Supplemental Materials

Proposition 1. *The distribution of the counts is given as follows:*

$$P(\mathbf{n}_{1:H};\pi_\theta) = \mathbb{I}(\mathbf{n}_{1:H} \in \Omega_{1:H}) P(\langle \mathrm{n}^{\mathrm{txn}}_1(z_d, z_{\mathrm{src}}, \tau)\forall\tau\rangle)$$
$$\prod_{t=1}^{H-1} P(\tilde{\mathrm{n}}_t \mid \mathrm{n}^{\mathrm{nxt}}_t; \boldsymbol{\beta}_t = \pi_t(\mathbf{n}_t)) \times P(\mathrm{n}^{\mathrm{nxt}}_t \mid \mathrm{n}^{\mathrm{arr}}_t) \tag{1}$$

in which the next zone count distribution is:

$$P(\mathrm{n}^{\mathrm{nxt}}_t \mid \mathrm{n}^{\mathrm{arr}}_t) = \prod_z \mathrm{Mul}(\mathrm{n}^{\mathrm{arr}}_t(z),\ \alpha(z'|z)\forall z'\forall z')$$

The travel-time count distribution given output of traffic control π is:

$$P(\tilde{\mathrm{n}}_t \mid \mathrm{n}^{\mathrm{nxt}}_t; \boldsymbol{\beta}_t = \pi_t(\mathbf{n}_t)) = \prod_{z,z'} \mathrm{Mul}(\mathrm{n}^{\mathrm{nxt}}_t(z,z'), p^{\mathrm{nav}}(\tau \mid z, z'; \beta^{zz'}_{t+1})\forall\tau)$$

And $\Omega_{1:H}$ is the set of consistent count tables satisfying count consistency constraints:

$$\mathrm{n}^{\mathrm{arr}}_{t+1}(z) = \sum_{z' \in Z} \bigg[\mathrm{n}^{\mathrm{txn}}_t(z', z, \tau = t+1) +$$
$$\tilde{n}_t(z', z,\ \tau = t+1) \bigg], \forall z \tag{2}$$
$$\mathrm{n}^{\mathrm{txn}}_{t+1}(z, z', \tau) = \mathrm{n}^{\mathrm{txn}}_t(z, z', \tau) + \tilde{\mathrm{n}}_t(z, z', \tau), \forall z, z', \tau > t+1 \tag{3}$$
$$\tag{4}$$

Proof. We can compute the count distribution by summing up the distributions of all joint trajectories $\boldsymbol{s}_{1:H}, \boldsymbol{a}_{1:H} \sim \mathbf{n}_{1:H}$ satisfying a given counts:

$$
\begin{aligned}
P(\mathbf{n}_{1:H}\ ; \pi_\theta) &= \sum_{\boldsymbol{s}_{1:H}, \boldsymbol{a}_{1:H} \sim \mathbf{n}_{1:H}} P(\boldsymbol{s}_{1:H}, \boldsymbol{a}_{1:H}) \\
&= \mathbb{I}(\mathbf{n}_{1:H} \in \Omega_{1:H}) \prod_{t=1}^{H} \sum_{\boldsymbol{s}_t, \boldsymbol{a}_t \sim \mathbf{n}_t} \prod_{m=1}^{M} \Big(\prod_{z,z'} \alpha(z'|z)^{\mathbb{I}(s_t^m = \langle z,\phi,\phi \rangle, a_t^m = z')} \\
&\quad \times \prod_{z,z',\tau} p^{\text{nav}}(\tau | z, z', \boldsymbol{\beta}_t = \pi_t(\mathbf{n}_t))^{\mathbb{I}(s_t^m = \langle z,\phi,\phi \rangle, a_t^m = z', s_{t+1}^m = \langle z, z', \tau \rangle}\Big) \\
&= \mathbb{I}(\mathbf{n}_{1:H} \in \Omega_{1:H}) \prod_{t=1}^{H} \Big(\sum_{\boldsymbol{s}_t, \boldsymbol{a}_t \sim \mathrm{n}^{\text{nxt}}} \prod_{m=1}^{M} \prod_{z,z'} \alpha(z'|z)^{\mathbb{I}(s_t^m = \langle z,\phi,\phi \rangle, a_t^m = z')} \\
&\quad \times \sum_{\boldsymbol{s}_t, \boldsymbol{a}_t \sim \mathrm{n}^{\text{txn}}} \prod_{m=1}^{M} \prod_{z,z',\tau} p^{\text{nav}}(\tau | z, z', \boldsymbol{\beta}_t = \pi_t(\mathbf{n}_t))^{\mathbb{I}(s_t^m = \langle z,\phi,\phi \rangle, a_t^m = z', s_{t+1}^m = \langle z, z', \tau \rangle)}\Big) \qquad (5)
\end{aligned}
$$

