AIAA SciTech Forum
8-12 January 2024, Orlando, FL
AIAA SCITECH 2024 Forum

10.2514/6.2024-0529

Modeling and Detection of Cyber-Attacks in UAV Swarms using a 2D-LWR Model and Gaussian Processes

Abhishek Kashyap*, Animesh Chakravarthy†, Kamesh Subbarao‡
The University of Texas at Arlington, Arlington, TX 76019

David Casbeer§, Isaac Weintraub¶, Brandon Hencey‖
Air Force Research Laboratory, Wright Patterson AFB, OH 45433

This paper considers a class of cyber-attacks attacking a swarm of Unmanned Aerial Vehicles (UAVs). Our focus is on scenarios wherein an attacker may hack into a subset of vehicles in the swarm and create subtle changes in their parameters. These hacked vehicles (referred to as malicious vehicles) are subsequently able to modify the behavior of the overall swarm. The swarm comprising the mix of malicious and normal vehicles is modeled using a system of coupled Partial Differential Equations (PDEs) in a two-dimensional LWR model. We develop a methodology that combines Gaussian Processes (GP) with this two-species 2D PDE model, and use this method for detecting the presence of such malicious vehicles in the swarm. A Bayesian Optimization scheme is employed to determine the optimal choice of basis and kernel functions that constitute the GP. Simulation results demonstrate that this detection architecture performs successful detection of the malicious vehicles, and also their mode of attack on the traffic.

I. INTRODUCTION

The study of potential cyber-attacks in different domains is an active area of research. Given that systems are becoming more interconnected, cyber physical systems that operate infrastructure and/or plants can make these assets more vulnerable and open to different attack vectors. Any failure will impact safety, with associated financial and societal ramifications. Many scenarios wherein cyber-attacks can occur are reported in the literature. These include, for example, smart grid attacks [1], attacks on gas transmission and distribution networks [2], large-scale process engineering plants [3], water networks, Unmanned Aerial Vehicles [4], and automobiles [5]. The detection of such attacks is an area of considerable research interest [6–9].

In this paper, our focus is on detecting cyber-attacks performed in swarms of UAVs. The class of cyber-attacks considered are those wherein an attacker hacks into the guidance and control program of a carefully chosen subset of vehicles within a UAV swarm, and converts them into vehicles with malicious intent. These malicious vehicles then perform a series of subtle changes in the way they interact with the other vehicles in the swarm, with the eventual objective to degrade the mission effectiveness of the swarm. We model the swarm by using partial differential equations (PDEs).

The use of PDEs to model multi-vehicle systems has a fairly long history, with the earliest origins being the Lighthill-Whitham Richards (LWR) model [10], which was used to model automotive traffic flows on highways. The LWR model is a first order equation governing the spatio-temporal evolution of density of vehicles, and represents the conservation of cars on a highway. It also has applicability in two-dimensional multi-vehicle systems including swarms of UAVs.

A preliminary method which used gas-kinetic based PDEs to model such attacks was presented in [11]. In [12], the authors employed a two-species one-dimensional LWR model for automotive traffic on a highway. The two species are normal and malicious vehicles, and the malicious vehicles (arbitrarily distributed among the normal vehicles) may perform *subtle* speed and/or lane changes, with an objective to force the traffic to either slow down, or speed up, and thereby attain an equilibrium velocity-density relationship which is different from that desired by the normal

*PhD student, Mechanical and Aerospace Engineering, University of Texas at Arlington, Arlington, TX
†Professor, Mechanical and Aerospace Engineering, University of Texas at Arlington, Arlington, TX
‡Professor, Mechanical and Aerospace Engineering, University of Texas at Arlington, TX
§Technical Area Lead, Control Science Center, AFRL
¶Electronics Engineer, Control Science Center, AFRL
‖Senior Mechanical Engineer, Aerospace Systems Directorate, AFRL

DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited. AFRL-2023-6124

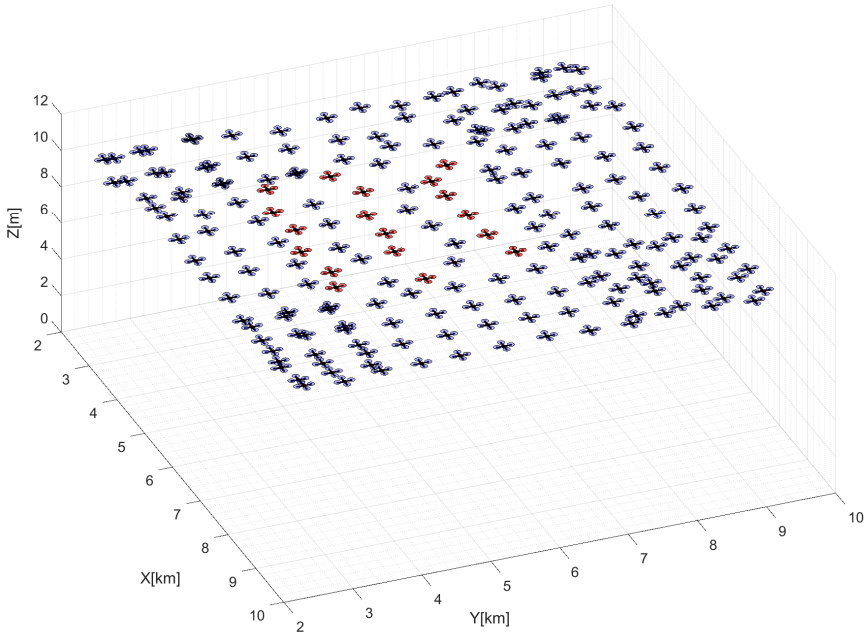


Fig. 1 UAV swarm comprising of normal (blue) and malicious (red) UAVs

vehicles. This paper advances the work of our previous paper [12] in the following ways: (a) Our framework now considers cyber-attacks occurring in a UAV swarm operating in a 2-D environment at constant altitude, by considering a two-species model involving normal and malicious vehicles, wherein the malicious vehicles seek to disrupt the equilibrium velocity-density profile of the normal vehicles. (b) A linearized analysis of the propagation of perturbations in the malicious and normal vehicles is performed and this helps to determine the direction of travel of the maximal percentage of malicious vehicles. (c) The two-species 2D LWR model is then integrated with a Gaussian Process (GP) framework, which is used to perform detection of the malicious vehicles in the swarm. This detection framework utilizes readings from multiple stationary sensors overseeing the swarm, that are used to train the GPR model. The trained GPR model is then used by a mobile sensor (which may be another UAV flying over the swarm at a higher altitude) moving in the direction of the wave of malicious vehicles (as predicted by the linearized analysis) to estimate the percentage of malicious vehicles in the swarm.

The rest of the paper is organised as follows. Section II defines the problem statement. In Section III, a brief description of the 2D LWR model for the single-species and two-species cases, respectively, are presented. An analysis of the linearized PDE models is performed in Section IV. In Section V, the overall solution architecture for attack detection is presented. A description of the Gaussian Process Regression (GPR) model and its implementation is given in Section VI. The efficacy of the methodology is demonstrated by simulations in Section VII.

II. MOTIVATION AND PROBLEM STATEMENT

Consider a swarm of UAVs flying at a constant altitude as shown in Fig 1. The swarm comprises a mix of normal vehicles (depicted in blue) and malicious vehicles (depicted in red). The malicious vehicles intend to modify the overall flow of vehicles in the swarm. During such an attack, it can be important to determine the location and density of the malicious vehicles in the swarm.

Let $\rho(x, y, t)$ represent the average density of vehicles (in terms of vehicles per unit area) and $U(x, y, t)$, $V(x, y, t)$ represent the average velocities of the vehicles along the x and y directions, respectively. Let $x \in [0, L_x]$ and $y \in [0, L_y]$, where L_x, L_y represent the dimensions of the 2D rectangular domain. Let $\rho_M(x, y, t)$ represent the density of malicious vehicles and $\rho_N(x, y, t)$ represent that of normal vehicles, such that $\rho_M + \rho_N = \rho$. Let p represent a parameter influencing the velocities U and V - it governs the attack mode of the malicious vehicles. Assume there is a single

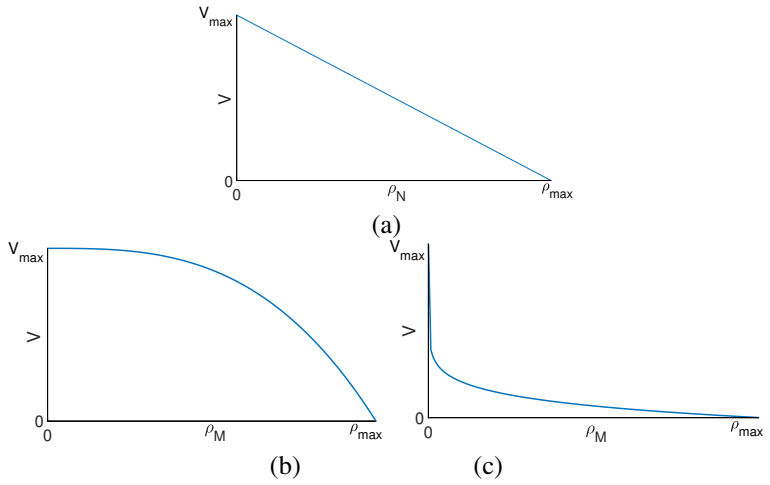


Fig. 2 Fundamental diagram between density and velocity for normal and malicious vehicles: (a) V **vs.**ρ_N**, (b)** V **vs.** ρ_M $(p > 1)$**, (c)** V **vs.** ρ_M $(p < 1)$

