Performance Assessment of Reinforcement Learning Policies for Battery Lifetime Extension in Mobile Multi-RAT LPWAN Scenarios

Martin Stusek, Pavel Masek, *Member, IEEE*, Dmitri Moltchanov, Nikita Stepanov, Jiri Hosek, *Senior Member, IEEE*, and Yevgeni Koucheryavy, *Senior Member, IEEE*

***Abstract*—Considering the dynamically changing nature of the radio propagation environment, the envisioned battery lifetime of the end device (ED) for massive machine-type communication (mMTC) stands for a critical challenge. As the selected radio technology bounds the battery lifetime, the possibility of choosing among several low-power wide-area (LPWAN) technologies integrated at a single ED may dramatically improve its lifetime. In this paper, we propose a novel approach of battery lifetime extension utilizing reinforcement learning (RL) policies. Notably, the system assesses the radio environment conditions and assigns the appropriate rewards to minimize the overall power consumption and increase reliability. To this aim, we carry out extensive propagation and power measurements campaigns at the city-scale level and then utilize these results for composing real-life use-cases for static and mobile deployments. Our numerical results show that RL-based techniques allow for a noticeable increase in EDs' battery lifetime when operating in multi-RAT mode. Furthermore, out of all considered schemes, the performance of the weighted average policy shows the most consistent results for both considered deployments. Specifically, all RL policies can achieve 90 % of their maximum gain during the initialization phase for the stationary EDs while utilizing less than 50 messages. Considering the mobile deployment, the improvements in battery lifetime could reach 200 %.**

***Index Terms*—LPWAN; Multi-RAT; End-device lifetime; Energy consumption optimization; Reinforcement learning**

I. Introduction

Recently introduced heterogeneous networks, i.e., 5G mobile systems, are envisioned to embrace multiple radio access technologies (RATs) targeting to effectively manage completely redefined requirements of wireless data transmissions for diverse types of EDs [1]. In this context, the low-power wide-area networks (LPWANs) enable industrial communication, which has not been previously considered due to insufficient technical parameters [2]. Notably, as the first LPWAN technologies reached the market during the last decade, it took them time to mature to the point where the industry companies were aware of these new opportunities. But there is not just a single LPWAN technology as the silver bullet to cover all the communication scenarios [3].

The concept of multi-RAT is reshaping the possible use cases for all verticals of the 5G and beyond 5G (B5G), i.e., enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (mMTC) [4]–[6]. Combining multiple communication technologies enables new types of device design [7]. However, the scope of today's IoT-ready devices mainly pertains to utilizing a single RAT. On the one hand, this helps reduce the price tag of the devices, but on the other hand, it limits the possible communication scenarios – as it relies on a single LPWAN technology [8], [9]. Although the increased complexity of the multi-RAT devices influences the price of the devices, it may positively impact the main operational parameter – the lifetime of the battery-powered device.

The input assumption in this work builds upon the hypothesis that the propagation conditions between the EDs and the base station (BS) operating the LPWAN technologies may drastically change over the EDs' lifetime [10]. Various changes in the surrounding environment (whether it is indoor or outdoor and caused by nature or human-initiated), including construction works, infrastructural changes, and weather conditions, may drastically affect the communication parameters in the frequency bands the LPWAN technologies operate in. Furthermore, one may expect drastic variations in propagation conditions in new IoT use-cases such as assets tracking on the move, where EDs are installed on vehicles, such as cars, busses, drones, etc. In both conventional deployments and new use-cases, the combined demands for message delivery delay of less than 10 s and message loss probability as low as 10% need to be satisfied according to ITU-R [11].

In this paper, we consider and evaluate the utilization of multiple RATs at a single ED aiming to dynamically switch between utilized LPWAN technologies and employ the most suitable one in terms of the radio propagation conditions. Since the propagation conditions are not known in advance, selecting the best RAT has to be conducted automatically in response to environmental changes. One of the tools that allow for a dynamic adaptation is reinforcement learning (RL), where the system continuously assesses the environment and assigns the weights to different options by attempting to maximize its reward. In our case, the reward is related to ED power consumption. We evaluate the proposed approach by considering stationary and mobile, two distinctly different deployment conditions. The input parameters are determined by utilizing power consumption laboratory results combined with field measurement campaigns and propagation conditions.

Notably, the application of RL policies while considering

M. Stusek, P. Masek, and J. Hosek are with Brno University of Technology, Faculty of Electrical Engineering and Communications, Dept. of Telecommunications, Brno, Czech Republic. Email: {lastname.firstname}@vut.cz

D. Moltchanov and Y. Koucheryavy are with Tampere University, Unit of Electrical Engineering, Tampere, Finland. Email: {firstname.lastname}@tuni.fi

N. Stepanov is an independent researcher, Email: nikitaleos29@gmail.com

the utilization of mobile, battery-powered, and performance-limited LPWA deployments represents a novel approach as it was expected that the increased power consumption diminishes the potential benefits of multi-RAT implementation. As the power consumption of the LPWA devices is likely to be at the scale of mA (while transmitting data), the correct implementation of selected communication technology can dramatically change the way the device is going to last and provides its functionality. As presented in this paper, even the most straightforward approach may significantly improve performance over the single-RAT solution. In particular, the implementation based on cumulative average allows for keeping the computational demands low without negatively impacting the battery life.

Our main contributions can be summarized as follows:

- Formalization of RL-based approaches for battery lifetime extension of LPWAN EDs supporting multi-RAT capabilities in case of energy consumption and city-scale propagation and mobility conditions.
- Detailed numerical comparison campaign for real-time tracking applications in close-to-reality deployment scenario on a city-scale with realistic street configurations and empirical propagation measurements.
- Design of RL-based operations to extend the battery lifetime. The performance of the weighted average policy shows the most consistent results while Thompson sampling outperforms ε-greedy, weighted average, and UCB options in the stationary scenarios by exploiting up to 99.5% of the theoretical gains (when the best interface is selected).

The rest of this text is organized as follows. In Section II, we provide the background and motivation for our study, as well as sketch the details of the utilized evaluation methodology. Then, in Section III, we discuss the results of our propagation and energy consumption measurements campaigns. Next, the RL algorithms for battery lifetime extension are introduced in Section IV. Numerical results are then discussed in Section V. Finally, conclusions are drawn in the last section.

II. Background and Utilized Methodology

We start this section with a description of the envisioned methodology and application scenarios. Then, the overview of the current situation of multiple RATs utilization in a single ED is given. Finally, the LPWAN technologies are further detailed.

A. Methodology and Considered Scenarios

In this work, we evaluate the performance of RL-based policies for two different types of EDs deployments. First, we aim to capture the time-dependent characteristics of a communication channel over multiple months. Our previous work shows that signal fluctuation may be as high as 30 dB even for stationary nodes [12]. Such a high signal variation may lead to increased power consumption, decreased service reliability, and a higher number of dropouts. We aim to overcome this issue by employing a multi-RAT solution complemented by suitable RL policies. The main idea behind this approach is to select the optimal radio interface for message transmission, i.e., the one with the lowest power consumption at a given moment. Practically, such a solution can diminish the negative impacts of multi-RAT solutions, including higher power consumption while ensuring increased reliability.

For the second scenario, mobile EDs are intended for asset tracking purposes. Input data for this scenario is derived from the large-scale measurement campaign conducted in the city of Brno, the Czech Republic. However, even such a large-scale measurement campaign provides only a discrete set of measurement points. Hence, scattered data interpolation methods must be employed to derive nearly continuous traces required for precise EDs tracking. The remaining steps of the mobile scenario are identical to the stationary deployments. However, for mobile EDs deployments, the signal fluctuations are expected to be significantly higher. Also, the rate of signal level changes is much faster compared to the stationary EDs deployment. Hence, selecting the optimal radio interface for the mobile deployments represents a more complex task than in the case of stationary nodes.

To assess the multi-RAT performance from the perspective of power consumption, we utilize the following methodology, see Fig. 1. First, we characterize the considered LPWAN technologies' energy efficiency in the range of operational radio conditions to obtain detailed power characteristics. In the second step, we conduct two measurement campaigns for (i) stationary and (ii) mobile EDs deployments. Further, the resulting data from the first campaign is used to identify the associated time-dependent propagation models using a doubly stochastic Markov chain framework. Finally, we formulate and apply the multi-armed bandit (MAB) RL framework by employing the developed models and energy consumption.

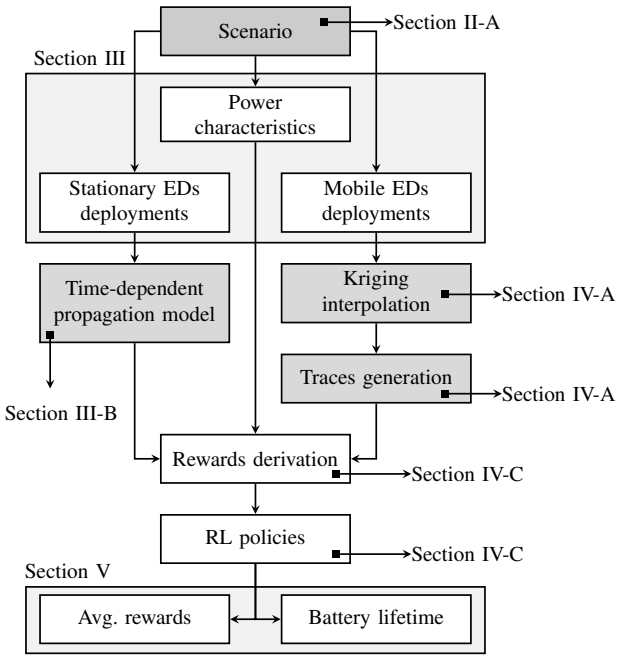


Fig. 1: Overview of our proposed methodology.

Notably, the coverage map in the case of mobile EDs in the city-scale scenario represents a critical part of the

developed methodology. First, scattered data from the city-scale measurement campaign are converted to a uniform grid with high resolution by employing the Kriging interpolation algorithm. Then a single location close to the center of the measured area representing the central point (warehouse) is selected from which all the ED traces are routed. In the final step, we apply the same MAB RL framework by combining the energy consumption assessments with interpolated RSRP/RSSI coverage map to assess its performance in mobile deployment.

B. Multi-RAT functionality

The premise of multi-RAT builds upon the idea of combining multiple communication interfaces into a single device; see Table I where the LPWAN technologies in question are listed. However, the practical implementation and overall functionality may be cumbersome. Moreover, the selection of individual communication interfaces influences the overall functionality of the multi-RAT device, i.e., power consumption, communication range, and cost. Hence, the choice of communication technologies that complement each other is essential.

