IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Short-Term Vehicle Traffic Prediction for Terahertz Line-of-Sight Estimation and Optimization in Small Cells

HARBIL ARREGUI[1], ANDONI MUJIKA[1], ESTíBALIZ LOYO[1], GORKA VELEZ[1], MICHAEL T. BARROS[2], and OIHANA OTAEGUI[1]

[1]Intelligent Transport System and Engineering department, Vicomtech, Donostia - San Sebastian, 20009 Spain
[2]Telecommunication Software and Systems Group, Waterford Institute of Technology, Ireland

Corresponding author: Harbil Arregui (e-mail: harregui@vicomtech.org).

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n. 671625, project CogNet)

ABSTRACT Significant efforts have been made and are still being made on short-term traffic prediction methods, specially for highway traffic based on punctual measurements. Literature on predicting the spatial distribution of the traffic in urban intersections is, however, very limited. This work presents a novel data-driven prediction algorithm based on Random Forests regression over spatio-temporal aggregated data of vehicle counts inside a grid. The proposed approach aims to estimate future distribution of V2X traffic demand, providing a valuable input for a dynamic management of radio resources in small cells. Radio Access Networks (RAN) working in the terahertz band and deployed in small cells are expected to meet the high-demanding data rate requirements of connected vehicles. However, terahertz frequency propagation has important limitations in outdoor scenarios, including distance propagation, high absorption coefficients values and low reflection properties. More concretely, in settings such as complex road intersections, dynamic signal blockage and shadowing effects may cause significant power losses and compromise the quality of service for some vehicles. The forthcoming network demand, estimated from the regression algorithm is used to compute the losses expected due to other vehicles potentially located between the transmitter and the receiver. We conclude that our approach, which is designed from a grid-like perspective, outperforms other traffic prediction methods and the combined result of these predictions with a dynamic reflector orientation algorithm, as a use case application, allows reducing the ratio of vehicles that do not receive a minimum signal power.

INDEX TERMS Wireless networks, Vehicular and wireless technologies, Radio access networks, Intelligent transportation systems, Antennas and propagation

I. INTRODUCTION

CURRENT trends in Intelligent Transport Systems (ITS) are making use of the forthcoming full vehicle to everything (V2X) capabilities. New technologies such as 5G are under intensive research, to offer ultra-high data rates (minimum peak data rates of 20 Gbit/s for downlink and 10 Gbit/s for uplink) and energy-efficient communication networks in a near future [1]. Higher throughputs per device are even expected in the future beyond the fifth generation (B5G) mobile networks [2]. These data rates require to operate at carrier frequencies of several gigahertz or more [3].

The terahertz frequency range, which falls at 0.1-10 THz lying between the microwaves and the infrared, is expected to satisfy the data rate requirements of next-generation wireless communication networks [4], [5]. Nevertheless, electromagnetic waves operating at this high spectrum present challenging implementation implications. First and foremost, Line of Sight (LoS) must exist between the transmitter and receiver to guarantee maximum data rate. The reason is that, unlike lower frequencies with higher wavelengths, terahertz signals cannot penetrate most obstacles, regardless of their material type. Secondly, any terahertz signal reflected on a non-smooth surface has connection breaker absorption and dispersion losses. And thirdly, the signal degrades immensely with distance as it propagates through air, given the attenuation caused by its molecular absorption.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2910225, IEEE Access

These challenges have attracted researchers to find new solutions that can be built on current infrastructure in order to facilitate terahertz-spectrum communications. These new solutions will require novel integrated software and hardware systems that complement the current infrastructure, to enable very reliable links transmitting at high data rates within short time windows. Network degradation and delays can potentially impact safety and thus, signal availability must be ensured. However, at the same time, the energy efficiency of the network is key.

In this sense, any opportunity to predict the demand of V2X communication needs seems promising to accordingly provision the network resources dynamically.

With this paper, we aim to describe a concrete use case: employing short term vehicle traffic prediction towards terahertz LoS estimation for orienting a set of reflectors to cover blockage signal areas. Therefore, we present a multidisciplinary three-stage approach focused on V2X communications on complex urban intersections. First, we introduce a novel grid-based short-term prediction method to estimate the forthcoming traffic demand, according to the number of vehicles observed and mapped into a grid. Second, this traffic prediction is transformed into a Non-Line-of-Sight (NLoS) probability map. And third, we describe and evaluate an optimization algorithm for orientable dielectric mirror reflectors, as new infrastructure components, which can be used to redistribute the beam to overcome obstacles between the receiver and the transmitter according to the NLoS probability map.

A. PAPER CONTRIBUTIONS

Our main contribution is the design and evaluation of a data-driven short-term vehicle traffic prediction method for Terahertz LoS estimation. The method infers the forthcoming spatial distribution of the vehicles inside a road intersection. The objectives and contributions of this paper are:

- The development of a small-cell vehicle density model to represent vehicle connectivity demand patterns in complex roundabouts and intersections.
- The adaptation and application of a Random Forest regression algorithm to our vehicle density model representation for short-term prediction and its evaluation under several design parameters.
- For each subcell, a method for the statistical estimation of received signal power at terahertz frequency bands, based on the predicted vehicle density.

These contributions are then completed with a use-case application: how to set up configurable reflector orientations to overcome dynamic blockage and shadowing effects by means of virtual LoS, based on the short-term predictions provided. The approach consists of a scenario with a terahertz-band small-cell station with a set of reflectors in fixed locations, with direct LoS among them when there are no vehicles. The orientation of these reflectors can be changed timely. A system model description is depicted in Fig. 1. Our hypothesis is that applying a data-driven short-term vehicle traffic prediction, we will be able to anticipate the communication needs and dynamically configure reflectors

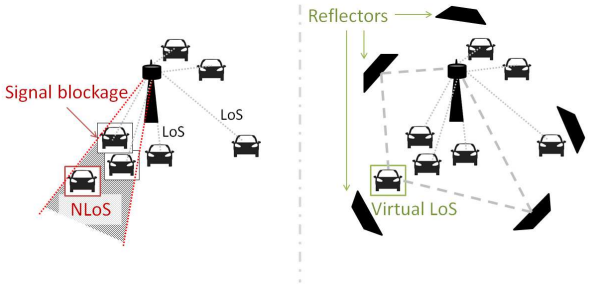


FIGURE 1. System model description. One of the vehicles is in a Non Line of Sight (NLoS) situation (left) due to signal blockage created by other vehicles. This situation is solved by the use of reflectors (right), which create a Virtual LoS.

- to optimize the direction of beams to the areas with the highest demand,
- to reduce shadowed areas without LoS.

In our approach, we make some assumptions:

1) Accurate geopositioning of vehicles is available. For instance, according to [6], sub-meter positioning accuracy with outage probabilities converging to zero could be achieved thanks to cooperative positioning in 5G networks.
2) It is considered that the communication bandwidth demand is proportional to the vehicle volume.
3) The road intersection or roundabout is covered by one small-cell.
4) The number of reflectors and their location is predefined
5) Small-scale channel propagation details, such as delay distributions and queue models, are obviated.

This paper proceeds in II with a description of relevant related research. Section III discusses our methodology. In Section IV, the experimental setup is explained in order to facilitate the reproducibility of the presented research. The main results are given in Section V and they are later discussed in Section VI. To end with, conclusions are summarized in Section VII.

II. RELATED WORK

The Connected Vehicle paradigm have promoted a bloom of research on methodologies that provide several degrees of automated control of all kinds. In fact, [7] showed that the automated control of the vehicle traffic can be done at multiple scales. At the same time, from 5G onwards, networks are expected to strengthen the capabilities of automating the communication network resources. Our paper contributes towards the opportunities to control the communication network at the radio level. To do so, we present a data-driven short-term traffic forecasting approach, aimed at making use of the fine-grained spatio-temporal features of the traffic micro-dynamics inside an urban small cell. This

will help in making decisions for the following few minutes and anticipate to changes in Radio Access Network (RAN) demand.

First, we proceed discussing previous literature on existing short-term traffic forecasting techniques to understand the need of novel methodologies. Then, we address the challenges being studied on terahertz communications, as one of the most promising frequency bands for 5G.

A. SHORT-TERM VEHICLE TRAFFIC PREDICTION

Traffic prediction is still a fruitful research area. From applications such as the design of infrastructures capable of handling the travel demand in an efficient manner, to the management of the resources to improve traffic safety or avoid congestion situations, the use of diverse sensing techniques to measure vehicle movements along the road network is well-known.

One of the deepest literature analysis on short-term traffic forecasting was done by [8]. In this work, authors find that the effort in most previous works has gone in 1) using data from motorways and freeways, 2) employing univariate statistical models, 3) predicting traffic volume or travel time and 4) using data collected from single point sources. Then, they present a list of 10 challenges, related relevant literature and future research directions on each of these challenges. More recently, in the paper published by [9], 130 research papers between 1984 and 2016 were summarized, with a special focus on spatial considerations. We invite the reader to check both review articles for a detailed understanding of the literature evolution.

From the methodological point of view, two big separate approach groups are found in the literature:

- Classical statistical methods, lead by the use of Auto-Regressive Integrated Moving Average (ARIMA) family of parametric models and their variants [10], [11], [12].
- Non-parametric data-driven methods making use of computational intelligent approaches where large datasets are required [13], [14], [15].

In the following, we give a short overview of them.

Inside the first group, the objective of [10] is the detection of outliers, with 15-min aggregated flow series, following a seasonal autoregressive integrated moving average plus generalized autoregressive conditional heteroscedasticity (SARIMA + GARCH) structure. Authors split this structure in three components: a Seasonal Integrated Moving Average, a short term Kalman Filter and a GARCH filter. Authors in [11] present an hybrid approach to explain the different components of the freeway traffic flow, combining a spectral analysis for intra-day periodic trends, an ARIMA model to predict the deterministic component in the residual part, and a GJR-GARCH model for the volatility part of the traffic flow.

Inside the second group, the work conducted by [13] uses 15-minute aggregated data to predict freeway traffic flow, based on a non-parametric method: an enhanced K-Nearest Neighbor (KNN) algorithm with Weighted Euclidean distance metric. Their method provides multiple time step forecasts (one hour and a half in steps of 15 minutes). They compare performance against the methods analyzed in [12], but spatial features are not taken into consideration. Connected Vehicle technology with Artificial Intelligence is used by [14], the generated data is sent to the edge devices (e.g., roadside units) for further processing. Other authors, such as [15], analyze the combination of several primary machine learning algorithms with satisfactory results.