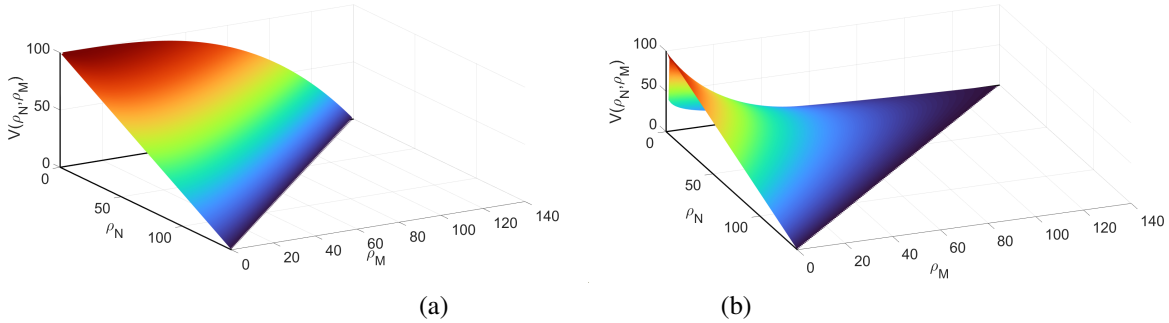


Fig. 3 Fundamental diagram for varying ρ_N **and** ρ_M **a)** $p > 1$ **b)** $p < 1$

mobile sensor which has the ability to measure the values of ρ, U, V, along its trajectory in the spatial (x, y) domain. The objective is to determine a suitable trajectory of the UAV, in conjunction with a methodology by which it can use the measurements along its trajectory to determine the percentage of malicious vehicles $\rho_M\% \equiv \rho_M/\rho \times 100$ and p.

III. PDE SWARM MODEL

The 1D single-species LWR model [10] has been extensively used in the literature due to its simplicity [13–15]. In [12], a 1D two-species LWR model was introduced, wherein the malicious species (M) interacts with the normal species (N) and influences the velocity of the normal vehicles based on their mode of attack as well as their own relative density. In the subsequent sections, we first present the $2D$ LWR model for the $2D$ UAV swarm comprising of normal vehicles alone, and we then extend our $1D$ two-species traffic model to a $2D$ two-species UAV swarm model.

A. 2D LWR Single-Species Model

Extending the $1D$ LWR traffic model to two-dimensions, the flow of vehicles in the swarm is modeled in terms of its average density $\rho(x, y, t)$ (in vehicles/km^2) and average vehicle velocities $U(x, y, t)$ and $V(x, y, t)$ along the x and y axes, respectively, as follows:

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho U)}{\partial x} + \frac{\partial (\rho V)}{\partial y} = 0 \tag{1}$$

Here, $U(x, y, t)$ and $V(x, y, t)$ are defined as:

$$\begin{aligned} U(x, y, t) &= U_{max}\left(1 - \frac{\rho(x, y, t)}{\rho_{max}}\right) \\ V(x, y, t) &= V_{max}\left(1 - \frac{\rho(x, y, t)}{\rho_{max}}\right) \end{aligned} \tag{2}$$

where, ρ_{max},U_{max} and V_{max} are the maximum density and maximum velocities in the x and y directions respectively. Fig 2(a) shows the above relationship between density and velocity along the y-axes direction. A similar relationship exists between density and velocity in the x direction. Thus it is assumed that the normal vehicles desire to maintain a linear velocity-density relationship, such that their speed decreases linearly with increasing density of vehicles in the swarm.

B. 2D LWR Two-Species Model

The $2D$ LWR model for the two-species system is as follows:

$$\begin{aligned} \frac{\partial \rho_N}{\partial t} + \frac{\partial(\rho_N U)}{\partial x} + \frac{\partial(\rho_N V)}{\partial y} &= 0 \\ \frac{\partial \rho_M}{\partial t} + \frac{\partial(\rho_M U)}{\partial x} + \frac{\partial(\rho_M V)}{\partial y} &= 0 \end{aligned} \tag{3}$$

where, $\rho_N(x, y, t)$ and $\rho_M(x, y, t)$ represent the average densities of the normal and malicious vehicles, respectively, in the swarm.

Along lines similar to the $1D$ LWR two-species LWR model given in [12], it is assumed that the malicious species of vehicles seek to alter the linear velocity-density relationship shown in Fig 2(a). Depending on their mode of attack, the malicious vehicles may either speed up the vehicles in the swarm (as demonstrated in Fig 2(b), or slow them down (as demonstrated in Fig 2(c)). When all vehicles are malicious, they adopt a velocity-density relationship as shown in Figs 2(b),(c). When a subset of the vehicles in the swarm are malicious, then the quantities $U(x, y, t)$ and $V(x, y, t)$ have the following representation:

$$\begin{aligned} U(x, y, t) &= U_{max}\left(1 - \left(\frac{\rho_N(x, y, t) + \rho_M(x, y, t)}{\rho_{max}}\right)^{p(x,y,t)}\right) \\ V(x, y, t) &= V_{max}\left(1 - \left(\frac{\rho_N(x, y, t) + \rho_M(x, y, t)}{\rho_{max}}\right)^{p(x,y,t)}\right) \end{aligned} \tag{4}$$

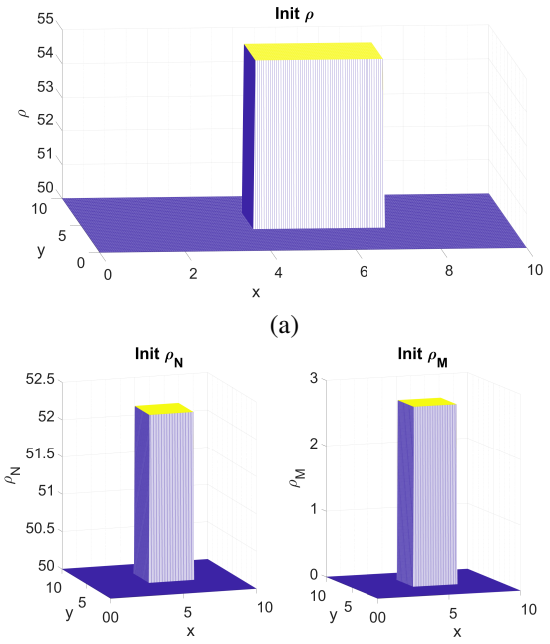
where $\rho_N(x, y, t)$, $\rho_M(x, y, t)$ represent the average densities of the normal and malicious vehicles respectively. The quantity $p(x, y, t)$ is given by $p(x, y, t) = 1 + k(\rho_M(x, y, t)/(\rho_M(x, y, t) + \rho_N(x, y, t)))$ where k can be either positive or negative. When there are no malicious vehicles (that is, $\rho_M = 0$), then $p = 1$ and the two-species case reduces to the single-species case where all vehicles are normal.

Fig. 3(a),(b) shows the equilibrium velocity V_e as a function of (ρ_N, ρ_M) for $p > 1$ and $p < 1$, respectively. When $p > 1$ (Fig 3(a)), as ρ_M% increases, the equilibrium velocity becomes higher compared to a single species case with the same total density. On the other hand, when $p < 1$ (Fig 3(b)), then as ρ_M% increases, the equilibrium velocity is lower for the two-species case compared to the single-species case for the same total density.

C. Comparison of spatio-temporal evolution of density in the Single and Two-species Models

In order to understand the behaviour of the single- and two-species $2D$ LWR models, a 10 km × 10 km area is chosen as the domain. On this domain, an initial distribution of normal vehicles shown in Fig 4(a), is considered. Here, all the vehicles are at an equillibrium density of 50 vehicle/km^2, with a small perturbation of normal vehicles in the middle of the domain ($x \in [4, 7]\ km$, $y \in [4, 7]\ km$). The LWR single species equation (Eq 1) is then used to simulate the flow of vehicles on this area and the density distribution at 129.6 s is shown in Fig. 5. In this scenario $p = 1$ and velocity and density follow a linear relationship.

