly, we find the most model divergence in the "cyber"-layers, where computation and decision occur, while most data exchange and physical interaction layers remain unchanged. This consistency is probably because gathering, actuation, and data transport are a joint function of all four analyzed papers. When integrating future analyses with other studies, we expect changes in the upper architecture layers only. Secondly, the changing focus of the discussed models highlights aspects of a heterogeneous system. It proves that the new multi-domain architecture inherits many, if not all, characteristics of the involved domains. For example, Cloud-security issues are not a typical concern for traditional control-oriented CPS. Thirdly, vulnerabilities, threats, and attacks may alter definition, range, and weight depending on the application domain. The taxonomy table and Sect. 5.2 show how similar threat or attack names can have different definitions and applications that the domain of origin might influence. It is thus reasonable to pre-define and clarify all taxonomy before reaching conclusions. However, the resulting multi-domain taxonomy is a product of the involved reference models' role, layer, and attack allocation. Each new system-of-systems analysis requires thus repeating or refining the present analysis.

Future work will test and extend the results of this analysis. Through a second study case, we will analyze the change and variability of detected issues. Simultaneously, on-site tests will help validate the extent and risks of the vulnerabilities involved.

**Acknowledgements** We thank Systems S.r.l for the funding and support of this project under the "Industry 4.0 for the Smart-* (I4S)" project.

## References

1. Adelantado F, Vilajosana X, Tuset-Peiro P, Martinez B, Melia-Segui J, Watteyne T (2017) Understanding the limits of LoRaWAN. IEEE Commun Mag 55(9):34–40. https://doi.org/10.1109/mcom.2017.1600613
2. Alguliyev R, Imamverdiyev Y, Sukhostat L (2018) Cyber-physical systems and their security issues. Comput Ind 100:212–223. https://doi.org/10.1016/j.compind.2018.04.017
3. Ashibani Y, Mahmoud QH (2017) Cyber physical systems security: analysis, challenges and solutions. Comput Secur 68:81–97. https://doi.org/10.1016/j.cose.2017.04.005
4. Augustin A, Yi J, Clausen T, Townsley W (2016) A study of LoRa: long range & low power networks for the internet of things. Sensors 16(9):1466. https://doi.org/10.3390/s16091466
5. Bellido-Outeirino FJ, Flores-Arias JM, Domingo-Perez F, Gil-de Castro A, Moreno-Munoz A (2012) Building lighting and automation through the integration and of DALI and with wireless sensor networks. IEEE
6. Bolbot V, Theotokatos G, Bujorianu LM, Boulougouris E, Vassalos D (2019) Vulnerabilities and safety assurance methods in cyber-physical systems: a comprehensive review. Reliab Eng Syst Saf 182:179–193. https://doi.org/10.1016/j.ress.2018.09.004
7. Committee LAT (2017) LoRaWAN™ backend interfaces 1.0 specification. Tech. rep., LoRa Alliance
8. Committee LAT (2017) LoRaWAN™ specification 1.1. Tech. rep., LoRa Alliance
9. Committee LAT (2020) LoRaWAN™ regional parameter specification 1.0.2. Tech. rep., LoRa Alliance
10. Contenti C (2002) Digitally addressable DALI dimming ballast. IEEE
11. Demetri S, Zúñiga M, Picco GP, Kuipers F, Bruzzone L, Telkamp T (2019) Automated estimation of link quality for LoRa: a remote sensing approach. In: Proceedings of the 18th international conference on information processing in sensor networks—IPSN '19. ACM Press. https://doi.org/10.1145/3302506.3310396
12. DiiA (2018) Dali-2: the differences and new version of the Dali standard. Tech. rep, Digital Illumination Interface Alliance. https://www.digitalilluminationinterface.org/data/downloadables/5/4/1711_technical-note-dali-2-the-new-standard.pdf
13. Fehri CE, Kassab M, Abdellatif S, Berthou P, Belghith A (2018) LoRa technology MAC layer operations and research issues. Procedia Comput Sci 130:1096–1101. https://doi.org/10.1016/j.procs.2018.04.162
14. Ghena B, Adkins J, Shangguan L, Jamieson K, Levis P, Dutta P (2019) Challenge: unlicensed lpwans are not yet the path to ubiquitous connectivity. In: The 25th annual international conference on mobile computing and networking, MobiCom '19. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3300061.3345444
15. Hallmans D, Sandström K, Nolte T, Larsson S (2015) Challenges and opportunities when introducing cloud computing into embedded systems, pp 454–459. https://doi.org/10.1109/indin.2015.7281777
16. Han S, Xie M, Chen HH, Ling Y (2014) Intrusion detection in cyber-physical systems: techniques and challenges. IEEE Syst J 8(4):1052–1062. https://doi.org/10.1109/jsyst.2013.2257594
17. Hein PF (2001) Dali—a digital and addressable lighting and interface for lighting and electronics. In: Industry applications conference, 2001. Thirty-sixth IAS annual meeting. Conference record of the 2001 IEEE, vol 2. IEEE, pp 901–905
18. Hevner AR (2007) A three cycle view of design science research. Scand J Inf Syst 19(2):4
19. Hofer F (2018) Architecture, technologies and challenges for cyber-physical systems in Industry 4.0—a systematic mapping study. In: 12th ACM/IEEE international symposium on empirical software engineering and measurement (ESEM). https://doi.org/10.1145/3239235.3239242
20. Lee J, Bagheri B, Kao HA (2015) A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. Manuf Lett 3:18–23. https://doi.org/10.1016/j.mfglet.2014.12.001