The multi-RAT device characteristics are strongly influenced by their operation modes. Therefore, from the perspective of communication interfaces operations, multi-RAT devices can be divided into three separate groups [8], [13]:

- *Parallel:* These devices are capable of communicating over multiple RATs simultaneously. Modern smartphones are typical representatives of this group by providing, e.g., cellular and Wi-Fi aggregation to improve communication throughput and reliability. Contrary, the main drawbacks of this solution are increased power consumption (both average and peak) and increased requirements on computational resources. Thus, the device must simultaneously control multiple data pipelines and radio transceivers, which also yields higher costs for such devices.
- *Selective:* Even though the device supports multi-RAT operations, only a single one is employed at a time. Nevertheless, such a device still provides higher connection versatility due to differences in spatial coverage of respective communication technologies. In areas with multi-technology coverage, the device may perform a selection of the most suitable technology based on the best power consumption, communication latency, data rate, or message loss. Furthermore, in comparison with the parallel operation mode, selective multi-RAT devices are cheaper to implement due to less stringent computational requirements. A part of the circuitry can be shared among radio interfaces, leading to further cost reductions.
- *Combined:* This operational mode combines two former groups and supports a certain level of parallel operations but does not utilize all available radio interfaces to the full extent. In practice, such a device may maintain network synchronization using all available RATs while it transmits data over a single interface. From the perspective of complexity and power consumption, sequential devices reside in-between the groups mentioned above. Thus, this particular approach is optimal for applications requiring a certain quality of service with devices constrained by their energy consumption.

TABLE I: Key parameters of LPWAN technologies [14]–[19].

	LoRaWAN	**Sigfox**	**NB-IoT**
Coverage (MCL)	157 dB	162 dB	164 dB
Technology	Proprietary	Proprietary	Open LTE
Spectrum	Unlicensed	Unlicensed	Licensed
Frequency	433, 868, 915 MHz	868, 915 MHz	700–2100 MHz
Bandwidth	125, 250, 500 kHz	100, 600 Hz	200 kHz
Max. EIRP UL	14 dBm[1]	14 dBm[1]	23 dBm
Max. EIRP DL	27 dBm[1]	14 dBm[1]	23 dBm
Downlink data rate	0.25-21.9 kbps[2]	0.6 kbps	0.5-27.2 kbps
Uplink data rate	0.25-11 kbps[2]	0.1-0.6 kbps	0.3-62.5 kbps[3]
Max. payload UL	242 B	12 B	1600 B
Max. payload DL	242 B	8 B	1600 B
Battery lifetime	10+ years	10+ years	10+ years
Module cost	6 $	3 $	12 $
Security	AES-128	AES-128	LTE Security

[1] The value is relevant for EU.
[2] 50 kbps for FSK modulation.
[3] 3GPP Release 13.

As the primary goal of this work is to extend the battery lifetime, we put all our effort into implementing the multi-RAT "selective" mode based on RL policies. It should allow us to achieve a relatively good battery lifetime without the need for complex computationally demanding algorithms.

C. Selected LPWAN Technologies

We chose three LPWAN technologies providing publicly available services with country-scale coverage in the Czech Republic for our consideration. These technologies include NB-IoT, Sigfox, and LoRaWAN. Albeit targeting the market of mMTC, all selected representatives possess unique properties and mechanisms ensuring long-range communication with low power consumption.

1) NB-IoT: This cellular technology represents one of the first LPWAN standards operating in the licensed spectrum. NB-IoT builds on top of the legacy LTE systems with which it shares a significant amount of infrastructure and numerology. For its operation, NB-IoT occupies a 200 kHz block in one of 13 frequency bands (another 4 bands in Rel. 14 and 7 additional in Rel. 15) in the range from 700 to 2100 MHz. In order to decrease the overall complexity of the system, NB-IoT supports only half-duplex transmission with frequency division duplex (FDD).

Aside from the reduced complexity, NB-IoT employs two energy conservation mechanisms, namely power saving mode (PSM) and extended discontinuous reception (eDRX), to extend battery lifetime. In addition, the utilization of licensed bands allows for transmission power of up to 23 dBm (two additional power classes limited to 20 and 14 dBm). The maximum message size is 1600 B due to the size of the service data unit (SDU) of the packet data convergence protocol (PDCP). NB-IoT also reflects the LTE numerology in the physical resource block (PRB) structure. The whole NB-IoT bandwidth fits into a single PRB of 180 kHz with 15 or 3.75

(only in UL direction) kHz sub-carrier spacing. For UL, single-carrier frequency multiple access (SC-FDMA) in combination with $\pi/2$-BPSK or $\pi/4$-QPSK modulation is used. In contrast, DL relies on orthogonal frequency multiple access (OFDMA) with a single QPSK modulation scheme.

The key property of LPWAN technologies, i.e., long-range communication with a link budget of 164 dB (+20 dB compared with LTE), is achieved mainly via repetitions. In the case of NB-IoT, both UL transmission and random access preamble can be transmitted up to 128 times. Moreover, DL provides an even more generous number of 2048 repetitions [18], [19].

2) LoRaWAN: It is representative of LPWAN technologies utilizing license-exempt frequency bands. These bands include traditional 433, 868, and 915 MHz frequencies with additional region-specific bands such as 500 and 780 MHz. In addition, LoRaWAN physical layer builds upon a proprietary long-range (LoRa) modulation which provides impressive variability. The spreading factor (SF) parameter allows adjusting the modulation's robustness, directly influencing the throughput and sensitivity. In total, six SF values ranging from SF7 to SF12 can be used. From the perspective of transmission time, each increase in SF approximately doubles the time-on-air (ToA) of the symbol

The LoRAWAN medium access control (MAC) technology is an open standard. In Europe, it defines sixteen channels in the 868 MHz band with a bandwidth of 125 or 250 kHz. Furthermore, the utilization of an unlicensed band imposes duty-cycle regulation of 1 % with a maximum radiated power of 14 dBm. Hence, LoRaWAN message size is limited from 51 B for SF12 with a maximum of 242 B with SF7.

In the 915 MHz (US) band, LoRaWAN can operate with 72 channels, each occupying a bandwidth of up to 500 kHz. A higher number of channels allows LoRaWAN to utilize frequency hopping between messages. Nevertheless, it also limits the dwell time to 400 ms per channel (with no subsequent transmission in less than 20 s), allowing for a maximum SF10. On top of that, listen before talk (LBT) is utilized for Japan and South Korea. Thus, the device must verify that the whole frequency band is free of signals stronger than -80 dBm [16], [17].

3) Sigfox: Represents another license-exempt LPWAN technology operating within the industrial, scientific, and medical (ISM) spectrum around the center frequency of 868 or 915 MHz. Sigfox occupies 200 kHz bandwidth dependent on geographical region. In most radio configurations (RC), each differential binary-phase shift keying (D-BPSK) coded message in UL covers 100 Hz of the total bandwidth. The only exceptions are the US and Latin American regions, where the message occupies 600 Hz of the bandwidth. Such an ultra narrowband (UNB) modulation provides an excellent signal-to-noise ratio (SNR) with extended coverage in order of tens of kilometers. However, the main drawback of this solution is the limited throughput of 100 or 600 bps (based on the used bandwidth).

As in the case of other unlicensed LPWAN technologies, the operation in the European ISM region restricts the maximum duty cycle to 1 % of the total time. Furthermore, combined with a throughput of 100 bps, it limits the Sigfox technology to 140 UL messages with a maximum size of 12 B per day. Transmissions in DL direction are restricted even more, with only four messages carrying 8 B payloads per day. The situation is different for regions utilizing the 600 Hz message bandwidth as the datagram is broadcasted three times using three different frequencies (frequency hopping). Also, the transmission's ToA must not reach 400 ms with a 20 s back-off period. For Japan and South Korea, the LBT technique must be employed [14], [15].

D. Reinforcement Learning

Machine learning (ML) algorithms play an essential role in the upcoming 5G systems. The basic premise is that, by observing and learning system behavior, which represents the current state of the communication channel, one can achieve improved power efficiency during the message transmission. However, the dynamic radio conditions rule out a significant share of ML algorithms from consideration (e.g., supervised learning and other methods based on static models). On the other hand, RL can overcome this issue by interacting with the environment.

Over the recent years, RL gained momentum along with other ML algorithms as enablers for computationally demanding tasks in modern heterogeneous communication systems. Notably, RL finds application in 5G networks allowing for efficient resource utilization (radio, mac scheduling), network slicing, and band allocation. However, due to the immense complexity of these tasks, deep RL is used predominantly [20]–[22]. MAB learning is often utilized for less demanding tasks, mainly for handover predictions and communication scheduling [23], [24]. However, the latest MAB-focused works also cover intelligent link configuration for 5G mMTC, millimeter wave beamforming adaptation, and mobile edge computing with task offloading [25]–[28].

In general, RL algorithms consider agents interacting with the environment by observing its state and executing actions. Notably, these actions trigger an environmental response recognized as a reward. Recurring action-reward events consequently allow exploring system behavior. Explicitly, it represents a search for the optimal action strategy. It must be noted that the environment response informs about the efficiency of the current action but not if it was the best available action [29].

All interactions of RL with the environment take place in a discrete-time, $t = 0, 1, 2, \ldots, n$. It must be noted the actual value of the time step can not be based on any common rule, but it is usually defined as a compromise between complexity and accuracy. At each run, an agent receives information about the state of the environment $S_t \in S$, and employs this information to select an action, $A_t \in A$, which yields a measurable reward, $R_t \in R$. Such an interaction model is often referred to as a Markov decision process (MDP). Notably, the agent's ultimate goal is to maximize the cumulative reward $\sum_{t=1}^{n} R_t$ with actions following the decision policy π. Also, the RL methods specify how the agent adapts its policy based on previous experience. In theory, this prior experience leads to an optimal decision policy π^* with maximum long-term accumulated reward [29].

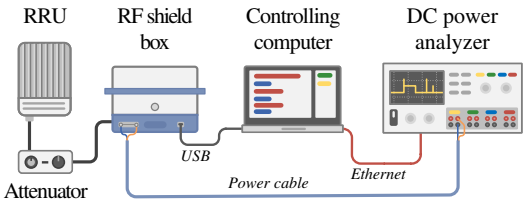


Fig. 2: Workplace for NB-IoT energy measurements.