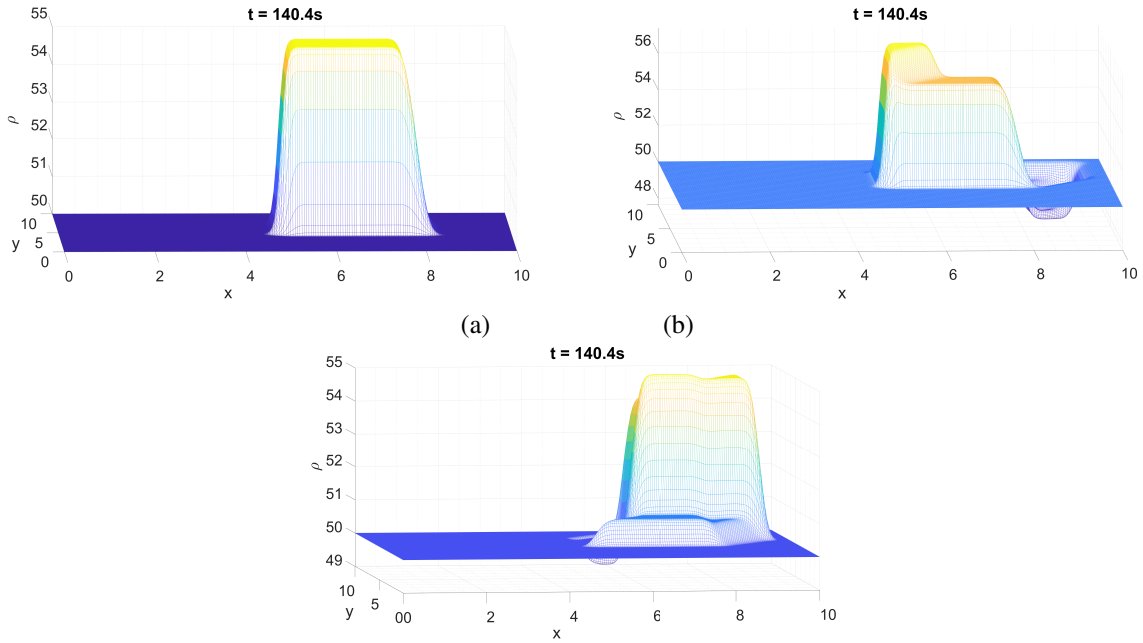
Next, a two-species scenario is considered where 5% of the normal vehicles in the perturbation of the initial single-species density distribution are replaced by malicious vehicles such that the total density distribution is identical to



(a)

(b)

Fig. 4 Initial density distributions for (a) Single species (b) Two species



(a)

(b)

(c)

Fig. 5 Density distributions at 108 s **(a) Single species (p = 1) (b) Two species** $(p < 1)$ **(c) Two species case** $(p > 1)$

that used for the single-species simulation. Fig. 4(b) shows this initial density distribution for the normal and malicious vehicles. The LWR two-species equation (Eq 3) is then used to simulate the swarm for this scenario with the parameter $p < 1$. Fig 5 (b) shows the density distribution at 129.6 s. Comparing the density distribution plots for both cases, it can be seen that though both perturbations propagate towards the $(10, 10)$ corner, the perturbation for the single-species case more or less retains the uniform shape at the initial conditions while that for the two-species case is uneven, with a higher density of vehicles at the back. Increased percentage of malicious vehicles can cause a higher density change at the back and also cause shocks impeding the flow of vehicles.

Next, changing the value of p so the scenario corresponds to $p > 1$ for the two-species case and again simulating the flow of vehicles, we get the density distribution at 129.6 s as shown in Fig 5(c). Again it can be seen that the density distribution for this case is quite different compared to the single-species case with $p = 1$. It is evident that the perturbation has propagated faster to the $(10, 10)$ corner, when compared to the single-species case.

IV. Determining Propagation of Perturbations in the Two-Species PDE Model using a Linearized Analysis

In this section, we perform a linearized analysis of the PDE model given in (3)-(4), and perform a wave evolution analysis. Assume a density distribution of vehicles as shown in Fig. 4 with equilibrium density of normal vehicles ρ_{Ne} and initial perturbations in the malicious and normal vehicles of $\Delta\rho_{M0}$ and $\Delta\rho_{N0}$, respectively, with these perturbations occurring at $x \in [x_{10}, x_{20}]$, $y \in [y_{10}, y_{20}]$. The linearized PDEs are as follows:

$$\begin{aligned}&\frac{\partial \Delta\rho_N}{\partial t} + \left[\left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right) U_{max}\right] \frac{\partial \Delta\rho_N}{\partial x} + \left[\left(-\frac{\rho_{Ne}}{\rho_{max}}\right) U_{max}\left(1 + k\log\left(\frac{\rho_{Ne}}{\rho_{max}}\right)\right)\right] \frac{\partial \Delta\rho_M}{\partial x} \\ &+ \left[\left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right) V_{max}\right] \frac{\partial \Delta\rho_N}{\partial y} + \left[\left(-\frac{\rho_{Ne}}{\rho_{max}}\right) V_{max}\left(1 + k\log\left(\frac{\rho_{Ne}}{\rho_{max}}\right)\right)\right] \frac{\partial \Delta\rho_M}{\partial y} = 0\end{aligned} \tag{5}$$

$$\frac{\partial \Delta\rho_M}{\partial t} + \left[\left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right) U_{max}\right] \frac{\partial \Delta\rho_M}{\partial x} + \left[\left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right) V_{max}\right] \frac{\partial \Delta\rho_M}{\partial y} = 0 \tag{6}$$

Writing (Eq)5) and (6) as a system of equations in matrix form, we have;

$$\begin{aligned}&\begin{bmatrix} \frac{\partial \Delta\rho_N}{\partial t} \\ \frac{\partial \Delta\rho_M}{\partial t} \end{bmatrix} + \mathbf{M_x} \begin{bmatrix} \frac{\partial \Delta\rho_N}{\partial x} \\ \frac{\partial \Delta\rho_M}{\partial x} \end{bmatrix} + \mathbf{M_y} \begin{bmatrix} \frac{\partial \Delta\rho_N}{\partial y} \\ \frac{\partial \Delta\rho_M}{\partial y} \end{bmatrix} = 0 \\ &\begin{bmatrix} \frac{\partial \Delta\rho_N}{\partial t} \\ \frac{\partial \Delta\rho_M}{\partial t} \end{bmatrix} + \begin{bmatrix} \left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right) U_{max} & \left(-\frac{\rho_{Ne}}{\rho_{max}}\right) U_{max}\left(1 + k\log\left(\frac{\rho_{Ne}}{\rho_{max}}\right)\right) \\ 0 & \left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right) U_{max} \end{bmatrix} \begin{bmatrix} \frac{\partial \Delta\rho_N}{\partial x} \\ \frac{\partial \Delta\rho_M}{\partial x} \end{bmatrix} \\ &+ \begin{bmatrix} \left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right) V_{max} & \left(-\frac{\rho_{Ne}}{\rho_{max}}\right) V_{max}\left(1 + k\log\left(\frac{\rho_{Ne}}{\rho_{max}}\right)\right) \\ 0 & \left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right) V_{max} \end{bmatrix} \begin{bmatrix} \frac{\partial \Delta\rho_N}{\partial y} \\ \frac{\partial \Delta\rho_M}{\partial y} \end{bmatrix} = 0\end{aligned} \tag{7}$$

The initial conditions for the perturbations are given by

$$\begin{bmatrix} \Delta\rho_N(x,0) \\ \Delta\rho_M(x,0) \end{bmatrix} = \begin{cases} \begin{bmatrix} \Delta\rho_{N0} \\ \Delta\rho_{M0} \end{bmatrix}, & x_{10} \le x \le x_{20}, y_{10} \le y \le y_{20} \\ 0, & \text{elsewhere} \end{cases} \tag{8}$$

To solve the above system of PDEs, the equations need to be transformed into a diagonal form. Hence, the matrices

$\mathbf{M_x}$, $\mathbf{M_y}$ are diagonalized as $\mathbf{M_x} = \mathbf{S}\mathbf{J_x}\mathbf{S^{-1}}$ and $\mathbf{M_y} = \mathbf{S}\mathbf{J_y}\mathbf{S^{-1}}$ respectively, where

$$\mathbf{J_x} = \begin{bmatrix} \left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right) U_{max} & 0 \\ 0 & \left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right) U_{max} \end{bmatrix}, \quad \mathbf{J_y} = \begin{bmatrix} \left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right) V_{max} & 0 \\ 0 & \left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right) V_{max} \end{bmatrix}$$
$$\mathbf{S} = \begin{bmatrix} 1 & -1 - k \log\left(\frac{\rho_{Ne}}{\rho_{max}}\right) \\ 0 & 1 \end{bmatrix}, \quad \mathbf{S^{-1}} = \begin{bmatrix} 1 & 1 + k \log\left(\frac{\rho_{Ne}}{\rho_{max}}\right) \\ 0 & 1 \end{bmatrix} \tag{9}$$

Transforming the variables $\mathbf{\Delta\rho} = \begin{bmatrix} \Delta\rho_N & \Delta\rho_M \end{bmatrix}^T$, as $\tilde{\mathbf{\Delta\rho}} = \boldsymbol{S^{-1}}\mathbf{\Delta\rho}$, we get

$$\begin{aligned} \tilde{\mathbf{\Delta\rho}} = \begin{bmatrix} u \\ v \end{bmatrix} &= \begin{bmatrix} 1 & 1 + k \log\left(\frac{\rho_{Ne}}{\rho_{max}}\right) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta\rho_N \\ \Delta\rho_M \end{bmatrix} \\ &= \begin{bmatrix} \Delta\rho_N + \left[1 + k \log\left(\frac{\rho_{Ne}}{\rho_{max}}\right)\right] \Delta\rho_M \\ \Delta\rho_M \end{bmatrix} \end{aligned} \tag{10}$$

Thus, the PDEs governing the transformed variables are given by

$$\begin{aligned} &\begin{bmatrix} \frac{\partial u}{\partial t} \\ \frac{\partial v}{\partial t} \end{bmatrix} + \begin{bmatrix} \left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right) U_{max} & 0 \\ 0 & \left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right) U_{max} \end{bmatrix} \begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial x} \end{bmatrix} \\ &+ \begin{bmatrix} \left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right) V_{max} & 0 \\ 0 & \left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right) V_{max} \end{bmatrix} \begin{bmatrix} \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial y} \end{bmatrix} = 0 \end{aligned} \tag{11}$$