We can simplify the summation over joint trajectory by multinomial distribution as follows:

$$
\sum_{\boldsymbol{s}_t, \boldsymbol{a}_t \sim \mathrm{n}^{\text{nxt}}} \prod_{m=1}^{M} \prod_{z,z'} \alpha(z'|z)^{\mathbb{I}(s_t^m = \langle z,\phi,\phi \rangle, a_t^m = z')} = \prod_{z} \mathrm{Mul}(\mathrm{n}_t^{\text{arr}}(z),\ \alpha(z'|z) \forall z' \forall z') \qquad (6)
$$

and

$$
\begin{aligned}
&\sum_{\boldsymbol{s}_t, \boldsymbol{a}_t \sim \mathrm{n}^{\text{txn}}} \prod_{m=1}^{M} \prod_{z,z',\tau} p^{\text{nav}}(\tau | z, z', \boldsymbol{\beta}_t = \pi_t(\mathbf{n}_t))^{\mathbb{I}(s_t^m = \langle z,\phi,\phi \rangle, a_t^m = z', s_{t+1}^m = \langle z, z', \tau \rangle)} \\
&\qquad = \prod_{z,z'} \mathrm{Mul}(\mathrm{n}_t^{\text{nxt}}(z, z'), p^{\text{nav}}(\tau \mid z, z'; \beta_{t+1}^{zz'}) \forall \tau) \qquad (7)
\end{aligned}
$$

By replacing (6) and (7) into (5), we have (1). □

Theorem 2. *The traffic control objective in (5) in main paper can be computed by expectation over counts*

$$
V(\pi_\theta) = \sum_{t=1}^{H} \mathbb{E}_{\boldsymbol{s}_{1:t}, \boldsymbol{a}_{1:t}}[r(\mathbf{n}_t) | \boldsymbol{a}_t, \boldsymbol{s}_t; \pi_\theta] = \sum_{t=1}^{H} \mathbb{E}_{\mathbf{n}_{1:t} \in \Omega_{1:t}}\big[r(\mathbf{n}_t | \pi_\theta)\big]
$$

Proof. Let $\boldsymbol{s}_t$ and $\boldsymbol{a}_t$ represent the joint-state and joint-action of all the agents at time step t

$$
V(\pi_\theta) = \sum_{t=1}^{H} \mathbb{E}_{\boldsymbol{s}_{1:t}, \boldsymbol{a}_{1:t}}[r(\mathbf{n}_t) | \boldsymbol{a}_t, \boldsymbol{s}_t; \pi_\theta] \qquad (8)
$$

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{s}_{1:t}, \boldsymbol{a}_{1:t}}[r(\mathbf{n}_t) | \boldsymbol{a}_t, \boldsymbol{s}_t; \pi_\theta] && (9) \\
&= \sum_{(\boldsymbol{s}_{1:t}, \boldsymbol{a}_{1:t})} P(\boldsymbol{s}_{1:t}, \boldsymbol{a}_{1:t}; \pi_\theta) \cdot [r(\mathbf{n}_t) | \boldsymbol{a}_t, \boldsymbol{s}_t; \pi_\theta] && (10) \\
&= \sum_{(\boldsymbol{s}_{1:t}, \boldsymbol{a}_{1:t})} f(\mathbf{n}_{1:t}; \pi_\theta) \cdot [r(\mathbf{n}_t) | \boldsymbol{a}_t, \boldsymbol{s}_t; \pi_\theta] && (11)
\end{aligned}
$$

where $f(\mathbf{n}_{1:t};\pi_\theta) = \prod_t \left(\prod_{z,z',\tau}\left[\alpha(z'|z)p^{\text{nav}}(\tau|z,z',\boldsymbol{\beta}_t=\pi_t(\mathbf{n}_t))\right]^{\tilde{\text{n}}_t(z,z',\tau)}\right)$ is a function only depending on the counts.
Notice that in (11), the expected immediate reward at time step t only depends on the count $\tilde{\text{n}}_t(\cdot,\cdot,\cdot)$ that arises from the joint state and action $(\boldsymbol{s}_t,\boldsymbol{a}_t)$. So instead of summing over all the joint state-action trajectories $(\boldsymbol{s}_{1:t},\boldsymbol{a}_{1:t})$, we can sum over the space of all possible counts

$$\mathbb{E}_{\boldsymbol{s}_{1:t},\boldsymbol{a}_{1:t}}[r(\mathbf{n}_t)|\boldsymbol{a}_t,\boldsymbol{s}_t;\pi_\theta] = \sum_{\mathbf{n}_{1:t}\in\Omega_{1:t}} P(\mathbf{n}_{1:t})\cdot[r(\mathbf{n}_t)|\pi_\theta](\text{from Proposition 1}) \tag{12}$$

$$= \mathbb{E}_{\mathbf{n}_{1:t}\in\Omega_{1:t}}[r(\mathbf{n}_t)|\pi_\theta] \tag{13}$$