III. Measurement Campaigns

The cornerstone of both intended scenarios is knowledge of LPWAN modules' energy consumption under various radio conditions. To this aim, we conducted a thorough power consumption measurement campaign of all LPWAN modules deployed on the multi-RAT prototype. All measurements were realized in a controlled laboratory environment to achieve the most coherent results.

Though the power measurement results are shared for both stationary and mobile EDs scenarios, the remaining campaigns were conducted differently. The stationary EDs deployment measurements covered two multi-RAT prototypes transmitting messages with a constant period in the time span of multiple months. Contrary, the mobile EDs were transferred to a particular location, conveyed a pre-defined number of datagrams, and moved to the next point.

A. Energy Consumption Measurements

The measurement campaign covered three selected LPWAN technologies, i.e., NB-IoT, LoRaWAN, and Sigfox, in the range of suitable signal levels. For example, in the case of NB-IoT, it represents signal levels from −68 to −133 dBm, covering all extended coverage level (ECL) classes from ELC 0 to ECL 2.

As depicted in Fig. 2 the desired signal levels were achieved via a step attenuator positioned between the remote radio unit (RRU) and the measured NB-IoT module placed in the radio frequency (RF) shield box. The electric current consumption was measured via power analyzer Agilent N6705A with the power sampling period of 0.0248 ms allowing to capture even the shortest power consumption peaks. Together with the known supply voltage of 3.3V and samples timestamps, acquired current consumption samples' served as an input of trapezoidal integration. The product of this operation is the exact value of energy consumption in Joules.

The whole process was repeated ten times for twelve different signal levels. Notably, all LPWAN modules were set to transmit 12 B UL messages without acknowledgment to be in line with the limitation of Sigfox technology. In the case of LoRaWAN, we conducted identical measurements for SF12 and SF9. Naturally, the highest SF allows achieving the most extended communication range and sensitivity at the expense of increased power consumption and duty cycle limitation (DC). Prolonged ToA resulting in a low data rate of SF12 is the main reason we also consider SF9 with worse sensitivity. With our intended 30 s transmission period of 12 B messages, SF9 represents the highest suitable SF. The latter SF9 is also the highest SF not violating 400 ms ToA policy for 12 B payload from the perspective of the US region [17].

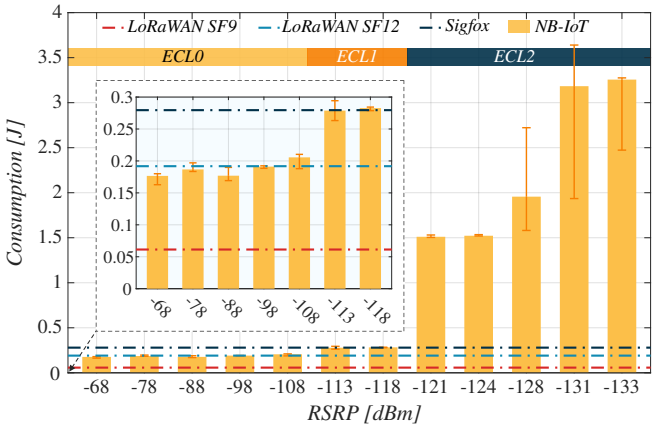
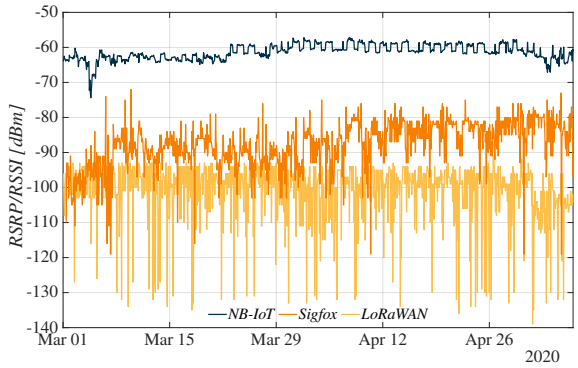


Fig. 3: Results of power consumption measurements.

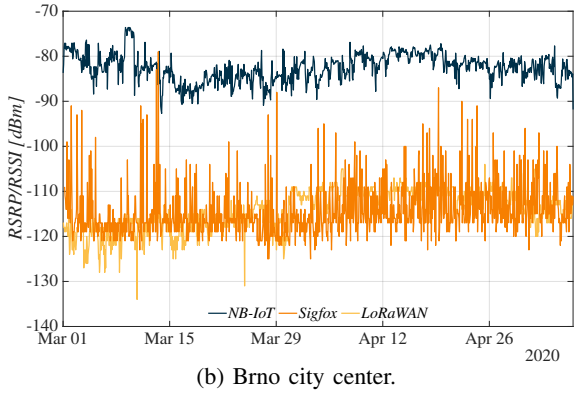
As a representative of NB-IoT technology, we selected the communication module SARA N210 produced by the company uBlox, which implements the Rel. 13 of 3GPP standard. The module was set to utilize the highest power class with a maximum transmission power of 23 dBm. It is worth mentioning that each measurement cycle captures actual payload transmission with tracking area update (TAU) exchange. This combination represents the real-world scenario the best, as TAU information must be transferred after each cell handover and follows after awakening from PSM. Next, the module designated as Microchip RN2483 represented LoRaWAN technology. Finally, for Sigfox, we utilized the S2-LP communication module produced by STMicroelectronics. Both formerly mentioned modules were set to use the maximum transmission power of 14 dBm.

The measurement results shown in Fig. 3 verify that the power consumption of LoRaWAN and Sigfox technologies is constant for all intended signal levels. However, further analysis reveals that Sigfox power consumption is even higher than in the case of LoRaWAN utilizing SF12. This difference is mainly caused by a long ToA interval of Sigfox messages that are repeated three times. Thus, the total time of Sigfox transmission equals 6.24 s for each message. On the other hand, for LoRaWAN with SF12, the total ToA is less than 1.5 s. For SF9, it is slightly above 0.2 s. However, from the perspective of Sigfox, it is still an impressive result as even with a four times longer transmission period, the power consumption is only about 30 % higher.

The most interesting results are observed for NB-IoT. Although the consumption for the whole ECL 0 class is nearly constant with minimal fluctuation, it starts to rise in ECL 1. But, in ECL 2, the power consumption starts growing nearly exponentially. From the signal level of −128 dBm also the whiskers representing the 5th and 95th percentile vary significantly. This phenomenon is primarily caused by message retransmissions, which play a crucial role under poor signal conditions causing massive energy consumption growth and variations in the results. A side-by-side comparison of the ECL classes reveals that the consumption in ECL 2 can be 15 times higher as compared to ECL 0.



(a) Brno University of Technology.



(b) Brno city center.

Fig. 4: Measured signal levels in stationary conditions.

B. Stationary Nodes Measurement Campaign

The ultimate goal of this campaign was to evaluate the time-dependent signal characteristics in different types of the urban environment. The first sensor, deployed near the Brno city center (BCC), represented a typical urban scenario of the European city with mid-rise buildings. The second unit was located at Brno University of Technology (BUT) premises on the outskirts of the town. This scenario represented a typical suburban environment with a less dense urban development.

Both devices were set identically to convey datagrams with a 12 B payload every hour. Notably, messages were transmitted via all interfaces in sequential order to minimize interference among radio interfaces. It was crucial as both LoRaWAN and Sigfox technology operate in 868 MHz ISM spectrum with NB-IoT occupying the neighboring 800 MHz frequency band. As in the case of power consumption measurements, all modules were allowed to utilize the highest available transmission power, i.e., 14 dBm for LoRaWAN and Sigfox, and 23 dBm in the case of NB-IoT. On top of that, LoRaWAN used the SF12 and coding rate of 4/5, which is the setup with the most extended communication range available in the European region [17].

The whole campaign spanned over two months, during which we collected more than 1400 messages from each interface and both measuring units. All measurements were conducted in a publicly available consumer network with a multi-gateway setup reflecting the real-world condition where a sensor can connect to the best available base station (which may change during the measurements). Considering the radio evaluation parameters, for LoRaWAN and Sigfox, we used RSSI as it represents the only available evaluation metric. On the other hand, the more complex NB-IoT supports the RSRP metric, which provides more accurate estimations by excluding interference from the remaining antenna sectors.

The cursory analysis of measurement results depicted in Fig. 4 verifies the general premise of signal propagation in urban and suburban areas. The samples from the suburban area indicate better results by 10 to 20 dB compared to the urban ones. In the case of NB-IoT, the more detailed analysis verifies the previous findings as the fluctuation of the BUT sensor is at least two times lower. Interestingly, LoRaWAN results in the urban environment display occasional short bursts of improved signal values. The suburban scenario shows similar behavior but in the opposite direction (decreased signal strength). Moreover, this finding is also valid for the Sigfox sensor in the city outskirt.

The rationale for this behavior is related to the core property of LoRaWAN and Sigfox technologies – multi-gateway reception. In other words, the conveyed message can be received by multiple gateways in the sensor's reach. The internal network mechanism filters the redundant datagrams and keeps only one copy of the message. Based on the acquired results, the messages are filtered based on the SNR value, which does not have to correlate with the RSSI metric. Hence, the short peaks representing lower RSSI values may stand for the samples with the best SNR.

C. City-Wide Measurement Campaign

For the non-stationary mobile scenario, we conducted a second measurement campaign focused on the city-wide coverage assessment. Again, all the measurements were performed using the same multi-RAT prototype with settings identical to the former campaign. However, the measurement procedure was different. Initially, the device was transferred to a specific location and positioned approximately one meter above the ground level, away from any building or constructions causing outage conditions. When the testing unit was powered up, it conveyed ten messages with a payload of 12 B. The duration between each message was 30 s. As in the previous case, the messages were transmitted sequentially to minimize inter-technology interference.