21. Lezzi M, Lazoi M, Corallo A (2018) Cybersecurity for Industry 4.0 in the current literature: a reference framework. Comput Ind 103:97–110. https://doi.org/10.1016/j.compind.2018.09.004
22. Lin J, Yu W, Zhang N, Yang X, Zhang H, Zhao W (2017) A survey on internet of things: architecture, enabling technologies, security and privacy, and applications. IEEE Internet Things J 4(5):1125–1142. https://doi.org/10.1109/jiot.2017.2683200
23. Lu Y (2017) Industry 4.0: a survey on technologies, applications and open research issues. J Ind Inf Integr **6**:1–10. https://doi.org/10.1016/j.jii.2017.04.005
24. Majed S, Ibrahim S, Shaaban M (2014) Energy smart grid cyber-threat exposure analysis and evaluation framework. In: Proceedings of the 16th international conference on information integration and web-based applications & services—iiWAS '14. ACM Press. https://doi.org/10.1145/2684200.2684308
25. Moteff J (2005) Risk management and critical infrastructure protection: assessing, integrating, and managing threats, vulnerabilities and consequences. Tech. rep., Congressional Research Service—The Library of Congress
26. Ryoo J, Kazman R, Anand P (2015) Architectural analysis for security. IEEE Secur Privacy 13(6):52–59. https://doi.org/10.1109/msp.2015.126
27. Semtech (2015) LoRaSX 1276/7/8/9 datasheet
28. Semtech (2017) SX 1301 datasheet. Tech. rep., Semtech
29. Subramanian N, Zalewski J (2016) Quantitative assessment of safety and security of system architectures for cyberphysical systems using the NFR approach. IEEE Syst J 10(2):397–409. https://doi.org/10.1109/jsyst.2013.2294628
30. Subramanian N, Zalewski J (2018) Safety and security analysis of control chains in SCADA using the NFR approach. IFAC-PapersOnLine 51(6):214–219. https://doi.org/10.1016/j.ifacol.2018.07.156
31. Systems NNA. Ul20x0—LoRaWAN™ luminaire controller
32. Varga P, Plosz S, Soos G, Hegedus C (2017) Security threats and issues in automation IoT. In: 2017 IEEE 13th international workshop on factory communication systems (WFCS). IEEE. https://doi.org/10.1109/wfcs.2017.7991968