★ 网络天下 ★

数据通信路由技术与验证算法技术论坛

基于联邦学习的无线流量预测

张传亭
山东大学
山东省无线通信技术重点实验室

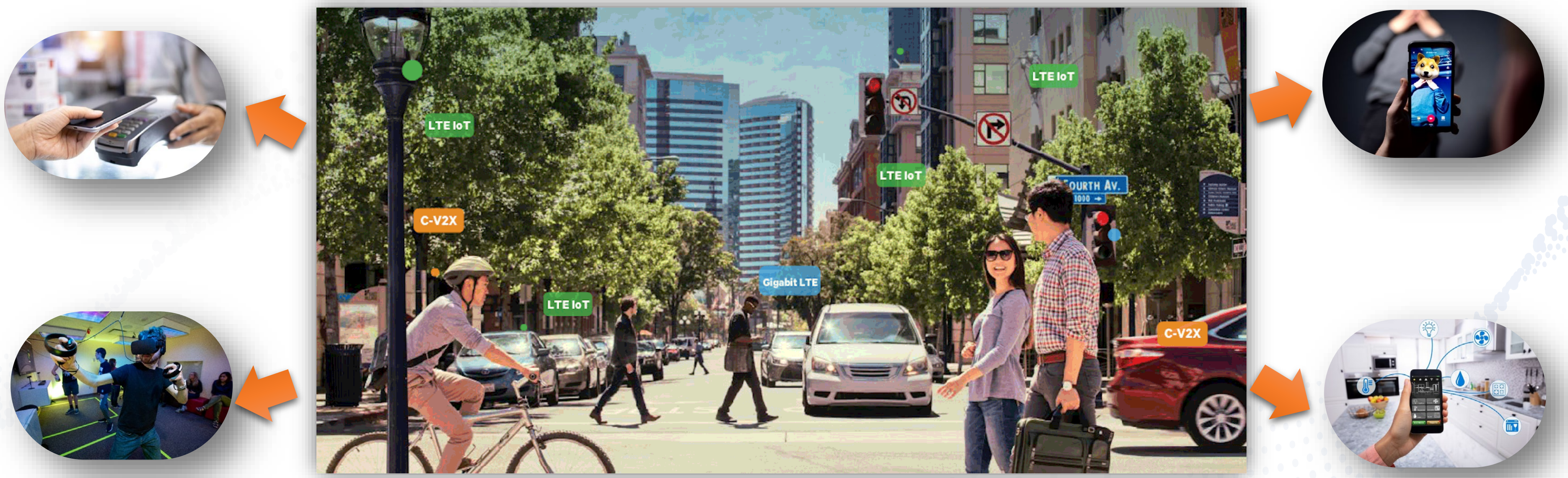
SHANDONG UNIVERSITY 1901

NetAI
Networked Intelligence Lab

Content

Wireless Big Data Analysis Matters

Modeling, analyzing, and predicting wireless service traffic in the communication systems are pivotal to achieving network intelligence.