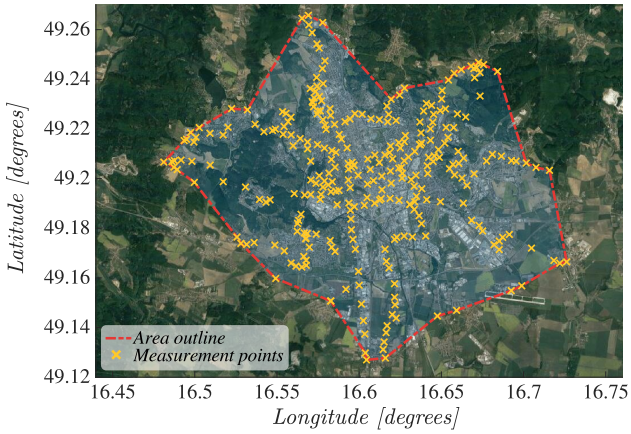
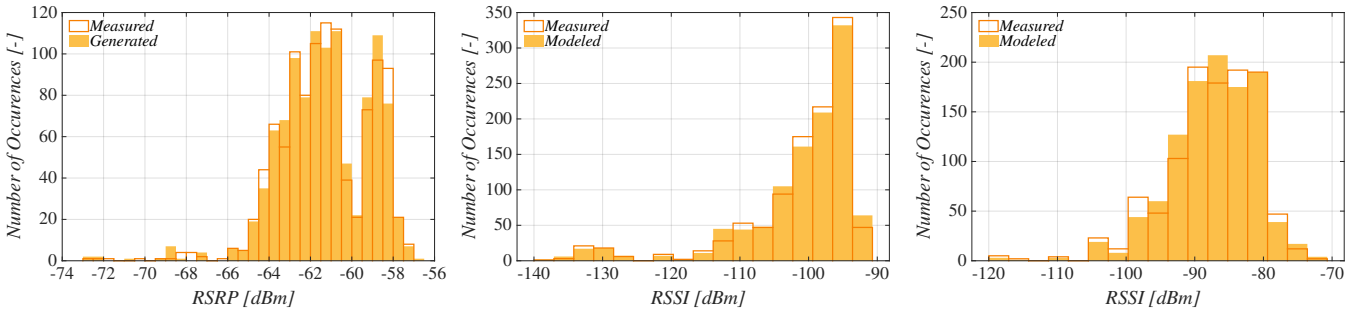


Fig. 5: Locations of measurement points.



(a) NB-IoT (b) Sigfox (c) LoRaWAN

Fig. 6: Relative frequency histogram of RSRP/RSSI samples.

Overall, the measurement campaign covered over 300 unique measurement points in the city of Brno and its outskirts; see Fig. 5. From the perspective of the geographical size, it represents an area spanning over 12 km north to south and 24 km east to west. Notably, the selected measurement spots followed the public transport stops, as one of the target applications falls in the smart transportation field.

By concentrating on the overall number of successfully served points, we observe that all the LPWAN technologies provided satisfactory results. From this perspective, the most robust connectivity offers NB-IoT with only three unserved locations. Sigfox, with 10 dropped points, lies in the middle, followed by LoRaWAN with 16 unserved places. These values represent the number of sites from which no message has been received. From the cumulative packet delivery ratio (PDR) perspective, the decline in LoRaWAN performance is even more pronounced. On the other hand, the PDR of NB-IoT and Sigfox is almost equal. Numerically expressed, it represents the PDR of 0.958 for NB-IoT, 0.947 in the case of Sigfox, and finally, 0.83 for LoRaWAN.

Similar to the previous metrics, NB-IoT also provides the best results in terms of signal levels, with an average RSRP of -76 dBm. On the other hand, the two remaining technologies display significantly lower values close to -100 dBm. More specifically, RSSI values were -112 dBm for Sigfox and -98 dBm in the case of LoRaWAN. These significant differences in signal strength are related to the differences in network topology, primarily due to differences in BSs density. Logically, NB-IoT provides the densest BSs infrastructure among all three technologies, with an average distance to the nearest BS not exceeding 0.52 km. Contrary, the most sparse deployment is provided by Sigfox, with a 3.45 km average BS-ED separation. Finally, LoRaWAN lies between these two, with an average distance of 1.86 km.

IV. Reinforcement Learning for Battery Lifetime

Though the ultimate goal of both scenarios is to achieve the maximum battery lifetime by using RL policies, preprocessing of input data differs significantly. In the case of time-dependent modeling (stationary EDs), the input signal samples are expanded by using a doubly stochastic Markov chain framework. For the non-stationary EDs, a sufficient input data set was generated using interpolation.

The remaining steps of the RL process were the same for both scenarios. First, the input data was used to derive the MAB rewards based on the signal level and corresponding power consumption. Moreover, in the case of non-stationary EDs, rewards were extended to cover also the dropout probabilities. Finally, we applied MAB RL policies with the primary goal of achieving the most extended battery lifetime possible.

A. Stationary Deployment Scenario

Albeit we collected more than 1400 messages from each technology, it is not sufficient for considered RL algorithms. Hence, we resorted to the development of a time-dependent model building upon doubly stochastic Markov chains. This model allows generating samples of unlimited length statistically equivalent to the original dataset.

At the first step of the model derivation, we identified the classes of models suitable for RSRP/RSSI samples approximation. To this aim, we observed the first- and second-order characteristics of samples, i.e., histogram and autocorrelation function (ACF). The resulting histograms and ACF from the BUT unit are depicted in Fig. 6 and Fig. 7, respectively. As one may observe, the histogram of relative frequencies has a specific structure that no available distribution can accurately model. Going further, ACF is characterized by near-exponential decay, which eventually approaches zero. Based on these findings, we can make the following conclusions: (i) the considered RSSI/RSRP process is ergodic in nature, (ii) doubly stochastic Markov chain framework (also known as the hidden Markov model) represents a promising candidate

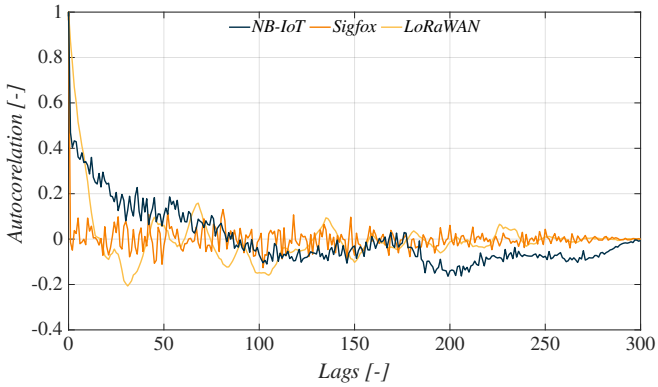


Fig. 7: ACFs for considered measurements.

for the target assessment [30]–[32]. Notably, the doubly stochastic Markov model offers a balance between the modeling accuracy and the simplicity of the fitting algorithm.

The next step in parametrizing the doubly stochastic Markov model is to determine the number of model states N. Then the transition probabilities $p_{ij},\ i,j = 1,2,\ldots,N$ from the current state i to the next one j and the conditional probability mass function (PMFs) associated with each state $f_i(j),\ i = 1,2,\ldots,N,\ j \geq 0$ can be derived. We used the procedure based on the kernel density estimation (KDE) function; however, one may use an arbitrary algorithm to fit the double stochastic Markov model from statistical data. The basic premise of utilizing the KDE function is to cluster the input data and derive the number of states N from the resulting curve [33]. Notably, this is a two-step process that includes (i) estimation of kernel distribution, i.e., non-parametric representation of probability density function (PDF), and (ii) data clusterization based on local maxima evaluation. The PDF is derived as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{1}$$

where n represents the number of samples, h is the bandwidth, x denotes the actual value, and x_i stands for random samples from an unknown distribution. As the Kernel density estimator K, we used the normal kernel function, which is evaluated at equally-spaced points x_i, covering the whole input data set. Notably, each local maximum of the resulting curve represents one boundary of the Markov chain state.

Once the number of states is determined, we proceed to derive the transition probabilities $p_{ij},\ i,j = 1,2,\ldots,N$ and the PMFs associated with each state, $f_i(j),\ i = 1,2,\ldots,N,\ j \geq 0$ by using conventional statistical approaches. Having defined the boundaries between individual states of the Markov chain, we calculate the number of state transitions for the particular value of i and j, i.e., changes between the previous and current value in the trace. Finally, the number of state changes is divided by the input trace length to obtain transition probabilities.

Aiming to assess the proposed time-dependent model's performance, we generated an entirely new data set containing ten thousand samples using the developed model. We used histogram and ACF of both statistical and model data for the performance comparison, as depicted in Figs. 6 and 8. As one can observe, the samples generated by the module provide a tight match with the measured data. Also, the exponentially decaying behavior of the ACF is well captured for all considered LPWAN technologies. Notably, the χ goodness-of-fit test performed with the significance level of 0.05 verifies that both measured and generated samples statistically belong to the same distribution. These findings allow us to conclude that the proposed approach can be utilized for time-dependent RSRP/RSSI modeling of LPWAN technologies.

B. Mobile Deployment Scenario

The process of samples derivation for mobile EDs vastly differs from the stationary nodes. For mobile ED, it represents a two-step process. First, measured data must be interpolated, and then the ED tracks are derived from the location coordinates.

1) RSRP/RSSI Model: The city-wide measurement campaign included over 300 unique measurement points. Though it represents an extensive data set, it is still not sufficient for signal coverage predictions for the arbitrarily selected path. To overcome this issue, we employed an interpolation algorithm to fill the missing data. The resulting evenly spaced grid with 50 m resolution is sufficient for tracks planning.

Specifically, we used the Kriging interpolation algorithm, which belongs to the class of geostatistical methods. These methods can generate the predictions surface and also provide an assessment of the interpolation accuracy by itself. According to the literature, Kriging accuracy is highest when spatially correlated distance or directional bias is present in the data [34]. The selection of the Kriging algorithm is based on our previous research, where this method provided the most promising results in terms of signal interpolation accuracy.

Similarly to other interpolation methods, according to the Kriging algorithm, the value of the predicted point is calculated as a weighted sum, i.e.,

$$x_q = \sum_{i=0}^{N} \lambda_i \cdot z_i, \tag{2}$$

where N is the number of neighbors surrounding the sample, z_i is the point value, and λ_i represents the weight of each element. However, the process of weights derivation differs as it depends not only on the distance to the sample point but also on the overall spatial arrangement. To this aim, the first step of the Kriging procedure is the creation of an experimental semi-variogram, given as

$$\gamma(h) = \frac{1}{2N(h)}\sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2, \tag{3}$$

where $N(h)$ is the number of pairs separated by the distance of h and $z(x_i)$ is the value of the input point [35].