Literature about comparisons between these big groups of methodologies is prolific. For example, the work done by [16] summarizes the bibliography from both approaches, with Neural Network (NN) techniques as the main representatives of the second group. In their work, authors state that the common approach in literature to compare NN performance for time-series prediction is to test their accuracy against classic ARIMA models, but very few provide clear evidence. The work by [8] states that parametric methods perform relatively poorly under unstable traffic conditions and complex road settings. Authors in [17] compare, as well, data-driven models (Neural Networks and Bayesian Networks) with a Seasonal Auto Regressive Moving Average (SARMA) time-series model, for short-term congestion forecast.

The problem stated in our work needs to handle very fast vehicle positioning inputs coming simultaneously with small spatial granularity in Euclidean Space, instead of single point or lane-level measurements. Existing methods in the literature do not provide high spatial-temporal resolution in traffic prediction. Several works have stressed the importance of making use of time-aggregated traffic measurements, even though many systems are able to collect data in short intervals (few seconds), to overcome the strong variability of traffic parameters. We use the concept of *temporal granularity* when referring to this minimal time aggregation units. Authors in [18] use 5 minutes, [19] states the 15-min interval as the best prediction interval and [20] recommended intervals not to be lower than 10 minutes. However, the key limitation of the existing methods for us is that they are all conceived for the domain of transportation planning and mobility where, in general, the basic spatial unit is the traffic link. To our knowledge, no other works have studied a fine-grain spatial traffic distribution inside an intersection for V2X communications optimization.

Suitable traffic representation and prediction methods, like the one described in this work, will undoubtedly impact on the prediction of blockage in wireless networks for connected vehicles. This will open a wide range of possibilities to optimize dedicated links based on geo-location and mobility patterns for adequate beamforming and beamsteering.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2910225, IEEE Access

B. LINK MODEL AND LOS COMPUTATION IN TERAHERTZ COMMUNICATIONS

The most cited technology for providing high data rate satisfying 5G demands is terahertz band communication [4]. This band is also under research for vehicular communications in B5G systems [2]. Proposed models such as [21]–[23], present an in-depth study into the different peculiarities that are uniquely found in the terahertz band. These models, however, have scalability issues as well as other limitations that include multiple propagation effects. Works like [24]–[26] have proposed more complete and advanced models using ray tracing techniques for the band $(0.06 - 1\text{THz})$. Other works like [27], have provided other types of advancements using reflector technology, which is indicated to solve the Non-Line of Sight (NLoS) problem. Also in [28], authors have used directional antennas for the same purpose.

The molecular absorption, refraction, reflection and diffraction losses were quantified in all these works. The budget analysis technique is used in previous studies for both millimeter-wave and terahertz communication and also terahertz [29]. Efforts in terahertz link modeling have been more concentrated towards indoor communications, while, outdoors signal propagation modeling is still being carried more in the millimeter-wave communications frequency [27]. Following the simple regression model analysis obtained from [30], [31], authors in [27] presented the blockage analysis for terahertz communication through the concept of the *probability of LoS*.

Some works have focused on coverage optimization. For example, in [32], authors have used ray-tracing and evolutionary algorithms such as a Genetic Algorithm (GA) and Dynamic Differential Evolution (DDE) to minimize a set of multiple transmitting antennas and determine their location to maximize the received power. But the dynamism of the evolving demand localization and the dynamic blocked zones are not accounted for in any of the mentioned works.

In fact, the problem of determining the time-varying availability of LoS remained unsolved in the literature. In addition, as far as we know, works about optimization in the case of reflector angles in this domain have not been published. Thus, we think that the inclusion of this application as a Use Case to present the full solution is of interest.

III. PROPOSED APPROACH

Location of vehicles inside a given small cell area is mostly obtained with positioning nodes techniques, and 5G includes alternatives with availability of LoS and with NLoS [33]. Some methods are able to use a single station, [34], or they may work cooperatively, as in [6]. Other possible options are the inclusion of additional metadata encoded inside the communication packet headers, or the use of external sensors such as computer vision systems. However, the major challenge still lies in predicting the progression of the traffic over time and identify areas within a given small cell that are likely to have increased signal shadowing effects.

In order to provide a good solution to the above-mentioned challenge, our approach covers the following three stages:

- Vehicle density modeling: to represent the scenario from the spatial subdivision of the scene, its temporal evolution (progression) and the data structures needed to handle vehicle counts.
- Short-term vehicle traffic prediction: to estimate forthcoming demand situations based on recent measures and data-driven models built from historical data.
- Radio coverage and LoS calculation: based on the predicted traffic, to estimate the signal power received at different locations inside the small-cell area. It includes losses due to shadows created by other potential vehicles.

Afterwards, we propose a workflow model to integrate these three stages into a continuous process where control actions can be taken to manage resources anticipating future situations, e.g., adjust the dynamic orientation of a set of reflectors for coverage optimization based on the short-term traffic prediction. Our aim is to design, develop and test this short-term prediction algorithm to estimate the future spatial distribution of the traffic inside a road intersection. We assume that a road crossing or a roundabout can be covered by one terahertz small-cell station. We assume, as well, that the small-cell station receives or calculates the exact location of each of the connected vehicles that are inside its area of action, as they move.

Our approach can be implemented in the network with a cross-layer network management system model that incorporates both the traffic prediction and reflectors' manipulation. A flowchart of the solution is depicted in Fig. 2. It is comprised with two main blocks: the Network Management and Dynamic Reflector Orientation. In the first block, measurements of the vehicles temporal evolution are transformed into a vehicle density model (the concepts of sn and asn, shown in the diagram are formally defined in III-A). Then the short-term traffic prediction is obtained, which will lead to the computation of coverage and line-of-sight spots. These results are then sent to the second block. In our use case, this information is converted into an optimal set of reflector angles through a genetic algorithm optimization process. The reflector actuator system is responsible for sending the new angle configuration for each reflector.

A. VEHICLE DENSITY MODELING

To model the spatial and temporal variations, the road intersection space is discretized as a cartesian grid, so that vehicle counts can be spatially aggregated by smaller subcells (Fig. 3). The following variables represent the local small-cell area in a 2-dimensional Euclidean Space:

- L_X: width (in meters) of each subcell
- L_Y: height (in meters) of each subcell
- N_X: number of subcells in X axis
- N_Y: number of subcells in Y axis

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2910225, IEEE Access

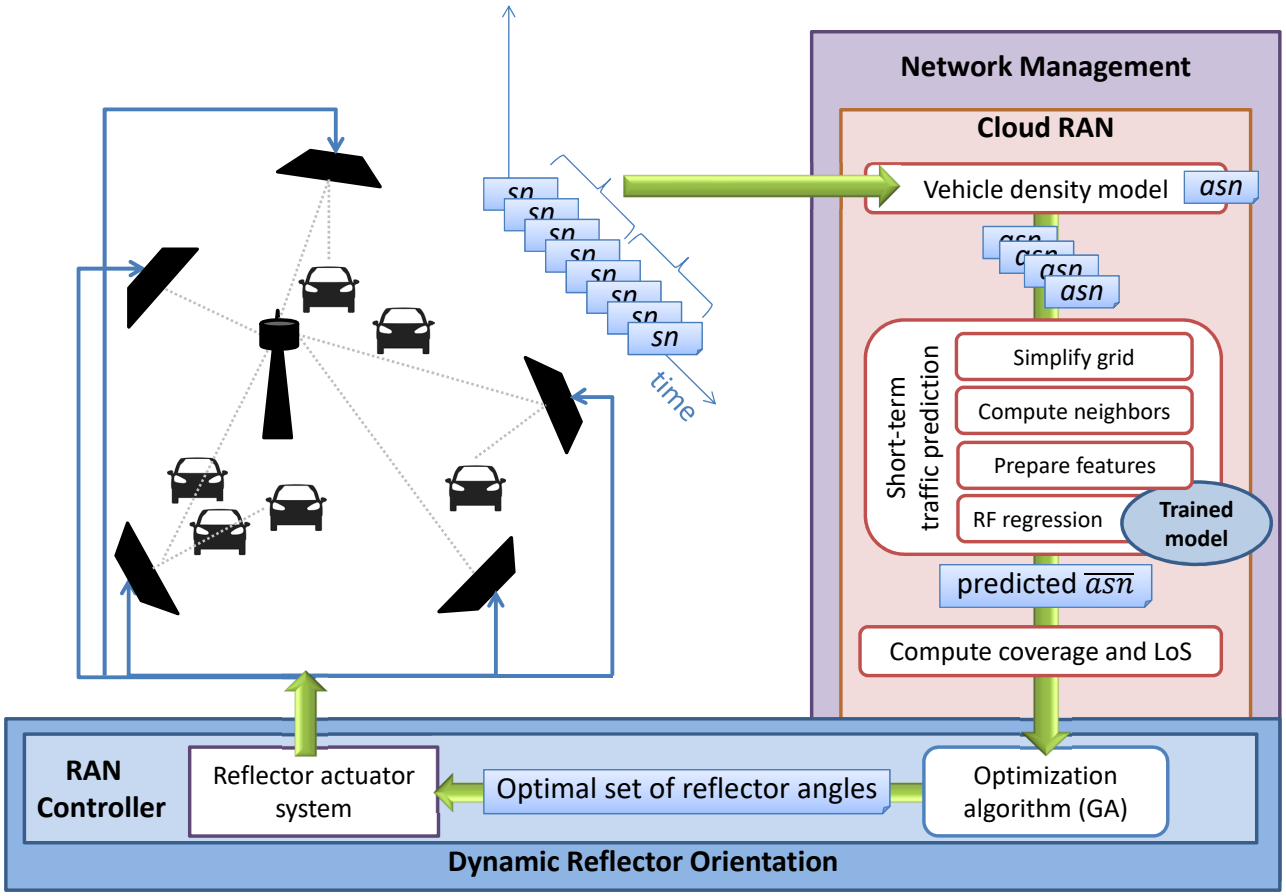


FIGURE 2. Flow diagram of the cross-layer network management system model. It is comprised with two main blocks: the Network Management and Dynamic Reflector Orientation.