Using the above expression, the value function can be computed as :

$$V(\pi_\theta)=\sum_{t=1}^{H}\mathbb{E}_{\boldsymbol{s}_{1:t},\boldsymbol{a}_{1:t}}[r(\mathbf{n}_t)|\boldsymbol{a}_t,\boldsymbol{s}_t;\pi_\theta]=\sum_{t=1}^{H}\mathbb{E}_{\mathbf{n}_{1:t}\in\Omega_{1:t}}\left[r(\mathbf{n}_t|\pi_\theta)\right] \tag{14}$$

□

Theorem 3. *The vehicle-based value function can be computed by collective expectation over the counts as follows:*

$$V_t^{zz'}(\pi_\theta^{zz'}) = \mathbb{E}_{\mathbf{n}_{1:H}}\Big[\sum_{\tau>t}\tilde{\text{n}}_t(z,z',\tau)V_t^{\mathbf{n}}(z,z',\tau)\Big|\pi_\theta\Big] \tag{15}$$

in which $V_t^{\mathbf{n}}(z,z',\tau)$ is the average accumulated reward of newly arrived vessels at z at time t going to z' computed based on the realized counts $\mathbf{n}_{1:H}$ as follows:

$$R_t^{\mathbf{n}}(z,z',\tau)=\sum_{\tau"=\,t}^{\tau}-C(z,\text{n}_{\tau"}^{\text{tot}}),\forall\tau\in[t+t_{\min}^{zz'},\ \ t+t_{\max}^{zz'}] \tag{16}$$

$$V_t^{\mathbf{n}}(z,z',\tau) = R_t(z,z',\tau)+\gamma\cdot V_\tau^{\mathbf{n}}(z') \tag{17}$$

$$V_t^{\mathbf{n}}(z,z') = \frac{\sum_{\tau=t+t_{\min}^{zz'}}^{t+t_{\max}^{zz'}} V_t^{\mathbf{n}}(z,z',\tau)\cdot\tilde{n}_t(z,z',\tau)}{\sum_{\tau=t+t_{\min}^{zz'}}^{t+t_{\max}^{zz'}}\tilde{n}_t(z,z',\tau)} \tag{18}$$

$$V_t^{\mathbf{n}}(z) = \frac{\sum_{z'}\text{n}_t^{\text{nxt}}(z,z')\cdot V_t^{\mathbf{n}}(z,z')}{\sum_{z'}\text{n}_t^{\text{nxt}}(z,z')}, \tag{19}$$

where $R_t^{\mathbf{n}}(z,z',\tau)$ is the reward accumulated by a vessel when it is still in zone z between time t and τ; $V_\tau^{\mathbf{n}}(z,z')$ is the average accumulated reward of a vessel which started crossing z to z' from time t. $V_\tau^{\mathbf{n}}(z')$ is the average accumulative reward of a vessel newly arrived at z' at time τ.

Proof. Based on exchangeability of vessels regard to the count, we can apply theorem 4 from [4] to have

$$P(s_{1:T}^m,a_{1:T}^m,\mathbf{n}_{1:T}) = P(\mathbf{n}_{1:T};\pi_\theta)\prod_{1:T}\prod_{z,z',\tau}\Big(\frac{\tilde{n}_t(z,z',\tau)}{\text{n}_t^{\text{arr}}(z)}\Big)^{\mathbb{I}(s_t^m=\langle z,\phi,\phi\rangle,a_t^m=z',s_{t+1}^m=\langle z,z',\tau\rangle)} \tag{20}$$

We denote the current zone of a vessel m at time t to be z_t^m. The individual value of a vessel crossing (z, z') from time t to τ can be computed as

$$\begin{aligned}
&\mathbb{E}_{s^m_{1:H}, a^m_{1:H}, \mathbf{n}_{1:H}}[\mathbb{I}(s_t^m = \langle z, \phi, \phi\rangle, a_t^m = z', s_{t+1}^m = \langle z, z', \tau\rangle) \sum_{t'=t:H} r_t^m] \\
&= \mathbb{E}_{\mathbf{n}_{1:H}}\Bigg[\sum_{s^m_{1:H}, a^m_{1:H}} \mathbb{I}(s_t^m = \langle z, \phi, \phi\rangle, a_t^m = z', s_{t+1}^m = \langle z, z', \tau\rangle) \\
&\prod_{1:T}\prod_{\bar{z}, \bar{z}', \bar{\tau}} \Big(\frac{\tilde{n}_t(z, z', \bar{z})}{\mathrm{n}_t^{\mathrm{arr}}(z)}\Big)^{\mathbb{I}(s_t^m = \langle \bar{z}, \phi, \phi\rangle, a_t^m = \bar{z}', s_{t+1}^m = \langle z, z', \bar{z}\rangle)} \sum_{t'=t:H} -C(z^m, \mathrm{n}_t^{\mathrm{tot}})\Bigg]
\end{aligned} \tag{21}$$

Similar to theorem 6 from [4], to compute the expression inside expectation in (21), we can construct an auxiliary MDP for individual vessel m with $p^{\mathrm{nav},\mathbf{n}}(\tau, z, z') = \frac{\tilde{n}_t(z, z', \tau)}{\mathrm{n}_t^{\mathrm{nxt}}(z, z')}$, $\alpha_t^{\mathbf{n}}(z'|z) = \frac{\mathrm{n}_t^{\mathrm{nxt}}(z, z')}{\mathrm{n}_t^{\mathrm{arr}}(z)}$ and $r_t^{\mathbf{n}}(z) = -C(z, n_t^{\mathrm{tot}})$. The value function $V^{\mathbf{n}}$ for the auxiliary MDP can be obtained by using Bellman equations as per (16)- (19).