The initial conditions for the transformed variables are given by

$$\begin{bmatrix} u_0 \\ v_0 \end{bmatrix} = \begin{cases} \begin{bmatrix} \Delta\rho_{N0} + \left[1 + k \log\left(\frac{\rho_{Ne}}{\rho_{max}}\right)\right] \Delta\rho_{M0} \\ \Delta\rho_{M0} \end{bmatrix}, & x_{10} \le x \le x_{20}, y_{10} \le y \le y_{20} \\ 0, & \text{elsewhere} \end{cases} \tag{12}$$

Solving the above system of PDEs (11), (12) using the method of characteristics, the equations governing the transformed variables are given by:

$$\begin{aligned} u(x,t) &= \begin{cases} u_0, & x_{10} + U_{max}\left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right) \le x \le x_{20} + U_{max}\left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right), \\ & y_{10} + V_{max}\left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right) \le y \le y_{20} + V_{max}\left(1 - \frac{2\rho_{Ne}}{\rho_{max}}\right) \\ 0, & \text{elsewhere} \end{cases} \\ v(x,t) &= \begin{cases} v_0, & x_{10} + U_{max}\left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right) \le x \le x_{20} + U_{max}\left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right), \\ & y_{10} + V_{max}\left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right) \le y \le y_{20} + V_{max}\left(1 - \frac{\rho_{Ne}}{\rho_{max}}\right) \\ 0, & \text{elsewhere} \end{cases} \end{aligned} \tag{13}$$

Finally, the equations governing the propagation of the perturbations $\Delta\rho_N(x,t)$ and $\Delta\rho_M(x,t)$ are given by the following transformation $\mathbf{\Delta\rho} = \boldsymbol{S}\tilde{\mathbf{\Delta\rho}}$.

$$\Delta\rho_N(x,t) = A + B \tag{14}$$

where

$$A = \begin{cases} \Delta\rho_{N0} + w\Delta\rho_{M0}, & x_{10} + U_{max}\left(1 - \dfrac{2\rho_{Ne}}{\rho_{max}}\right)t \le x \le x_{20} + U_{max}\left(1 - \dfrac{2\rho_{Ne}}{\rho_{max}}\right)t, \\ & y_{10} + V_{max}\left(1 - \dfrac{2\rho_{Ne}}{\rho_{max}}\right)t \le y \le y_{20} + V_{max}\left(1 - \dfrac{2\rho_{Ne}}{\rho_{max}}\right)t \\ 0, & \text{elsewhere} \end{cases}$$

$$B = \begin{cases} -w\Delta\rho_{M0}, & x_{10} + U_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)t \le x \le x_{20} + U_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)t, \\ & y_{10} + V_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)t \le y \le x_{20} + V_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)t \\ 0, & \text{elsewhere} \end{cases}$$

$$w = \left[1 + k \log\left(\frac{\rho_{Ne}}{\rho_{max}}\right)\right]$$

Thus, the perturbation $\Delta\rho_N$ is composed of two parts A and B: A has positive amplitude and propagates with velocity components $U_{max}\left(1 - \dfrac{2\rho_{Ne}}{\rho_{max}}\right)$ along the x direction and $V_{max}\left(1 - \dfrac{2\rho_{Ne}}{\rho_{max}}\right)$ along the y direction. B has negative amplitude and propagates with velocity components $U_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)$ in x direction and $V_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)$ in y direction respectively.

$$\Delta\rho_M(x,t) = \begin{cases} \Delta\rho_{M0}, & x_{10} + U_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)t \le x \le x_{20} + U_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)t, \\ & y_{10} + V_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)t \le x \le y_{20} + V_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)t \\ 0, & \text{elsewhere} \end{cases} \tag{15}$$

Thus, the perturbation $\Delta\rho_M$ has positive amplitude and propagates with velocity component $U_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)$ along the x direction and velocity component $V_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)$ along the y direction, the same wave velocity as Part B of $\Delta\rho_N$. Also in both perturbations, the wave velocities are functions of the equilibrium density ρ_{Ne} and swarm properties $U_{max}, V_{max}, \rho_{max}$ and the attack parameter k only influences the amplitude of the perturbations in normal vehicles alone.

Interpretation of Solutions

In order to physically interpret these results, let us consider the following two $1D$ traffic cases, where there is an initial equilibrium distribution of normal vehicles and a small perturbation comprising of malicious and normal vehicles as shown in Fig 6 and Fig 7. In the first case, the equilibrium density is less than the critical density $\rho_{crit} = 0.5\rho_{max}$, where ρ_{max} is taken as 140 vehicles/km^2. Fig 8 shows the propagation of the perturbations in $\rho = \rho_N + \rho_M$, ρ_N and ρ_M in space and time. The corresponding figures on the right hand side show the respective top views. It can be seen that in the plot for ρ_N, the perturbation $\Delta\rho_N$ is composed of two parts. Part A is a positive perturbation which moves with the wave velocity $U_{max}\left(1 - \dfrac{2\rho_{Ne}}{\rho_{max}}\right)$ in the x direction while Part B is a negative perturbation which moves with the wave velocity $U_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)$ in the x direction. Also both wave velocities are positive and the perturbation move towards the $x = 10km$ side. In the plot for ρ_M, it is seen that the perturbation $\Delta\rho_M$ is a positive perturbation which which moves with the wave velocity $U_{max}\left(1 - \dfrac{\rho_{Ne}}{\rho_{max}}\right)$ in the x direction. Since the wave velocity for $\Delta\rho_M$ is the same as for Part B, it also moves towards the $x = 10km$ side.

In the second case, the equilibrium density is greater than the critical density $\rho_{crit} = 0.5\rho_{max}$. Fig 9 shows the propagation of the perturbations in $\rho = \rho_N + \rho_M$, ρ_N and ρ_M in space and time. The corresponding figures on the right

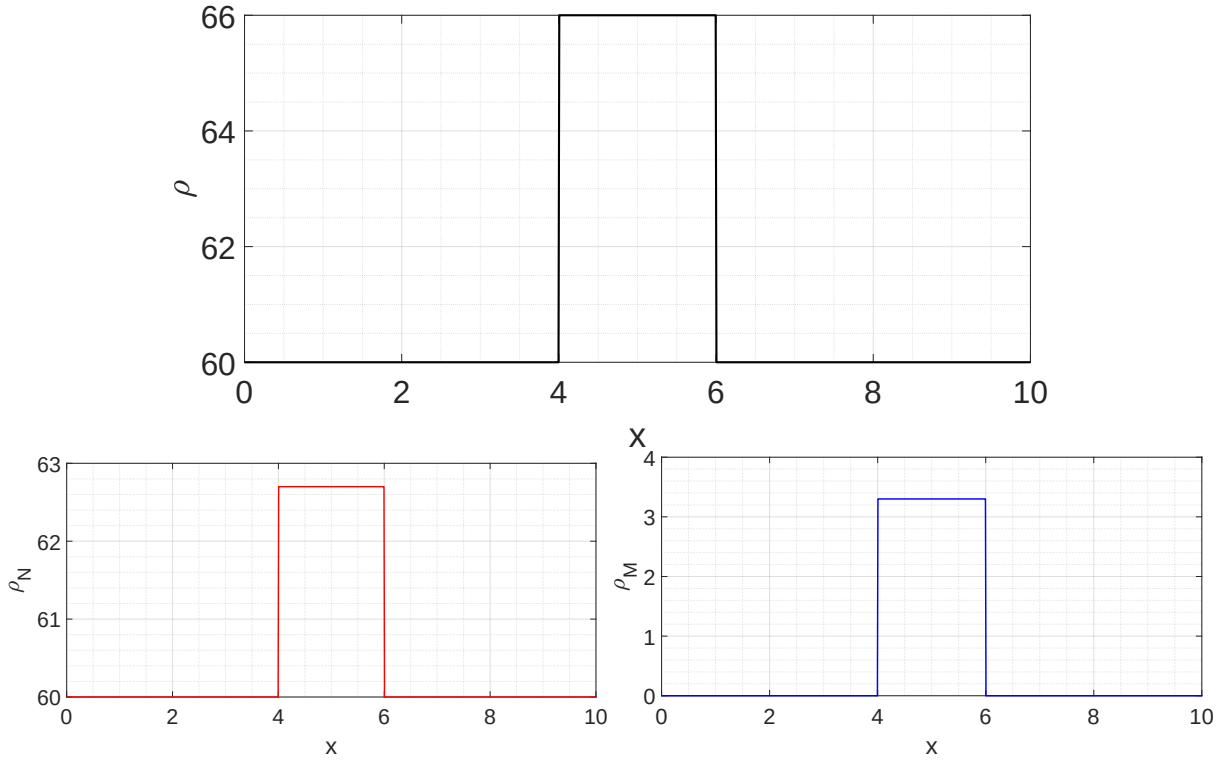


Fig. 6 Initial (ρ, ρ_N, ρ_M) distribution for the linearized analysis (Case 1)

