Sparse Vehicular Sensor Networks for Traffic Dynamics Reconstruction

Eduardo del Arco, Eduardo Morgado, Mihaela I. Chidean, Julio Ramiro-Bargueño, Inmaculada Mora-Jiménez, and Antonio J. Caamaño

***Abstract*—In this paper, we propose the use of an ad-hoc wireless network formed by a fraction of the passing vehicles (sensor vehicles) to periodically recover their positions and speeds. A static roadside unit (RSU) gathers data from passing sensor vehicles. Finally, the speed/position information or space–time velocity (STV) field is then reconstructed in a data fusion center with simple interpolation techniques. We use widely accepted theoretical traffic models (i.e., car-following, multilane, and overtake-enabled models) to replicate the nonlinear characteristics of the STV field in representative situations (congested, free, and transitional traffic). To obtain realistic packet losses, we simulate the multihop ad-hoc wireless network with an IEEE 802.11p PHY layer. We conclude that: 1) for relevant configurations of both sensor vehicle and RSU densities, the wireless multihop channel performance does not critically affect the STV reconstruction error, 2) the system performance is marginally affected by transmission errors for realistic traffic conditions, 3) the STV field can be recovered with minimal mean absolute error for a very small fraction of sensor vehicles (FSV) $\approx 9\%$, and 4) for that FSV value, the probability that at least one sensor vehicle transits the spatiotemporal regions that contribute the most to reduce the STV reconstruction error sharply tends to 1. Thus, a random and sparse selection of wireless sensor vehicles, in realistic traffic conditions, is sufficient to get an accurate reconstruction of the STV field.**

***Index Terms*—Vehicular ad hoc networks, space-time velocity, geospatial analysis, combinatorial optimization.**

I. Introduction

TRADITIONALLY, the study of traffic congestion has been performed using sensors (cameras, magnetic loops, CW radar, etc) placed at specific points on the road, reporting velocity and occupation, generally every minute. Fixed sensors present several drawbacks, such as their deployment and maintenance costs or accuracy [1]. In some cases, it is necessary to perform a high density sensor deployment, e.g., the *variable message signs* in the M42 Birmingham that require a density greater than 6 loops/km/lane to work properly [2].

Manuscript received July 10, 2014; revised November 7, 2014 and February 20, 2015; accepted April 9, 2015. Date of publication May 12, 2015; date of current version September 25, 2015. This work was supported in part by the Autonomous Community of Madrid under Project S2013/MAE-2835 and in part by the Spanish Ministry of Economy and Competitiveness under Project TEC2013-48439-C4-1-R. The work of M. I. Chidean was supported by the FPU Research Grant AP2012-2981 from the Spanish Ministry of Education, Culture, and Sports. The Associate Editor for this paper was M. Zhou.

The authors are with the Department of Signal Theory and Communications, Rey Juan Carlos University, Fuenlabrada 28943, Spain (e-mail: antonio.caamano@urjc.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2015.2424671

Emergent wireless Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication technologies allow the composition of microscopic traffic information through the car-gathered data or Floating Car Data (FCD). They are implemented either by adapting existing technologies like 2G/3G, WiMAX or LTE (Long Term Evolution), or by using dedicated standards such as IEEE 802.11p or WAVE.

Due to the ubiquity of smartphones equipped with accurate Global Navigation Satellite System (GNSS) receivers, cellular networks are the obvious choice. The popularity of these terminals has led to the emergence of public and private real time traffic services using position and speed of the users to suggest alternative routes. Examples of such services are INRIX [3], Waze [4], Google Traffic [5] or public real-time traffic information services offered by traffic agencies like Spain's Traffic Authority [6].

The performance of these systems relies only on the voluntarily contributed additional information (contributed data), e.g., destination. The INRIX service requires entering the destination and it explicitly performs user tracking [7], [8]. Additionally, a merit figure is not provided since "[. . . if] the number or density of probe vehicles in a region of the road network is low there may not be sufficient information available to reliably determine traffic conditions" [7]. Waze also requires the introduction of the destination address and to "passively contribute [with] traffic road data" [4]. Further, Waze and Google applications show serious vulnerabilities both in the preservation of the anonymity of users and in the resistance to the injection of traffic data to generate artificial traffic jams [9].

Some of the problems in confidentiality and anonymity may be partially circumvented [10] by means of Virtual Trip Lines, which obfuscate individual spatial information. Though providing both privacy and performance measurements, it results in a higher-than-real density estimation when the average speed decreases. A near-optimal solution to this problem is achieved by means of submodular optimization methods [11]. However, the best results with these methods require the a priori definition of a source-destination model. An extremely low sampling frequency for position and speed (one sample per day) can be used in order to preserve privacy of drivers [12]. However, this sparse approach is only useful to calculate long term statistics of a road system usage, not to estimate short-term congestion or even to predict travel times.

In addition, it remains the problem of frequent signaling from different vehicles, which may lead to the congestion of the uplink in mobile cells [13]. A reduction of up to 95% in the cellular

network load can be achieved by using V2V communications [14]. Therefore, V2V and V2I communications are to be the preferred means for passing information from vehicles to traffic information processing centers.

On the other hand, in the area of wireless massive data gathering and processing, huge strides have been taken toward scalable architectures. Wireless Sensor Networks (WSN) have evolved from simple planar networks composed of but a few static sensors [15] to large-scale, self-organized, mobile multimedia data gathering networks [16], [17]. In order to search for scalable mobile architectures to be used in vehicular traffic data gathering, WSN represent a paradigm worth of exploration.

The present work addresses the problem of reconstructing the spatio-temporal traffic dynamics using data from sensor vehicles. In order to recreate realistic traffic density conditions, we simulate a typical speed pattern of individual vehicles using widely accepted microscopic traffic models. These vehicles are able to set up an ephemeral wireless network with V2V and V2I communication capabilities to communicate their position and speed. The transmission uses IEEE 802.11p in a realistic wireless channel such that data loss may occur. We assume that only a fraction of those vehicles provide data to the Data Fusion Center. From the received data, *spatio-temporal congestion diagrams* or *spatio-temporal developments of speed* can be rendered for a given road segment and time interval. For the sake of simplicity, we will refer to them as space-time-velocity field or *STV field*.

This work will answer the following questions:

- What is the relationship among the fraction of sensor vehicles, the temporal sampling frequency and the quality of the rendered STV field?
- What is the impact of road traffic conditions on the wireless ad-hoc network performance?
- What is the impact of the wireless multi-hop ad-hoc network performance on the quality of the information retrieved?
- Why the sensor vehicle data acquisition and STV field rendering is an appropriate approach to reconstruct the traffic dynamics?

We will provide comprehensible figures of merit that relate clearly the fraction of total traffic and the temporal sampling frequency with the STV field reconstruction. We will also provide a Space-Time (ST) field with the performance of wireless multi-hop ad-hoc network in terms of End-to-End Packet Loss. We will show that, even with a low quality wireless channel, there are enough data to render a high quality STV field. Finally, using submodular optimization techniques [18], [19], we will show that data from ST areas that greatly reduce the error variance in the STV reconstruction are those where sensor vehicles are prone to be found.

The remainder of this paper is organized as follows: next section details the system model, with basic definitions, scope and assumptions. The used traffic models, the wireless communication model, the reconstruction method and the basis for the submodular analysis of the STV field are presented in Section III. Experimental results are shown in Section IV. Finally, Section V collects the main conclusions.

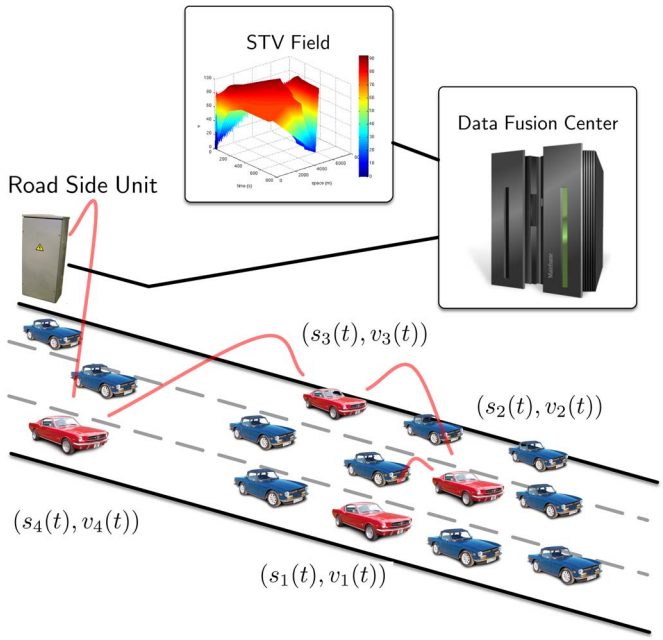


Fig. 1. Outline of the Vehicular Sensor Network along with the roadside and Data Fusion Center infrastructure. Blue cars mark the Passive Vehicles while red cars mark the Sensor Vehicles. Red lines indicate the relay channel used to communicate the quasi-synchronous measurements of vehicles position (s_α) and velocity (v_α) at sampling time t. These data are delivered to the Data Fusion Center to create the STV field.

II. System Model

We propose a WSN paradigm as starting point for the design and development of the FCD collecting and processing system. A WSN is a network of distributed autonomous devices measuring and monitoring environmental conditions cooperatively. Measurements from the sensor nodes can either be processed locally or relayed through multi-hop wireless channel to the network sink, where final processing and presentation are performed. The system described in this work is a Vehicular Sensor Network (VSN) where the sensed variables are the position $s_\alpha(t)$ and velocity $v_\alpha(t)$ of each vehicle α at time t (see Fig. 1).