□

Theorem 4. *The vehicle-based policy gradient for $\pi^{zz'}$ is*

$$\begin{aligned}
&\nabla_\theta V_1^{zz'}(\pi_\theta^{zz'}) = \mathbb{E}_{\mathbf{n}_{1:H}}\Big[\sum_{t=1:H} \sum_{\tau=t+t_{\min}^{zz'}}^{t+t_{\max}^{zz'}} \tilde{\mathrm{n}}_t(z, z', \tau) \times \\
&\big[(\tau - t - t_{\min}^{zz'}) \cdot \nabla_\theta \log(\pi_\theta^{zz'}(\mathbf{n}_t)) + (t_{\max}^{zz'} - (\tau - t)) \cdot \nabla_\theta \log(1 - \pi_\theta^{zz'}(\mathbf{n}_t))\,\big] V_t^{\mathbf{n}}(z, z', \tau)\Big]
\end{aligned} \tag{22}$$

Proof. For each zone pair $\langle z, z'\rangle$ we have,

$$\begin{aligned}
&\nabla_\theta V_1^{zz'}(\pi_\theta^{zz'}) \\
&= \sum_{\mathbf{n}_1} \nabla_\theta\Big[P(\mathbf{n}_1|\pi_\theta) \sum_{\mathbf{n}_{2:H}} P(\mathbf{n}_{2:H}|\mathbf{n}_1, \pi_\theta) \sum_{\tau>1} \tilde{\mathrm{n}}_t(z, z', \tau) V_1^{\mathbf{n}}(z, z', \tau)\Big] && (23) \\
&= \sum_{\mathbf{n}_1}\sum_{\tau>1} \tilde{\mathrm{n}}_t(z, z', \tau) V_1^{\mathbf{n}}(z, z', \tau) \nabla_\theta P(\mathbf{n}_1|\pi_\theta) \\
&+ \sum_{\mathbf{n}_1} P(\mathbf{n}_1|\pi_\theta) \nabla_\theta\Big[\sum_{\mathbf{n}_{2:H}} P(\mathbf{n}_{2:H}|\mathbf{n}_1, \pi_\theta) \sum_{\tau>2} \tilde{\mathrm{n}}_t(z, z', \tau) V_2^{\mathbf{n}}(z, z', \tau)\Big] && (24) \\
&\text{continues to unroll the terms, we have:} && (25) \\
&= \sum_{t=1}^{H}\sum_{\mathbf{n}_{1:t}}\sum_{\tau>t} \Big[\tilde{\mathrm{n}}_t(z, z', \tau) V_t^{\mathbf{n}}(z, z', \tau) \nabla_\theta P(\mathbf{n}_t|\pi_\theta)\Big] && (26) \\
&\text{using the log-trick, we have:} && (27) \\
&= \sum_{t=1}^{H}\sum_{\mathbf{n}_{1:t}}\sum_{\tau>t} \Big[\tilde{\mathrm{n}}_t(z, z', \tau) V_t^{\mathbf{n}}(z, z', \tau) P(\mathbf{n}_t|\pi_\theta) \nabla_\theta \log P(\mathbf{n}_t|\mathbf{n}_{t-1}, \pi_\theta)\Big] && (28)
\end{aligned}$$

Notice that

$$P(\mathbf{n}_t|\mathbf{n}_{t-1},\pi_\theta) = P(\tilde{\mathrm{n}}_t \mid \mathrm{n}_t^{\mathrm{nxt}}; \boldsymbol{\beta}_t = \pi_t(\mathbf{n}_t)) \times P(\mathrm{n}_t^{\mathrm{nxt}} \mid \mathrm{n}_t^{\mathrm{arr}}) \mathbb{I}(\mathbf{n}_{t-1:t} \in \Omega_{t-1}) \tag{29}$$

in which Ω_{t-1} is the count space satisfying constraints (2), (3). Using (29) into (28), we have