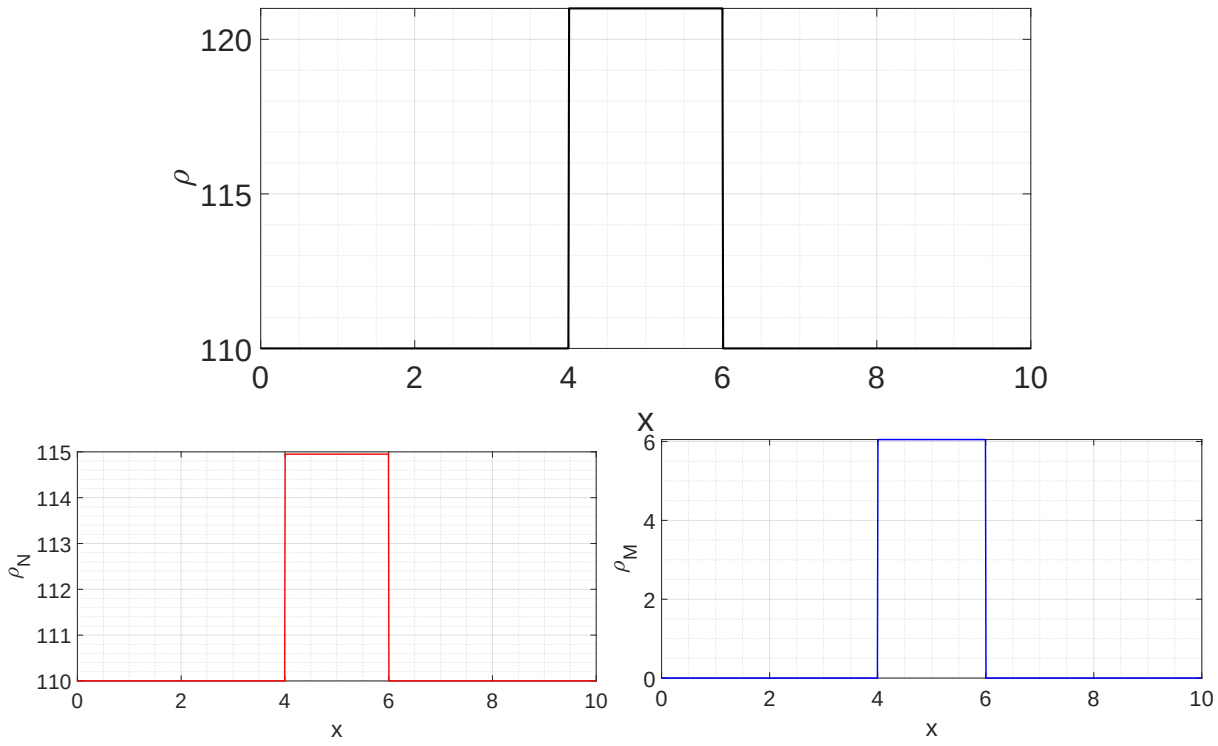


Fig. 7 Initial density (ρ, ρ_N, ρ_M) distribution for the linearized analysis (Case 2)

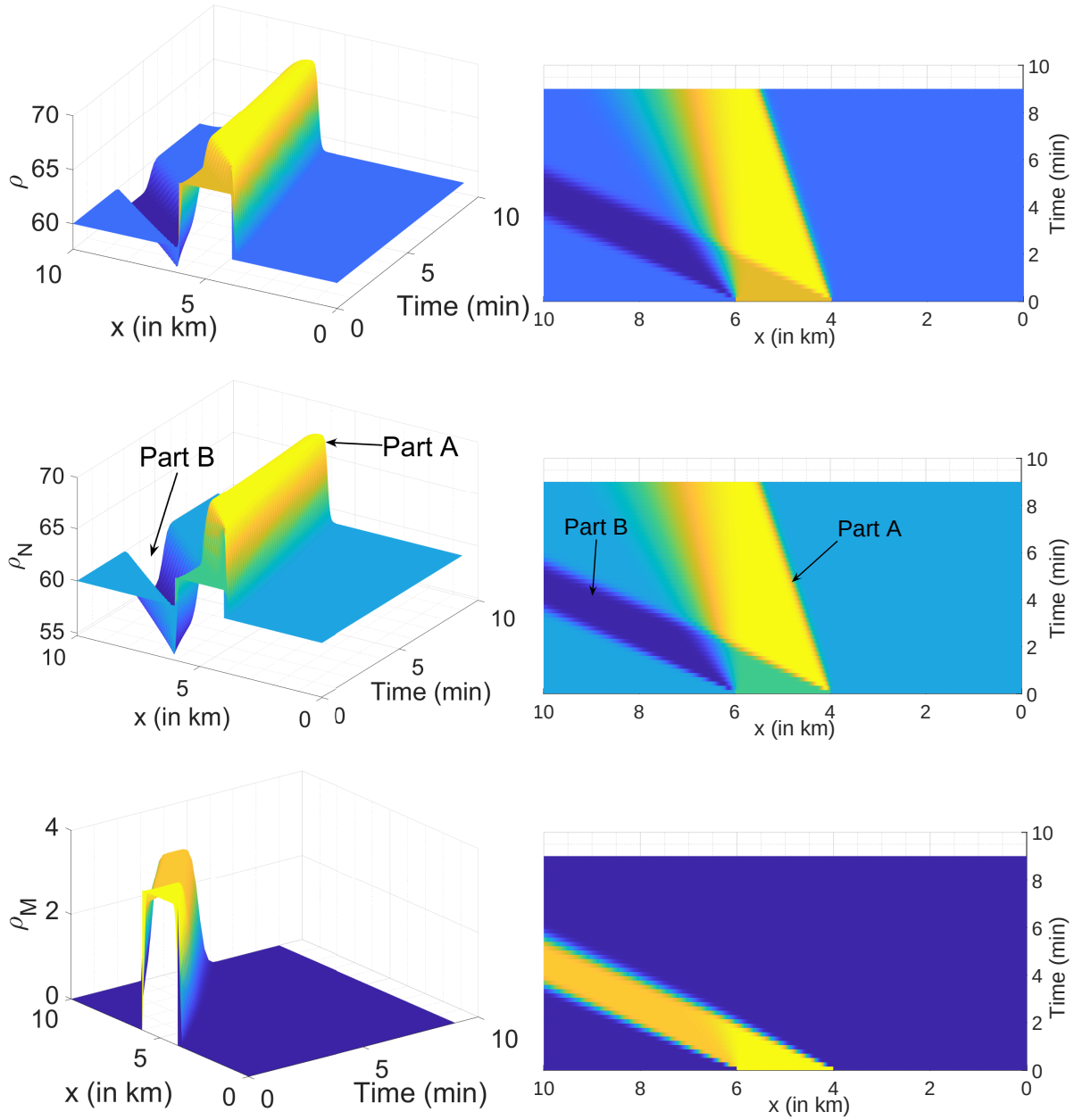


Fig. 8 Propagation of the perturbations in density (ρ, ρ_N, ρ_M) for Case 1. Figures in the right column show the top view of the figures in the left column to illustrate the propagation of perturbations