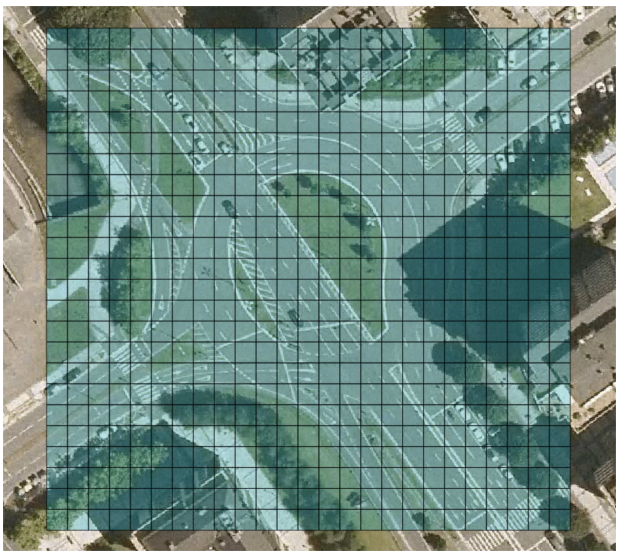


FIGURE 3. Example of grid over a roundabout in an urban scenario. Subcell size fits 1 or 2 cars maximum.

Thus, the small-cell area, which covers $(N_X L_X) \times (N_Y L_Y)\ m^2$, can be represented as a $N_X \times N_Y$ matrix of subcells. Each subcell can then be identified by a pair (x, y), where $x = 1, 2, ...N_X$ and $y = 1, 2, ...N_Y$. Each vehicle is matched into a single subcell at each snapshot; the one where the vehicle occupies the biggest area. That is, if a vehicle is located at a boundary among different subcells, it is matched to the subcell that hosts the largest part of the vehicle surface inside. Depending on the size of the subcells, more than one vehicle can be assigned to the same subcell. As a statistical approach, and for further aggregations, this is valid enough.

When describing the method, we prefer not to state fixed subcell sizes, because we believe that can be a further design decision, depending on the size and geometry of the intersection lanes. Moreover, a compromise between prediction (and coverage optimization) accuracy, as well as computation times must be ensured. We suggest sizes where 1-2 average-sized vehicles could be accommodated. Specifically, this allows making a good statistical representation of vehicles that are stopped at the intersection entries and moving along internal lanes. We think that this size meets a balance because of its relatively small spatial grain.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2910225, IEEE Access

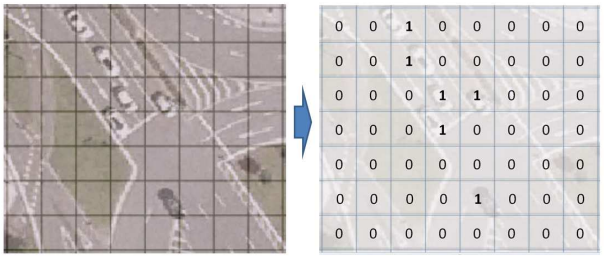


FIGURE 4. Vehicle positions observed in a snapshot, matched to grid subcells. If a vehicle is located at a boundary among different subcells, it is asigned to the subcell that hosts the largest part of the vehicle area inside.

A time-level discretization is also needed: the *snapshot*. This snapshot is understood as the current set of vehicle locations at a given time step, mapped into the grid (shown in Fig. 4).

In real applications, communication events from node vehicles are expected to be asynchronous, but the snapshot concept remains valid to represent the events occurred between two consequent time instants. Therefore, we define t_{step} as time sampling resolution and T_a as duration of a snapshot aggregation. The reason to aggregate snapshots is to obtain more stable measurements from a highly dynamic movement of vehicles. If t_{step} is short enough, the presence of the vehicle in all the subsequent subcells can be observed along the trajectory, at least, as it traverses the intersection. To ensure this, for $speed_{max}$ as the maximum vehicle speed inside the intersection, we can set a conservative threshold $t_{step} < min(L_X, L_Y) \cdot speed_{max}$.

From each observation or simulation run of length T, we obtain: a total number of T/t_{step} snapshots, $\{sn_i\}_{i=1}^{T/t_{step}}$, where each snapshot sn_i is a $N_X \times N_Y$ matrix representing the number of vehicles in each of the subcells at that instant; a total number of T/T_a aggregated snapshots, $\{asn_t\}_{t=1}^{T/T_a}$, where each one is a $N_X \times N_Y$ matrix representing the sum aggregation of vehicles in each of the subcells during the last aggregation period. Thus, we represent vehicle density by asn_t:

$$asn_t = \sum_{i=1+(t-1)(T_a/t_{step})}^{t(T_a/t_{step})} sn_i \qquad (1)$$

where $1 <= t <= T_a/t_{step}$.

Additional geometric alternatives: When designing a grid-like subdivision for our vehicle density modeling, we would also like to consider some other alternatives. For instance, we open the possibility where the vehicle density model relies on circular or hexagonal patterns instead of square grids. We see this approach specially promising in circular-shaped intersections such as roundabouts. Nevertheless, matching vehicle positions onto these alternative representations is computationally more complex.

B. SHORT-TERM VEHICLE TRAFFIC PREDICTION

The ordered sequence of snapshots and aggregated snapshots represent time-series of traffic volume at each subcell.

Given the current time t, having observed the previous pr aggregated snapshots, we want to predict a future aggregated snapshot with a prediction horizon of value p. Therefore, we use the measures from prior times $(asn_t, asn_{t-1}, ..., asn_{t-pr})$ as input variables. This can be formulated as a regression problem, where the output variable is asn_{t+p}.

Each asn element is n-dimensional, with $n = N_X \times N_Y$, since each subcell is processed in parallel as a single regression problem.

As presented in Section II, one of the most recent approaches for short-term traffic prediction, [13], makes uses of an enhanced KNN algorithm. Authors compare their proposal successfully to other alternatives in literature. Using their algorithm to apply into our problem, would have been a valid first intuition. However, this presents some inconveniences:

- KNN methods are *lazy learners* since they do not create any model, and at every inference, they need to look for the K nearest neighbors in the full training search space.
- KNN is hard to parallelize
- In our setting, multiple subcells in the grid need to be analyzed at the same inference process

A balance between accuracy and a fast response needs to be accounted when approaching a solution for vehicular communications. In this sense, a kNN algorithm-based inference may not be the most efficient one regarding computational requirements for the long term. Therefore, we propose to evaluate and compare its accuracy and computation times, with another kind of machine learning technique. To do so, we develop a new prediction alternative, based on Random Forest Regression [35], which is easy to parallelize. Random Forests, ensemble of decision trees, are able to handle thousands of input variables without variable selection, since different random subsets of features are chosen to split on at each tree node. They are good at generalizing with new input data if they fall inside a known data range. Two main parameters must be tuned: the number of trees, and the maximum depth.

The method consists of the following steps:

Grid simplification: The resolution of the grid compromises the speed of the calculation. But the topology of the road confines the greatest amount of vehicles only to some subcells. We have defined a percentile threshold th_{limit}, which represents the minimum number of total vehicle observations in a subcell for further consideration inside the algorithm; otherwise, we discard it. Consequently, we define a simplification ratio $0 \leq \gamma \leq 1$. The objective is to improve the computation costs with very little impact in accuracy. Another reason to consider grid simplification is that the function of each subcell inside the grid might be different.

While some mainly represent traffic that is approaching the intersection, others represent the main connection nodes inside the intersection, which are affected more by the effects of the signalling. For example, when a traffic light is red, or the congestion level is high, subcells at these locations will observe high vehicle densities. At the same time, when a traffic light turns green, the subcell may be released immediately descending the vehicle density in the corresponding asn.

Random Forest Regression: Random Forest (RF) regression has been used for its simplicity and ease of configuration. RF has been proved as a strong machine learning technique that provides good accuracy. Forests are ensembles of decision trees, making use of *bagging*, i.e., bootstrapped aggregation, which means that each tree is trained with random subsets of the input training records, and outputs are aggregated. The outputs of the different decision trees are averaged to obtain the final prediction value. The main difference between RF and other decision tree ensembles is that, in RF, the features considered when building each tree node are also selected at random.

Our regression model needs to fit the output value of $asn_{xy,t+p}$ based on a set of $n_{feat} = pr + 1$ features:

$$asn_{xy,t+p} = f(asn_{xy,t}, asn_{xy,t-1}, ..., asn_{xy,t-pr}) \quad (2)$$

Neighbor-order (Π): We propose a cell neighbor condition to find the relationship between adjacent subcells. The neighbor condition can be measured by the Neighbor-order parameter. A pair (x', y'), where $x' = 1, 2, ...N_X$ and $y' = 1, 2, ...N_Y$, is said a Π-order neighbor of pair (x, y), if and only if the following two conditions are met:

$$\begin{aligned} |x' - x| &\leq \Pi \\ |y' - y| &\leq \Pi \end{aligned} \quad (3)$$

The value of Π must be chosen with caution; in general, the higher the value, the higher is the increase in the dimensionality of the input space leading to higher variances in the output of the learning algorithm. However, one of the benefits of using RF is that its random feature selection makes it robust to handle high dimensional datasets efficiently. Computationally, using neighbors is a costly operation for model training, since it increases significantly the number of observed features used by the regression algorithm: $n_{feat} = (pr + 1)(2\Pi + 1)^2$.

Computational complexity analysis: A RF regressor is based on a combination of multiple decision trees. With n_{train} number of training examples, n_{feat} number of features, num_{tree} trees and a fixed maximum tree depth max_{depth}, the computational complexity of training can be represented as $\mathcal{O}(n_{train} \cdot n_{feat} \cdot max_{depth} \cdot num_{tree})$. For the full cell, since we are now able to discard some subcells due to grid simplification, we get: $\mathcal{O}(\gamma \cdot N_x \cdot N_y \cdot n_{train} \cdot n_{feat} \cdot max_{depth} \cdot num_{tree})$.

When predicting, a binary decision needs to be done in each tree node, and thus, the computational complexity is reduced to $\mathcal{O}(\gamma \cdot N_x \cdot N_y \cdot max_{depth} \cdot num_{tree})$.

C. RADIO COVERAGE AND LOS COMPUTATION

Communications in connected vehicles can be bi-directional but, for simplicity, we take the small-cell station as the transmitter, and the vehicle node as the receiver.

In our procedure, first, we need to compute the probability of finding an obstacle (a vehicle) in each subcell. Then, following a link-budget analysis based on a 2.5D ray-tracing model [36], we analyze the power received at a given subcell. We say that we use a 2.5D model because we consider some heights, without building a full 3D model. For the link-budget calculation, we must include Non-Line of Sight losses, which depend on the probabilities we have computed. Each step is described as follows:

Obstacle Finding Probability: The probability P_{o_k} represents the probability of finding at least one vehicle (obstacle) in a subcell at a certain instant. We propose a method to compute this value using the vehicle density prediction, which provides the number of vehicle observations expected in that subcell during one aggregated snapshot. So, being c the capacity of a subcell, i.e., the number of average-sized vehicles that can occupy a subcell at the same time due to its size, $m = \overline{asn_{t,xy}}$ the prediction obtained for the subcell at aggregation period t and $s = T_a/t_{step}$ the number of time steps in the aggregated snapshot, then:

$$P_{o_k} = \begin{cases} 1 - \frac{\binom{c\cdot(s-1)}{m}}{\binom{c\cdot s}{m}}, & m \leq c \cdot (s-1) \\ 1, & m > c \cdot (s-1) \end{cases} \quad (4)$$

where $\binom{A}{B}$ denotes a binomial coefficient.