$$\begin{aligned}
&\nabla_\theta V_1^{zz'}(\pi_\theta^{zz'}) \\
&= \sum_{t=1}^{H} \sum_{\mathbf{n}_{1:t}} \sum_{\tau>t} \Big[\tilde{\mathrm{n}}_t(z,z',\tau) V_t^{\mathbf{n}}(z,z',\tau) P(\mathbf{n}_t|\pi_\theta) \nabla_\theta \log P(\tilde{\mathrm{n}}_t \mid \mathrm{n}_t^{\mathrm{nxt}}; \boldsymbol{\beta}_t = \pi_t(\mathbf{n}_t)) \\
&= \mathbb{E}_{\mathbf{n}_{1:H}} \Big[\sum_{t=1:H} \sum_{\tau=t+t_{\min}^{zz'}}^{t+t_{\max}^{zz'}} \tilde{\mathrm{n}}_t(z,z',\ \tau) \times \\
&\big[(\tau - t - t_{\min}^{zz'}) \cdot \nabla_\theta \log(\pi_\theta^{zz'}(\mathbf{n}_t)) + (t_{\max}^{zz'} - (\tau - t)) \cdot \nabla_\theta \log(1 - \pi_\theta^{zz'}(\mathbf{n}_t))\,\big] V_t^{\mathbf{n}}(z,z',\tau)\Big]
\end{aligned}$$

□

1 Real Data experimental results for additional 10 days :

Day	Hour 4			Hour 5			Hour 6			Hour 7			Hour 8			Avg. Travel Time(0.6C)		
	Unsch.	**Sch.**	**SD**	**Unsch.**	**Sch.**	**SD**	**Unsch.**	**Sch.**	**SD**	**Unsch.**	**Sch.**	**SD**	**Unsch.**	**Sch.**	**SD**	**Unsch.**	**Sch.**	**SD**
11	3	3.21	0.5	4	2.95	0.98	7	1.68	0.82	6	0	0	3	1.19	0.66	207.96	107.35	1.44
12	3	2.7	0.74	5	2.21	0.86	7	4.31	0.61	6	4.47	1.31	1	4.46	1.06	206.25	153.7	2.04
13	10	6.21	0.83	8	3.1	1.15	10	1.63	0.73	3	2.04	0.69	0	4.04	0.94	213.09	125.87	1.45
14	4	3.97	1.03	4	2.47	0.75	5	2.35	0.77	2	4.1	0.93	3	6.47	1.24	222.05	125.76	1.45
15	6	3.83	0.81	9	4.79	0.77	8	1.65	0.78	4	7.33	1	0	8.94	1.28	209.4	125.99	1.46
16	5	1.59	0.62	8	1.94	0.63	9	0.63	0.66	2	0.17	0.4	1	3	0.8	207.63	107.73	2.23
17	8	8.38	0.66	8	2.9	0.92	8	2.2	0.77	2	3.1	0.98	3	5.44	1.37	199.9	148.51	1.37
18	2	3.71	1	12	2.66	0.89	11	3.49	1	2	1.97	0.93	0	3.62	0.97	200.27	159.67	1.42
19	5	2.57	0.86	7	3.63	1.14	5	2.15	1.08	3	6.42	1.11	1	7.28	1.18	195.04	103.05	1.87
20	3	1.19	0.5	2	1.65	0.57	1	1.5	0.75	0	4.1	1.07	1	5.28	1.25	192.67	117.41	1.58

2 Synthetic Data Experimental Setup :

The settings of all instances are as follows. We use a graph with 23 edges, for each edge, minimum travel time is set to 2 sec and maximum travel time is uniformly sampled from [10sec, 20sec]. Each vessel's arrival time at the starting edge is uniformly sampled from [1sec, 20sec], each vessel consumes one unit of resource when traversing an edge, for all experiments delay penalty w_d = 1 and horizon = 200. For each setting, we generate 5 instances and average values are reported.

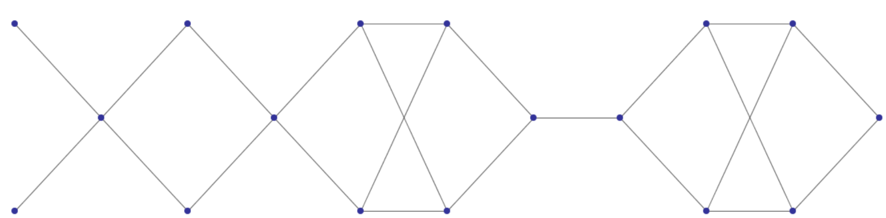


Figure 1: Graph of synthetic data experiments

2.1 DDPG Baseline :

As mentioned in the main document we use a DDPG algorithm [3] as one of our baselines. We learn a critic function $\tilde{V}_w^{zz'}(\mathrm{n}_t^{\mathrm{tot}}, \mathrm{n}_t^{\mathrm{nxt}}, \pi_t^{zz'})$ to estimate the vessel-based value of a waterway zz' given the counts $\mathrm{n}_t^{\mathrm{tot}}, \mathrm{n}_t^{\mathrm{nxt}}$ and traffic control action $\beta_t^{zz'} = \pi_t^{zz'}(\mathrm{n}_t^{\mathrm{tot}})$. For each sampled $\mathbf{n}_t$, the DDPG critic is updated by empirical vessel-based value as:

$$w^{new} \leftarrow w^{old} - \alpha_{lr} \sum_{z,z'} \nabla_w [\tilde{V}^{zz'}(\mathrm{n}_t^{\mathrm{tot}}, \mathrm{n}_t^{\mathrm{nxt}}, \pi_t^{zz'}) - \mathrm{n}_t^{\mathrm{nxt}}(z,z') \cdot V_t^{\mathbf{n}}(z,z')]^2$$