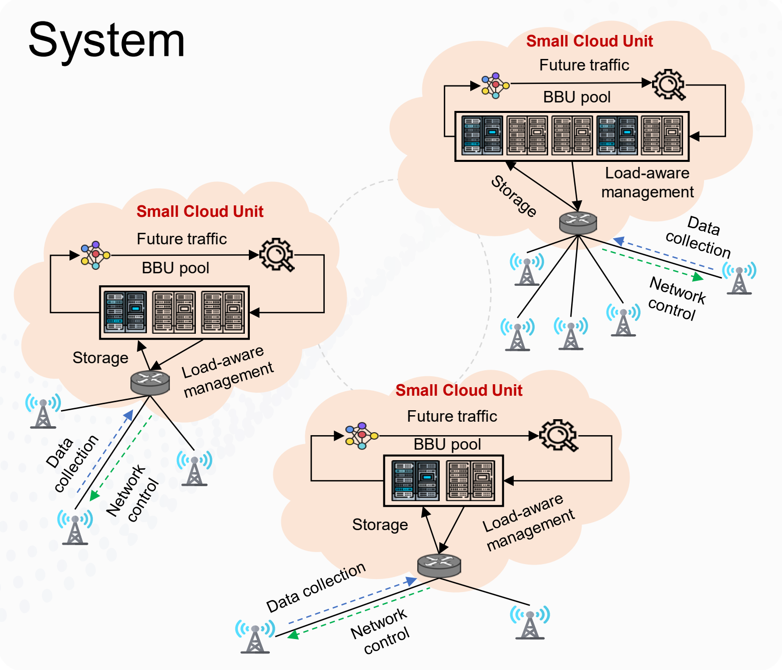


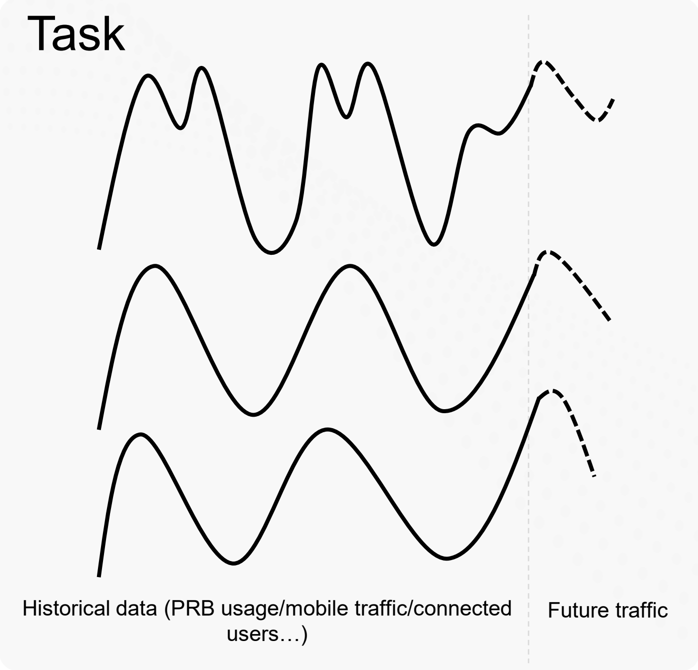
Wireless Big Data Analysis Matters

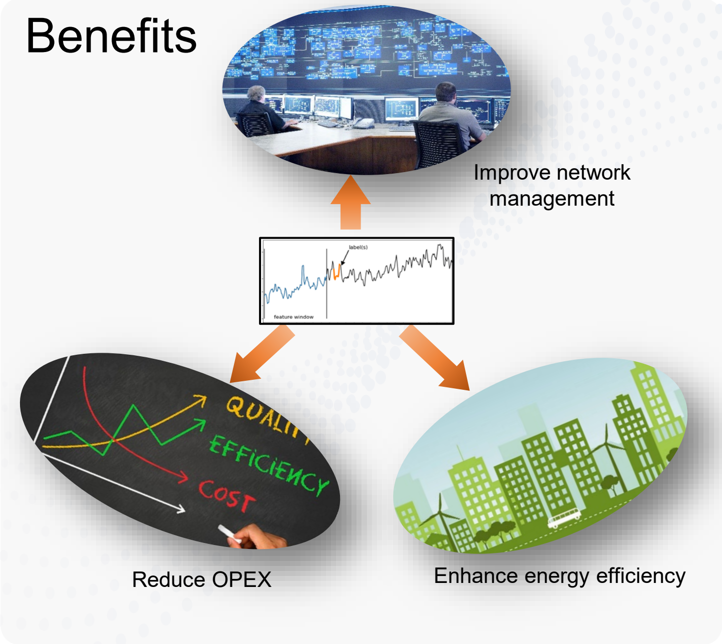
- Wireless traffic prediction is crucial in future AI-empowered communication systems, with prediction we can:
  - **Improve network management** through dynamic congestion control
  - **Reduce operating expenditure** by accurate radio resource purchase
  - **Enhance energy efficiency** by intelligent BS on/off

Wireless Traffic Prediction

Predicting communication system's future traffic (**short- or long-term**) based on **historical wireless traffic** and **cross-domain data** by designing **novel AI algorithms** under **various application scenarios**.







Existing Methods and Challenges

- Centralized methods, e.g., ST-DenseNet[1] and STC-Net[2]
  - Need to transfer raw data to datacenter to learn a generalized model
  - Consume lots of bandwidth
  - May have high latency for mission-critical tasks
  - Involve no cooperation from multiple MNO due to data privacy
- Fully distributed methods, e.g., Gaussian Process Regression[3]
  - Could not capture spatial dependences among different BSs/cells/regions
  - May have limited data, especially in places with newly deployed infrastructures
  - Involve no cooperation also due to data privacy

1. C. Zhang, H. Zhang, D. Yuan and M. Zhang, "Citywide Cellular Traffic Prediction Based on Densely Connected Convolutional Neural Networks," in *IEEE Communications Letters*, vol. 22, no. 8, pp. 1656-1659, Aug. 2018
2. C. Zhang, H. Zhang, J. Qiao, D. Yuan and M. Zhang, "Deep Transfer Learning for Intelligent Cellular Traffic Prediction Based on Cross-Domain Big Data," *in IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1389-1401, June 2019
3. Y. Xu, F. Yin, W. Xu, J. Lin and S. Cui, "Wireless Traffic Prediction With Scalable Gaussian Process: Framework, Algorithms, and Verification," in *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1291-1306, June 2019

What Do We Need for Wireless Traffic Prediction

❑We need a model that can
- Capture both **spatial** and **temporal** dependencies
- Be trained/**deployed at the edge** to reduce latency
- **Without transferring data** from local to datacenter
- Collaborate between **multiple MNOs** to fully release the power of data