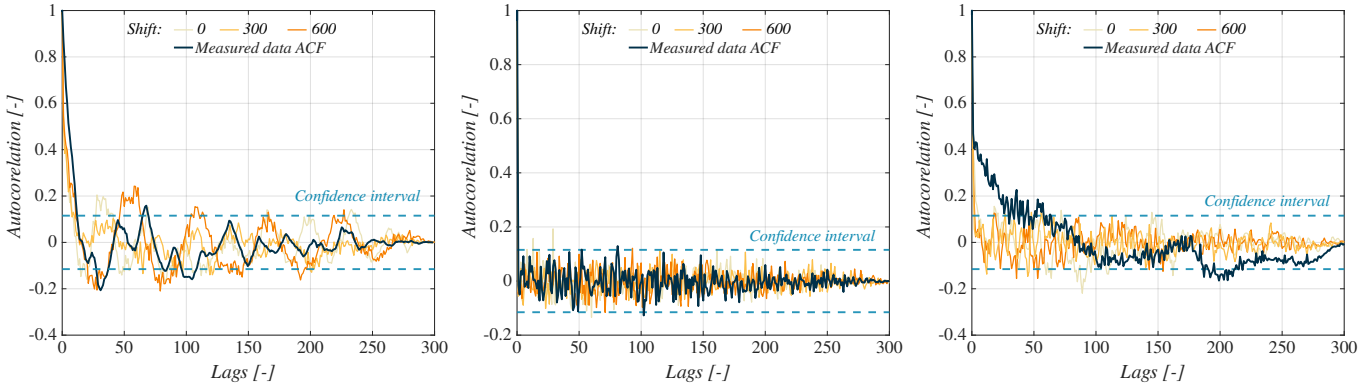
The experimental semi-variogram represents only a discrete set of points, which is insufficient for our purposes. To this aim, we interlaced the experimental semi-variogram with a Spherical model, which is defined as

$$\gamma(h) = \begin{cases} c_o + c_1\left[\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right] & \text{for} \quad 0 < h < a \\ c_o + c_1 & \text{for} \quad h \geq a \end{cases}, \tag{4}$$

where a is the range, c_0 represents nugget variance, and $c_0 + c_1$ stands for the sill. In the last step of the interpolation, we apply the Ordinary Kriging to derive the value of desired points [36]. Such a system is stated as

$$\begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} C_{11} & \cdots & C_{1n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C_{n1} & \cdots & C_{nn} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} C_{10} \\ \vdots \\ C_{n0} \\ 1 \end{bmatrix}, \tag{5}$$

where λ_i represents the point weight, μ is the Lagrange parameter, and C_{1n} denotes the covariance between the location



(a) NB-IoT (b) Sigfox (c) LoRaWAN

Fig. 8: Comparison of ACFs of modeled and empirical data.

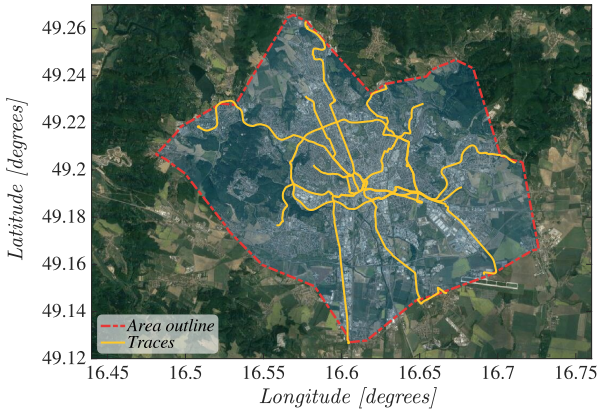


Fig. 9: Illustration of the routes for delivery application.

of sample points x_1 and x_n. Next, the covariance for the C_{1n} is computed as follows

$$C_{1n} = Cov(x_1 - x_n) = C(0) - \gamma(x_1 - x_n), \qquad (6)$$

where $C(0)$ represents the semi-variogram sill; finally, λ is the semi-variogram output for points x_1 and x_n.

2) Service Delivery Model: Once the signal coverage map is obtained, we are in a position to proceed with defining tracks for the mobile EDs. Thus, in a web map application, we created several routes intersecting the city of Brno in all possible directions, mimicking the delivery and transit paths. These tracks, depicted in Fig. 9, were designed to start, end, or cross a single transit point representing the central warehouse. Finally, all constructed tracks were exported to GPS exchange format (GPX) for the subsequent processing steps. Notably, the GPX data was exported with points resolution ranging between 30 and 100 m providing sufficient correlation with the interpolation grid of the coverage map.

Continuing the process, we set the initial position of the mobile ED to the central warehouse. From this point on, the ED followed the random path from the list of available tracks. After it reached the end of the selected way, it returned to the central warehouse and randomly chose another route. Notably, the ED was set to transmit a single 12 B message every 30 s. Also, the ED speed was not assigned to a specific value, but it ranged from 15 to 50 km/h to provide a city-like traffic pattern. The signal value for the intended locations was generally derived as follows. When the ED reached the point of message transmission, which was derived using the actual ED speed and distance to the previous point, we extracted the signal value from the closest cell of the interpolated coverage map. Notably, we constructed two coverage maps, one created using the highest signal values and the second from the lowest signal levels. These two maps allowed us to extend the soundness of data further, as the samples are uniformly generated from the range of maximum and minimum values at each location. This process is repeated until all tracks are not covered, and in total, it is relaunched 100 times.

C. Machine Learning Formalization

1) Solution Methods: One of the most popular RL approaches is called Q-learning. This solution implies that an agent learns the action-value function $Q^{\pi}(S, A)$ corresponding to the expected accumulated reward when action A is taken in the state S relying on the RL policy π. However, the computational and memory requirements make the calculation practically impossible when the state-action space becomes too large. To overcome this issue, deep RL uses machine learning models such as neural networks to approximate the immense state-action space. The following action is then determined by the maximum output of the underlying Q-network [29], [37].

Notably, in this work, the main focus is given to the MAB problem. It simplifies the Q-learning problem as the agent can choose only one of K actions at a given time instance t, whereas the environment's state S_t remains unchanged for all time steps. It makes all the successive time steps independent and identically distributed (IID). Utilization of the MAB approach further simplifies the role of the time step as it represents only how many actions have been taken to this point. In our model, the action is taken every 30 seconds representing the message transmission period. Nevertheless, the primary goals stay the same, i.e., choose actions such that the total reward received within a certain number of time instances is maximized. This simplified approach can be formalized as $A_r \in 1, 2, 3 \ldots, K$ with a reward $R_t \sim N(\mu, \sigma^2)$ where k is the action taken [29], [38].

Further, the bandit is allowed to pull a single arm at each time instance t, i.e., represents a single action bandit with finite

action space. Finally, it must be noted that due to the dynamic characteristics of the radio propagation environment also, the reward distribution is non-stationary. Hence RL policies have to pay a cost for this fact [38].

The rewards are based on the amount of consumed energy during the data transmission. Thus, the technology requiring the least amount of energy is rewarded with a value of 1. The second and third technology is awarded 0.5 and 0, respectively. Notably, the consumed energy is directly derived from the sample's signal levels. To this aim, results from the energy consumption measurements from Section III-A were used.

For the mobile EDs, the rewards generation is extended to incorporate the probability of message loss. When the signal level or SNR drops under a particular value, the message is considered lost. In terms of signal strength, it means RSRP less than -135 dBm for NB-IoT and RSSI under -142 dBm in the case of Sigfox. For LoRaWAN, the borderline RRSI values are delineated by -129 dBm for SF9 and -137 dBm in the case of SF12. Considering the noise values, the NB-IoT signal-to-interference plus noise ratio (SINR) limit is around -8 dB. For Sigfox and LoRaWAN with SF12, the minimum SNR is as low as -20 dB. Nevertheless, the minimum value of SNR for LoRaWAN using SF9 is only -12 dB. Thus, when any technology's signal/SNR decreases below these thresholds, it is associated with zero reward [14], [17], [39].

On top of that, the second mobile nodes scenario encompasses the message losses based on the measurement statistics. Based on the PDR values introduced in section III-C, the probability of successful message derivation is further reduced by the loss probability accordingly. The possibility of message loss is taken from the uniform distribution following the PDR values as mentioned above.

D. Reinforcement Learning Policies

As mentioned in the previous section, the process of selecting the action A_t with the ultimate goal of achieving the maximum cumulative reward is driven by RL policies. In our work, we tested four RL policies: ε-greedy, weighted average, UCB, and Thompson sampling.

1) ε-greedy: This policy represents the simplest method of addressing the MAB problem. Its operation is divided into the exploration and exploitation phase. In the exploration phase, the algorithm randomly selects the arms to pull with the probability given by the parameter ε. This approach helps ε-greedy policy to overcome issues with the local-optimum solution and discover the arm with the actual highest rewards. In the remaining time, i.e., exploitation phase $(1-\varepsilon)$, the algorithm attempts to gain the highest rewards by pulling the same arm repeatedly. To select the optimal radio interface, the algorithm keeps statistics of average rewards from each arm. However, the statistics are calculated incrementally to save computational resources (due to the possibility of implementation on power-restricted devices). Using the incremental averaging algorithm, the value of action Q_{k+1} is defined as

$$Q_{k+1} = Q_k + \frac{1}{k+1}\left[r_{k+1} - Q_k\right], \tag{7}$$

where k is the order of current step, r_{k+1} is the current reward, and Q_k represents the average of the first k actions [40].

2) Weighted Average: In the case of a non-stationary environment, the ε-greedy policy provides unsatisfactory results, as all samples in statistics have the same weight. In other words, the influence of each action on the resulting average is identical. Nevertheless, in the dynamically changing environment, it is logical to increase the impact of the more recent actions compared to more distant ones. To this aim, the weighted average policy uses step-size constant α, which controls samples' weight. With this modification, the iterative average is calculated as

$$Q_{k+1} = Q_k + \alpha\left[r_{k+1} - Q_k\right], \tag{8}$$

where weight parameter α $(0 < \alpha \leq 1)$ controls action significance, Q_k is average of k previous actions, and r_{k+1} represents current reward [40].

3) Upper Confidence Bound: Instead of relying on the selection of arbitrary action in the exploration phase with constant probability, UCB policy changes its exploration-exploitation ratio as it gathers more knowledge about the environment. In the beginning, UCB focuses primarily on exploration when the actions tried the least number of times are preferred. However, over time UCB moves towards exploitation, selecting the actions with the highest estimated rewards.

With UCB, the value of the $k+1$ action Q_{k+1} is given as

$$Q_{k+1} = Q_k + c\sqrt{\frac{\ln k}{N_k}}, \tag{9}$$

where Q_k is the estimated value of the action at time step k, N_k is the number of times the arm has been selected prior to time k, and c represents a confidence value controlling the exploration level.