Then, vessel-based policy gradient for this DDPG is computed as follows

$$\theta^{new} \leftarrow \theta^{old} + \alpha_{lr} \sum_{z,z'} \nabla_\theta \tilde{V}^{zz'}(\mathrm{n}_t^{\mathrm{tot}}, \mathrm{n}_t^{\mathrm{nxt}}, \pi_t^{zz'})$$

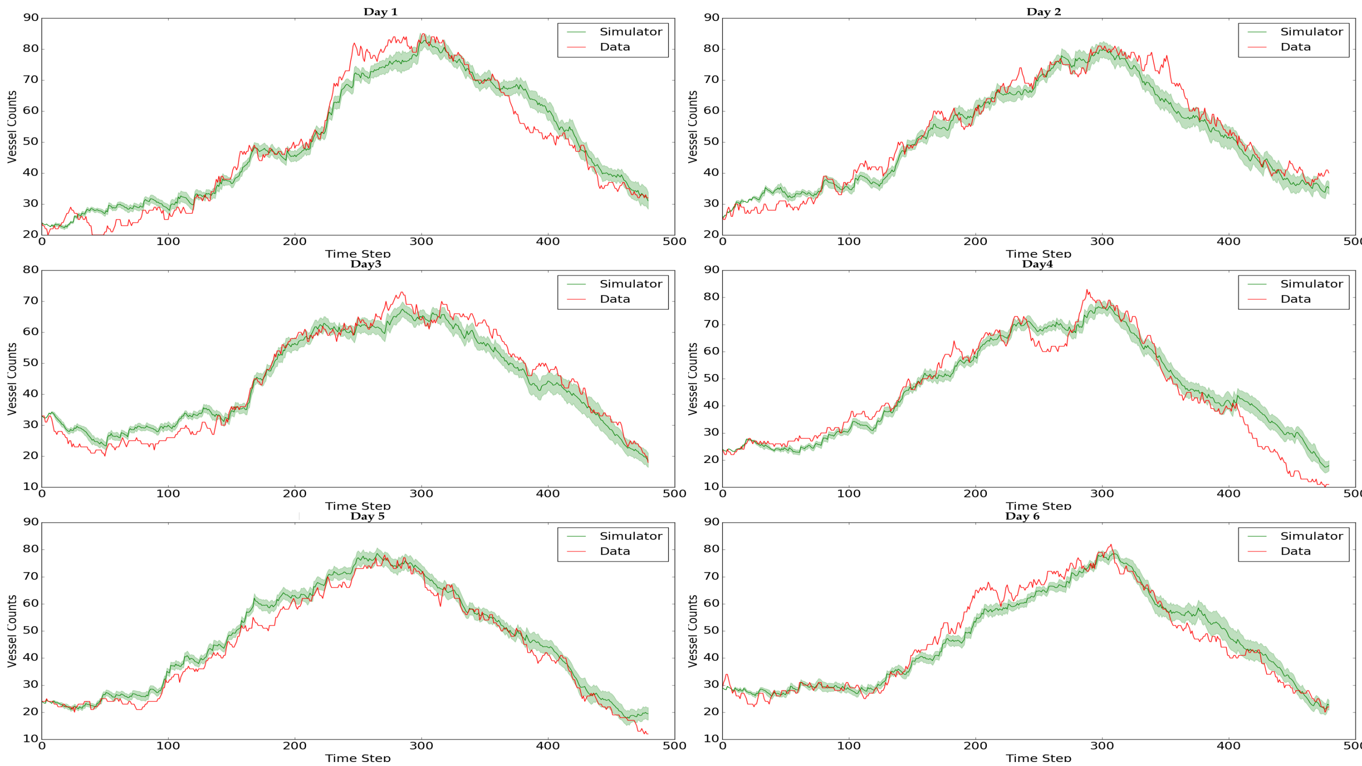
2.2 Implementation Details :

For Vessel-PG policy network π_θ, we use a one big neural network with each zone as one sub-network which are segregated from one another. Input to the big network is $\mathrm{n}_t^{\mathrm{tot}}$, we then apply masking on the input vector so that each zone(sub-network) receives only the count information of $\mathrm{n}_t^{\mathrm{tot}}(z)$ and neighboring zone count $\mathrm{n}_t^{\mathrm{tot}}(z')$. For each subnetwork, we use 2 hidden layer, each layer with hidden nodes = total zones and each zone sub-network outputs a vector $\langle \beta_t^{zz'} \rangle_{\forall z'}$ and $\beta_t^{zz'} \in [0,1]$. For each hidden layer, we use $\tanh$ activation and for the output layer sigmoid activation is used on each unit so that we get the value between [0, 1]. Layer-norm [1] is applied before each hidden layer and output layer. We use Adam optimizer [2] with learning rate 1e-3. All of our model are implemented on pytorch [5]. Same network architecture is also used for both DDPG policy and critic network, and PG.