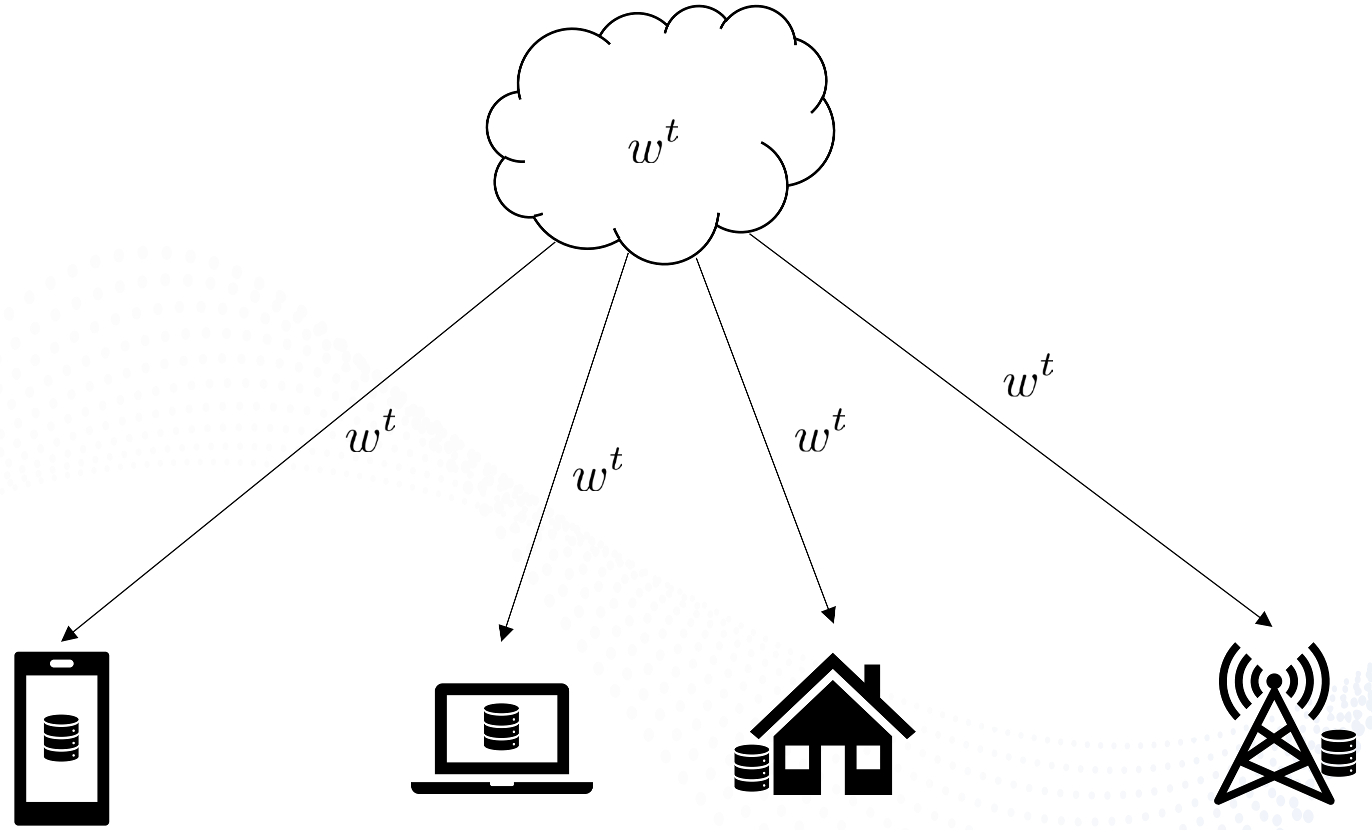
❑Federated learning can fulfill the above requirements
- Temporal dependencies are modeled by local model, spatial dependencies are captured through model aggregation
- Can be trained/deployed at the edge sever
- No need to transfer raw data, just model
- Can be readily shared among different MNOs

Federated Learning Basics

Central server

Round t

Local clients



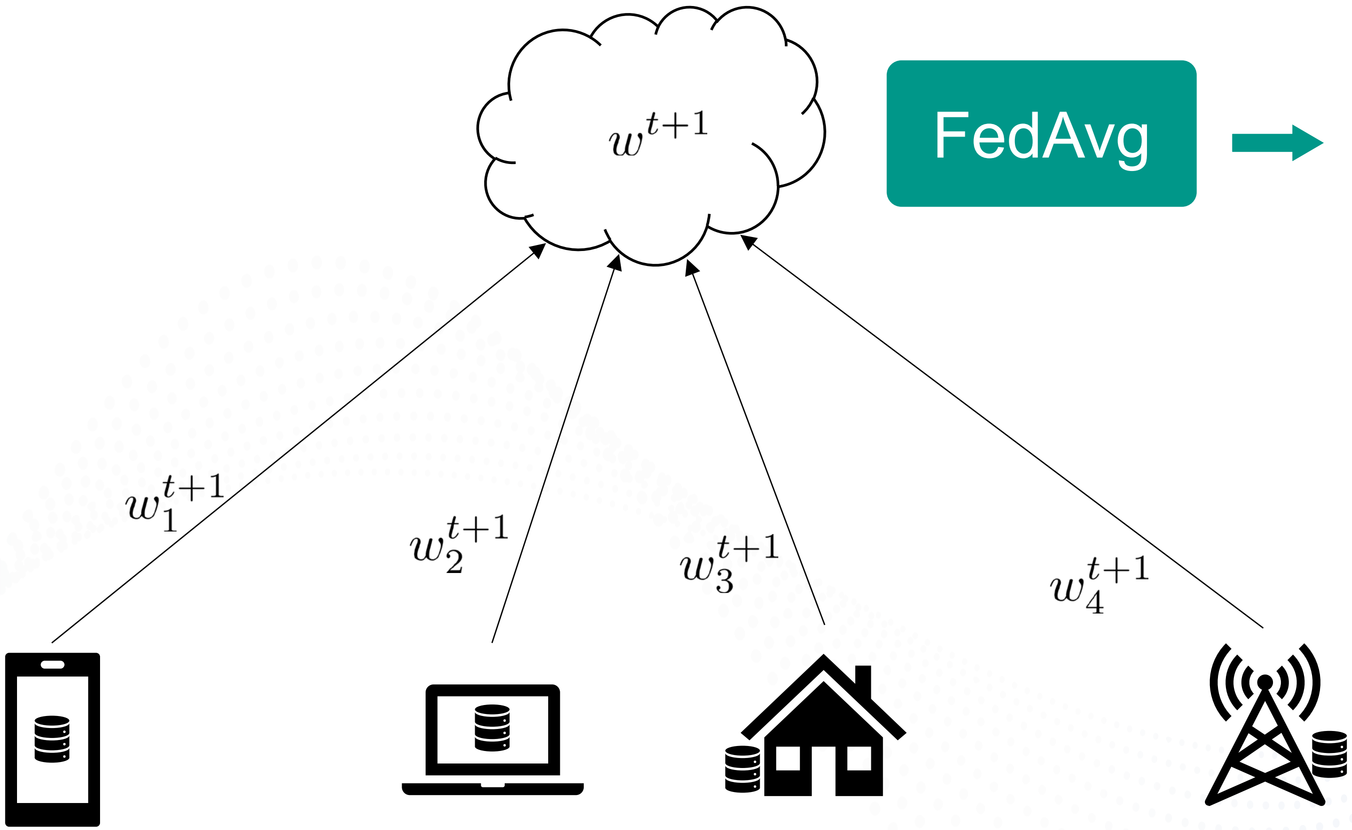
Update w^t using local data and get new local model w_k^{t+1}