We assume that only a small part of the vehicles belong to the VSN, which we name sensor vehicles. In addition to the sparsity of monitored trajectories, we also need to consider a sampling period T_s (which will be assumed constant). The sensed trajectories are not continuous functions of space, time and velocity but individual samples of continuous trajectories at regular time intervals. We may now define the different VSN functional elements:

- Sensor Vehicles (SV): They record navigation related data and forward them to other sensor vehicles in a V2V scheme (multi-hop wireless network). SV are embedded in traffic and are indistinguishable from Passive Vehicles (PV). The union of both SV and PV sets forms the Ground Set (Ω), the whole set of vehicles. The Fraction of Sensor Vehicles (FSV) in a stretch of the highway is defined as the ratio between cardinalities (i.e., the number of vehicles) of the SV subset and the Ground Set Ω.
- Road Side Units (RSUs): They collect data from the SV by means of V2I communications. These are the destination

TABLE I
NOTATION

Symbol	Meaning
Ω	*Ground Set*
$\|\Omega\|$	*Ground Set* cardinality: Number of vehicles
$SV \subset \Omega$	Set of Sensor Vehicles
$PV \subset \Omega$	Set of Passive Vehicles
$(s_\alpha(t), v_\alpha(t))$	Space and velocity samples of the α vehicle in t
X_Ω	Ground Truth: Space, time and velocity samples of Ω
X_{SV}	Space, time and velocity samples of SV
X_{PV}	Space, time and velocity samples of PV
$X_{\widetilde{SV}}$	X_{SV} without samples lost due to wireless channel
$\|X_\Omega\|$	Number of samples of X_Ω
$\|X_{SV}\|$	Number of samples of X_{SV}
$\|X_{PV}\|$	Number of samples of X_{PV}
$FSV = \frac{\|SV\|}{\|\Omega\|}$	Fraction of Sensor Vehicles
T_s	Sampling period
$\Omega = \bigcup_{n=1}^{\|\Omega\|} \alpha_n$	Ground set partitioned by vehicles
ρ_{RSU}	Road Side Unit density

nodes in the multihop segment of the VSN. RSUs do not add new data but route the samples from SVs to the Data Fusion Center. We consider that RSUs are regularly spaced along the highway, being characterized by their density (ρ_{RSU}) in units per kilometer.

- Data Fusion Center (DFC): The VSN sink of samples collected from the SV set (X_{SV}). It is used to construct the *STV field*.

For the sake of clarity, a summary of previous notation is presented in Table I.

III. MODELS AND METHODS

Our hypothesis is that it is possible to render an accurate estimation of the ground truth STV field from X_{SV} and a small number of SV. In order to evaluate the quality of the reconstructed STV field, the impact of the FSV and T_s have to be tested. To perform a fair comparison, we will consider a linear interpolation model constructed with X_{SV} (design set) and evaluated using X_{PV}.

Since the sets Ω and PV (and its respective samples X_Ω and X_{PV}) are not available in a real world scenario, a microscopic traffic simulation is necessary to obtain data with the required granularity. In Section III-A we describe the model considered in this work.

We present in Section III-B the wireless multi-hop communications model with realistic wireless channels. The network quality is characterized in terms of probability of data loss. The reduced design set is defined as $X_{\widetilde{SV}}$ (see Table I).

We also outline in Section III-C the methods to evaluate the reconstruction quality in a scenario with complex traffic dynamics for different values of parameters ρ_{RSU}, FSV and T_s. Finally, a submodular analysis of the STV field is performed in Section III-D.

A. Traffic Model

We require the generation of traffic data with plausible macroscopic features taking into account the individual movement of vehicles. Two mature models were implemented for this purpose, the *Intelligent Driver Model* (IDM) [20] for longitudinal kinematics and its transversal complement for lane and road changes *Minimizing Overall, Braking decelerations Induced by Lane changes* (MOBIL) [21]. Both models are car-following and accident-free, with continuous states, discrete time and continuous space outputs.

TABLE II
IDM PARAMETERS USED IN SIMULATION

Parameter	(units)	Value
a	(m/s^2)	1
b	(m/s^2)	1.65
T	(s)	1.6
v_0	(km/h)	f_{v_0}
s_0	(m)	2
l	(m)	5
δ		4
$v_{\max}$	(km/h)	120

TABLE III
M30 RING ROAD VERSUS SIMULATION FREE TRAFFIC STATISTICS

Statistical Moments	M30	Simulation
$\mu = E[v/v_{\max}]$	0.92	0.92
$\sigma = E[(v/v_{\max} - \mu)^2]^{1/2}$	0.086	0.054
$\zeta = E[(v/v_{\max} - \mu)/\sigma)^3]$	-2.17	-1.59
$\kappa = E[(v/v_{\max} - \mu)/\sigma)^4]$	12.47	11.50

The IDM controls the individual acceleration of a given vehicle (with a desired velocity) depending on the distance to the following car and their relative velocity. Table II summarizes the IDM parameters and the values considered in this work. The behavior of the IDM model is almost completely determined by parameters a (maximum acceleration), b (desired braking deceleration) and T (safety time headway between cars) [20]. The selected values are realistic, e.g., a car in free traffic would take $\approx$35 s and $\approx$500 m to reach a *desired velocity* of 120 km/h. The other IDM parameters are s_0 (minimum separation between cars), l (length of car) and δ (smoothness of the acceleration profile).

The desired velocity v_0 is usually assumed fixed to the maximum legal speed $(v_{\max})$ in the road and it is the same for all drivers. However, to account for a realistic approximation of traffic flow in free conditions, in this work we consider v_0 as a random variable (r.v.) with probability density function f_{v_0}. We set f_{v_0} to be as similar as possible as that of the traffic flow measured in congestion-free conditions and in a flow conservative section (no ramps nor intersections) of a freeway. Actual vehicle velocities were collected from loops embedded in the tarmac of the M30 ring road of Madrid (Spain) and were used to estimate a phenomenological desired velocity distribution. A piecewise-linear cumulative distribution function was obtained from the phenomenological one to implement a nonparametric random number generator for the desired velocities (in our simulation) using the procedure described in [22]. The numerical results of the procedure are summarized in Table III.

The second column in Table III presents values of the actual statistics for the M30 ring road. The value of the first moment (μ) indicates that the actual average speed is slightly lower than $v_{\max}$. The second moment (σ) indicates the dispersion of average speeds. The third moment (ζ) measures the symmetry

TABLE IV
MOBIL PARAMETERS USED IN SIMULATION

Parameter	(units)	N	E1	E2
p		0.3	0	0
b	(m/s^2)	2.5	2.5	4

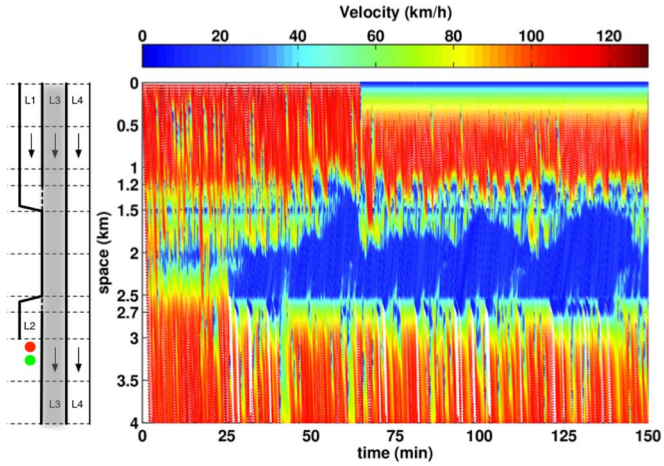


Fig. 2. Simulated scenario (left) and corresponding STV field (right) for $L3$. Several congestion features can be observed: 1) A bottleneck, present throughout the whole simulation, near the $L1$ on-ramp; 2) Another bottleneck appears due to vehicles attempting to exit $L3$ by the already congested $L2$ off-ramp; 3) $L2$ congestion increases with time, and the jam propagates upstream in $L3$, crossing the 1) and 2) bottlenecks; 4) By reducing the $L3$ inflow, the jam retracts (min 62, Km 2); however the aforementioned bottlenecks remain; 5) Upstream jam boundary oscillates by inflow and outflow differences. However, the downstream jam boundary (Km. 2.5–2.7) remains fixed throughout the whole simulation, thus generating bursty traffic.

of the distribution: a negative value means that there are more vehicles with desired velocity below $v_{\max}$. The high kurtosis (κ) indicates that values away from the average are rare, which is typical on a highway. The third column in Table III lists the resulting velocity statistics for free traffic segments in the present simulation.

Regarding MOBIL [21], it controls the transverse dynamics of vehicles, i.e., lane changes as well as the vehicle inputs and outputs by on-ramps and off-ramps. This model determines the change of lane for each vehicle by means of a decision-maker (see Table IV). The decision-maker uses both a security criteria b (which affects the maximum allowed deceleration of the new vehicle behind in the objective lane) and a politeness factor p (which balances the gain of acceleration between the vehicle and the new follower). In this paper, we define three possible lane-change behaviors: N (normal), $E1$ (egotistic) and $E2$ (extreme). By default, a driver behaves normally (N). When a vehicle must change lanes abruptly, either to go to an off-ramp or to continue driving from a stuck position, $E1$ or $E2$ substitute the normal driver behavior. Behaviors $E1$ and $E2$ only take into account self-gains but not at the cost of safety.