Direct Coverage computation: The power received directly from the transmitter in a given subcell (x, y) is represented as $R_{tr,xy}$, and computed with the following formula [27]:

$$R_{tr,xy} = 141.7 - \alpha(f, r) \quad (5)$$

where f is the frequency of the signal, r is the distance from the transmitter to the center of the corresponding subcell and $\alpha(f, r)$ is the LoS path-loss.

The path-loss computation takes into account the spreading, the molecular absorption, reflections and the scattering. To do so, we specifically want to compute the effect of the obstacles that interfere with the LoS between the transmitter and the receiver (Fig. 5). Thus, NLoS losses substitute scattering and reflection losses. Thus, the path-loss is:

$$\alpha(f, r) = 10 \log_{10} \Psi(f, r) + 10 \log_{10} \beta(f, r) + L_{NLoS} \quad (6)$$

where $\Psi(f, r) = \left(\frac{c}{4\pi f r}\right)^2$ and $\beta(f, r) = e^{-\frac{1}{2}k(f)r}$ are quantifications of losses due to spreading and molecular absorption respectively, with c being the speed of light in vacuum and $k(f)$ being the molecular absorption itself [27].

Non-Line of Sight losses: L_{NLoS} is the loss due to other cars located between the transmitter and the receiver. L_{NLoS} is computed with the cascaded knife edge method [37], where

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2910225, IEEE Access

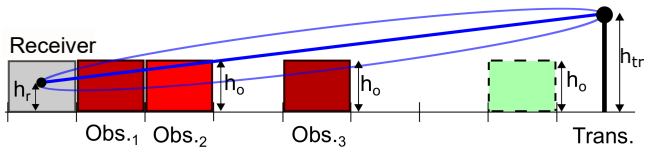


FIGURE 5. Direct LoS is obstructed by several obstacles (vehicles) found between transmitter and the receiver. These obstacles are generalized to the size of a subcell and a height h_o. We calculate the shadowing weight of each obstacle at each time, based on the estimated vehicle density ($\overline{asn}_{xy,t} > 0$), but only the three nearest obstacles from the receiver are considered (shaded in red). The subcell closest to the transmitter (in dashed lines), also expects a positive density, but its effect is neglected in this method.

losses due to up to the nearest three obstacles, $\{J(v_{o_k})\}_{k=1}^{3}$, are summed in the following way:

$$L_{NLoS} = \begin{cases} J(v_{o_2}) + T(J(v_{o_1}) + J(v_{o_3}) + C), & v_{o_2} > -0.78 \\ 0, & v_{o_2} \leq -0.78 \end{cases} \quad (7)$$

where $T = 1 + e^{-\frac{J(v_{o_2})}{6}}$, $C = 10 + 0.04r$ and v_{o_k} are dimensionless parameters dependent on the heights of obstacles, h_o, transmitter, h_{tr}, and receiver, h_r and distances among them [37]. We refer readers to that work for further details on how v_{o_k} is computed. Note that we kept the naming T here to follow the cited reference, but it must not be mixed with the T observation run time, previously defined in our vehicle density model.

The loss due to each obstacle v_{o_k}, is computed as follows:

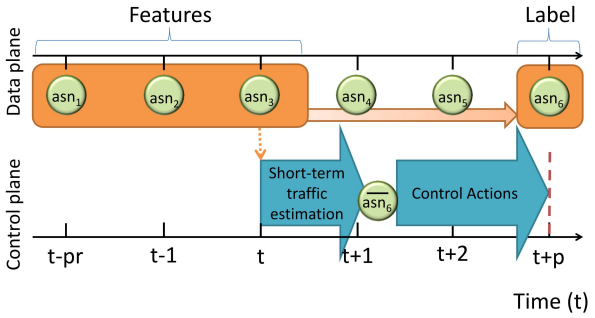
$$J(v_{o_k}) = P_{o_k} \cdot \left(6.9 + 20\log_{10}\left(\sqrt{(v_{o_k} - 0.1)^2 + 1} + v_{o_k} - 0.1\right)\right) \quad (8)$$

D. WORKFLOW MODEL

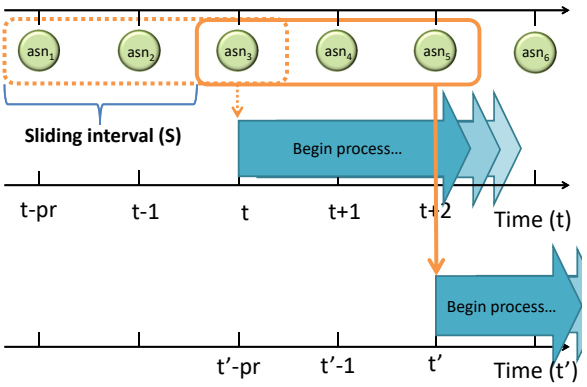
The overall workflow of our approach is represented in Fig. 6. The aggregated snapshots are processed in a streaming way. A sliding interval of size S, represents how often control actions are expected to be taken. As explained previously, at time instant t, current asn and pr previous asn matrices are processed to obtain the $t + p$-th element. The sliding interval, represents the next t' value when the process will be repeated: $t' = t + S$. The prediction horizon p is key, since it sets the available time to make the computations for the traffic estimation and perform any control actions designed to improve the efficiency of the RAN based on that estimation, e.g., a reconfiguration of a set of reflectors using an optimization algorithm. An important factor that needs to be considered when dimensioning the parameters is that if $S < p$, more than one process will need to be executed in parallel.

E. APPLICATION USE CASE: ORIENTING REFLECTORS FOR OPTIMIZING ANTENNA COVERAGE

The use-case application of the traffic prediction presented in our work is the optimization of a terahertz small-cell antenna coverage, by dynamically adapting the orientation



(a) Single process



(b) Sliding window process

FIGURE 6. asn matrices are processed in stream. Prediction horizon p limits the available time for the short-term prediction and any control actions that need to be taken (a). Computation can be performed upon arrival of each aggregated snapshot matrix, but this is likely to be very inefficient. In this example (b), a sliding interval of $S = 2$ is shown.

of a set of reflectors. The proposed approach includes a set of J reflectors, whose locations are fixed. The actual selection of the number of reflectors, as well as the selection of locations for the small-cell antenna and the reflectors, are out of the scope of this work, but we find that building a model for a reasonable geometric deployment based on the intersection's physical characteristics could be of interest. However, to install them at a suitable height over the traffic, a cost-effective alternative is to reuse existing traffic lights, streetlight poles and similar infrastructure.

The orientation of the reflectors is variable, represented by ω_j parameter. The angle of each reflector is independent from others. Due to the fluctuation of the number of vehicles tracing different trajectories, the wireless signal demand and signal shadows may vary with time from one location to another inside the same cell. We seek to optimize the set of ω_j values to ensure the best coverage according to the estimated vehicle flow: maximize the received power by each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2910225, IEEE Access

vehicle, and ensure a minimum power for most vehicles. Consequently, we use the outcome of the RF predictions to feed a reflector angle orientation optimization process based on a Genetic Algorithm (GA). GAs are well-known heuristic search and optimization tools [38].

Before detailing the algorithm, we need to define the effects of the use of reflectors, and the objective function that the GA will try to optimize:

Reflector Coverage computation: As long as the subcell is in the scope of the reflectors with ω_j orientation, to compute the total received power in subcell (x, y), $R_{tot,xy}$, we have to add the power received directly from the transmitter, $R_{tr,xy}$, and the power that comes from each of the J installed reflectors:

$$R_{tot,xy} = R_{tr,xy} + \sum_{j=1}^{J} R_{j,xy} \tag{9}$$

$R_{j,xy}$ is computed the same way as $R_{tr,xy}$ but we need to replace distance r in Eq. 6 with the sum of the distance from the transmitter to the reflector, r_{tr}, and the distance from the reflector to the receiver, r_{rr}. Consequently, L_{NLoS} must be computed using the correspondent sum of distances (Eq. 7).

$$\begin{aligned} R_{j,xy} = 141.7 - (10\log_{10}\Psi(f, r_{tr} + r_{rr}) + \\ +10\log_{10}\beta(f, r_{tr} + r_{rr}) + L_{NLoS}) \end{aligned} \tag{10}$$

Objective Function: Once we know how to compute the power in each subcell, $R_{tot,xy}$, for a given set of ω_j, the aim is to optimize the coverage of the whole scene. For that, we propose an objective function that consists of the mean power received by each vehicle:

$$\overline{R} = \frac{\sum_{x=1}^{N_X}\sum_{y=1}^{N_Y} R^*_{tot,xy} asn_{t,xy}}{\sum_{x=1}^{N_X}\sum_{y=1}^{N_Y} asn_{t,xy}} \tag{11}$$

Note that $asn_{t,xy}$ is not available in the optimization process and therefore $\overline{asn_{t,xy}}$ is used. By using the mean $\overline{R}$ value, there is a risk of orientating all the signal strength to the areas with high vehicle density and let areas with low vehicle density without any signal coverage. Therefore, we define $R^*_{tot,xy}$, as the power received in that subcell, but truncated with a threshold, TH, and penalized with a given value, P:

$$R^*_{tot,xy} = \begin{cases} R_{tot,xy}, & R_{tot,xy} > TH \\ P \cdot (R_{tot,xy} - TH), & R_{tot,xy} \leq TH \end{cases} \tag{12}$$

With this penalization, we procure a minimum power of TH dBm while the maximum coverage is sought.

Genetic Algorithm: Each individual in the GA encodes a configuration of reflectors and their chromosomes are the orientation angles of reflectors themselves, $\{\omega_1, \omega_2, ..., \omega_J\}$. The GA must be initialized: the first population can be generated by assigning chromosomes for all individuals randomly or by inserting a specific individual. E.g., as a starting point, a predefined configuration that obtains good coverage in different scenarios can be used. If we consider that the traffic will not change considerably in the following minutes, the current configuration of reflectors can also be used to initialize one of the individuals of the population. Two individuals are recombined by taking half of the angles by random from one of the parents and taking the remaining half of the angles from the other parent. Mutation increases or decreases the value of the 25% of the angles, at random.