Federated Learning Basics

Central server

Round t

Local clients



FedAvg → $w^{t+1} = \sum_k w_k^{t+1}$

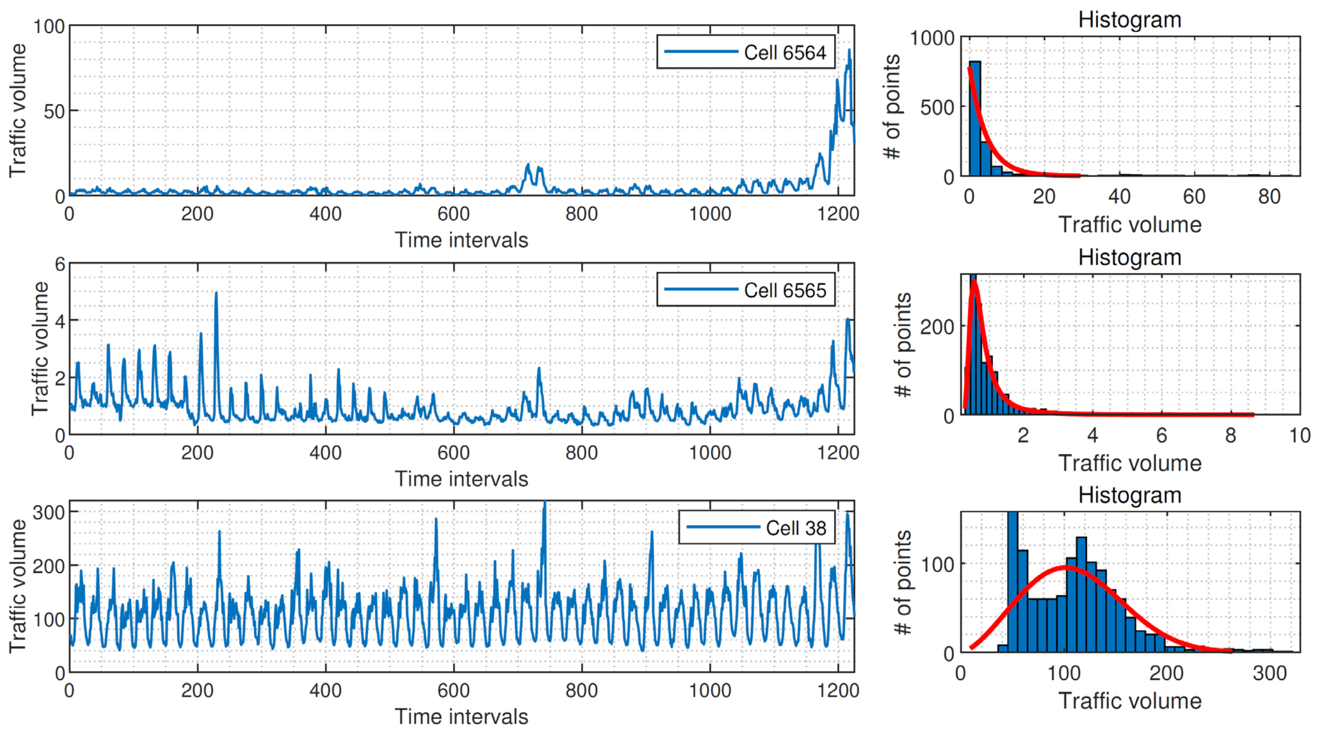
Iterate the above process until end, then we get the final model