The simulated scenario shown at the left panel in Fig. 2 is composed by a main roadway with two lanes, $L3$ and $L4$, on-ramp $(L1)$, off-ramp $(L2)$ and traffic lights that avoid the fast emptying of the off-ramp. Fig. 2 represents the simulation output X_Ω for the $L3$ lane, rendered in a STV field. The simulator introduces vehicles through lanes $L1$, $L3$ and $L4$ with an input flow of $Q = 720$ veh/hour/lane, gradually increasing the flow till the maximum flow per lane. The maximum theoretical flow is limited by the timehead T. However, the inner dynamics of the congestion limits the sustainable maximum flow per lane, which was in average 1700 veh/hour/lane. From now on, we will use samples from $L3$ as the Ground Truth.

B. Wireless Channel Model

SV forward their location, time and velocity (X_{SV}) to the RSUs by means of a multihop wireless network. Let us consider a route with H sensor vehicles and the destination RSU, i.e., $H + 1$ nodes, with Decode and Forward (DAF) relaying strategy [23]. Fig. 3 shows the considered scheme, where the node labeled as 0 corresponds to the source sensor vehicle and the node labeled as H corresponds to the destination (which is always a RSU). The main feature of the VSN is the ability to deliver samples to the DFC. Therefore, a convenient measure of channel performance is the failure in the samples delivery. This merit figure is the End-to-End probability of packet decoding error from nodes 0 to H, named as P_H. The objective of this section is to provide a method to perform the calculation of P_H.

We consider a Non-Line-Of-Sight situation (Rayleigh fading) due to the low FSV scenario. We also consider that Doppler spread (caused by differences in speed between successive relays) is given in any single-hop. In these conditions, each single-hop has a different statistical behavior. On the other hand, the DAF scheme exploits channel coding to achieve best performance. It has been shown that this scheme gets best end-to-end performance than analog relaying (Amplify and Forward, AAF) [23]. Therefore, the calculation of P_H requires analyzing each single-hop separately in order to obtain the uncoded Average Bit Error Rate (ABER) and the coded Average Packet Loss Rate (APLR) for the hth hop.

In the following, we show how to accurately take into account the effects of the relative movement among SV (or with respect to RSUs) in the performance of the wireless communication system. We use the IEEE 802.11p physical layer as base system for wireless communications between SV. This is an Orthogonal Frequency Division Multiplexing (OFDM) system with M-Quadrature Amplitude Modulation (M-QAM) and coherent detection.

In an OFDM system, relative movement of the transmitter and the receiver results in the greatest impairment such a system can suffer. This is due to the fact that the available frequency bandwidth is broken down into smaller subchannels that are required to remain orthogonal (zero inter-channel correlation) throughout the communication. Doppler shifting results in a destruction of such orthogonality between subchannels. The following is a description of the structure of the Inter-Carrier Interference (ICI) and its effect in the quality of the wireless links between probe channels. This is done by evaluating the symbol error probability of the M-QAM modulation symbols per subchannel. That is why we will not describe the full calculation of the ICI effect on the OFDM but only summarize the most relevant features of the end result. Restricting our attention to a single OFDM symbol interval d_i, we can observe

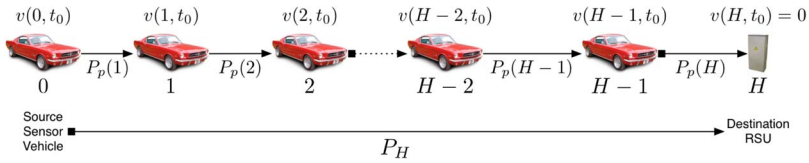


Fig. 3. Multi-hop route with $H+1$ nodes for a given time t_0 (discrete time slot of 1s). The node labeled 0 is the source sensor vehicle, and node H is the RSU (static). The packet payload is single sample of space-time and speed. End-to-End Packet Loss Probability is a function of single-hop $P_p(h)$ and therefore of single-hop $P_b(h)$.

that the output of the received signal in the correlator tuned to the ith subcarrier can be written as

$$\hat{d}_i = \beta_i(t_0)d_i + \frac{NT}{2\pi j}\sum_{k=1\neq i}^{N}\frac{\beta'_k(t_0)d_k}{k-i} + n_i \tag{1}$$

where $\beta_k(t) = w_1 + jw_2$ represents the fading time varying response of the mobile channel, N is the number of subcarriers in the OFDM system (52 in 802.11p) and T is the symbol duration of the OFDM symbol (8 μs in 802.11p). The first term in (1) represents the desired signal, the second one is the ICI term, and n_i is the additive noise. It has been shown in [24] that (1) can be rewritten in terms of an imaginary r.v. $Z = u + jv$ which accurately follows the statistics of the ICI contribution, such that:

$$\hat{d}_i = \beta_i(t_0)d_i + \frac{NT}{2\pi j}Z + n_i. \tag{2}$$

Let A_i be the event of making an error in the real or imaginary part of the ith subchannel. Noting that the real and imaginary parts of $\beta_i(t_0)$ are independent zero mean Gaussian r.v. with variance equal to $1/2(f_W(w) = (1/\sqrt{\pi})e^{-w^2})$, we can write the approximate (averaging through all subchannels) M-QAM OFDM Symbol Error Probability as

$$P_s \approx P_s^K = \frac{1}{N}\sum_{i=1}^{N}P_s^K[A_i] \tag{3}$$

where the symbol error probability for each subchannel can be written as

$$P_s^K[A_i] = \prod_{x=u,v,w_1,w_2}\int_{-\infty}^{\infty} dx\; P_s\left[A_i|Z,\beta_i(t_0)\right] \times f^K_{Z_r,Z_i}(u,v)f_W(w_1)f_W(w_2) \tag{4}$$

and the conditional symbol error probability can be written in terms of the statistics of the aforementioned r.v. $\beta_i(t_0)$ as

$$P_s\left[A_i|Z,\beta_i(t_0)\right] = 1 - \left(\left(1 - P_s^R\left[A_i|Z,\beta_i(t_0)\right]\right) \times \left(1 - P_s^I\left[A_i|Z,\beta_i(t_0)\right]\right)\right). \tag{5}$$

From (3) and assuming the use of Gray code (where the coding of adjacent symbols differs in only one bit), the ABER (P_b) can be approximated as:

$$P_b \approx \frac{P_s}{S}$$

where S is the number of bits in the modulation (i.e., 4 in the case of 16-QAM). The calculation of (4) has to be performed numerically, where $f^K_{Z_r,Z_i}$ is the bivariate density function of the real and imaginary parts of the ICI using a two-dimensional Gram-Charlier series, truncated at a total order K [24]. When the average SNR (E_b/N_0) is large, the ICI is the limiting factor in performance at any speed and for any N sub-carriers. Thus, the speed difference $\Delta v(h,t) = v(h,t) - v(h-1,t)$ between adjacent relaying sensor vehicles is the only variable being considered in the wireless channel model, keeping the SNR per bit, modulation, channel number, code gain and packet length unchanged. This mechanism is responsible for differentiated statistical behavior for single-hops.

The DAF scheme explores Forward Error Correction (FEC) capabilities of IEEE 802.11p. Let be the packet length L bits and error correction capability of e bits in l adjacent bits. Then, $P_p(h)$, is a function of the ABER $P_b(h)$ and can be computed as follows:

$$P_p(h) = 1 - \left[1 - \sum_{k=e+1}^{l}\binom{l}{k}\left(1 - P_b(h)\right)^{l-k}P_b(h)^k\right]^{\frac{L}{l}}. \tag{6}$$

Finally, we are able to compute P_H as the end-to-end APLR in a digital wireless relaying network:

$$P_H = 1 - \prod_{h=1}^{H}\left(1 - P_p(h)\right). \tag{7}$$

We assume the IEEE 802.11p parameters (16-QAM, $E_s/N_0 = 9$ dB in a power control loop, 5 GHz radio channel and 1/2 coding rate) and calculate the APLR for a given set of vehicle speeds that are sufficiently representative of the variations in the traffic flow.

C. STV Reconstruction and Error Evaluation

We want to characterize the quality of the information gathered by the VSN given FSV and T_s. In this section the models used to generate the traffic flow and its parameters are knowingly ignored, whereas only X_{SV} is considered as design set. Thus, the rendering process of a STV field in a specific space-time region should be faced with the unique knowledge of some samples from X_{SV}. We consider the STV field as a 2-D surface in 3-D space. The velocity surface is a graph of a function $f : s \times t \subset \mathbb{R} \times \mathbb{N} \longrightarrow \mathbb{R}^+$. Since we only have access to the value of f at X_{SV}, it is necessary to find an

interpolator of f to approximate the velocity at other points in the road domain, (e.g., in a highway off-ramp at rush hour).

A robust reconstruction of the STV field requires the testing of different configurations of SV, subject to FSV and T_s, along with error calculation with respect to X_{PV}. For a traffic scenario containing $|\Omega|$ vehicles, the number of distinct subsets that can be formed (the k-combinations of its elements) is given by the binomial coefficient. We are only interested in the SV distributions that typically occur in actual traffic. The SV selection process in the simulation follows a Poisson process with respect to the inter-arrival times. Thus, we ensure that the selection of SV among Ω for an individual configuration follows, on average, a uniform distribution.