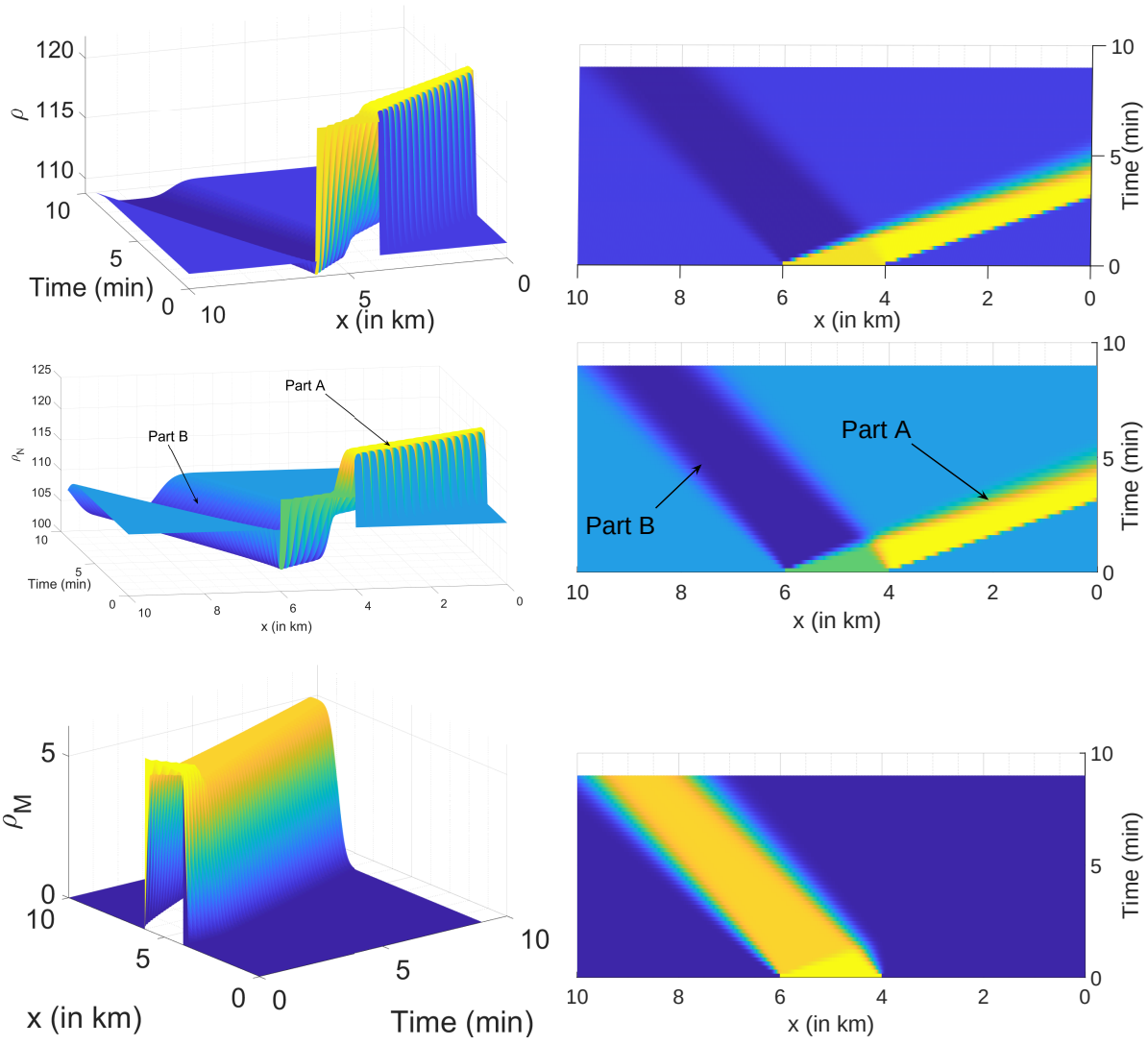


Fig. 9 Propagation of the perturbations in density (ρ, ρ_N, ρ_M) for Case 2. Figures in the right column show the top view of the figures in the left column to illustrate the propagation of perturbations

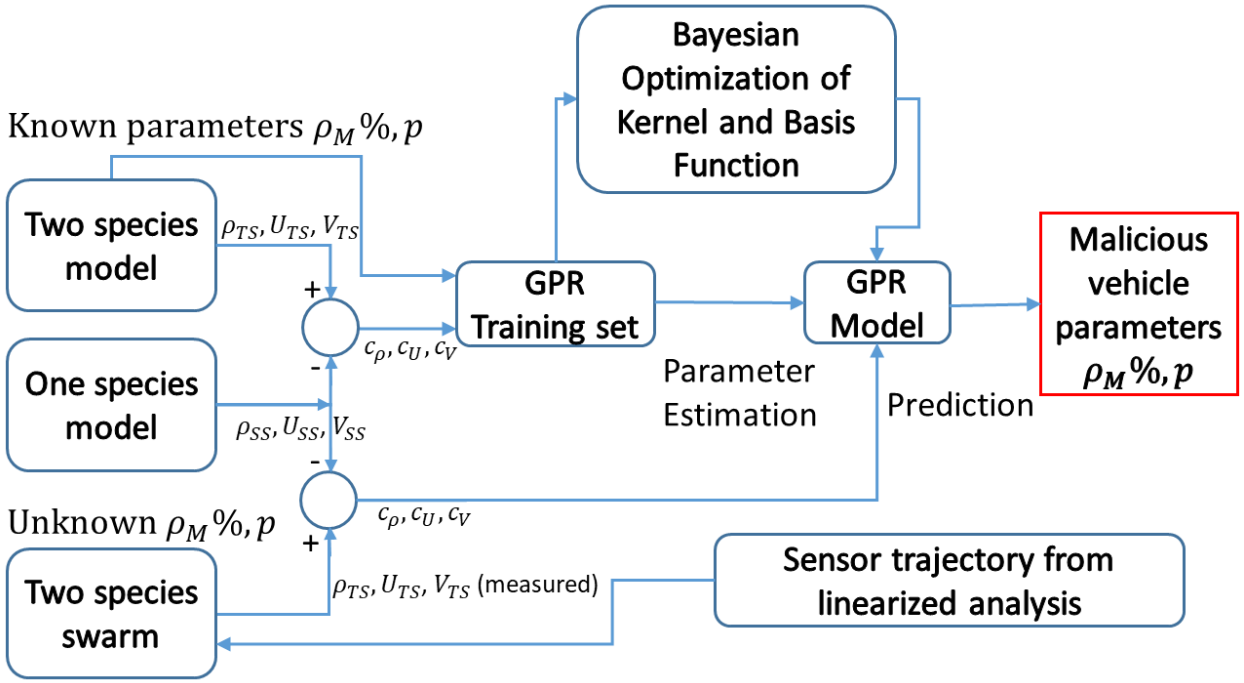


Fig. 10 Flowchart of the Algorithm

hand side show the respective top views. Similar to Case 1, the perturbation $\Delta\rho_N$ is composed of two parts. Part A is a positive perturbation which moves with the wave velocity $U_{max}\left(1-\frac{2\rho_{Ne}}{\rho_{max}}\right)$ in the x direction while Part B is a negative perturbation which moves with the wave velocity $U_{max}\left(1-\frac{\rho_{Ne}}{\rho_{max}}\right)$ in the x direction. However since $\rho_{Ne} > \rho_{crit}$, wave velocity of Part A is negative and hence it move towards the $x = 0km$ side while Part B of the perturbation has positive wave velocity and move towards the $x = 10km$ side. In the plot for ρ_M, it is seen that the perturbation $\Delta\rho_M$ is a positive perturbation which which moves with the wave velocity $U_{max}\left(1-\frac{\rho_{Ne}}{\rho_{max}}\right)$ in the x direction which is positive similar to Case 1. Thus, the wave velocity for the propagation of malicious vehicles is always positive and its magnitude is a function of the equilibrium density. Also, the amplitude of the perturbation remains the same with time. This analysis demonstrates that if a mobile sensor travels with a velocity equal to the wave velocity of $\Delta\rho_M$ for the initial point of the perturbation detection, it can detect and track the maximum percentage of malicious vehicles in the swarm.

V. SOLUTION ARCHITECTURE

We now discuss an architecture to detect the percentage density of the malicious vehicles ($\rho_M\%$) as well as the attack mode p. The architecture is shown in Fig 10. Let ρ_{SS} and ρ_{TS} represent the total densities for a single-species baseline case and two-species case respectively, and $U_{SS}, U_{TS}, V_{SS}, V_{TS}$ be correspondingly defined. Define quantities $c_\rho(t)$, c_U and $c_V(t)$, where

$$\begin{aligned} c_\rho(t) &\equiv \rho_{TS}(\bar{x}_i, \bar{y}_i, t) - \rho_{SS}(\bar{x}_i, \bar{y}_i, t) \\ c_U(t) &\equiv U_{TS}(\bar{x}_i, \bar{y}_i, t) - U_{SS}(\bar{x}_i, \bar{y}_i, t) \\ c_V(t) &\equiv V_{TS}(\bar{x}_i, \bar{y}_i, t) - V_{SS}(\bar{x}_i, \bar{y}_i, t), i = 1, \cdots, N \end{aligned} \tag{16}$$

where, $(\bar{x}_i, \bar{y}_i), i = 1, \cdots, N$ represent the locations of the stationary sensors. As shown in Fig 10, the quantities c_ρ, c_U and c_V are computed using simulated models of the single-species and two-species scenarios (with both simulations being performed from the same initial condition), and these are then used to create the GPR model by determining the optimal kernel and basis functions for the GP model and then estimating these optimal functions. From this GPR model,

predictions of the unknown $\rho_M\%$ and p can be made as the mobile sensor moves along the trajectory of the maximum $\rho_M\%$ as determined from the wave velocity analysis.

VI. GAUSSIAN PROCESS REGRESSION MODELS

GPR Models are non-parametric kernel-based Bayesian regression and probability distribution models [16] and have been widely used in regression problems, where the relationship between the input and the output are difficult to model using a single function. One of the advantages of using these models is that the confidence bounds of the regressed outputs can be calculated.

A. Training Data and the Gaussian Process Model

Consider a training set $\mathcal{D}$ with n observations, $\mathcal{D} = \{(\mathbf{x}_i, y_i)|i = 1, ..., n\}$, where $\mathbf{x}_i$ denotes a $d-$dimensional input vector and y_i denotes a scalar output or response variable. For this work, $\mathbf{x}_i$ refers to $c_\rho(t)$, c_U and $c_V(t)$ for a particular sensor location at a particular time and y_i refers to $\rho_M\%$, p. Concatenating all the n input vectors, a $d \times n$ input matrix $\mathbf{X}$ and the response variable as a $n \times 1$ vector $\mathbf{Y}$, the functional mapping $\eta(.)$ links the input set $\mathbf{x_i}$ to output y_i. In the GPR model, $\eta(\mathbf{x})$ is modelled in the form [17]:

$$y(\mathbf{x}) = \underbrace{\mathbf{h}(\mathbf{x})^\intercal \beta}_{M(\mathbf{x};\beta)} + \underbrace{f(\mathbf{x})}_{\mathcal{V}(\mathbf{x},\mathbf{x}';\sigma^2,\theta)} \quad . \tag{17}$$

Here, $\beta = (\beta_1, \ldots, \beta_s)^\intercal$ is a s-dimensional vector of the basis function coefficients to be estimated while $\mathbf{h}(\mathbf{x})$ represents the chosen basis function. The term $f(\mathbf{x})$ is a GP with zero mean and a kernel function $k(\mathbf{x}, \mathbf{x}'; \theta)$. σ^2 is defined as the error variance between the observed and predicted values. The choice of the basis function and the kernel function play an important role in the accuracy of the GPR model and an optimal choice of kernel and basis functions can be determined by using a Bayesian optimization method as discussed in [18, 19]. Once the optimal kernel and basis functions are determined, the following GP prior can be used to represent $\eta(.)$

$$\eta(\mathbf{x})|\beta, \sigma^2, \theta \ \sim \ \mathcal{GP}\left(M(\mathbf{x};\beta), \mathcal{V}(\mathbf{x}, \mathbf{x}^{'}; \sigma^2, \theta)\right) \tag{18}$$

where $M(\mathbf{x};\beta) \equiv \mathbb{E}[\eta(\mathbf{x})]$ denotes the mean function and $\mathcal{V}(\mathbf{x}, \mathbf{x}'; \sigma^2, \theta) \equiv cov[\eta(\mathbf{x}), \eta(\mathbf{x}')]$ denotes the covariance function $\mathcal{V}(\mathbf{x}, \mathbf{x}'; \sigma^2, \theta) = k(\mathbf{x}, \mathbf{x}'; \theta) + \sigma^2 I_n$.