Send w_k^{t+1} to the central server for model aggregation

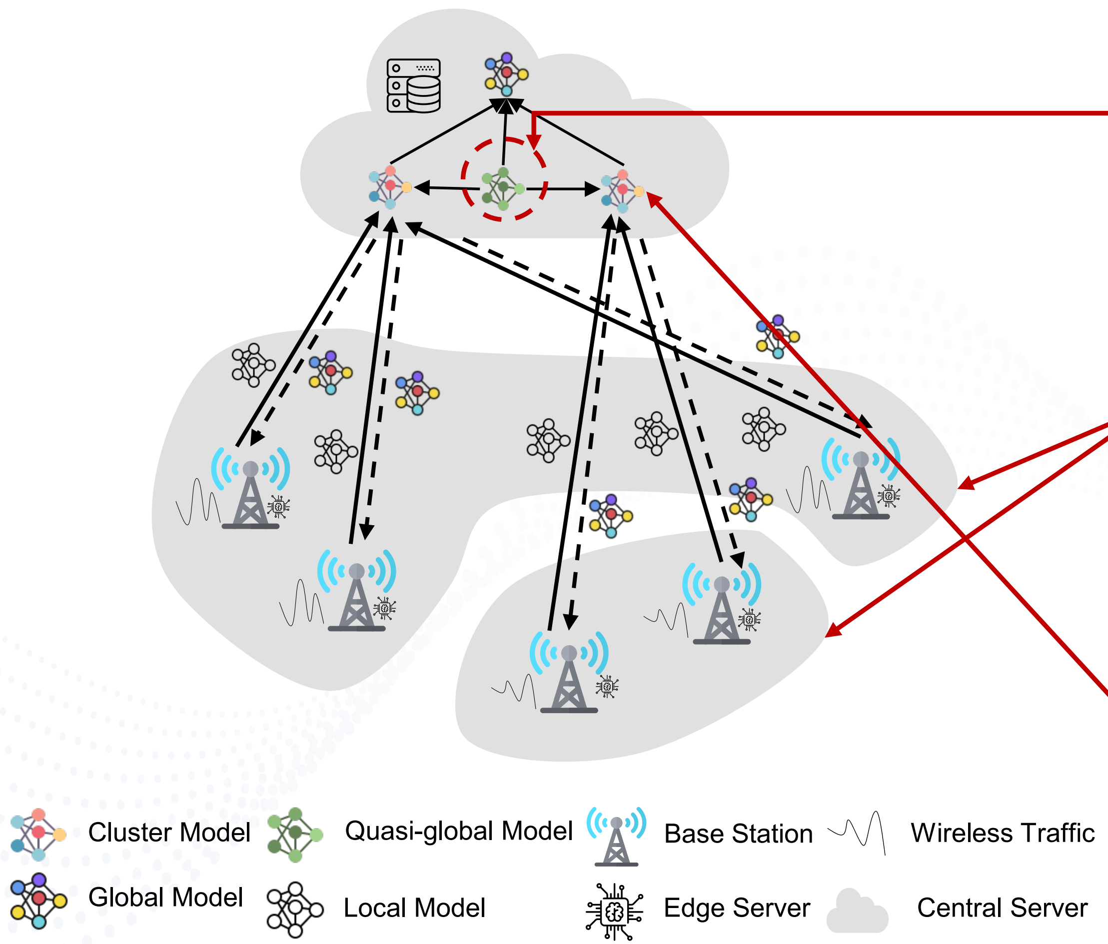
Content

FedAvg For Wireless Traffic Prediction

- It works, but suffers from precision problem
  - Wireless traffic data are highly heterogenous, different places have different traffic patterns
  - Simple average of local model to produce the global one generalize not well
- Motivation
  - Train a well-generalized global model by reducing heterogeneity of wireless traffic



System Model and Workflow



Data augmentation

Each BS calculates its ***hourly-average traffic value per week*** and send the $\varphi\%$ data to central, where a ***quasi-global model*** is trained