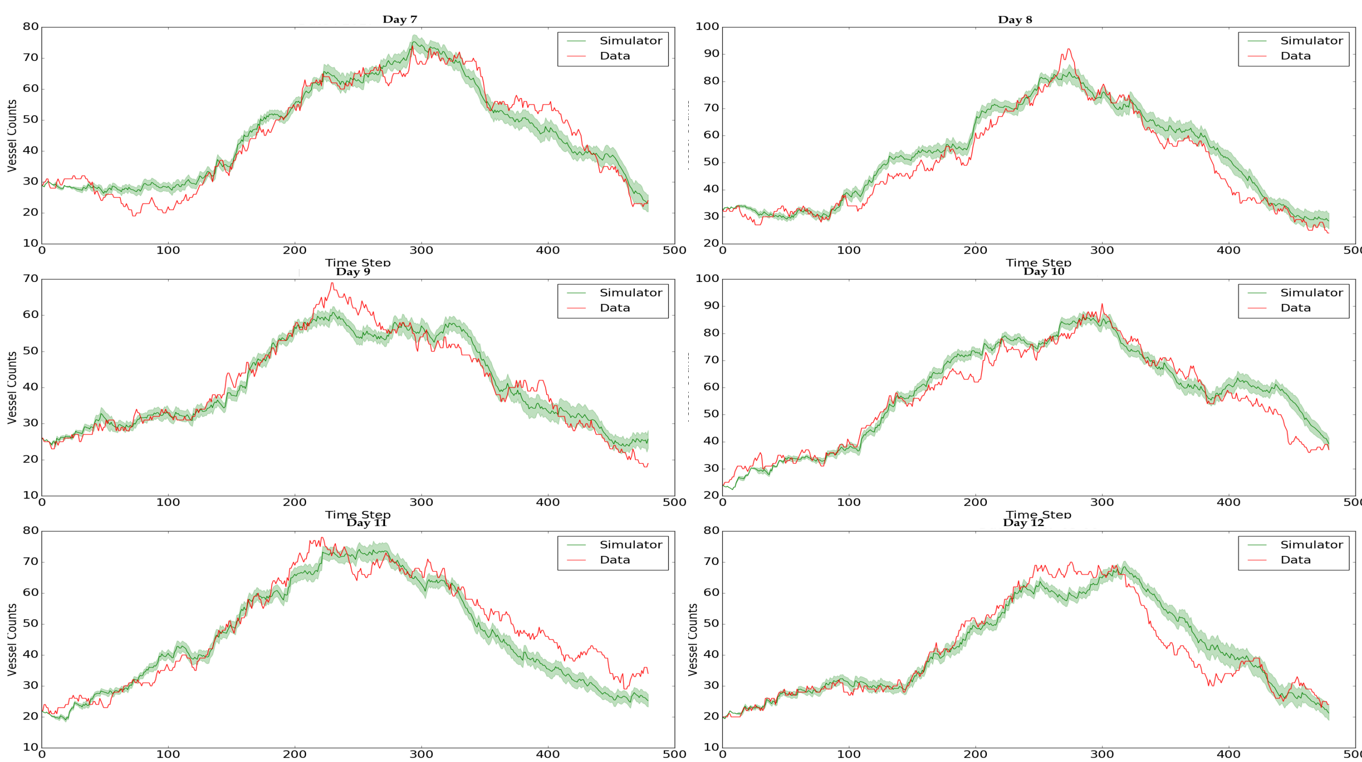
Algorithm 1:

Initialize network parameters $\theta^{\text{Vessel-PG}}$, θ^{PG} ,θ^{DDPG} for actor and w^{DDPG} for critic.
$\alpha \leftarrow$ actor learning rate
$\alpha' \leftarrow$ critic learning rate
repeat
  Sample count vectors $\mathbf{n}_{1:H} \sim P(\boldsymbol{n}_{1:H}; \pi_\theta)$
  Compute empirical individual values using (16) - (19)
  Update critic as :
  $w_{new}^{DDPG} \leftarrow w_{old}^{DDPG} - \alpha' \sum_{z,z'} \nabla_w [\tilde{V}^{zz'}(\mathrm{n}_t^{\mathrm{tot}}, \mathrm{n}_t^{\mathrm{nxt}}, \pi_t^{zz'}) - \mathrm{n}_t^{\mathrm{nxt}}(z,z') \cdot V_t^{\mathbf{n}}(z,z')]^2$
  Update actor as :
  $\theta_{new}^{\text{Vessel-PG}} \leftarrow \theta_{old}^{\text{Vessel-PG}} + \alpha \sum_{z,z'} \nabla_\theta V^{zz'}(\mathrm{n}_t^{\mathrm{tot}}, \mathrm{n}_t^{\mathrm{nxt}}, \pi_t^{zz'})$
  $\theta_{new}^{DDPG} \leftarrow \theta_{old}^{DDPG} + \alpha \sum_{z,z'} \nabla_\theta \tilde{V}^{zz'}(\mathrm{n}_t^{\mathrm{tot}}, \mathrm{n}_t^{\mathrm{nxt}}, \pi_t^{zz'})$
  $\theta_{new}^{PG} \leftarrow \theta_{old}^{PG} + \alpha \nabla_\theta R$
until ***convergence***
return $\theta^{\text{Vessel-PG}}$, θ^{PG} ,θ^{DDPG}, w^{DDPG}