The STV field reconstruction is addressed by applying interpolation techniques with the finite set X_{SV} as design data. Two types of interpolation techniques can be identified, those that use a model for f, and those that are parameter-free. If a model for f is assumed, it is necessary to adjust the model parameters, thus making the technique highly specific to the addressed problem. Though the microscopic traffic models referred above could be used as a model, it is preferable to avoid any model-specific interpolation in order to prevent overfitting. Therefore, we do not make any prior assumption on the model and the velocity surface is reconstructed using only X_{SV}. We use a local and exact interpolation technique well known in geostatistics: the Triangulated Irregular Network (TIN), which does not require any adjustable parameter. It is based on the generation of a mesh of *Delaunay* irregular triangles such that the vertex of these triangles are formed by X_{SV}. Then, each velocity in X_{PV} is estimated using the TIN interpolation. To the best of our knowledge, this technique has never been used to estimate velocities of vehicles in the context of traffic flow.

The *Delaunay* triangulation of X_{SV} maximizes the minimum angle over all triangulations of X_{SV} [25]. The resulting triangles maximize nearest local contributions, i.e., "long and skinny" triangles are avoided (if possible). Furthermore, *Delaunay* triangulation minimizes the roughness (integral of the Sobolev semi-norm of the gradient) of the resulting surface, for any set of scattered data [26]. This property holds no matter what the actual height data is (in our case, velocity). The consequence of the previous properties in this problem is that space-time trajectories of different vehicles never cross if sufficient data points are available. If not, unphysical results follow for the reconstructed trajectories, i.e., two cars may occupy the same space at the same time. Thus, an additional error measure based on the lower-bound on the data density could be obtained. However, the reconstruction of individual trajectories, apart from privacy concerns [12], [10], is out of the scope of the present work.

The aforementioned procedure is applied for all triangles which cover the space-time domain (See Fig. 4). The function can be evaluated only inside the convex hull of the whole triangulation. The error evaluation is now straightforward using X_{PV}, which is provided by the traffic flow simulation (or the actual traffic data).

Let $\mathbf{x}_p \in X_{PV}$, we define the error ε:

$$\varepsilon(\mathbf{x}_p) = \varepsilon(s_p, t_p) = v(\mathbf{x}_p) - \hat{v}(\mathbf{x}_p). \tag{8}$$

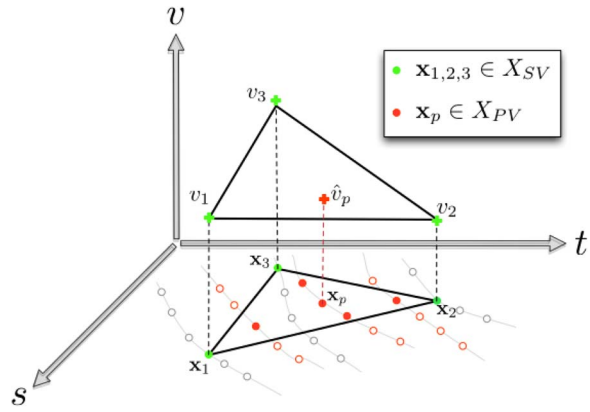


Fig. 4. Triangle in the Triangular Irregular Network. Green dots (triangle vertices) are space and time samples with an associated velocity (green cross). These samples are in X_{SV}. Red dots and their associated velocities are in X_{PV}. The red cross is not the actual value of the velocity in $\mathbf{x}_p$, but the estimation $f_{\mathcal{T}}(\mathbf{x}_p) = \hat{v}_p(\mathbf{x}_p)$.

In order to evaluate the error and its variation for each FSV and T_s pair, K realizations of SV on the Ω set are used. Each sample $(\mathbf{x}, v) \in X_\Omega$ is selected $M(\mathbf{x})$ times in the K different realizations.

We can also calculate the error $\varepsilon(\mathbf{x})$ between the actual $v(\mathbf{x})$ and the estimated $\hat{v}(\mathbf{x})$ which can be obtained from $\{\hat{v}_k(\mathbf{x})\}_{k=1}^{K}$, which in turn come from the K SV sets. Thus, the error for each space-time pair $\mathbf{x} = (s, t)$ is:

$$\varepsilon(\mathbf{x}) = v(\mathbf{x}) - \frac{1}{M(\mathbf{x})} \sum_{k=1}^{M(\mathbf{x})} \hat{v}_k(\mathbf{x}). \tag{9}$$

We can also compute the corresponding variance of $\varepsilon(\mathbf{x})$ as:

$$\sigma^2_{\hat{v}}(\mathbf{x}) = \frac{1}{M(\mathbf{x})} \sum_{k=1}^{M(\mathbf{x})} \hat{v}_k^2(\mathbf{x}) - \left[\frac{1}{M(\mathbf{x})} \sum_{k=1}^{M(\mathbf{x})} \hat{v}_k(\mathbf{x}) \right]^2. \tag{10}$$

The comparison of the reconstruction performance for several FSV and T_s is achieved by using the Mean Absolute Error (MAE), which avoids massive error cancellations:

$$\text{MAE} = \frac{1}{|X_\Omega|} \sum_{i=1}^{|X_\Omega|} |\varepsilon(\mathbf{x}_i)|. \tag{11}$$

By using the methodology of [27] and [28], we propose Algorithm 1 to calculate both (10) and (11).

Algorithm 1 Error Computation Over K SV Realizations

Input Ω, X_Ω, FSV, T_s, K
Output $\varepsilon(\mathbf{x})$, $\sigma^2_{\hat{v}}(\mathbf{x})$
  $M(\mathbf{x}) \leftarrow 0 \; \forall \mathbf{x} \in \Omega$ ▷ Vector initialization of visiting index
  $\mu_{\hat{v}}(\mathbf{x}) \leftarrow 0 \; \forall \mathbf{x} \in \Omega$ ▷ Vector initialization of mean
  $\sigma^2_{\hat{v}}(\mathbf{x}) \leftarrow 0 \; \forall \mathbf{x} \in \Omega$ ▷ Vector initialization of variance
  for $k = 1$ to K **do**
    perform random selection of SV / $|SV| = FSV|\Omega|$
    PV is selected as complementary of SV
    perform subsampling of $\mathbf{x} \in X_{SV}$ / $\mathbf{x} = (s, T_s t)$

perform convex hull of X_{SV}
perform Delaunay triangulation D
$f_D \leftarrow \{f_\mathcal{T}\}_1^D$
for all $\mathbf{x}_p \in X_{PV}$ and convex hull of X_{SV} **do**
$M(\mathbf{x}_p) \leftarrow M(\mathbf{x}_p) + 1$
$\delta(\mathbf{x}_p) \leftarrow f_D(\mathbf{x}_p) - \mu_{\hat{v}}(\mathbf{x}_p)$
$\mu_{\hat{v}}(\mathbf{x}_p) \leftarrow \mu_{\hat{v}}(\mathbf{x}_p) + \frac{\delta(\mathbf{x}_p)}{M(\mathbf{x}_p)}$
$\sigma^2_{\hat{v}}(\mathbf{x}_p) \leftarrow \sigma^2_{\hat{v}}(\mathbf{x}_p) + \delta(\mathbf{x}_p) \cdot [f_D(\mathbf{x}_p) - \mu_{\hat{v}}(\mathbf{x}_p)]$
end for
end for
for all $\mathbf{x} \in \Omega$
$\varepsilon(\mathbf{x}) \leftarrow v(\mathbf{x}) - \mu_{\hat{v}}(\mathbf{x})$
$\sigma^2_{\hat{v}}(\mathbf{x}) \leftarrow \frac{\sigma^2_{\hat{v}}(\mathbf{x})}{M(\mathbf{x})-1}$
end for

D. Submodular Function Optimization

In addition to the evaluation of the reconstruction error, it is desirable to quantify how far the reconstruction is from the optimal solution. In this section we will use submodular function optimization [18] to calculate this deviation from optimality. Intuitively, we try to determine those ST pairs such that the error variance over the test set is minimized. There is bound to be a ST region where the gain in variance reduction is the highest. As we incorporate additional ST pairs into the estimation of the STV reconstruction, additional gains are to be expected but they are likely to dwindle (diminishing returns). If this reduction in additional gains is monotonic, the utility function (error variance over the test set) is said to exhibit *submodularity*. We will exploit this property in order to answer whether there is an optimal sampling rate and FSV.

Given FSV and T_s, let us divide the ST-field into observed and unobserved locations, being X_{SV} and X_{PV}, respectively. Similar to the error and variance defined in the previous section, we define a utility function $F(\cdot)$ or *expected reduction of variance* at the PV locations:

$$F(SV) = \mathsf{VAR}(X_{PV}) - \int P(\mathbf{x}_{SV})\mathsf{VAR}(X_{PV}|X_{SV} = \mathbf{x}_{SV})d\mathbf{x}_{SV} \quad (12)$$

where $\mathsf{VAR}(X_{PV}|X_{SV} = \mathbf{x}_{SV})$ is defined as

$$\frac{1}{|PV|}\sum_{\alpha \in PV} \mathbb{E}\left[(X_\alpha - \mathbb{E}[X_\alpha|\mathbf{x}_{SV}])^2 |\mathbf{x}_{SV}\right]. \quad (13)$$

In order to find the best SV subset maximizing (12), a combinatorial optimization problem must be solved

$$SV^* = \max_{SV \subseteq \Omega} F(SV) \text{ s.t. } SV \in \mathfrak{F} \quad (14)$$

where $\mathfrak{F} \subseteq 2^\Omega$ is a collection of feasible subsets of Ω, i.e., all sets of size at most k: $\mathfrak{F} = \{SV \subseteq \Omega : |SV| \leq k\}$. Note that this is a NP-complete problem. However, under certain conditions, it is possible to achieve a near optimal solution exploiting the submodular properties of some functions and data sets.