The parameter Q_k represents the exploitation part. In this phase, the action that currently has the highest estimated reward will be the chosen one. Conversely, the second part represents the exploration phase driven by the parameter c. If the action has not been selected often or not at all, then N_k will be small. As a result, it will lead to significant uncertainty, making this action more likely to be chosen. However, the uncertainty decreases with each action's selection, making it less likely to be selected in the exploration phase. Notably, when the action is not selected, its uncertainty will grow slowly due to logarithmic dependency. Conversely, the certainty proliferates with each selection as the increase in N_k is linear. Thus, as time progress, the exploration gradually decreases as the second part of (9) goes to zero [38], [41].

4) Thompson Sampling: Unlike the other RL policies described in this text, Thomson sampling is a probabilistic algorithm based on Bayesian ideas. The sampling in its name denotes that it picks samples from a probability distribution for each arm. Usually, a Beta distribution for each arm based on its number of attempts and successful rewards throughout history is used. In other words, a random sample from the posterior Beta distribution is taken at each iteration, and one

with the maximum value is chosen. The value of $k+1$ action Q_{k+1} is sampled from the Beta distribution defined as

$$Q_{k+1} = \beta(N_k^1 + 1, N_k^0 + 1), \tag{10}$$

where N_k^1 is the number of times the action got a successful reward prior actual round, whereas N_k^0 denotes the opposite cases, i.e., when the reward was zero. This approach allows the Thomson sampling to balance the exploration-exploitation dilemma.

V. Numerical Results

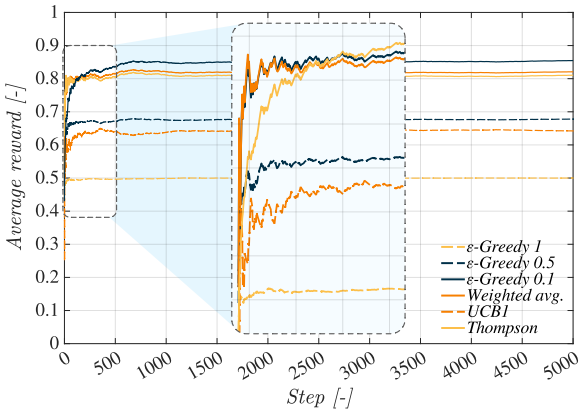
In this section, we present the result for both stationary and mobile EDs scenarios utilizing the RL policies mentioned above. The stationary nodes results include average rewards as well battery lifetime expectancy. Subsequent mobile EDs scenarios add simulations with different SFs for LoRaWAN technology and extend the results with outages probabilities.

A. Stationary End Devices

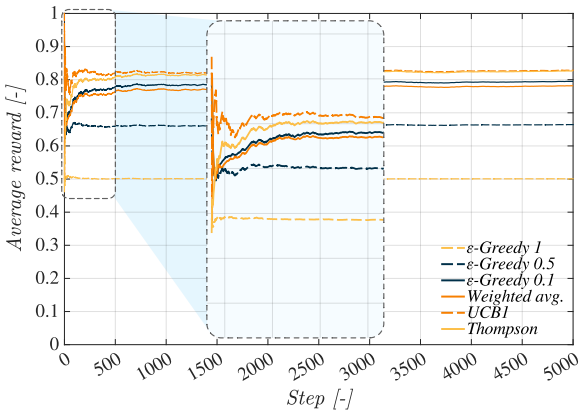
At the first step of the RL policies assessment, we focused on their ability to select the best radio interface and achieve maximum average rewards. To this aim, we generated input data set of 5000 samples using the proposed time-dependent model based on Markov chains. Moreover, we launched the RL policies 200 times to improve statistical confidence and averaged the results throughout all realizations. The same time-dependent model was then used to generate a greatly extended dataset intended for battery lifetime predictions.

1) Average Rewards: As depicted in Fig. 10, the simulations were conducted for both locations, i.e., BUT campus and BCC unit. Notably, the analysis of the BUT unit shows surprising results. Albeit being the more advanced RL policy, the UCB algorithm displays the second-worst results in terms of average rewards, hardly overcoming a value of 0.65 at maximum. The UCB policy probably exploited the local optimum solution, which crippled its results. The only worse policy is ε-greedy 1, which operates in exploration mode all the time, providing a random selection of interfaces. On the other hand, the ε-greedy 0.5 shows slightly better results than the UCB policy. It suggests that exploitation is better for the current scenario than exploration, as the ε-greedy 0.1 provides the second-best performance. Notably, the weighted average RL policy with $\alpha = 0.2$ provides satisfactory results as it holds third place with average rewards of more than 0.8 in its steady-state. However, the first place with an average reward of almost 0.85 belongs to Thompson sampling.

In the case of the BCC, the position of the UCB RL policy is entirely different as it represents the best-performing algorithm. Surprisingly, the UCB and weighted average policies indicate overshoots in the initial phase of the algorithm runs (first 100 steps, i.e., time instances t of the conducted action). The UCB even achieves the average reward of 1, which represents the maximum achievable value. This behavior is most likely caused by the exploration phase with a certain level of serendipity in selecting the correct radio interface. However, with subsequent selections, the average rewards decrease to expected levels. Regarding the remaining RL policies, their



(a) BUT campus



(b) Brno city center

Fig. 10: Average rewards of RL policies.

performance is similar to the BUT unit. Hence, the Thompson sampling (holding second place) currently represents the most reliable RL policy.

Lastly, it is essential to mention that all RL policies can achieve 90 % of their maximum average rewards during the initialization phase utilizing less than 50 messages. With the UCB algorithm, it is possible to reach this value with only 25 messages. Though providing one of the best results in the exploitation state, the convergence time of Thompon sampling may be slightly longer. This fact is pronounced especially in the case of the BUT unit, where the transition time is roughly two times longer than for other policies. However, it must be noted that the best performing RL policies allow for exploiting up to 85 % of the theoretical maximum average reward defined by the value one.

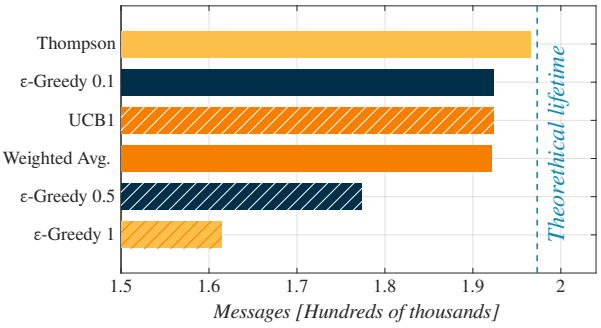
2) Expected Battery Lifetime: In the following assessment step, we evaluated the expected battery lifetime of multi-RAT devices. For this scenario, we considered a commonly utilized lithium battery LS 14500, produced by the Saft company. This primary cell provides a nominal voltage of 3.6 V and a capacity of 2.6 Ah, which equals the total charge of 33696 J. Due to the extensive battery capacity, the input sample data set had to be extended to up to two million samples.

As in the previous case, the battery lifetime expectancy results depicted in Fig. 11 reveal the most critical findings for the BUT unit. At first sight, the battery lifetime prediction

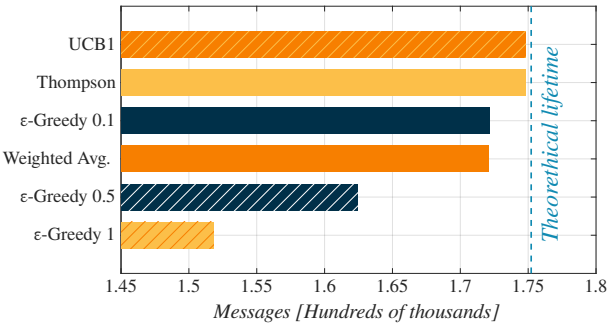
data show significant differences compared to the average rewards depicted in Fig. 10. Notably, the first two policies, i.e., Thompson sampling and ε-greedy 0.1, indicate expected results and hold the same place in both figures. However, the remaining UCB, weighted average, and ε-greedy 0.5 are in a completely different order. Most surprisingly, the UCB policy holds third place in terms of expected battery lifetime but is the next to last in the case of average rewards. Conversely, the performance of weighted average and ε-greedy 0.5 is underwhelming, which may seem counter-intuitive. Nevertheless, a more detailed analysis of the results reveals the reasons for such behavior.

Although the UCB policy average rewards are not great, it still manages to select the suboptimal interface represented by the second best-performing technology. Notably, LoRaWAN consumption is exceptionally close to the NB-IoT technology leaving only a tiny gap between them. Hence, by selecting the LoRaWAN or NB-IoT interfaces most of the time, UCB penalization of low rewards diminishes. On the other hand, more straightforward policies such as ε-greedy and weighted average select the radio interfaces in the exploration phase randomly, levering the Sigfox arm, which displays the highest power consumption. This blind selection is the main reason for the performance drop of these two policies in terms of battery lifetime. From the perspective of the best performing RL policies, Thompson sampling can exploit 99.5 % of the theoretical maximum. This value delineated by the dashed line represents the highest achievable battery lifetime when the best performing radio interface is selected for every transmission.

In the case of the BCC, the results are much more predictable. In fact, the battery lifetime prediction follows the

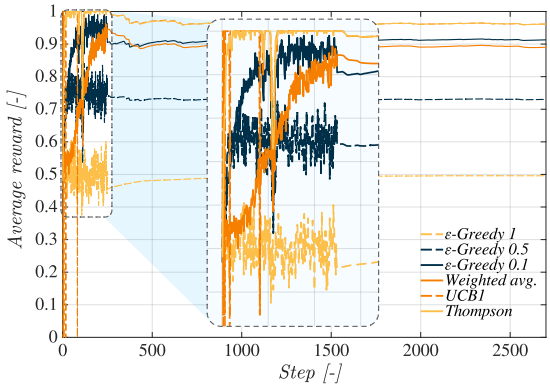


(a) BUT campus

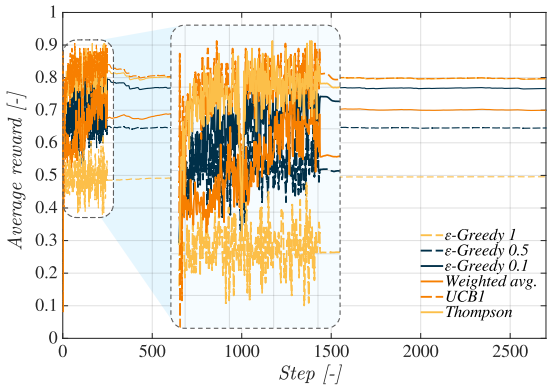


(b) Brno city center

Fig. 11: Battery lifetime expectancy for all studied policies.