The benefit of the GA is that it always retains a feasible result and the available time for the control actions can be dedicated to the optimization of generations. We propose the use of average vehicle densities to create the base populations and to use them as initializations. Then, at each new t', previously calculated populations are used.

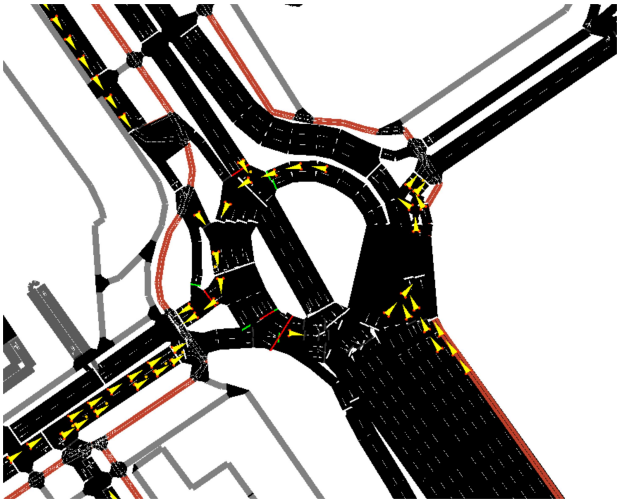


FIGURE 7. A screenshot extracted from SUMO [39] during simulation run. The small triangles represent vehicles.

IV. EXPERIMENTAL SETUP

We have adopted a vehicle traffic micro-simulation approach to generate the scenarios needed to test and validate the proposed methodology. In our work, we use the Simulation of Urban MObility (SUMO) [39] tool to generate datasets of vehicle traffic following different sets of flow demand. SUMO is purely microscopic and it performs a time-discrete stochastic car-following model simulation with a default step length of 1 second. In our setup, the time step has been set to $t_{step} = 0.1$ seconds and the total length of each simulation run has been set to $T = 3600$ seconds, thus a total of 36000 snapshots were generated for each simulation run. A screenshot of the scenario, extracted from a simulation run, can be seen in Fig. 7.

The network has been imported from OpenStreetMap (OSM) XML files [40], and prepared making use of the NETCONVERT tool, with special settings in order to import traffic light logics into the SUMO-routable network file. Trips and routes have been randomly generated using RANDOMTRIPS SUMO tool, tuning the repetition-rate, to enable two different setups of demand patterns and repeating each pattern 5 times, obtaining 10 hours of traffic simulation. (see Fig. 8).

In the proposed scenario the small-cell area has been

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2910225, IEEE Access

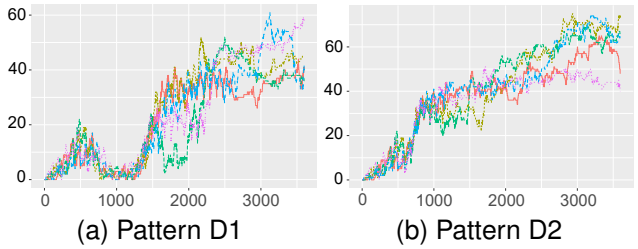


FIGURE 8. Evolution of the total vehicle volume time series in each 60-minute length simulation run for demand pattern D1 (a) and D2 (b). Five simulations for each demand, cover a total duration of 10 simulation hours.

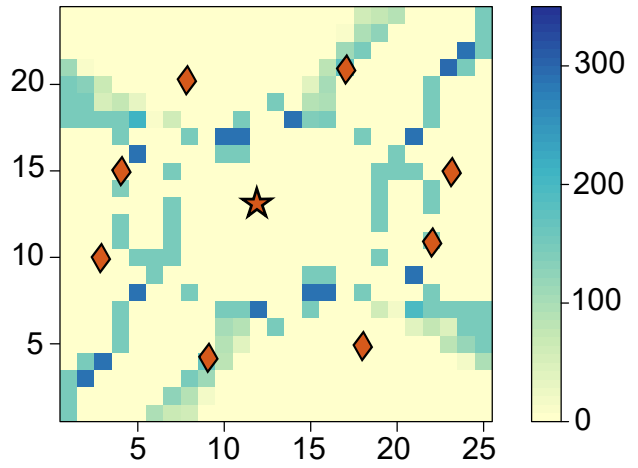


FIGURE 9. Vehicle density (asn) in each subcell during a 30 second aggregation of snapshots. Darker subcells represent a higher vehicle density; generally, they are located at the surroundings of the traffic lights and lane junctions. These subcells also suffer from higher variance. This is due to the number of vehicles that wait stopped at the same subcell for several seconds, and then they are released to continue the flow. In this figure, we include the location of the transmitter (star symbol) and the reflectors (rhomboids) used in our experimental setup.

divided in $N_x = 25$, $N_y = 24$, thus 600 subcells of $L_x = 5$ and $L_y = 5$ meters. The transmitter is set at the center of the roundabout, at a fixed height from the ground of $h_{tr} = 2$ meters.

Reflector system scenario parameters: To validate the approach, we design a reflector system scenario with a set of concrete experimental values which we proceed to describe. A total of eight reflectors with fixed locations are evenly distributed. Their location has been chosen in an arbitrary mode, following an intuitive deployment to cover the area efficiently. Fig. 9 shows the actual locations of the transmitter and all the reflectors in the roundabout scenario. The grid represents the spatial subdivision of the small cell, while the color in each subcell represents the variance of the aggregated snapshot. We choose to work at $f = 0.3$ THz frequency and we make use of the same transmitter power as [27]. To perform the measurements we have set the following parameters: the height of vehicles, and thus the height of the obstacles, is set to $h_o = 1.5$ meters [41] while the receiver is located at $h_r = 1$ meter high. Additionally, and since the size of the subcells is 5×5, we set the capacity to $c = 2$ vehicles. Regarding the GA parameters, the population has 20 individuals and 2000 generations are computed before stopping and taking the best reflector configuration so far. In each generation, two new individuals are created. A total of four angles are taken from a parent and the four remaining angles are taken from another parent to create the first individual. The second one is created the other way round. Mutation changes the orientation of two out of eight reflectors randomly.

We set the threshold $TH = 35$ dBm as the minimum amount of received power sought, and moreover, we have set an arbitrary penalization value of $P = 30$ for testing purposes. The selection of the most suitable set of parameters was out of the scope of this work.

Three different reflector configurations have been tested: no reflectors (NO-REF), a static *naive* approach with static reflectors orientated towards the most occupated area in couples (ST-REF) and the solution provided by the GA (GA-REF). In ST-REF, we orient reflectors to the inbound-outbound roads, since the vehicles in these roads are prone to suffer NLoS coverage due to the vehicles inside the roundabout.

V. RESULTS

A. PREDICTION ACCURACY

To assess the goodness of fit of our prediction algorithm, we choose the Root Mean Square Error (RMSE, Eq. 13) and the Mean Absolute Percentage Error (MAPE, Eq. 14). MAPE is unit-less and insensitive to changes in the magnitude of the forecasts (unlike, e.g., the RMSE). The only negative aspect of using MAPE is that its value is not defined when the real observed value is 0. However, it allows us making comparisons between aggregated snapshots of different T_a:

$$\text{RMSE}_{xy} = \sqrt{\sum_{i=1}^{N} \frac{(asn_{i,xy} - \overline{asn_{i,xy}})^2}{N}} \tag{13}$$

$$\text{MAPE}_{xy} = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{|asn_{i,xy} - \overline{asn_{i,xy}}|}{asn_{i,xy}} \right) \cdot 100 \tag{14}$$

where $asn_{i,xy}$ is the observed value in the x, y subcell, $\overline{asn_{i,xy}}$ is the forecast value, and N is the number of observations.

From the training data obtained according to the experimental setup, 75% of the dataset was used for training and 25% for testing, at random. The main results are presented in Table 1. By default, all tests are performed using $T_a = 30s$, $pr = 15$ and $p = 2$. It is relevant to mention that results are given for equal-sized groups of traffic volume ranges. In fact, different performance is observed according to the volume group. Columns named RMSE and MAPE represent the corresponding averaged error values. With the objective of validating the results, we have compared these errors against the ones obtained by a *naive* prediction (RMSE-naive and MAPE-naive), which consists on choosing the last observed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2910225, IEEE Access

TABLE 1. Average RMSE and MAPE errors obtained by the RF regression algorithm according to model parameters (number of trees and maximum depth) and volume groups. Comparison with the errors obtained by the *naive* approach. The improvement percentage value represents the ratio of subcells where the RF algorithm outperforms the *naive* algorithm. Average computation time for each subcell prediction.

Volume	numTrees - maxDepth	RMSE	RMSE-naive	MAPE	MAPE-naive	% improvement	time (ms)
Grp.1	100 - 15	7.01	17.37	NA	NA	30.3%	0.26
	250 - 15	6.97	17.30	NA	NA	30.0%	1.94
	500 - 15	6.98	17.49	NA	NA	30.0%	3.02
	750 - 15	6.38	16.79	NA	NA	30.4%	4.11
Grp.2	100 - 15	7.19	21.31	320.28	266.46	54.3%	0.30
	250 - 15	7.08	19.76	312.45	246.23	54.4%	1.97
	500 - 15	7.13	19.94	316.19	256.45	54.3%	3.05
	750 - 15	6.52	18.75	290.54	232.16	55.1%	4.18
Grp.3	100 - 15	8.07	21.95	160.56	162.10	60.7%	0.39
	250 - 15	7.88	21.24	156.36	157.15	60.6%	2.05
	500 - 15	7.94	22.14	159.16	162.24	61.6%	3.26
	750 - 15	7.75	20.60	154.18	154.00	60.8%	4.47
Grp.4	100 - 15	13.48	36.26	96.69	137.50	70.6%	0.50
	250 - 15	13.18	34.47	94.38	130.27	70.0%	2.13
	500 - 15	13.13	34.51	94.25	132.39	70.1%	3.56
	750 - 15	13.05	34.53	93.09	129.23	70.3%	4.88
Grp.5	100 - 15	28.59	56.73	35.02	54.08	68.8%	0.64
	250 - 15	28.52	56.59	34.79	53.72	69.4%	2.26
	500 - 15	28.42	56.82	34.31	53.11	68.9%	3.76
	750 - 15	28.48	56.06	34.46	52.53	68.0%	5.19
Grp.6	100 - 15	33.71	61.65	19.06	20.08	42.1%	0.71
	250 - 15	33.12	60.40	18.80	19.50	41.6%	2.38
	500 - 15	33.16	59.77	18.84	19.38	41.7%	3.90
	750 - 15	32.36	59.16	18.32	18.93	41.5%	5.41
Grp.1	100 - 20	7.01	17.64	NA	NA	NA	0.49
	250 - 20	6.97	17.36	NA	NA	NA	2.25
	500 - 20	6.96	17.56	NA	NA	30.0%	3.75
	750 - 20	6.99	17.82	NA	NA	29.7%	4.99
Grp.2	100 - 20	7.13	19.82	317.00	257.79	NA	0.51
	250 - 20	7.08	20.81	315.90	261.02	53.7%	2.28
	500 - 20	6.98	20.23	309.60	256.67	54.7%	3.79
	750 - 20	7.10	19.38	314.36	245.45	54.8%	5.05
Grp.3	100 - 20	8.41	23.59	166.16	171.40	60.9%	0.62
	250 - 20	7.91	21.94	156.85	165.13	62.9%	2.42
	500 - 20	7.85	22.36	155.24	165.85	63.0%	4.02
	750 - 20	8.04	22.02	159.62	162.56	NA	5.38
Grp.4	100 - 20	13.34	35.23	95.15	132.80	70.3%	0.79
	250 - 20	13.31	34.56	93.43	127.78	70.0%	2.64
	500 - 20	13.24	34.51	94.63	131.47	71.5%	4.33
	750 - 20	13.51	35.21	96.57	136.84	71.2%	5.73
Grp.5	100 - 20	28.80	57.32	34.83	53.38	67.7%	0.92
	250 - 20	28.78	56.73	35.11	53.66	68.7%	2.92
	500 - 20	28.70	56.98	35.15	53.23	68.2%	4.56
	750 - 20	28.51	55.89	34.88	53.20	69.0%	6.07
Grp.6	100 - 20	33.74	61.59	19.10	20.05	42.6%	0.95
	250 - 20	33.63	61.05	18.87	19.75	41.4%	3.09
	500 - 20	33.75	61.72	19.01	19.94	41.5%	4.74
	750 - 20	33.95	62.40	19.16	20.29	42.1%	6.20