TABLE V
SUBMODULARITY NOTATION

$\mathcal{V}_i$	Set of cells at time interval i
$\lvert\mathcal{V}_i\rvert$	Number of cells at time interval i
$\{\mathcal{V}_i\}_{i=1}^T$	Ground set partitioned by $\sum_{i=1}^T \lvert\mathcal{V}_i\rvert$ cells
$\mathcal{A} \subset \mathcal{V}_i$	Subsets of $\mathcal{V}_i$
$X_\mathcal{A} \in \mathcal{A}$	Samples of $\mathcal{A}$

The *submodularity* is a property of set functions $F : 2^\Omega \mapsto \mathbb{R}$, which assign a value $F(SV)$ to each subset $SV \subseteq \Omega$. If F is submodular, then the diminishing returns property is satisfied, i.e., for all $SV \subseteq SV' \subseteq \Omega$ and $\alpha \in \Omega \backslash SV'$, then

$$F(SV \cup \{\alpha\}) - F(SV) \geq F(SV' \cup \{\alpha\}) - F(SV'). \quad (15)$$

It has been shown that the expected reduction of variance has, in addition to the property of submodularity, the property of monotonicity [12], Section II, [29], Section V. This is the necessary condition to apply a lazy-greedy algorithm that results in a near-optimal solution. As a consequence, we can expect a lower bound of F for $|SV_k| = k$

$$F(SV_k) \geq (1 - \exp(-1)) \max_{SV:|SV| \leq k} F(SV). \quad (16)$$

Unfortunately, the sparse sampling of the STV field using sensor vehicles presents a major obstacle that must be overcome: different sensor vehicles are not ensured to provide the same number of samples or even to provide a reasonable number samples. Therefore, we cannot assume that the solution to our optimization problem is a choice of individual vehicles SV s.t. $k = FSV|\Omega|$. However, we can redefine the problem assuming that $\mathfrak{F}$ is a space-time partition of the whole ST region into spatio-temporal regions or *cells* (see Table V).

A traffic-meaningful division of the road lane under test is found as a set $\mathcal{V}$ of equally sized segment or cells:

$$|\mathcal{V}| = \left\lfloor \frac{L}{n_{MAX}(s_0 + l)} \right\rfloor \quad (17)$$

where L is the length of the road lane, s_0 and l are IDM parameters (see Table II) and n_{MAX} is the maximum number of vehicles per cell. Thus, we redefine the maximization problem (14) as:

$$\mathcal{A}^* = \max_{\mathcal{A} \subseteq \mathcal{V}} F(\mathcal{A}) \text{ s.t. } \mathcal{A} \in \mathfrak{F} \quad (18)$$

where $\mathcal{A}$ is a subset of cells and F is the expected reduction of variance

$$F(\mathcal{A}) = \mathsf{VAR}(X_\mathcal{V}) - \int P(\mathbf{x}_A)\mathsf{VAR}(X_\mathcal{V}|X_\mathcal{A} = \mathbf{x}_A)d\mathbf{x}_A. \quad (19)$$

The time dimension of the cell is related to the sampling rate. This period must be large enough to collect a sufficient amount of samples but short enough to be practical. Furthermore, it must avoid mixing two or more traffic states into the same cell. Let $\Sigma_{\mathcal{V},\mathcal{V}}$ be the covariance matrix of the velocity for each subset of $|\mathcal{V}|$ cells. We define the noise power as σ^2 and the modulus of the largest eigenvector $\mathbf{v}$ of $\Sigma_{\mathcal{V},\mathcal{V}}$ as $\|\mathbf{v}\|^2$. The stochastic self-adjoint matrix $\Sigma_{\mathcal{V},\mathcal{V}}$ experiences a *phase transition* for

$|X_{\mathcal{V}}|$ samples obtained from $|\mathcal{V}|$ sensors s.t. $|X_{\mathcal{V}}|/|\mathcal{V}| \geq \sigma^4/\|\mathbf{v}\|^4$ (Eq. 2.19 in [30]). During phase transition the largest eigenvalue "collapses" from noise to a signal subspace eigenvalue. A typical value of $\sigma^4/\|\mathbf{v}\|^4 \approx 4$ holds true for a wide range of signals; at least $|X_{\mathcal{V}}| \gtrsim 4 \cdot |\mathcal{V}|$ samples are needed to reliably extract the main signal eigenvector using $|X_{\mathcal{V}}|$ measurements from $|\mathcal{V}|$ cells. Therefore, if we ensure that the previous condition holds, i.e.,

$$|X_{\mathcal{V}}| \geq 4 \cdot |\mathcal{V}|. \tag{20}$$

We can ensure that the $\Sigma_{\mathcal{V},\mathcal{V}}$ is non-singular as at least one eigenvalue can be extracted. We will term this constraint as the *condition for covariance matrix phase transition*.

In the present work, we build $\Sigma_{\mathcal{V},\mathcal{V}}$ with a Sampling Rate of 1 Hz. If there are more than one vehicle in the cell at the same sampling instant, we perform a random selection among them. Missing data (no vehicle for a given sampling instant) are imputed to the average velocity of the whole cell. The amount of desired samples is

$$|X_{\mathcal{V}}| \geq 4\,\frac{L}{n_{MAX}(s_0 + l)}. \tag{21}$$

Therefore, the duration of each cell (in seconds) is

$$P = \frac{|X_{\mathcal{V}}|}{\text{Sampling Rate}}. \tag{22}$$

For a STV field of length L and duration t_{MAX}, we divide the field in $T = \lfloor t/P \rfloor$ data sets with corresponding covariance matrices $\{\Sigma^i_{\mathcal{V},\mathcal{V}}\}_{i=1}^{T}$.

We use a specific implementation of the lazy-greedy algorithm [31] for $\mathcal{A}^*$ and $F(\mathcal{A}^*)$ calculation. This algorithm allows us to explore the most relevant road regions for the STV reconstruction. These greedily selected spatio-temporal regions, $\mathcal{A}^*$, are closely related with the optimal cell choice. The optimal for $|\mathcal{A}| = k$ involves the calculation of $k-1, k-2, \ldots, k=1$ if F and $\mathfrak{F}$ are both submodular.

IV. Experimental Results

In this section we present the evaluation of the feasibility of multihop VSN and their effect on the STV field reconstruction performance. Thus, we start by considering that the effects of the wireless communication network are not relevant, i.e., the End-to-End Packet Loss Probability is deemed negligible (Section IV-A) and we examine the impact of varying T_s and FSV. Then, we proceed to introduce the wireless multi-hop channel to evaluate its effects in the STV field reconstruction (Section IV-B). Afterwards we will check the robustness of the VSN against different configurations of RSU, FSV and T_s (Section IV-C). We observe that for very low sensor vehicle densities, the performance in the STV field reconstruction is outstanding. But why? In Section IV-D we will show that the answer lies in the domain of the discrete and combinatorial optimization mathematics.

A. VSN Performance in an Ideal Channel

We have performed experiments with $FSV = [1, 30]\%$ and $T_s = \{1 \cdots 10, 15, 20, 25, 30, 45, 60\}$ seconds. Fig. 5 shows the X_{SV} samples in a STV field (single outcome) on $L3$, allowing us to make a qualitative performance of the VSN analysis. Two FSV and three sampling values for T_s are presented in Fig. 5.

For a small fraction of sensor vehicles, $FSV = 1\%$, the single trajectories are visible as well as the areas of the highway where the vehicle speed is lower. Note that some paths did not start at the beginning of the road (nor they did finish at the end of the lane). These trajectories appear from 1.2 km and disappear from 2.5 km. These vehicles merged on $L3$ from the on-ramp $L1$ and left $L3$ by the off-ramp $L2$. Some vehicles appeared and disappeared at arbitrary road places, since they changed of $L4$ to $L3$ or vice versa. As the sampling time was increased, other characteristics are visible. With $T_s = 5$s [Fig. 5(b)], it became more apparent that there is no synchronous operation among vehicles. With $T_s = 30$ s [Fig. 5(c)], we found a high dispersion of samples, and therefore it is difficult to segment a single trajectory. On the other hand, in the low speed areas, samples are spatially closer and the macroscopic structure of the congestion is visible.

With a tenfold increase ($FSV = 10\%$), all *Ground Truth* features were captured (compare with Fig. 2).

Fig. 6 corresponds to the output of Algorithm 1 for $FSV = 10\%$ and $T_s = 1$. Fig. 6(a) is the absolute error field $|\varepsilon(\mathbf{x})|$, (11); Fig. 6(b) shows the variance of velocity estimation with random selections of SV sets, $\sigma^2_{\hat{v}}(\mathbf{x})$.

A small variance in a certain ST pair indicates increased robustness against different design set configurations of SV (provided by a specific SV set). On the other hand, a higher variance indicates that the speed estimation at that ST pair is sensitive to the selection of the SV. The error variance is low (less than 5 $(\text{m/s})^2$) at the core of the traffic congestion. On the other hand, the error variance increases along the space region $[1, 1.5]$ km, matching with $L1$-$L3$ merging area. The spatial region above the upstream jam border indicated a high level of dependence in the SV occurrence. The region of burst traffic, located beyond the downstream border of the jam, had a medium grade of uncertainty (around 20 $(\text{m/s})^2$).