(a) Without losses.



(b) With losses.

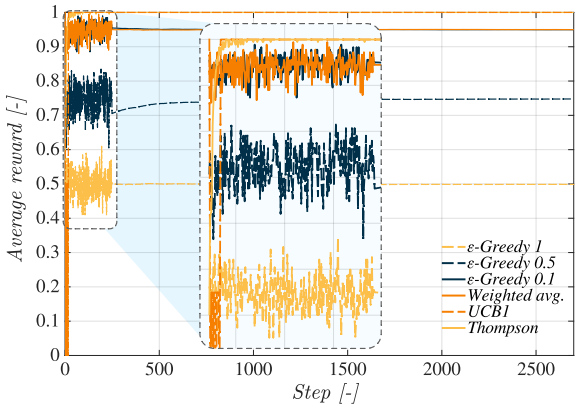
Fig. 12: Average rewards of RL policies for LoRaWAN SF9.

average rewards. Notably, the UCB algorithm outperforms Thompson sampling by a thin margin of 500 messages. The generally positive results of Thompson sampling make this policy a promising candidate for deploying in multi-RAT devices.

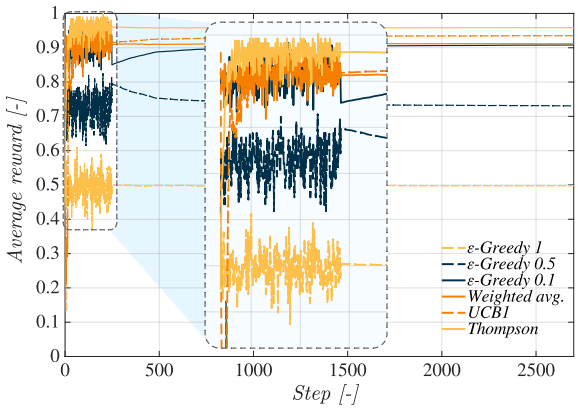
B. Mobile End Devices

For the mobile EDs, we conducted a similar set of tests as in the case of stationary units, i.e., we assessed the average rewards and evaluated expected energy consumption. Notably, the representation of energy consumption is slightly different as the presented data displays the consumed charge in J for the distance each node traveled on the considered traces. In other words, these results provide a side-by-side comparison of energy consumed by traversing the virtual trails, identical to all RL policies. Finally, the last difference covers the inclusion of message loss numbers.

We compare average rewards for two different LoRAWAN SFs in the first phase, namely SF9 and SF12. Although the SF12 allows LoRaWAN to achieve the most extended communication range possible, DC limitations hamper its usability for our scenario. In the case of the EU, it requires almost 150 s of radio silence after 12 B message transmission. However, it is not in line with our needed message period of 30 s. Furthermore, if we consider the US frequency band, the SF12 can not be used at all due to a dwell limitation of 400 ms; see Section II-C for more information. On top of that, we also

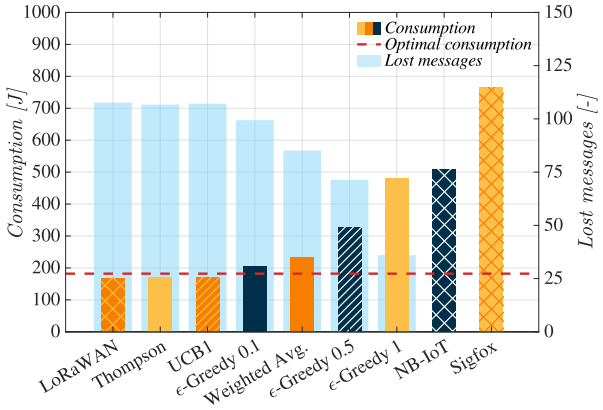


(a) Without losses.

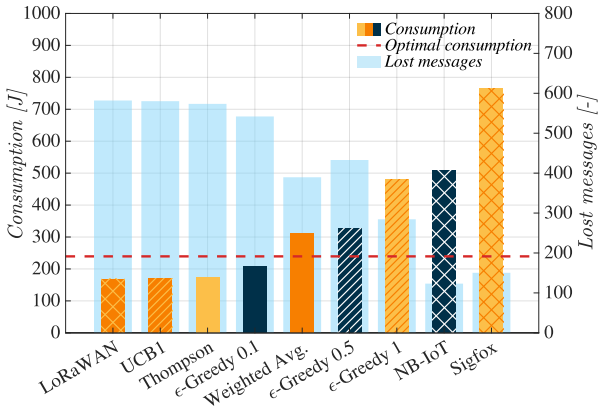


(b) With losses.

Fig. 13: Average rewards of RL policies for LoRaWAN SF12.



(a) Without losses.



(b) With losses.

Fig. 14: Average rewards of RL policies for LoRaWAN SF9.

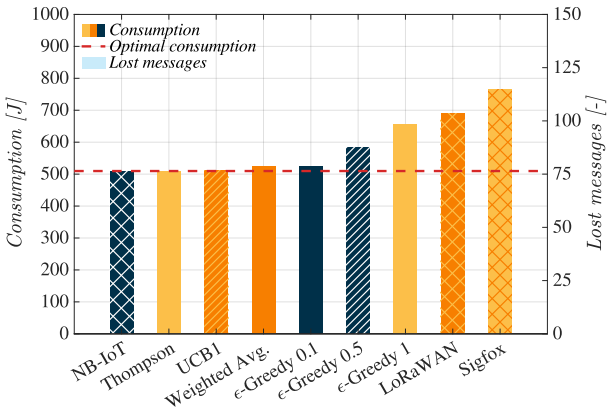
compare results from the perspective of including/excluding outages. In the figures, where the outages are excluded, the message loss can occur solely when RSRP/RSSI or SNR drops under a certain level; see Section IV-C. Contrary, the results that include message losses also cover the PDR statistics (see Section III-C for details) to reflect the possibility of an outage.

1) Average Rewards: A cursory analysis of average rewards for SF9 depicted in Fig. 12 reveals that the initial exploration phase indicates higher variation than the stationary nodes. However, when it reaches the steady-state, the results are comparable to stationary EDs. In the case of results without losses, the average rewards are even higher than for stationary EDs. Also, the order of individual policies is similar to the stationary units, with Thompson sampling and UCB occupying front positions. It is clear that the inclusion of outages brings some uncertainty, resulting in lower average rewards. Remarkably, the outages more pronouncedly influence advanced algorithms like Thomson sampling and UCB, which still provide better results, but the lead over simpler policies is smaller.

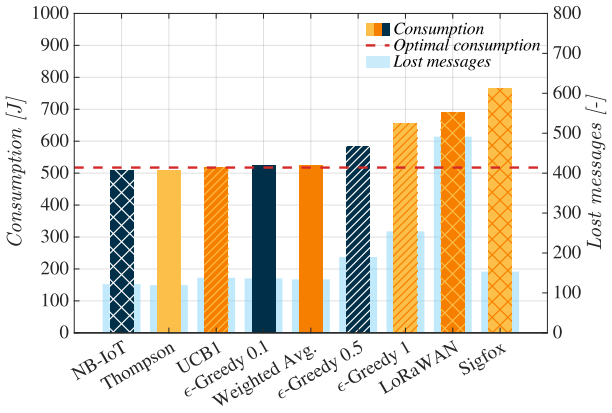
Proceeding further, one may conclude that the average rewards for SF12 depicted in Fig. 13 follow the same pattern as SF9 results. However, the level of rewards for both scenarios is significantly higher compared to SF9. It has a logical explanation, as the increased sensitivity of SF12 diminishes the possibility of message loss due to low RSSI/SNR levels, which is typical for SF9. Impressively, in the case of results without outages, the Thompson sampling and UCB policies nearly reach the maximum value of average rewards. These positive results are mainly caused by the fact that the energy consumption of NB-IoT is superior to LoRaWAN in SF12, with a very low loss probability. Hence, selecting the NB-IoT technology frequently represents an optimal solution for higher reliability scenarios where the private infrastructure is not a deciding factor. Furthermore, the average rewards for message outages are also high as the loss probability of NB-IoT is less than 5 %.

2) Energy Consumption: The energy consumption results depicted in Fig. 9 represent more than three thousand messages in a single run. However, the process was relaunched 100 times to improve accuracy by averaging. Notably, for the proper understanding of the results, it is essential to describe the meaning of the "Optimal consumption" line in Figs. 14 and 15. This dashed red line represents the energy consumption value if all the messages would be transmitted over the best interface or, via the technology, ensuring successful delivery in the case of an outage. It must be noted that this line does not have to correspond with the combination of transmission providing the lowest power consumption.

Analyzing the energy consumption results for SF9, see Fig. 14, it is clear that the LoRaWAN holds first place in terms of consumed energy. Also, the previously mentioned Thompson sampling and UCB provide comparable results, with



(a) Without losses.



(b) With losses.

Fig. 15: Average rewards of RL policies for LoRaWAN SF12.

expected consumption below the optimal value line. However, a more detailed analysis reveals that all these choices indicate a significant amount of outages. In the first case (Fig. 14a), the outages are caused by insufficient RSSI/SNR levels. Therefore the number of message losses is only marginal. Conversely, when the losses are included (Fig. 14b), the number of outages is significant. Due to the tendencies of Thompson sampling and UCB policies to exploit the currently optimal solution, they cannot react to these occasional dropouts. Surprisingly, the weighted average policy can handle this issue with relative success. In terms of energy consumption, it still provides better results than the two remaining LPWAN technologies, and it is in the middle of RL policies. On the other hand, it still allows decreasing message loss by approximately 20 % as compared to single LoRaWAN technology.

In the case of energy consumption results for SF12 depicted in Fig. 15, the situation is radically different. The Thompson sampling and UCB still provide the best results among all policies, but the first place with the lowest power consumption belongs to NB-IoT. The increased sensitivity of SF12 gives the impression of zero outages when the message losses are neglected. However, for the latter case, LoRaWAN displays a high amount of outages, even for higher SF. Notably, the performance of Thompson sampling is worth mentioning. It provides nearly identical energy consumption as NB-IoT, but it is able to decrease the number of outages slightly.

From the perspective of versatility, the performance of the weighted average policy should be noted. Though it does not perform the best, it still holds fourth place (considering only RL policies) with only marginally higher consumption over NB-IoT with a comparable number of outages. Moreover, the weighted average policy displays consistent results for both SFs regardless of loss probabilities inclusion.