value. The column named *% improvement* represents the ratio of subcells where the RMSE error of the RF algorithm is lower than RMSE-naive. In addition, the computation time is also given. NA values occur when MAPE is undefined, due to 0-values.

To strengthen the validity of our approach, comparing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2910225, IEEE Access

TABLE 2. Average RMSE and MAPE errors obtained by the enhanced KNN algorithm, for three different values of K. Average computation time for each subcell prediction.

Volume	K	RMSE	MAPE	time (ms)
Grp.1	5	7.17	NA	33.97
	10	7.34	NA	33.33
	15	7.03	NA	33.60
Grp.2	5	7.73	342.73	34.22
	10	7.59	338.12	33.69
	15	7.29	321.69	34.07
Grp.3	5	8.47	166.33	33.58
	10	8.07	159.32	33.17
	15	7.62	150.75	33.34
Grp.4	5	15.48	106.12	33.84
	10	14.64	102.40	33.50
	15	14.10	99.13	33.08
Grp.5	5	33.38	42.77	33.13
	10	33.73	41.75	32.81
	15	32.82	40.91	32.87
Grp.6	5	38.36	21.35	31.16
	10	38.34	21.31	30.77
	15	37.98	21.18	30.60

against the state of the art, tests with the enhanced-KNN algorithm [13] are also presented in Table 2. From the different parameters compared, the best result is obtained with $K = 15$.

The impact of adding one neighbor into the RF regression features is presented in Table 3. If we compare these results with the ones in Table 1, there is a subtle overall improvement in the accuracy of the prediction RMSE and MAPE values.

If we take the average *naive* predictions, and set its average RMSE values for each volume groups as a reference value 1, we can make a simple comparison chart, depicted in Fig. 10, representing the improvement ratio obtained by each algorithm: enhanced KNN (KNN), Random Forest (RF), and Random Forest with $\Pi = 1$ (RF1). The best improvement is provided by the RF1, and the worst (except for Grp.3) is provided by the KNN algorithm. All three perform best at volume groups Grp.2, Grp.3 and Grp.4. One of the conclusions that we get is that for extreme values (such as Grp.5 and Grp.6, which represent the highest volume groups), results are worse for the three algorithms. The main reason to explain this behavior is that, since the number of elements in all groups is equal, for these extreme values, the volume range is wider. Therefore, prediction accuracy is worse. But, in addition, what we demonstrated with these results is, that for these extreme cases, RF generalizes better than KNN.

The average results of the RF short-term traffic predictions have also been evaluated for different prediction horizons, ranging from $p = 2$ to $p = 6$, which are summarized in Fig. 11. As shown, the larger the prediction horizon, the larger is the error.

The RF regression algorithm obtains better results than the naive approach on most cases for subcells where the

TABLE 3. Average RMSE and MAPE errors, as well as percentage of subcells with improved predictions, obtained by the RF algorithm with $\Pi = 1$, $maxDepth = 15$

Volume	numtree	RMSE	MAPE	impr.(%)	time (ms)
Grp.1	100	6.66	NA	31.1%	2.10
	250	6.82	NA	30.9%	5.50
	500	6.80	NA	30.9%	7.42
	750	6.91	NA	30.5%	9.12
Grp.2	100	6.92	308.55	55.2%	2.16
	250	7.00	310.75	55.5%	5.65
	500	7.30	323.55	54.8%	7.63
	750	7.14	316.97	55.3%	9.42
Grp.3	100	7.58	149.44	64.7%	2.35
	250	7.53	149.87	63.8%	6.17
	500	7.64	151.45	63.8%	7.95
	750	7.59	151.79	63.5%	9.83
Grp.4	100	12.80	90.42	70.6%	2.64
	250	12.89	91.63	70.7%	6.76
	500	13.04	91.59	72.5%	8.71
	750	12.89	91.59	72.2%	10.58
Grp.5	100	27.45	33.25	67.9%	2.95
	250	27.39	33.68	69.1%	7.60
	500	27.22	33.26	69.0%	9.88
	750	27.14	33.42	68.8%	11.80
Grp.6	100	32.80	18.32	39.9%	3.09
	250	32.26	17.97	39.7%	7.51
	500	31.90	17.89	39.5%	10.00
	750	32.00	17.94	39.7%	12.05

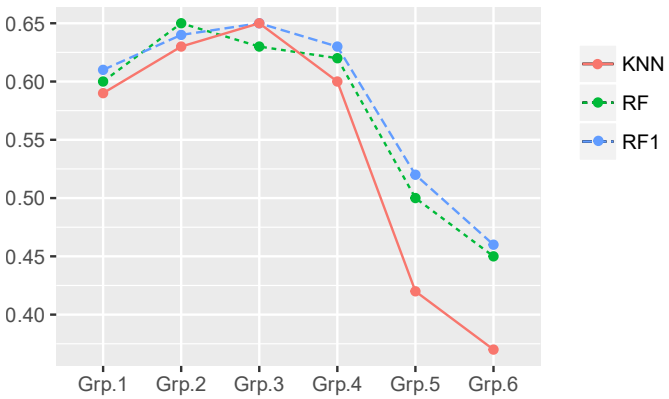


FIGURE 10. Improvement ratio of the RMSE value having the *naive* algorithm as reference value 1.

vehicle density does not represent extreme values, that is, very low or very high volume groups. When looking at the RMSE error, all the configurable parameters tested obtain a better performance. Moreover, when comparing our RF and the enhanced-KNN approach, except for the volume group Grp.3, RF outperforms both RMSE and MAPE metrics, while needing shorter computation times.

B. COMPUTATION TIME

For the given training set, we measure the time needed to make inferences. The time value shown in the tables,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2910225, IEEE Access

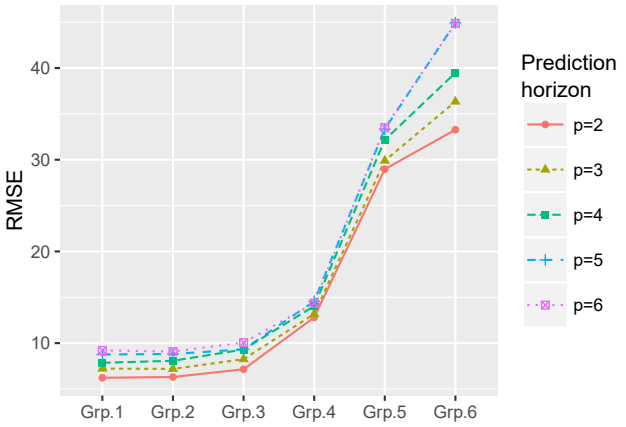


FIGURE 11. RMSE error for different volume groups and prediction horizons (p). Higher the prediction horizon, higher is the error. Please note that the effect of the increase of the RMSE for different volume groups is because this error depends on the magnitude.

represents the average time needed to compute the regression for a single subcell. Therefore, the prediction for the full grid, unless simplified, implies computing $N_X \times N_Y$ regressions.

On one hand, for the RF tuning parameters tested (number of trees and maximum depth), timing ranges between 0.26 and 6.20 ms (Table 1). The increase in any of these two parameters increases the computation time. If we extend the regression features including one-order neighbors, RMSE and MAPE errors are slightly smaller, but computation time needed gets worse.

On the other hand, the enhanced-KNN algorithm [13] takes 33.14 ms in average, by subcell, to compute an estimation (Table 2). The value of the chosen $K = \{5, 10, 15\}$ has little impact in the average value (± 0.28 ms).

Due to the grid-like representation of intersections, the computation cost of the regression algorithm is directly linked to the number of vehicles. In order to find ways to simplify the execution, we have analyzed the different behaviours of the subcells inside the grid.

C. SUBCELL ANALYSIS FOR GRID SIMPLIFICATION

We perform a hierarchical clustering to find groups among the different subcells, based on the measures we have proposed. The result is that the training data strongly support (with $p - value > 95\%$, using multiscale bootstrap resampling) two big groups of subcells, which are represented in Fig. 12 (a). We seek to check if the predictions obtained by the RF for subcells in each cluster are different. To show this, we represent the percentage of subcells where the RF outperforms the $naive$ algorithm, in Fig. 12 (b). With these results, on one hand, we can safely assume that for subcells belonging to Cluster 1, the RF algorithm is a better alternative most of the time, for volume groups Grp.2, Grp.3, Grp.4 and Grp.5. On the other hand, for subcells belonging to Cluster 2, the RF is a good choice only for volume groups Grp.4 and Grp.5. The prediction process can then be simplified.