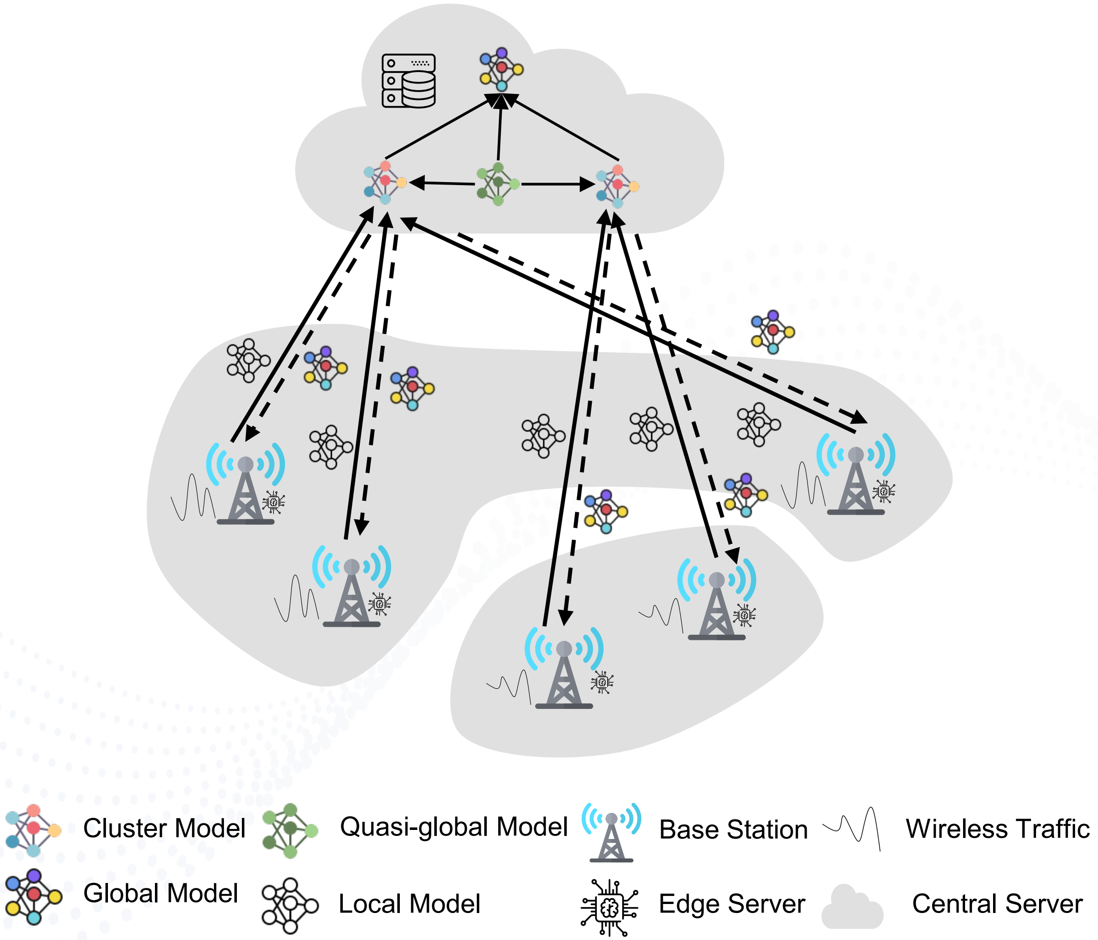
BS clustering

Central server ***cluster the BSs into different groups***, based on the augmented data and geolocation

Model training

Perform global model training using ***dual attention-based optimization*** scheme

Problem Formulation



# of clusters

# of BSs in cluster c

# of BSs

$$\min_{w\in\mathbb{R}^d} f(x) = \frac{1}{KC}\sum_{c=1}^{C}\sum_{k=1}^{K_c} F_{c,k}(w)$$

Loss function

parameters

# of data samples

Specific error function

$$F_{c,k}(w) = \sum_{i=1}^{N}\mathcal{L}(w; x_i, y_i)$$

$y_i = x_t$

$$x_i = \{x_{t-pL}, \cdots, x_{t-L}, x_{t-\tau}, \cdots, x_{t-1}\}$$

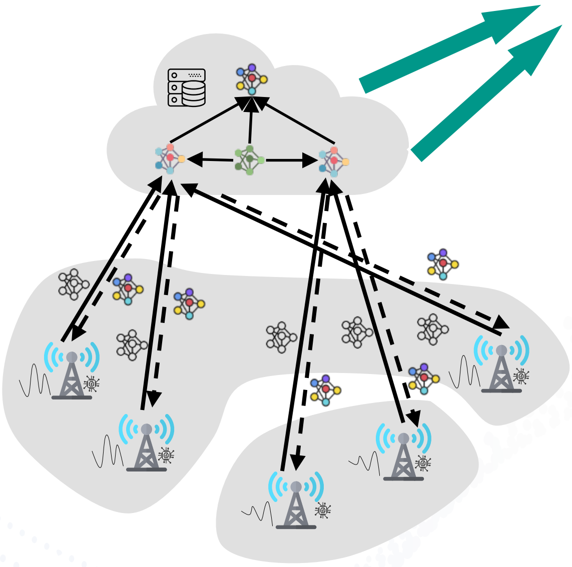
Length of periodicity

Period span

Length of closeness

Time index

Dual Attention Based Federated Optimization



$$\arg\min_{w^{t+1}} \left\{ \sum_{c=1}^{C} \frac{1}{2} \alpha_c \mathcal{L}(w^t, w_c^{t+1})^2 + \frac{1}{2} \rho \beta \mathcal{L}(w^t, w_Q)^2 \right\}$$

Local Attention

Layer-wise attention score computed via the distance between ***local*** model and the ***global*** model

Quasi-global Attention

Layer-wise attention score computed via the distance between ***quasi-global*** model and the ***global*** model

Intra-cluster update
Inter-cluster update

The global model has a minimum Euclidean distance to each local model (***enhance generalization***) and quasi-global model (***reduce heterogeneity***) in parameter space.