Fig. 6(c) is a comprehensive set of results for $P_H = 0$. Increasing the FSV (at constant T_s), the error decreases. For constant FSV and decreasing T_s, the MAE decreases till it reaches a minimum and then it starts to grow again (red curves). Only for values of FSV greater than 9% (blue curves), the MAE decreases monotonically with decreasing T_s. Thus, only for FSV values equal or greater than 9%, accurate reconstruction of the STV field can be ensured.

B. Realistic Wireless Channel

ST fields are useful to analyze the spatial and temporal distribution of data, in addition to the speed. In this section, we use ST fields to represent the average End-to-End packet loss P_H over multiple realizations. Fig. 7 shows the multi-hop wireless network performance, $P_H(\mathbf{x})$, for ρ_{RSU}=$\{4, 1, 0.25\}$ km^{-1} and $FSV = 10\%$. Each value on $\mathbf{x}$, $P_H(\mathbf{x})$, has been calculated and averaged for a representative number of realizations.

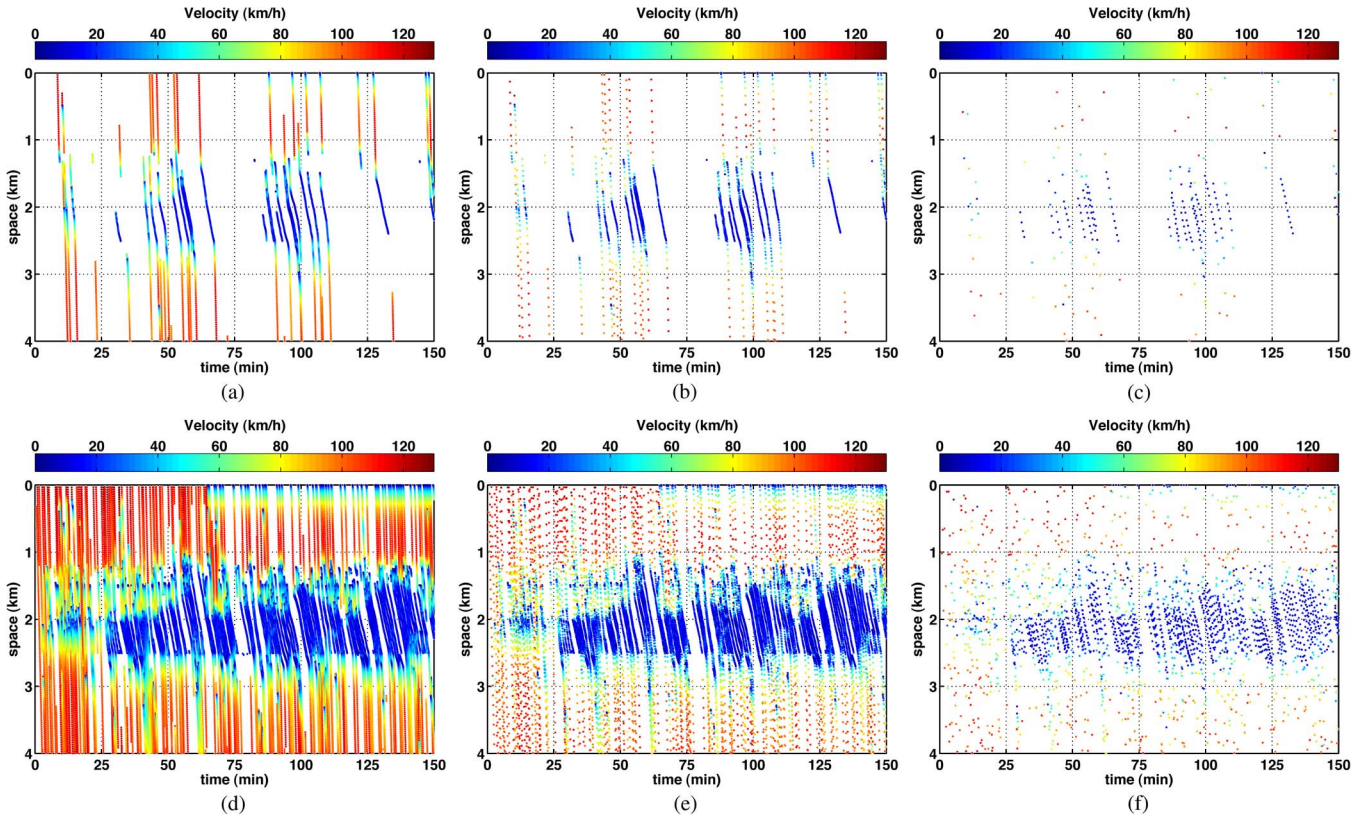
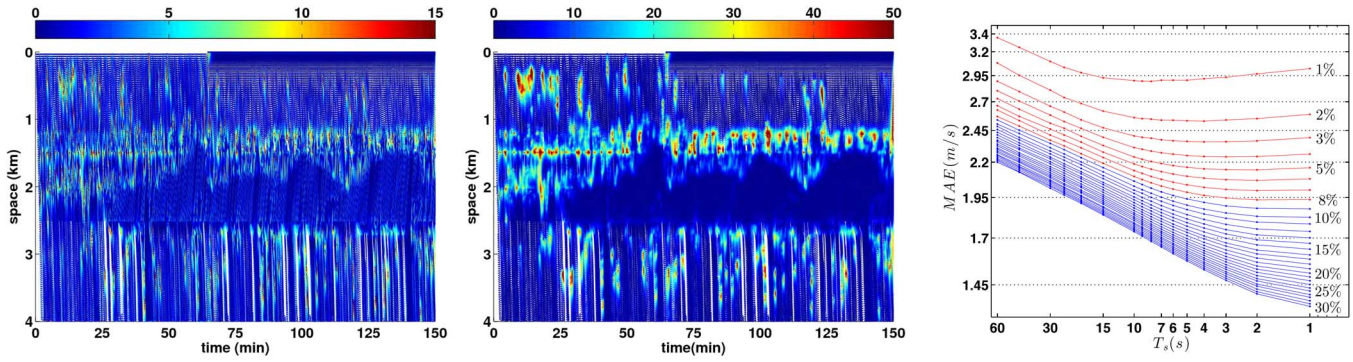


Fig. 5. Lossless spatio-temporal sampling on $L3$. Sampled Ω is composed by individual contributions of SV, X_{SV}. FSV has an impact in space and time resolution. Even with a very low FSV, it is possible to distinguish the upstream and downstream border of the jam. As the FSV increases, the high frequency spatio-temporal details are resolved (first row versus second row). As the sampling period T_s increases, the region with the scattered samples maps with free traffic and higher speeds. By contrast, samples in a congested region are closer to each other due to lower speeds and closely follow the macroscopic features of the jam. (a) $FSV = 1\%$, $T_s = 1$ s (b) $FSV = 1\%$, $T_s = 5$ s (c) $FSV = 1\%$, $T_s = 30$ s (d) $FSV = 10\%$, $T_s = 1$s (e) $FSV = 10\%$, $T_s = 5$ s (f) $FSV = 10\%$, $T_s = 30$ s.



(a) $|\varepsilon|$ (m/s) (b) $\sigma_{\hat{v}}^2$ $(m/s)^2$ (c) MAE (m/s)

Fig. 6. (a) Absolute error field $|\varepsilon(\mathbf{x})|$, (b) Variance in speed estimation $\sigma_{\hat{v}}^2$ with different SV sets, and (c) MAE in STV reconstruction versus T_s for a full sweep of FSV (ideal channel). MAE axis is logarithmic. For FSV values below 9% MAE reaches a minimum till it begins to grow for low values of T_s (red curves). This subtle performance degradation is due to a huge difference between the sides of the TIN triangles (long and skinny triangles). For FSV values above 9%, the MAE is monotonically decreasing (blue curves).

For $\rho_{RSU} = 4$ km^{-1} and $FSV = 10\%$ [Fig 7(a)], a flat (homogeneous) and good (low probability) field is obtained. This reveals that the majority of communications are carried out with a single-hop due to the low spacing between RSUs. Inside the jam, the relative velocities are low and transmission errors due to Doppler are negligible. If the ρ_{RSU} is decreased the network performance turns sensitive to the RSU location and the traffic dynamics. With $\rho_{RSU} = 1$ km^{-1} (Fig. 7(b) the P_H is higher in the middle of the jam and the free traffic zones due to multi-hop and Doppler effect, respectively. The congested regions contribute with more errors due to the larger communication chain (multi-hop effect) while the Doppler effect reveals in the free traffic zones. With $\rho_{RSU} = 0.25$ km^{-1} [Fig. 7(c)], there is only one RSU located in the core of the jam. We can appreciate the combined effect of large multi-hop chains and Doppler effect. The higher distance to RSU, the higher P_H (multi-hop).