VI. Conclusions

To improve the operational lifetime of end-user equipment in LPWAN systems under dynamically changing propagation conditions, we considered the employment of the multi-RAT approach at a single ED and the automatic selection between available RAT throughout its operation. To facilitate the radio selection process's dynamic adaptation, we proposed RL techniques. In this case, the system in question regularly determines the environment conditions and assigns the weights to different options attempting to achieve the maximum reward level. To assess the performance of the system in realistic conditions, we performed two large-scale measurement campaigns targeting power consumption and radio signal propagation. Furthermore, we verified the performance of the proposed schemes in diverse, realistic conditions by modeling both stationary deployment and city-wide delivery service conditions.

Our numerical results indicate that the considered RL-based techniques allow for a noticeable increase in EDs' lifetime when operating in multi-RAT mode. Out of all considered schemes, the performance of the weighted average policy shows the most consistent results for both stationary and mobile deployments. For the stationary EDs, the best performing RL policy, Thomspon sample, exploits up to 85 % of the theoretical gains. In the case of mobile EDs, the RL policies battery lifetime difference can be as high as 200 %. These findings are of particular importance in deployments with harsh or hard-to-reach conditions.

Acknowledgment

For the research, the infrastructure of the SIX Center was used. The described research was financed by the Technology Agency of the Czech Republic project No. TN01000007.

References

[1] K. Samdanis and T. Taleb, "The Road beyond 5G: A Vision and Insight of the Key Technologies," *IEEE Network*, vol. 34, no. 2, pp. 135–141, 2020.

[2] M. Marvi, A. Aijaz, and M. Khurram, "Toward an Automated Data Offloading Framework for Multi-RAT 5G Wireless Networks," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2584–2597, 2020.

[3] M. Stusek, D. Moltchanov, P. Masek, K. Mikhaylov, O. Zeman, M. Roubicek, Y. Koucheryavy, and J. Hosek, "Accuracy Assessment and Cross-Validation of LPWAN Propagation Models in Urban Scenarios," *IEEE Access*, vol. 8, pp. 154 625–154 636, 2020.

[4] D. Solomitckii, A. Orsino, S. Andreev, Y. Koucheryavy, and M. Valkama, "Characterization of mmWave Channel Properties at 28 and 60 GHz in Factory Automation Deployments," in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2018, pp. 1–6.

[5] N. Stepanov, D. Moltchanov, and A. Turlikov, "Modeling the NB-IoT Transmission Process with Intermittent Network Availability," in *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*. Springer, 2020, pp. 241–254.

[6] F. Wang and V. K. Lau, "Dynamic RAT Selection and Transceiver Optimization for Mobile Edge Computing Over Multi-RAT Heterogeneous Networks," *arXiv preprint arXiv:2108.08098*, 2021.

[7] K. Mikhaylov, M. Stusek, P. Masek, R. Fujdiak, R. Mozny, S. Andreev, and J. Hosek, "On the Performance of Multi-Gateway LoRaWAN Deployments: An Experimental Study," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2020, pp. 1–6.

[8] R. Mozny, M. Stusek, P. Masek, K. Mikhaylov, and J. Hosek, "Unifying Multi-Radio Communication Technologies to Enable mMTC Applications in B5G Networks," in *2020 2nd 6G Wireless Summit (6G SUMMIT)*. IEEE, 2020, pp. 1–5.

[9] V. Petrov, A. Samuylov, V. Begishev, D. Moltchanov, S. Andreev, K. Samouylov, and Y. Koucheryavy, "Vehicle-based Relay Assistance for Opportunistic Crowdsensing over Narrowband IoT (NB-IoT)," *IEEE Internet of Things journal*, vol. 5, no. 5, pp. 3710–3723, 2017.

[10] S. Kavuri, D. Moltchanov, A. Ometov, S. Andreev, and Y. Koucheryavy, "Performance Analysis of Onshore NB-IoT for Container Tracking During Near-the-Shore Vessel Navigation," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2928–2943, 2020.

[11] ITU-R, "Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface," International Telecommunication Union, M.2410-0, July 2017.

[12] M. Stusek, D. Moltchanov, P. Masek, S. Andreev, Y. Koucheryavy, and J. Hosek, "Time-Dependent Propagation Analysis and Modeling of LPWAN Technologies," in *2020 IEEE Globecom Workshops (GC Wkshps*, 2020, pp. 1–7.

[13] K. Mikhaylov, V. Petrov, R. Gupta, M. A. Lema, O. Galinina, S. Andreev, Y. Koucheryavy, M. Valkama, A. Pouttu, and M. Dohler, "Energy Efficiency of Multi-Radio Massive Machine-Type Communication (MR-MMTC): Applications, Challenges, and Solutions," *IEEE Communications Magazine*, vol. 57, no. 6, pp. 100–106, 2019.

[14] Sigfox, "Sigfox Connected Objects: Radio Specifications," Sigfox, Ref.: EP-SPECS, Rev.: 1.4, November 2019.

[15] A. Lavric, A. I. Petrariu, and V. Popa, "Long Range Sigfox Communication Protocol Scalability Analysis Under Large-Scale, High-Density Conditions," *IEEE Access*, vol. 7, pp. 35 816–35 825, 2019.

[16] LoRa Alliance™, "LoRaWAN™ 1.0.3 Specification," LoRa Alliance™, Final release, Beaverton, July 2018.

[17] LoRa Alliance®, "RP002-1.0.1® Regional Parameters," LoRa Alliance®, Final, Beaverton, February 2020.

[18] O. Liberg, M. Sundberg, E. Wang, J. Bergman, J. Sachs, and G. Wikström, *Cellular Internet of Things: From Massive Deployments to Critical 5G Applications*. Academic Press, 2019.

[19] A. Hoglund, X. Lin, O. Liberg, A. Behravan, E. A. Yavuz, M. Van Der Zee, Y. Sui, T. Tirronen, A. Ratilainen, and D. Eriksson, "Overview of 3GPP Release 14 Enhanced NB-IoT," *IEEE Network*, vol. 31, no. 6, pp. 16–22, November 2017.

[20] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond: A Deep Reinforcement Learning Based Approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4585–4600, 2021.

[21] F. Al-Tam, N. Correia, and J. Rodriguez, "Learn to Schedule (LEASCH): A Deep Reinforcement Learning Approach for Radio Resource Scheduling in the 5G MAC Layer," *IEEE Access*, vol. 8, pp. 108 088–108 101, 2020.

[22] F. Tang, Y. Zhou, and N. Kato, "Deep Reinforcement Learning for Dynamic Uplink/Downlink Resource Allocation in High Mobility 5G HetNet," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 12, pp. 2773–2782, 2020.

[23] Y. Shi, Q. Cui, W. Ni, and Z. Fei, "Proactive Dynamic Channel Selection Based on Multi-Armed Bandit Learning for 5G NR-U," *IEEE Access*, vol. 8, pp. 196 363–196 374, 2020.

[24] S. Ali, A. Ferdowsi, W. Saad, N. Rajatheva, and J. Haapola, "Sleeping Multi-Armed Bandit Learning for Fast Uplink Grant Allocation in Machine Type Communications," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 5072–5086, 2020.

[25] R. Karmakar, G. Kaddoum, and S. Chattopadhyay, "SmartCon: Deep Probabilistic Learning-Based Intelligent Link-Configuration in Narrowband-IoT Toward 5G and B5G," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 1147–1158, 2022.

[26] B. M. ElHalawany, S. Hashima, K. Hatano, K. Wu, and E. M. Mohamed, "Leveraging Machine Learning for Millimeter Wave Beamforming in Beyond 5G Networks," *IEEE Systems Journal*, vol. 16, no. 2, pp. 1739–1750, 2022.

[27] P. Wei, K. Guo, Y. Li, J. Wang, W. Feng, S. Jin, N. Ge, and Y.-C. Liang, "Reinforcement Learning-Empowered Mobile Edge Computing for 6G Edge Intelligence," *IEEE Access*, vol. 10, pp. 65 156–65 192, 2022.

[28] R. Zhou, X. Zhang, S. Qin, J. C. Lui, Z. Zhou, H. Huang, and Z. Li, "Online Task Offloading for 5G Small Cell Networks," *IEEE Transactions on Mobile Computing*, vol. 21, no. 6, pp. 2103–2115, 2022.

[29] F. Bach, R. Sutton, and A. Barto, "Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning Series)," 2018.

[30] T. Rolski, "Upper Bounds for Single Server Queues With Doubly Stochastic Poisson Arrivals," *Mathematics of Operations Research*, vol. 11, no. 3, pp. 442–450, 1986.

[31] S. B. Slimane and T. Le-Ngoc, "A Doubly Stochastic Poisson Model for Self-similar Traffic," in *Proceedings IEEE International Conference on Communications ICC'95*, vol. 1. IEEE, 1995, pp. 456–460.

[32] D. Moltchanov, Y. Koucheryavy, and J. Harju, "The Model of Single Smoothed MPEG Traffic Source Based on the D-BMAP Arrival Process with Limited State space," in *Proc. of ICACT*, 2003, pp. 55–60.

[33] Y. Huang, X. Chen, and W. B. Wu, "Recursive Nonparametric Estimation for Time Series," *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 1301–1312, Feb 2014.

[34] Q. Zhang and J. Wu, "Image Super-resolution Using Windowed Ordinary Kriging Interpolation," *Optics Communications*, vol. 336, pp. 140 – 145, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003040181400889X

[35] A. McBratney and R. Webster, "Choosing Functions for Semi-variograms of Soil Properties and Fitting Them to Sampling Estimates," *Journal of soil Science*, vol. 37, no. 4, pp. 617–639, 1986.

[36] M. R. Inggs and R. T. Lord, "Interpolating Satellite Derived Wind Field Data using Ordinary Kriging, with Application to the Nadir Gap," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 34, no. 1, pp. 250–256, January 1996.

[37] Y. Yu, T. Wang, and S. C. Liew, "Deep-Reinforcement Learning Multiple Access for Heterogeneous Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1277–1290, 2019.

[38] T. Lattimore and C. Szepesvári, "Bandit Algorithms," *Cambridge University Press (preprint)*, 2020.

[39] M. El Soussi, P. Zand, F. Pasveer, and G. Dolmans, "Evaluating the Performance of eMTC and NB-IoT for Smart City Applications," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–7.

[40] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.

[41] B. Hao, Y. Abbasi-Yadkori, Z. Wen, and G. Cheng, "Bootstrapping Upper Confidence Bound," *arXiv preprint arXiv:1906.05247*, 2019.