B. GPR Training and Detection of the Malicious Vehicles using GP Posterior

In the GPR training data, the quantities c_ρ, c_U, c_V, at different times, already defined in Section V, represent the input matrix $\mathbf{X}$ and the known values of the traffic parameters $\rho_M\%$, p represent the response variable $\mathbf{y}$. Since the GPR algorithm mentioned in this paper works only for scalar outputs, therefore two training datasets are constructed, one with $\rho_M\%$ as the output and other with p as the output.

After the optimal basis function $h(\mathbf{x})$ and the kernel function $k(\mathbf{x}, \mathbf{x}'; \theta)$ are chosen, then their parameters β, θ and σ^2, (represented by ϑ) are determined by maximising the likelihood of $P(P(\mathbf{y}|\mathbf{X}; \vartheta)$ using the training data [16]. The marginal log likelihood of $P(\mathbf{y}|\mathbf{X}; \vartheta)$ can be written as

$$\begin{aligned} \log P(\mathbf{Y}|\mathbf{X}; \vartheta) = &-\frac{1}{2}(\mathbf{Y} - \mathbf{H}\beta)^T {\mathbf{K}_\sigma}^{-1}(\mathbf{Y} - \mathbf{H}\beta) \\ &- \frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}_\sigma| \end{aligned} \tag{19}$$

where $\mathbf{H} = (\mathbf{h}(\mathbf{x_1})^\intercal, \ldots, \mathbf{h}(\mathbf{x_n})^\intercal)^\intercal$ are the basis functions of the GP evaluated at the n input points and $\mathbf{K}_\sigma$ is given by

$$\mathbf{K}_{\sigma_{i,j}} = k(\mathbf{x}_i, \mathbf{x}_j; \theta) + \Delta_{ij}\sigma^2, \ i, j = 1, \ldots, n$$

Initially, the algorithm estimates the value of β which maximises the log likelihood for a given θ, σ^2. This estimate for β is given by

$$\tilde{\beta}(\theta, \sigma^2) = \left[\mathbf{H}^T {\mathbf{K}_\sigma}^{-1}\mathbf{H}\right]^{-1} \mathbf{H}^T {\mathbf{K}_\sigma}^{-1}\mathbf{Y} \tag{20}$$

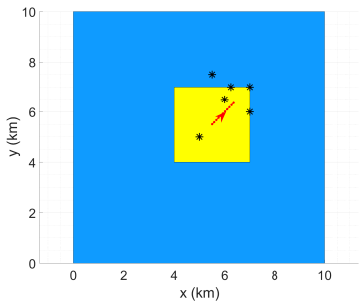


Fig. 11 GPR training and test scenario. The blue color represents the equilibrium density of the swarm at initial time. The yellow color represents the location of the perturbations. The black stars represent the location of the fixed sensors for GPR training. The red trajectory represents the trajectory of the moving sensor.

Substituting the estimated $\tilde{\beta}(\theta, \sigma^2)$ in the original (19), the $\beta-$profiled log-likelihood is obtained which depends on θ and σ^2 only. Next, the algorithm maximizes the above log-likelihood over the space of θ and σ^2 to obtain their estimates. This is done using an unconstrained gradient based optimisation solver in Matlab.

After the parameters of the GP model are estimated, they can be used to detect the malicious parameters $\mathbf{y}^*$, $(\rho_M\%^*/p^*)$, for a set of previously undefined time series input $\mathbf{x}^*$ (c_ρ^*, c_U^* and c_V^*) obtained by comparing the traffic measurements with the baseline single species model. Using methods given in [20], the posterior GP for prediction of the response variable , $\hat{\vartheta}_{MLE}$ is given by:

$$\eta(\mathbf{x})|\hat{\vartheta}_{MLE}, \mathcal{D} \sim \mathcal{GP}\left(M^*(\mathbf{x}; \beta, \theta), k^*(\mathbf{x}, \mathbf{x}'; \theta) + \sigma^2 I_n\right) \tag{21}$$

where

$$M^*(\mathbf{x}; \beta, \theta) = M(\mathbf{x}^*; \beta) + \mathbf{t}(\mathbf{x}^*)^\top \mathbf{K}_\sigma^{-1}(\mathbf{y} - \mathbf{H}\beta) \tag{22}$$

$$k^*(\mathbf{x}, \mathbf{x}'; \theta) = k(\mathbf{x}^*, \mathbf{x}^{*\prime}; \theta) - \mathbf{t}(\mathbf{x}^*)^\top \mathbf{K}_\sigma^{-1} \mathbf{t}(\mathbf{x}^{*\prime})^\top \tag{23}$$

$$\mathbf{t}(\mathbf{x}^*) = (k(\mathbf{x}^*, \mathbf{x}_1; \theta), \ldots, k(\mathbf{x}^*, \mathbf{x}_n; \theta)) \tag{24}$$

Here, the mean $M^*(\mathbf{x}; \beta, \theta)$ gives the mean predicted values of the response variable and $k^*(\mathbf{x}, \mathbf{x}'; \theta)$ computes the confidence bounds of the mean.

VII. SIMULATION RESULTS

To test the above methodology, a training dataset was initially created and used to train a GPR model. After the GPR models are created, a test two-species UAV swarm distribution is adopted and measurements are taken by the sensor moving along the trajectory of the maximum $\rho_M\%$ as predicted by the linearized analysis. The steps are outlined in the following subsections:

A. Creation of the GPR Training Data

For the baseline case of all vehicles normal, the initial density distribution was chosen as given in Fig 4(a) and the LWR single-species model (Eq 1) was used to simulate the flow of vehicles for 129.6 s or 2.16 min. Fig 5(a) shows the density distribution of the vehicles at final time. From the simulation data, measurements of ρ, U, V are collected at discrete time intervals of 10.8 s at six locations: $(5, 5)$, $(6, 6.5)$, $(6.25, 7)$, $(7, 7)$, $(7, 6)$ and $(5.5, 7.5)$ (shown by black stars in Fig 11) to make up the baseline data. Then, the two-species case was considered as shown in Fig. 4(b) such that the total density distribution remains the same ($\rho_N + \rho_M = \rho$). The LWR two-species traffic model (Eq 3) was then used to simulate the flow of vehicles again for 2.16 min (see Fig. 5(b)) and similar measurements of ρ, U and V were then taken at the same discrete times and at the same locations. These measurements were then compared to obtain c_ρ, c_U and c_V respectively. Fig 12 shows distribution of c_ρ at $t = 129.6\,s$. The above steps were then repeated three more times, taking different percentages of malicious vehicles for the two species model. These values of c_ρ, c_U and c_V, along with the known values of $\rho_M\%$ and p at those locations and times make up the training datasets for $\rho_M\%$ GPR model and p GPR model respectively.