3 Simulator Accuracy Plot:

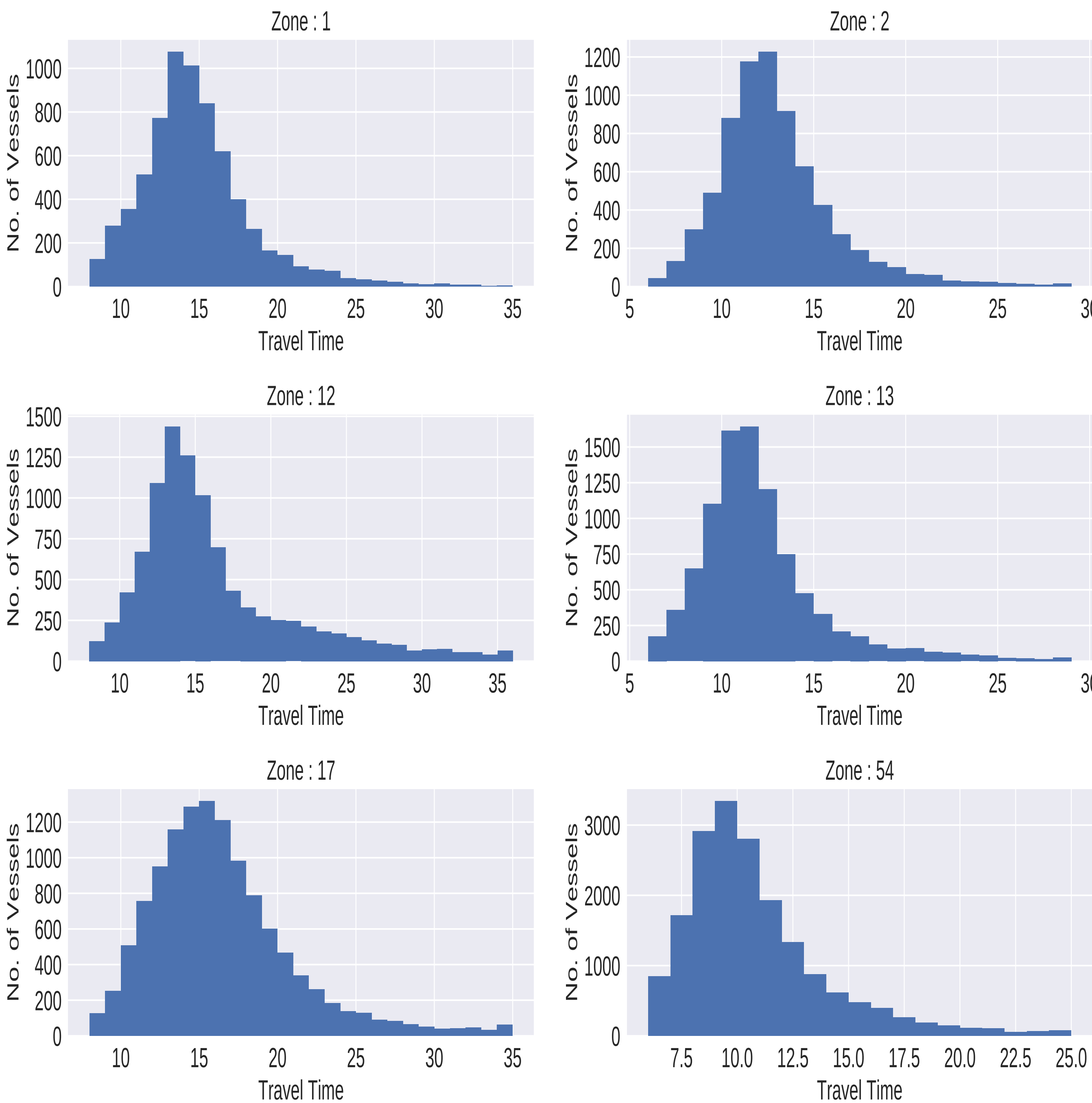


(a)



(b)

4 Real Data Travel Time of high traffic zones over 4 months period :



References

[1] Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[3] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[4] Duc Thien Nguyen, Akshat Kumar, and Hoong Chuin Lau. Collective multiagent sequential decision making under uncertainty. In *AAAI Conference on Artificial Intelligence*, pages 3036–3043, 2017.

[5] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.