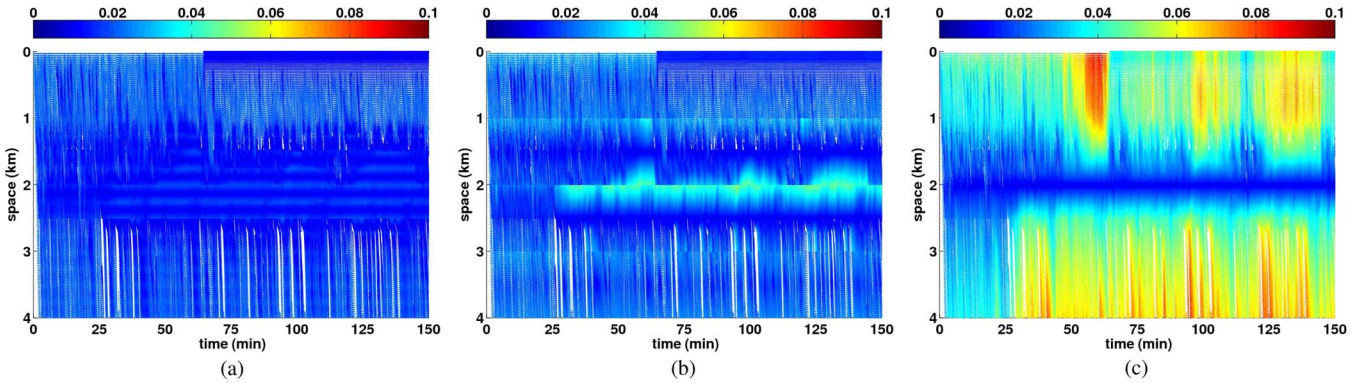


Fig. 7. ST representation of the probability of End-to-End packet loss or Packet Loss Field. With $FSV = 10\%$ and 0.25 km$^{-1} \leq \rho_{RSU} \leq 4$ km^{-1}, the packet loss probability is under 0.025. (a) $FSV = 10\%$, $\rho_{RSU} = 4$ km^{-1} (b) $FSV = 10\%$, $\rho_{RSU} = 1$ km^{-1} (c) $FSV = 10\%$, $\rho_{RSU} = 0.25$ km^{-1}.

TABLE VI
MAE (M/S) $P_H \neq 0$

T_s (s)	1	5	10	20	30
FSV=1%					
$\rho_{RSU} = 4km^{-1}$	3.0168	2.9028	2.9056	2.9948	3.1090
$\rho_{RSU} = 1km^{-1}$	3.0167	2.9055	2.9055	2.9951	3.1095
$\rho_{RSU} = 0.25km^{-1}$	3.0231	2.9086	2.9091	2.9974	3.1138
FSV=10%					
$\rho_{RSU} = 4km^{-1}$	1.8222	1.9017	2.0284	2.2032	2.3119
$\rho_{RSU} = 1km^{-1}$	1.8222	1.9021	2.0292	2.2040	2.3127
$\rho_{RSU} = 0.25km^{-1}$	1.8223	1.9066	2.0345	2.2098	2.3190

However, this probability is higher at some ST regions, specially in those located in the $0-1$ km and $50-70$ min intervals. This effect is caused by the upstream jam boundary oscillation (see Fig. 2), which augments the multi-hop chain length.

C. VSN Performance in a Realistic Wireless Channel

In this section, we apply the Algorithm 1 considering the effect of communications. The X_{SV} samples are taken out with probability $P_H(\mathbf{x})$, resulting in $X_{\widetilde{SV}}$. Table VI reflects results for $FSV = 1\%$ and $FSV = 10\%$. Note that the effect in the MAE is negligible, (approximately of $5 \cdot 10^{-4}$ m/s). It has not been possible to find any difference for FSV beyond 10%.

D. Conditions for Optimal VSN Performance

We show the results for $n_{MAX} = 20$ vehicles. As exposed in Section III-D, this maximum packaging involves $|\mathcal{V}| = 28$ cells per time slot and cell length of 142.28 m. In order to comply with the covariance matrix phase transition condition, the amount of samples is $|X_{\mathcal{V}}| = 112$. Thus, taking one sample per second, the duration of each cell is 112 s. Considering the whole simulation time (and discarding 150 s at the beginning due to lack of samples), the cells are grouped in 79 time slots. This allows to build 79 $\Sigma_{\mathcal{V},\mathcal{V}}$ covariance matrices of 28×28 elements. The total number of cells is 2212.

We use lazy greedy algorithm [31] in order to determine the cells that maximize (19) for a given cell budget. In Fig. 8 a ST-field of the selected cells for the different budgets is shown. The black cells are the first choice in the maximization process ($|\mathcal{A}| = 1$) and dimming shades of grey are used to indicate

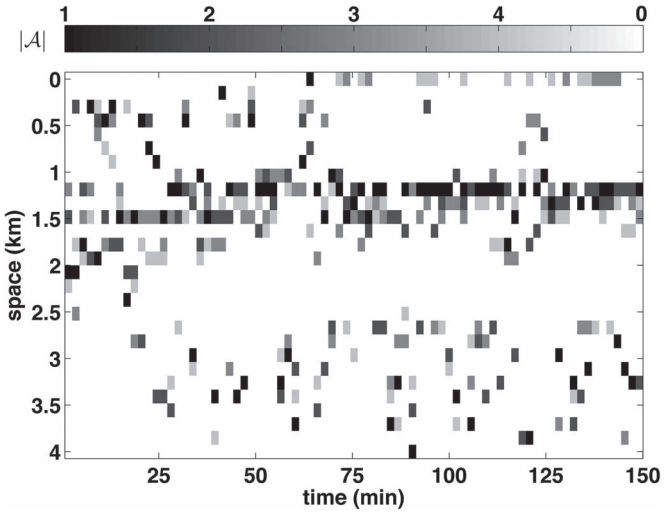


Fig. 8. Results of submodular analysis for $n_{MAX} = 20$. There are 79 covariance matrices, each of them is built with $|\mathcal{V}| = 28$ cells of about 140 m length and $|X_{\mathcal{V}}| = 112$ velocity samples (four times the rank of each $\Sigma_{\mathcal{V},\mathcal{V}}$) at one sample per second. The cell budget is $|\mathcal{A}| = 4$ on each time slot. The scale indicates the order in the lazy greedy selection process for $F(\mathcal{A})$ maximization. First choices (black cells) are located mainly at the upstream boundary.

second, third and fourth choices with increasing cell budgets, respectively. White cells are unselected ST regions with the above criteria. We can see that selected cells are not only located mainly at the upstream boundary of the jam but also at the lane merging ST regions, i.e., locations with the highest flow differences.

We represent the probability that at least one sensor vehicle is located in the selected cells ($P(|SV| \geq 1)$) in Fig. 9, according to their respective cell budgets ($|\mathcal{A}| = \{1, \ldots, 4\}$) and for different FSV. For a budget of 4 $\mathcal{A}^*$-ST regions per time slot, $P(|SV| \geq 1)$ suffers a sharp transition around a probe-vehicle density of $FSV = 9\%$. The interpretation is simple: Let us consider the best 4 ST regions in a given time interval and a mere one-in-ten proportion of sensor vehicles. Then, SV are almost surely sampling those regions with the highest reduction in the reconstruction error. In Fig. 6(a) and (b) it can be seen that the sources of error and speed variance are localized in the transitional zones between free and congested traffic. Thus, a $FSV = 9\%$ is needed to ensure that ST cells with the highest

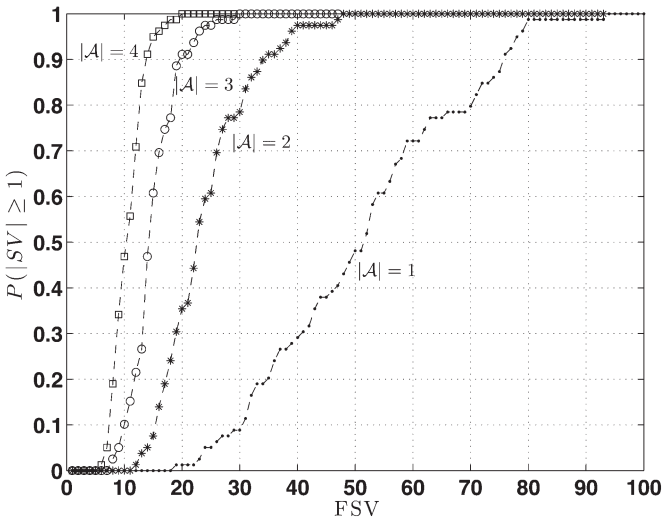


Fig. 9. $P(|SV| \geq 1)$ is the probability of the presence of at least one sensor vehicle in the shaded cells in Fig. 8, ranked in intensity.

flow differences are sampled by at least one sensor vehicle. We can conclude that it is not important that SV are found where car density is the highest (congestion) but where the highest flow differences are found.

V. Conclusion

At this point, we have opened a way quite clear to answer the questions proposed at the beginning of the work:

i) What is the relationship of FSV, T_s and the quality of the collected data? With $FSV = 9-10\%$ virtually all the phenomenology of traffic is captured, including the smaller transient ST regions. In these circumstances, it is possible to increase the sampling period T_s in regions with low speed variance (both free and heavily congested traffic regions). We can state that a FSV around 9–10% is sufficient to provide a high-quality traffic flow characterization.

ii) What is the relationship between the dynamics of traffic flow and the performance of multi-hop wireless network model? The VSN has two main sources of transmission errors: Doppler and multihop retransmission errors. When the traffic flow is fluid, the dominant source of transmission errors is the Doppler effect. In congested traffic, transmission errors are mainly caused by retransmissions from sensor vehicles to the RSUs.

iii) What is the influence of P_H in the reconstruction error of X_Ω? None of the traffic conditions stated above (free or congested) are relevant to reduce the error in the estimation of the STV field. Transitional traffic, where none of the two sources of transmission errors (Doppler and multihop errors) are dominant, is the key to reduce the estimation error. Thus, wireless transmission errors have a negligible effect in the reconstruction error.

iv) Why the sparse VSN and the STV field rendering succeed in capturing the traffic dynamics? We have found, by means of submodular optimization, the ST regions (cells) maximizing the expected reduction of the speed variance. We have calculated the probability of finding at least one sensor vehicle in these areas for varying density of sensor vehicles. We have found that such probability sharply increases with a mere 9% of the present vehicles periodically reporting their position, time and speed.