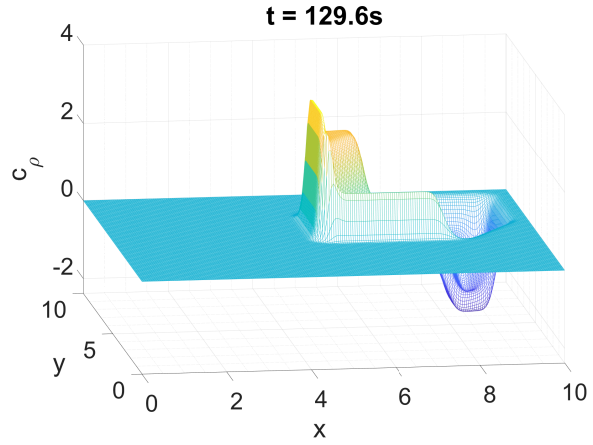
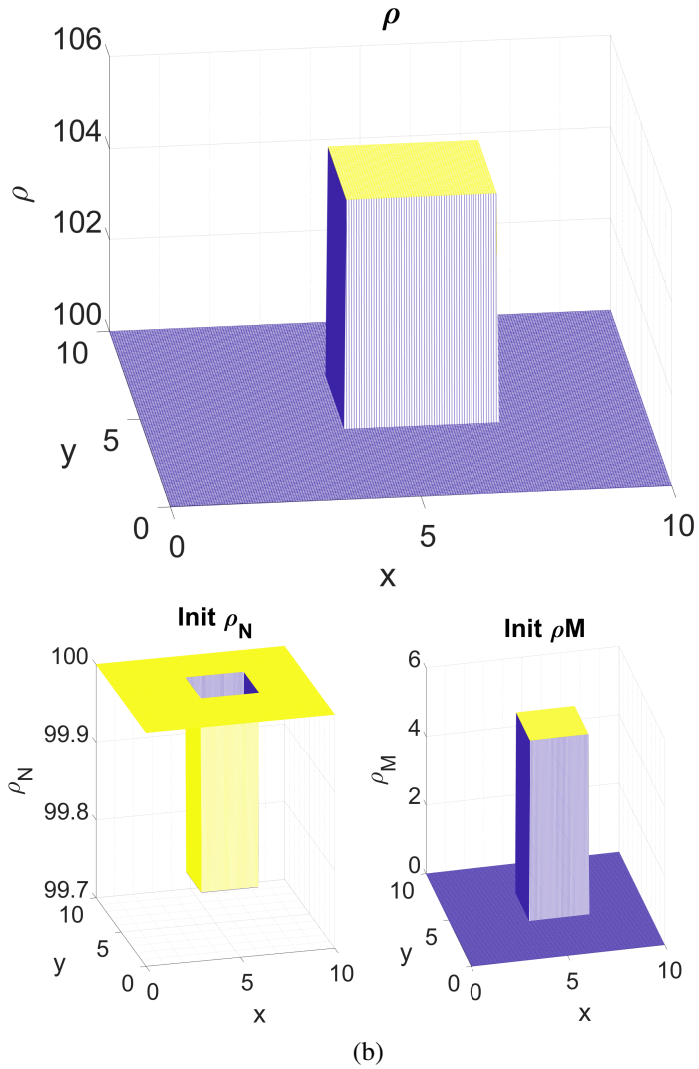


Fig. 12 Computed c_ρ distribution



(b)

Fig. 13 Initial density distributions for the test case

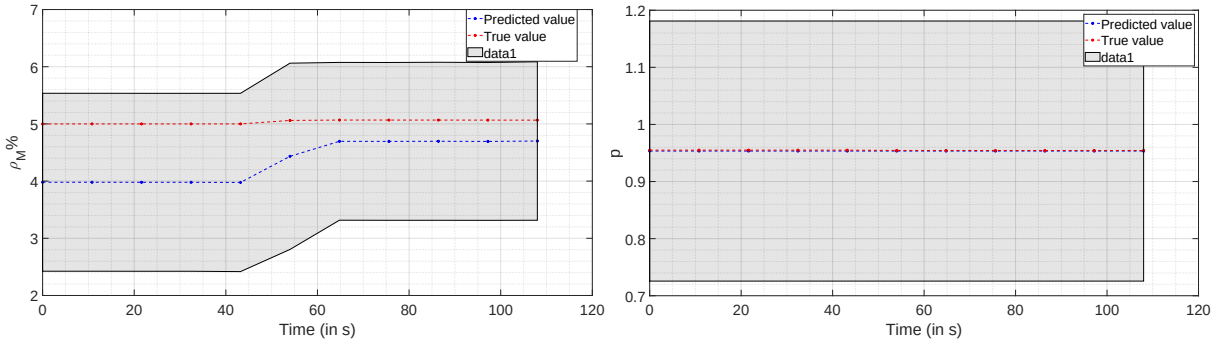


Fig. 14 Predictions of the Malicious Vehicle parameters

B. GPR Training and Prediction

A GPR model is chosen by initially applying Bayesian Optimization on the dataset to determine it's kernel and basis functions. The parameters of the chosen model are then optimised using a Parameter Optimization Algorithm discussed in Sec VI. After the GPR models are defined, their validity is tested by taking an initial two-species distribution as shown in Fig 13. Vehicle flow is simulated for 2.16 min and the values of c_ρ, c_U and c_V are computed after comparison with the baseline case at distinct time intervals of 10.8 s along the computed trajectory of the maximal percentage of ρ_M vehicles, as shown in Fig. 11. These values are then used to determine the ρ_M% and p from their GPR models. Fig 14 shows the predicted and true values of ρ_M% and p for each of these locations with the gray shaded region representing the 95% confidence bounds. In all the cases, the mean predicted values lie close to the true values, and the true values are contained within the confidence bounds.

VIII. CONCLUSIONS

We consider a class of cyber-attacks where an attacker may potentially hack into the autonomous software of a subset of vehicles in an UAV swarm flying at a constant altitude. The hacked vehicles, referred to as malicious vehicles, are assumed to be arbitrarily distributed among the normal vehicles. A 2D LWR two-species PDE model is employed to model the flow of normal and malicious vehicles in the UAV swarm. A linearized analysis governing the propagation of small perturbations in malicious and normal vehicle densities is performed in order to find the direction of the wave along which the maximum ρ_M% can be detected. This is then combined with a GPR framework to develop a detection scheme that determines both the fraction of the malicious vehicles, as well as the manner in which they influence the average velocity of the swarm. Simulations demonstrate the working of the methodology.

Acknowledgments

This material is based on research sponsored by the Air Force Research Laboratory under agreement number FA8650-20-2-5853. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory of the U.S. Government.

References

[1] Yan, Y., Qian, Y., Sharif, H., and Tipper, D., "A survey on cyber security for smart grid communications," *IEEE Communications Surveys & Tutorials*, Vol. 14, No. 4, 2012, pp. 998–1010.

[2] Bou-Harb, E., Fachkha, C., Pourzandi, M., Debbabi, M., and Assi, C., "Communication security for smart grid distribution networks," *IEEE Communications Magazine*, Vol. 51, No. 1, 2013, pp. 42–49.

[3] Chang, E. S., Jain, A. K., Slade, D. M., and Tsao, S. L., "Managing cyber security vulnerabilities in large networks," *Bell Labs technical journal*, Vol. 4, No. 4, 1999, pp. 252–272.

[4] Abbaspour, A., Yen, K. K., Noei, S., and Sargolzaei, A., "Detection of fault data injection attack on UAV using adaptive neural network," *Procedia computer science*, Vol. 95, 2016, pp. 193–200.

[5] Reilly, J., Martin, S., Payer, M., and Bayen, A. M., "Creating complex congestion patterns via multi-objective optimal freeway traffic control with application to cyber-security," *Transportation Research Part B: Methodological*, Vol. 91, 2016, pp. 366–382.

[6] Sargolzaei, A., Yazdani, K., Abbaspour, A., Crane III, C. D., and Dixon, W. E., "Detection and mitigation of false data injection attacks in networked control systems," *IEEE Transactions on Industrial Informatics*, Vol. 16, No. 6, 2019, pp. 4281–4292.

[7] Roy, T., and Dey, S., "Secure Traffic Networks in Smart Cities: Analysis and Design of Cyber-Attack Detection Algorithms," *2020 American Control Conference (ACC)*, IEEE, 2020, pp. 4102–4107.

[8] Kavousi-Fard, A., Dabbaghjamanesh, M., Jin, T., Su, W., and Roustaei, M., "An evolutionary deep learning-based anomaly detection model for securing vehicles," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[9] Wu, J., Luan, H., and Zhang, L., "Simultaneous State and Cyber-attack Estimation for Autonomous Vehicular Flow Models via Boundary Observers," *2021 American Control Conference (ACC)*, IEEE, ????

[10] Lighthill, M. J., and Whitham, G. B., "On kinematic waves II: A theory of traffic flow on long crowded roads," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, Vol. 229, No. 1178, 1955, pp. 317–345.

[11] Ghanavati, M., Chakravarthy, A., and Menon, P., "PDE-based analysis of cyber-attacks in vehicle swarms," *2018 IEEE Conference on Decision and Control (CDC)*, IEEE, 2018, pp. 1329–1334.

[12] Kashyap, A., Chakravarthy, A., and Menon, P. P., "Detection of Cyber-Attacks in Automotive Traffic Using Macroscopic Models and Gaussian Processes," *IEEE Control Systems Letters*, Vol. 6, 2022, pp. 1688–1693.

[13] Giorgi, F., Leclercq, L., and Lesort, J.-B., "A traffic flow model for urban traffic analysis: extensions of the LWR model for urban and environmental applications," *Transportation and Traffic Theory in the 21st Century*, Emerald Group Publishing Limited, 2002.

[14] Li, J., Chen, Q.-Y., Wang, H., and Ni, D., "Analysis of LWR model with fundamental diagram subject to uncertainties," *Transportmetrica*, Vol. 8, No. 6, 2012, pp. 387–405.

[15] Göttlich, S., Iacomini, E., and Jung, T., "Properties of the LWR model with time delay," *arXiv preprint arXiv:2003.12090*, 2020.

[16] Williams, C. K., and Rasmussen, C. E., *Gaussian processes for machine learning*, Vol. 2, MIT press Cambridge, MA, 2006.

[17] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P., "Design and analysis of computer experiments," *Statistical science*, Vol. 4, 1989.

[18] Snoek, J., Larochelle, H., and Adams, R. P., "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, Vol. 25, 2012.

[19] Gelbart, M. A., Snoek, J., and Adams, R. P., "Bayesian optimization with unknown constraints," *arXiv preprint arXiv:1403.5607*, 2014.

[20] Haylock, R., and O'Hagan, A., "On inference for outputs of computationally expensive algorithms with uncertainty on the inputs," *Bayesian Statistics*, Vol. 5, 1996, pp. 629–637.