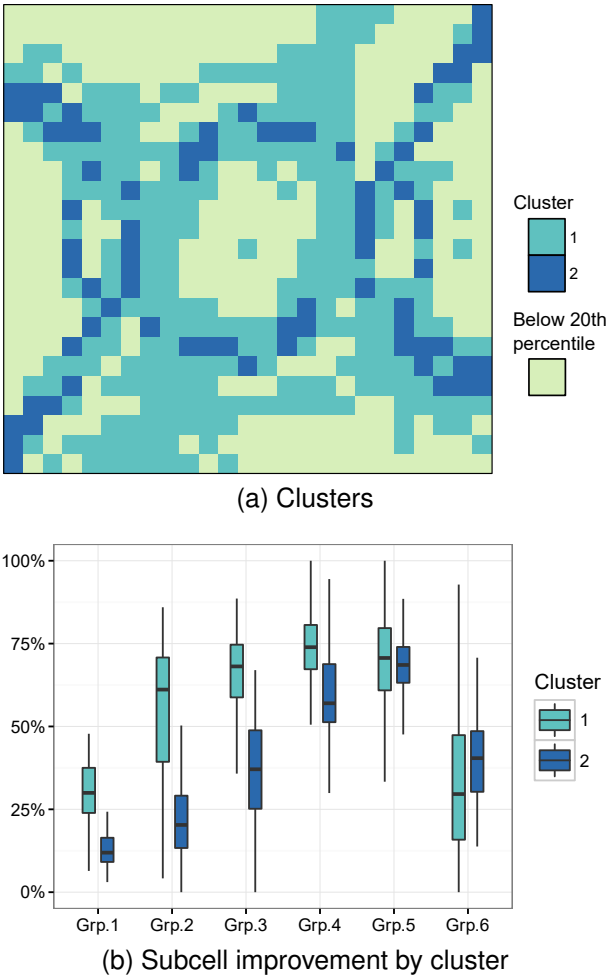


FIGURE 12. During the grid simplification, values below 20th percentile threshold are discarded; from the rest, the hierarchical clustering obtains two clear clusters of subcells with similar behaviour (a). Subcells of Cluster 1 show better results when applying the RF algorithm, than the *naive* method, improvement value over 50% (b).

D. RADIO COVERAGE WITH AND WITHOUT A REFLECTOR SYSTEM CONFIGURATION

Fig. 13 shows an illustrative comparison among radio coverage results in each subcell for a given traffic prediction, under different conditions. First, at time t, we collect the sequence of the previous aggregated snapshots, and we estimate a forthcoming one, p asn observations ahead. Based on that estimation ($\overline{asn_{t+p}}$), we compute the GA. Then, we compare the results for the three configurations, computing the signal power received by the real set of connected vehicles measured by asn_{t+p}, in each case.

The use of reflectors, even with the static approach, obtains a clear improvement of the received power. Moreover, this value increases drastically with the application of the GA. Results indicate that the number of vehicles that are covered by a power lower than 35 dBm is at least 13% lower with the *naive* approach and it can reach a 25% decrease. With the GA, the decrease comparing to the no reflector scenario is at least 53% and it can reach 99%.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2910225, IEEE Access

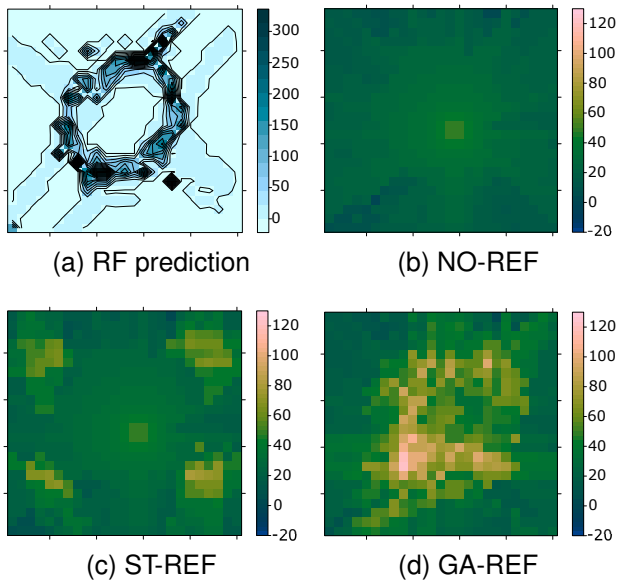


FIGURE 13. For a given asn prediction (a), we show the received power (dBm) for three configurations. The first one (b), represents the $\overline{R}$ value in the scenario without any reflectors. The second one (c), uses the *naive* static configuration of reflectors, and the third (d), uses the GA-based optimized configuration. In (b), we observe shadows in the upper and bottom areas of the grid, due to the high volume of vehicles at the roundabout entrances. In (c), the signal received in some regions which are distant from the transmitter increases. However, coverage improves more in (d), where shadowed zones are reduced and the areas with the highest expected demand are being supplied with high values of $\overline{R}$.

If we relax the threshold to 25 dBm, results are similar. The decrease is between 30% and 85% when comparing against the static approach, and between 39% and 100% against the GA.

Our GA-based optimization significantly outperformed the static configuration on all the tested cases. This is a substantial advancement compared to existing works on reflectors and their orientation.

VI. DISCUSSION

Overall, our approach has shown promising results to estimate future NLoS situations and improve the coverage in small cells to support V2X communications.

Our grid-like RF regression algorithm is an efficient prediction method in subcells where the vehicle density does not represent very high or very low values. Moreover, we observed that two types of subcells are supported by the data we used, and that the application of our method is favorable in one of the types. This leads to the possibility to simplify the prediction process, applying the regression algorithm only to a part of the subcells and ultimately reduce computational costs, leading to a fast estimation of NLoS. Thus, extra time can be used by the genetic algorithm on optimizing the orientations of reflectors.

As shown, the simple use of static reflectors improves the average power received by the vehicles. However, results have been greatly enhanced when using the dynamic configuration method that we have presented.

However, we would like to raise some discussion topics and some aspects that are still open for further research.

Traffic dynamics might be expected to vary when heterogeneous connected vehicles are involved (e.g., regular and automated vehicles with different levels of inter-vehicle interactions). The peculiarities of this kind of traffic were not addressed in our study, but it is a relevant research topic that deserves further analysis. Moreover, the lack of available real traffic data sets with both the temporal and spatial details limits our analysis. However, the effect of using generated traffic data in our prediction approach is expected to have low impact due to 1) small sampling periods are used to generate each snapshot, being able to observe the smallest changes between snapshots; and 2) many samples are aggregated in each aggregated snapshot, statistically smoothing the vehicle density.

We see room for improvements on the traffic prediction accuracy in the future; for example, combining different prediction methods. If so, nevertheless, additional computational needs should also be evaluated. Focusing on prediction accuracy under abnormal conditions is of great interest [15].

The reflector orientation use case has been limited to a single frequency inside the terahertz band (0.3 Thz), but the received power values may vary at higher bands. The size of the reflectors has been neglected, but this was explored in [27], and authors found out that, for outdoor communication, 100% of LoS is possible with plane dielectric mirrors larger than 1 meter size. In general, we found out that there are many ongoing research lines on different techniques and materials, such as graphene [42], to obtain configurable reflector surfaces.

In our experimental setup, locations of antenna and reflectors were set arbitrarily following an intuitive configuration based on intersection geometry. Shape and size of the intersection lanes, as well as traffic signaling such as road crossings, traffic lights and similar, should be considered to build a reasonable geometric model. This design decision undoubtedly can affect the solution performance. To propose the best location configuration, it is important to know where potential signal blockages happen due to stopped vehicles given traffic dynamics. Nonetheless, reusing existing infrastructure (streetlight poles or traffic lights) could be a cost-effective alternative. From our point of view, this reflector model itself would be worth a research paper on its own.

Additionally, reflections due to vehicle surfaces were out of the scope of the study, as well as other sources of rapid fluctuations in the received signal strength, over very short travel distances or short time durations. In general, these details are analyzed by small-scale propagation models. Our vehicle density model prediction and reflector optimization algorithms make use of aggregations in both space and time. Thus, we find that the proposed analysis is comprehensive only for large-scale approaches. In any case, we believe that our work could be a basis for other researchers that would explore and extend the models more in depth.

Even with exciting results towards the minimization of

shadowing effects, some concerns are raised for both LoS and NLoS calculations including other side-effects on terahertz signals, including diffraction and Doppler effects. This does not impact on the contribution of the paper, but opens doors for future work on communication optimization, with the possibility of using our approach to reconfigure antennas, and dynamically manipulating beamforming and beamsteering.

Another concern is surrounding the implementation of our proposed approach into existing communication infrastructures. We envision the most computationally expensive processes to be offloaded to the cloud, with a CloudRAN architecture. This raises questions about the signaling required between CloudRAN and RAN controllers. This challenge is an active research issue, for instance, on 5G infrastructure as a whole, and methods for fronthaul and backhaul signaling can be optimally designed not to interfere in network services performance [43]. Therefore, we see that our approach can be coupled to overall solutions of this type of advancements and this is not foreseen as a major issue.

VII. CONCLUSIONS

The safety of many critical applications relying on V2X, is highly dependent on the quality of service of the communication network. In general, providing new communication solutions requires novel advancements in network resource allocation based on computational intelligence approaches towards high data rate, highly reliable and high quality connections. Many solutions devoted to manage the RAN, will need to estimate the demand in advance and estimate areas with limited coverage. Based on this direction, in this work we have presented a data-driven solution to estimate the short-term spatial traffic distribution of connected vehicles inside urban intersections, and to use this information to maximize coverage inside a small cell.

We aimed to contribute to the provision of high quality communication links with a descriptive grid model to represent the density of these vehicles in space and time and a machine learning algorithm based on regression to estimate short-term vehicle density. Existing traffic prediction methods, which are focused on traffic links, are not suitable for the problem we present, which pursues the understanding of the fine-grain spatial distribution of traffic. After adapting and testing an existing kNN-based algorithm against a custom one based on Random Forests (RF), we have demonstrated that our RF approach is faster (in all volume groups) and more accurate (in all volume groups except one). We also have discovered the existence of two groups of subcells inside the grid, which enables the simplification of the prediction process, without compromising accuracy. This outcome allows estimating vehicle density predictions in real time.

Moreover, we have presented a method to relate the estimated vehicle density with the received coverage at different locations inside the intersection. Our study has shown the differences in the terahertz signal power received at each subcell position inside a grid, with a single transmitter antenna and no reflectors, with the use of static reflectors, and with the use of configurable reflectors.