In summary, we have shown that the use of wireless ad-hoc networks in the context of traffic flow for collecting floating car data is technically feasible, even with very small penetration rates of the system. The reconstruction of space-time-velocity field and, in consequence, the characterization of space-time traffic patterns for a given set of roads, is achieved with less than 9% of involved vehicles.

References

[1] G. Leduc, "Road traffic data: Collection methods and applications," *Working Papers Energy, Transp. Climate Change*, vol. 1, p. 55, 2008.
[2] H. Rehborn, S. Klenov, and J. Palmer, "An empirical study of common traffic congestion features based on traffic data measured in the USA, the UK, and Germany," *Phys. A: Statist. Mech. Appl.*, vol. 390, no. 23, pp. 4466–4485, Nov. 2011.
[3] *INRIX XD Traffic*. [Online]. Available: http://www.inrix.com
[4] *Waze*. [Online]. Available: https://www.waze.com
[5] *Google Maps for Business*. [Online]. Available: http://www.google.com/intl/en/enterprise/mapsearth/
[6] *Dirección General de Tráfico*. [Online]. Available: http://infocar.dgt.es/etraffic/
[7] A. Petrie, D. Jordan, and J. Burr, *Method and System for Collecting Traffic Data*, Patent App. EP20 120 754 045, 2014, [Online]. Available: https://www.google.com/patents/EP2727098A1?cl=en
[8] A. Balasundaram, K. Foreman, R. Nandivada, and K. Yee, Traffic forecasting, Patent App. PCT/US2013/035 229. 2014, [Online]. Available: https://www.google.com/patents/WO2013154901A8?cl=en
[9] T. Jeske, "Floating car data from smartphones: What Google and Waze know about you and how hackers can control traffic," in *Proc. BlackHat Eur.*, 2013, pp. 1–12.
[10] B. Hoh *et al.*, "Enhancing privacy and accuracy in probe vehicle-based traffic monitoring via virtual trip lines," *IEEE Trans. Mobile Comput.*, vol. 11, no. 5, pp. 849–864, May 2012.
[11] A. Krause, E. Horvitz, A. Kansal, and F. Zhao, "Toward community sensing," in *Proc. Int. Conf. IPSN*, 2008, pp. 481–492.
[12] A. Krause, R. Rajagopal, A. Gupta, and C. Guestrin, "Simultaneous optimization of sensor placements and balanced schedules," *IEEE Trans. Autom. Control*, vol. 56, no. 10, pp. 2390–2405, Oct. 2011.
[13] G. Araniti, C. Campolo, M. Condoluci, A. Iera, and A. Molinaro, "LTE for vehicular networking: A survey," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 148–157, May 2013.
[14] R. Stanica, M. Fiore, and F. Malandrino, "Offloading Floating Car Data," in *Proc. IEEE 14th Int. Symp. WoWMoM*, 2013, pp. 1–9.
[15] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Comput. Netw.*, vol. 38, no. 4, pp. 393–422, Mar. 2002.
[16] O. Younis, M. Krunz, and S. Ramasubramanian, "Node clustering in wireless sensor networks: Recent developments and deployment challenges," *IEEE Netw.*, vol. 20, no. 3, pp. 20–25, May/Jun. 2006.
[17] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, "Wireless multimedia sensor networks: Applications and testbeds," *Proc. IEEE*, vol. 96, no. 10, pp. 1588–1605, Oct. 2008.
[18] S. Fujishige, *Submodular Functions and Optimization*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2005.
[19] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies," *J. Mach. Learn. Res.*, vol. 9, pp. 235–284, Jun. 2008.
[20] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Phys. Rev. E, Statist. Phys., Plasmas, Fluids, Related Interdisciplinary Topics*, vol. 62, pp. 1805–24, Aug. 2000.
[21] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model MOBIL for car-following models," *Transp. Res. Rec.*, vol. 1999, no. 1, pp. 86–94, Jan. 2007.
[22] W. Kaczynski, L. Leemis, N. Loehr, and J. McQueston, "Nonparametric random variate generation using a piecewise-linear cumulative

distribution function," *Commun. Statist.—Simul. Comput.*, vol. 41, no. 4, pp. 449–468, Dec. 2011.
[23] E. Morgado, I. Mora-Jimenez, J. J. Vinagre, J. Ramos, and A. J. Caamano, "End-to-end average BER in multihop wireless networks over fading channels," *IEEE Trans. Wireless Commun.*, vol. 9, no. 8, pp. 2478–2487, Aug. 2010.
[24] T. Wang, J. Proakis, and E. Masry, "Performance degradation of OFDM systems due to Doppler spreading," *IEEE Trans. Wireless Commun.*, vol. 5, no. 6, pp. 1422–1432, Jun. 2006.
[25] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf, *Computational Geometry: Algorithms and Applications*, 3rd ed. Berlin, Germany: Springer-Verlag, 2008.
[26] S. Rippa, "Minimal roughness property of the Delaunay triangulation," *Comput. Aided Geometric Des.*, vol. 7, no. 6, pp. 489–497, Nov. 1990.
[27] D. E. Knuth, *Art of Computer Programming, Volume 2: Seminumerical Algorithms*, 3rd ed. Reading, MA, USA: Addison-Wesley, Nov. 1997.
[28] T. F. Chan, G. H. Golub, and R. J. LeVeque, "Algorithms for computing the sample variance: Analysis and recommendations," *Amer. Statist.*, vol. 37, no. 3, pp. 242–247, 1983.
[29] A. Das and D. Kempe, "Algorithms for subset selection in linear regression," in *Proc. 14th Annu. ACM STOC*, May 2008, pp. 45–54.
[30] B. Nadler, "Finite sample approximation results for principal component analysis: A matrix perturbation approach," *Ann. Statist.*, vol. 36, no. 6, pp. 2791–2817, 2008.
[31] A. Krause, "SFO: A toolbox for submodular function optimization," *J. Mach. Learn. Res.*, vol. 11, pp. 1141–1144, 2010.

Eduardo del Arco received the B.Eng. degree in electrical and electronic engineering from Glyndwr University, Wrexham, U.K., in 2008; the M.Eng. degree in telecommunication engineering from the University of Alcalá, Alcalá de Henares, Spain, in 2009; and the M.Sc. degree in telecommunication engineering from the Rey Juan Carlos University, Fuenlabrada, Spain, in 2013. He is currently working toward the Ph.D. degree with the Department of Signal Theory and Communications, Rey Juan Carlos University. His research interests include wireless sensor networks, vehicular communications and submodular optimization.

Eduardo Morgado received the degree in telecommunication engineering from the Universidad Carlos III de Madrid, Leganés, Spain, in 2004 and the Ph.D. degree from the Rey Juan Carlos University, Fuenlabrada, Spain, in 2009. He is currently an Associate Professor with the Department of Signal Theory and Communications, Rey Juan Carlos University. His research interests include signal processing for wireless communications with applications to ad hoc and sensor networks.

Mihaela I. Chidean received the B.Sc. degree in telecommunication engineering and computer systems engineering from the Rey Juan Carlos University (URJC), Fuenlabrada, Spain, in 2011 and the M.Sc. degree in multimedia and communications from the Carlos III University of Madrid, Leganés, Spain, in 2013. She is currently working toward the Ph.D. degree with the Department of Signal Theory and Communications, University Rey Juan Carlos in Madrid, Spain. Her research interests include body sensor networks with medical applications, physiological signal processing, dynamic routing, distributed signal processing, and wireless sensor networks for energy efficiency.

Julio Ramiro-Bargueño received the B.S. degree in physics, the Advanced Studies Diploma, and the Ph.D. degree in physics from the Universidad Autónoma de Madrid, Madrid, Spain, in 1991, 1993, and 1997, respectively. Since 2004, he has been an Associate Professor with the Department of Signal Theory and Communications, Rey Juan Carlos University, Fuenlabrada, Spain. His research interest is focused on wireless communications, wireless sensor networks, and new communications for vehicular applications.

Inmaculada Mora-Jiménez received the degree in telecommunication engineering from the Universidad Politécnica de Valencia, Valencia, Spain, in 1998, and the Ph.D. degree from the Universidad Carlos III de Madrid, Leganés, Spain, in 2004. She is currently an Associate Professor with the Department of Signal Theory and Communications, Rey Juan Carlos University, Fuenlabrada, Spain. Her main research interests include statistical learning theory, neural networks, and their applications to image processing, bioengineering, and communications.

Antonio J. Caamaño received the B.Sc. and M.Sc. degrees in theoretical physics in 1995 from the Universidad Autónoma de Madrid, Madrid, Spain, and the Ph.D. degree in telecommunications engineering from the Carlos III University of Madrid, Leganés, Spain, in 2003. Since 2003, he has been an Associate Professor with the Department of Signal Theory and Communications, Rey Juan Carlos University, Fuenlabrada, Spain. His main research interests lie in the fields of MANET optimization, bioengineering, and statistical signal processing.