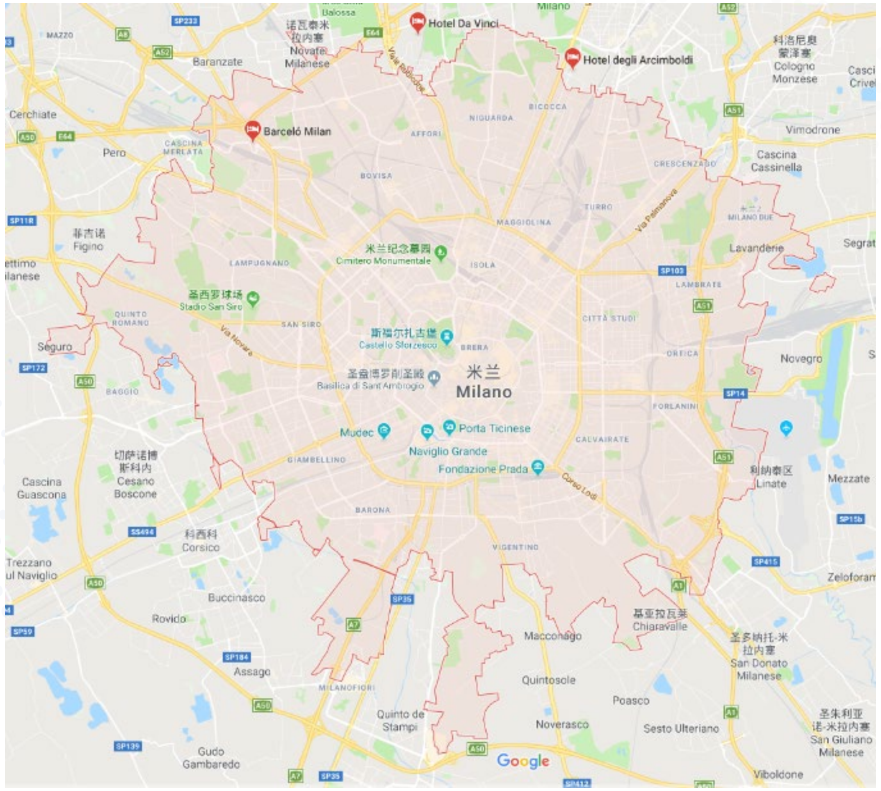
FedDA

Local update $$w_c^{t+1} = w_c^t - \eta \nabla \mathcal{L}(w_c^t; x, y)$$

Server update $$w^{t+1} = w^t - \gamma \{ \sum_{c=1}^{C} \alpha_c (w^t - w_c^{t+1}) + \rho\beta(w^t - w_Q) \}$$

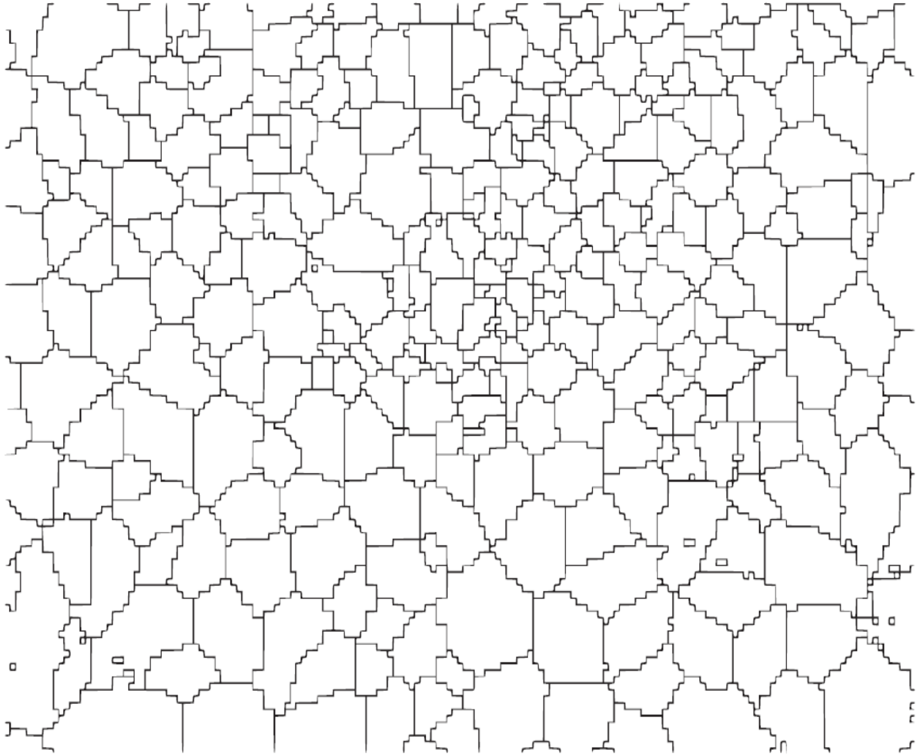
Dataset Explanation

- The wireless traffic data analyzed here comes from a large telecommunications service provider in Europe, Telecom Italia, as part of the “Big Data Challenge”