To date and to our knowledge, no previous works have addressed dynamic changes of the radio network based on traffic prediction, and as we have shown, the benefits of our approach are motivating. In our use case, we have used a genetic algorithm to decide optimal reflector configurations to ensure good signal quality providing a virtual line of sight. Based on a cross-layer network management approach, we have achieved the optimization of signal reflectors and increased the received wireless power in a connected vehicles scenario. The effects of signal shadowing were decreased sometimes up to 100% due to optimizations based on the traffic prediction results. This outcome is important to validate the idea of highly adaptable networks and their benefits to high-frequency communications, such as the terahertz-based approaches.

Future work is envisioned into four main directions. One of the research directions is to decide optimal sliding intervals and aggregation times to find a compromise between prediction accuracy according to traffic dynamics and the costs of frequent RAN control reconfigurations. The second one is to analyze a vehicle density model relying on circular patterns instead of square grids. We see this approach specially promising in circular-shaped intersections such as roundabouts. Another one is to design highly adaptable beamforming and beamsteering techniques towards the increase of the probability of line of sight connections. Last but not least, we envision to deepen on the quantification of benefits of the overall system by the relationship between signal power efficiency and link layer metrics such as bit error rate. We commit to investigate this relationship more in depth as well as other relationships that can highlight the initial findings of this paper.

REFERENCES

[1] ITU-R, "ITU-R M.[IMT-2020.TECH PERF REQ] - Minimum requirements related to technical performance for IMT-2020 radio interface(s)," International Telecommunication Union, Geneva, Draft new Report, Feb 2017.

[2] S. Mumtaz, J. M. Jornet, J. Aulin, W. H. Gerstacker, X. Dong, and B. Ai, "Terahertz communication for vehicular networks," IEEE Transactions on Vehicular Technology, vol. 66, no. 7, pp. 5617–5625, July 2017.

[3] M. Koch, Terahertz Communications: A 2020 vision. Dordrecht: Springer Netherlands, 2007, pp. 325–338.

[4] I. F. Akyildiz, J. M. Jornet, and C. Han, "Teranets: ultra-broadband communication networks in the terahertz band," IEEE Wireless Communications, vol. 21, no. 4, pp. 130–135, 2014.

[5] H. Elayan, O. Amin, R. M. Shubair, and M. Alouini, "Terahertz communication: The opportunities of wireless technology beyond 5g," in 2018 International Conference on Advanced Communication Technologies and Networking (CommNet), April 2018, pp. 1–5.

[6] A. Dammann, R. Raulefs, and S. Zhang, "On prospects of positioning in 5G," in 2015 IEEE International Conference on Communication Workshop (ICCW), June 2015, pp. 1207–1213.

[7] P. Kachroo, S. Agarwal, B. Piccoli, and K. ÃŰzbay, "Multiscale modeling and control architecture for v2x enabled traffic streams," IEEE Transactions on Vehicular Technology, vol. 66, no. 6, pp. 4616–4626, June 2017.

[8] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we are going," Transportation Research Part C: Emerging Technologies, vol. 43, Part 1, pp. 3 – 19, 2014, special Issue on Short-term Traffic Flow Forecasting.

[9] A. Ermagun and D. Levinson, "Spatiotemporal traffic forecasting: review and proposed directions," Transport Reviews,

vol. 38, no. 6, pp. 786–814, 2018. [Online]. Available: https://doi.org/10.1080/01441647.2018.1442887

[10] J. Guo, W. Huang, and B. M. Williams, "Real time traffic flow outlier detection using short-term traffic conditional variance prediction," Transportation Research Part C: Emerging Technologies, vol. 50, pp. 160 – 172, 2015, special Issue on Road Safety and Simulation.

[11] Y. Zhang, Y. Zhang, and A. Haghani, "A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model," Transportation Research Part C: Emerging Technologies, vol. 43, no. Part 1, pp. 65 – 78, 2014, special Issue on Short-term Traffic Flow Forecasting.

[12] J. Guo, W. Huang, and B. M. Williams, "Adaptive kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," Transportation Research Part C: Emerging Technologies, vol. 43, no. Part 1, pp. 50 – 64, 2014, special Issue on Short-term Traffic Flow Forecasting.

[13] F. G. Habtemichael and M. Cetin, "Short-term traffic flow rate forecasting based on identifying similar traffic patterns," Transportation Research Part C: Emerging Technologies, vol. 66, no. Supplement C, pp. 61 – 78, 2016, Advanced Network Traffic Management: From dynamic state estimation to traffic control.

[14] S. M. Khan, K. C. Dey, and M. Chowdhury, "Real-time traffic state estimation with connected vehicles," IEEE Transactions on Intelligent Transportation Systems, vol. PP, no. 99, pp. 1–13, 2017.

[15] F. Guo, J. W. Polak, and R. Krishnan, "Predictor fusion for short-term traffic forecasting," Transportation Research Part C: Emerging Technologies, vol. 92, pp. 90 – 100, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X18305527

[16] M. Karlaftis and E. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," Transportation Research Part C: Emerging Technologies, vol. 19, no. 3, pp. 387 – 399, 2011.

[17] G. Fusco, C. Colombaroni, and N. Isaenko, "Short-term speed predictions exploiting big data on large urban road networks," Transportation Research Part C: Emerging Technologies, vol. 73, pp. 183 – 201, 2016.

[18] J.-B. Sheu, "A stochastic modeling approach to dynamic prediction of section-wide inter-lane and intra-lane traffic variables using point detector data," Transportation Research Part A: Policy and Practice, vol. 33, no. 2, pp. 79 – 100, 1999.

[19] P. Babington, HCM 2010 : Highway Capacity Manual. Washington, D.C: Transportation Research Board, 2010.

[20] P. Vythoulkas, "Alternative approaches to short term traffic forecasting for use in driver information systems," Transportation and traffic theory, vol. 12, pp. 485–506, 1993.

[21] S. Priebe, M. Kannicht, M. Jacob, and T. Kürner, "Ultra broadband indoor channel measurements and calibrated ray tracing propagation modeling at THz frequencies," Journal of Communications and Networks, vol. 15, no. 6, pp. 547–558, 2013.

[22] H. J. Song, K. Ajito, A. Wakatsuki, Y. Muramoto, N. Kukutsu, Y. Kado, and T. Nagatsuma, "Terahertz wireless communication link at 300 GHz," in IEEE Topical Meeting on Microwave Photonics (MWP), 2010, pp. 42–45.

[23] T. Kürner and S. Priebe, "Towards THz communications-status in research, standardization and regulation," Jounal of Infrared, Milimeter and Terahertz Waves, vol. 35, no. 1, pp. 547–558, 2014.

[24] C. Han, A. O. Bicen, and I. F. Akyildiz, "Multi-ray channel modeling and wideband characterization for wireless communications in the terahertz band," IEEE Transactions on Wireless Communications, vol. 14, no. 5, pp. 2402–2412, 2015.

[25] C. Han, A. O. Bicen, and I. Akyildiz, "Multi-Wideband Waveform Design for Distance-Adaptive Wireless Communications in the Terahertz Band," IEEE Transactions on Signal Processing, vol. 64, no. 4, pp. 910–922, 2016.

[26] S. Priebe and T. Kürner, "Stochastic Modeling of THz Indoor Radio Channels," IEEE Transactions on Wireless Communications, vol. 12, no. 9, pp. 4445–4455, 2013.

[27] M. T. Barros, R. Mullins, and S. Balasubramaniam, "Integrated terahertz communication with reflectors for 5G small-cell networks," IEEE Transactions on Vehicular Technology, vol. 66, no. 7, pp. 5647–5657, July 2017.

[28] C. Han and I. F. Akyildiz, "Three-dimensional end-to-end modeling and analysis for graphene-enabled terahertz band communications," IEEE Transactions on Vehicular Technology, 2016.

[29] C. Jansen, S. Priebe, C. Moller, M. Jacob, H. Dierke, M. Koch, and T. Kürner, "Diffuse scattering from rough surfaces in THz communication channels," IEEE Transactions on Terahertz Science and Technology, vol. 1, no. 2, pp. 462–472, 2011.

[30] S. Sun, T. A. Thomas, T. S. Rappaport, H. Nguyen, I. Z. Kovacs, and I. Rodriguez, "Path loss, shadow fading, and Line-of-Sight probability models for 5G urban macro-cellular scenarios," in 2015 IEEE Globecom Workshops, Dec 2015, pp. 1–7.

[31] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," IEEE Transactions on Antennas and Propagation, vol. 61, no. 4, pp. 1850–1859, April 2013.

[32] S.-H. Liao, C.-C. Chiu, and M.-H. Ho, "Comparison of GA and DDE for optimizing coverage in indoor environment," International Journal of Communication Systems, vol. 27, no. 11, pp. 3232–3243, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/dac.2537

[33] A. Shahmansoori, G. Seco-Granados, and H. Wymeersch, Survey on 5G Positioning. Cham: Springer International Publishing, 2017, pp. 165–196.

[34] J. Talvitie, M. Valkama, G. Destino, and H. Wymeersch, "Novel algorithms for high-accuracy joint position and orientation estimation in 5G mmWave systems," in 2017 IEEE Globecom Workshops (GC Wkshps), Dec 2017, pp. 1–7.

[35] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[36] S. Hussain, "Efficient Ray-Tracing Algorithms for RadioWave Propagation in Urban Environments," Ph.D. dissertation, School of Electronic Engineering, Dublin City University, Dublin, Ireland, 9 2017.

[37] ITU-R, "Propagation by diffraction," International Telecommunication Union, Geneva, Draft new Report, 2009.

[38] K. Deb, "An introduction to genetic algorithms," Sadhana, vol. 24, no. 4, pp. 293–315, Aug 1999.

[39] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of SUMO - Simulation of Urban MObility," International Journal On Advances in Systems and Measurements, vol. 5, no. 3&4, pp. 128–138, December 2012.

[40] "Openstreetmap contributors," Retrieved from http://planet.openstreetmap.org, planet dump [Data file from: 2016-01-26].

[41] M. Boban, T. T. V. Vinhoza, M. Ferreira, J. Barros, and O. K. Tonguz, "Impact of vehicles as obstacles in Vehicular Ad Hoc Networks," IEEE Journal on Selected Areas in Communications, vol. 29, no. 1, 2011.

[42] Y. Zhang, Y. Feng, J. Zhao, T. Jiang, and B. Zhu, "Terahertz beam switching by electrical control of graphene-enabled tunable metasurface," Scientific Reports, vol. 7, no. 1, p. 14147, 2017. [Online]. Available: https://doi.org/10.1038/s41598-017-14493-8

[43] A. Garcia-Saavedra, J. X. Salvat, X. Li, and X. Costa-Perez, "WizHaul: On the Centralization Degree of Cloud RAN Next Generation Fronthaul," IEEE Transactions on Mobile Computing, 2018.