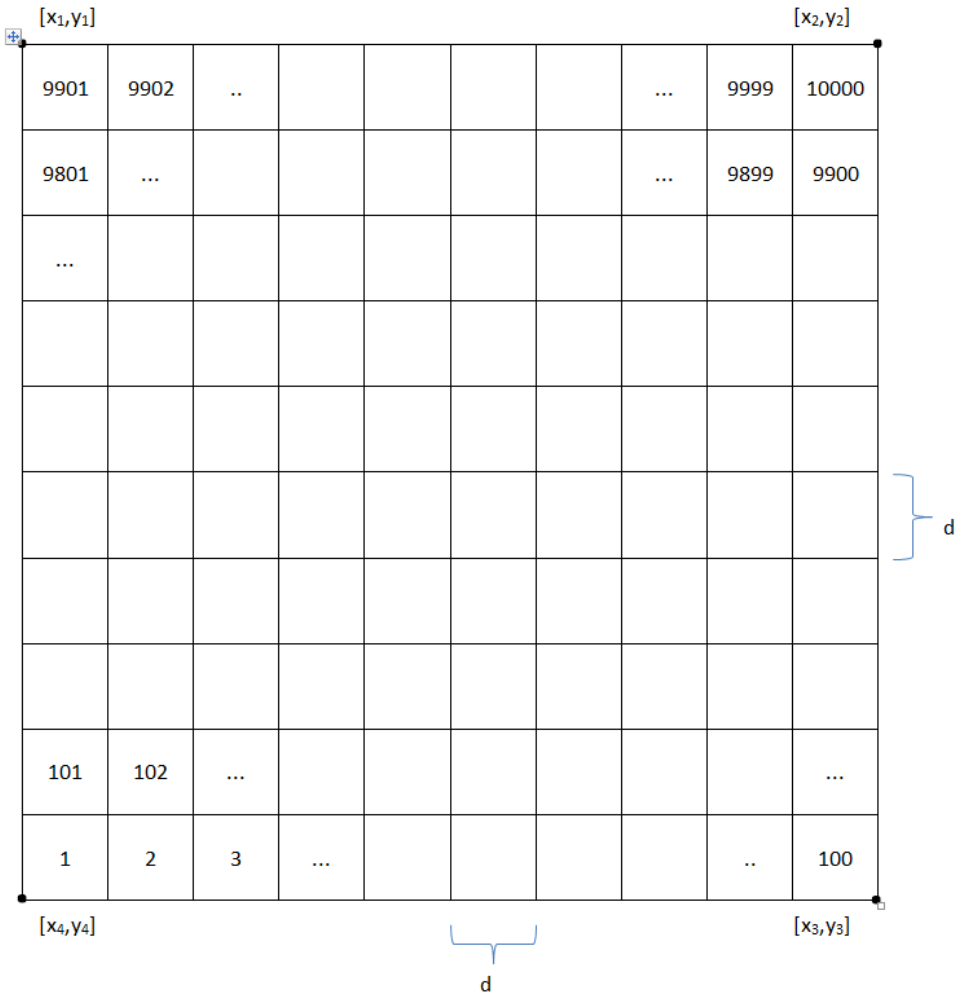
City boundary of Milan

BS coverage in the city of Milan

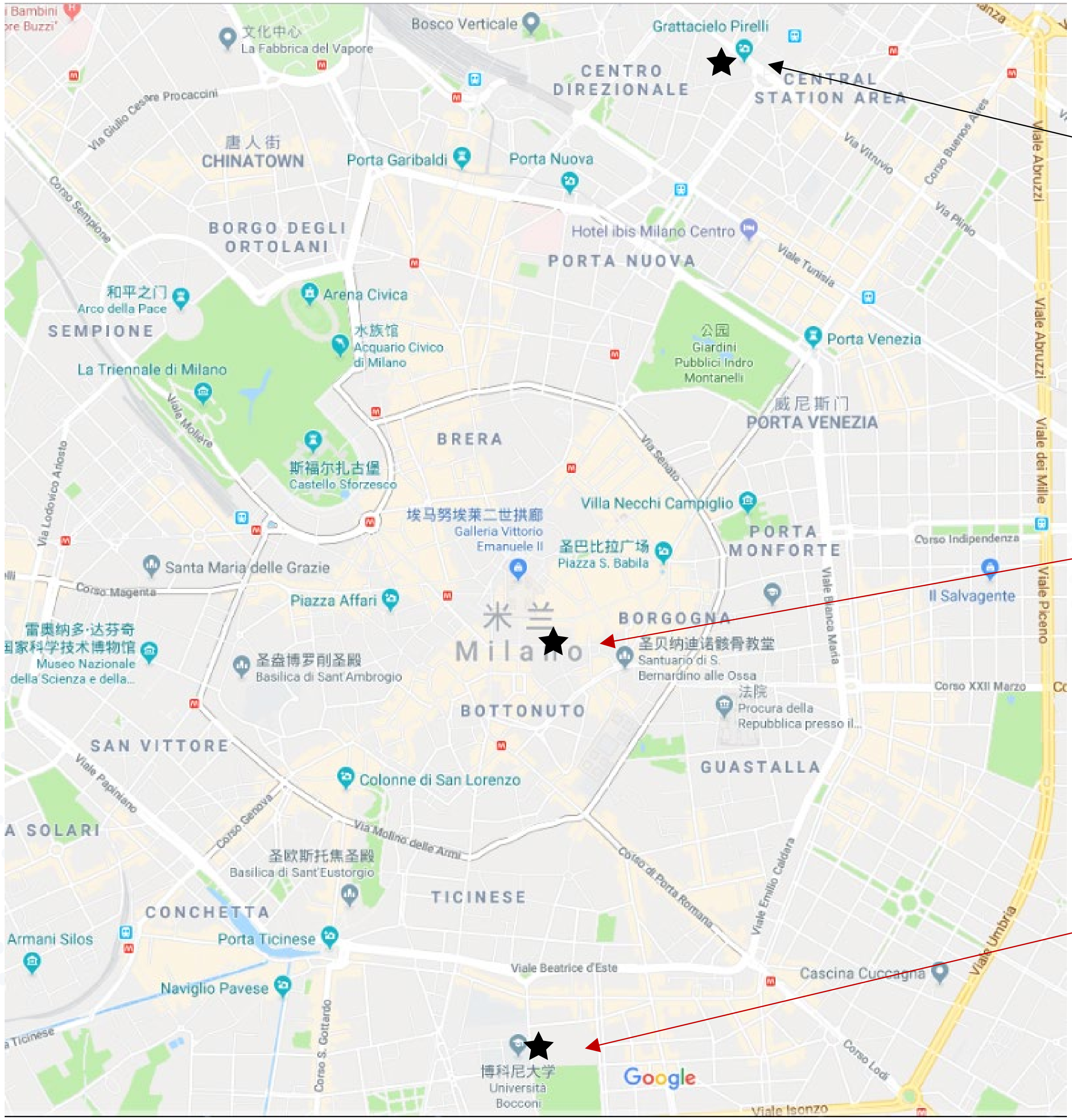


Data Introduction

- The city is divided into 100*100 cells and each cell covers an area of 235m * 235m
- Each cell's call detailed records (CDRs) of Telecom Italia were logged
  - SMS-In / SMS-Out
  - Call-In / Call-Out
  - Internet traffic
- Time granularity & span
  - 10 minutes
  - Two months from 2013-11-01 to 2014-01-01



Selected Areas for Spatiotemporal Analysis



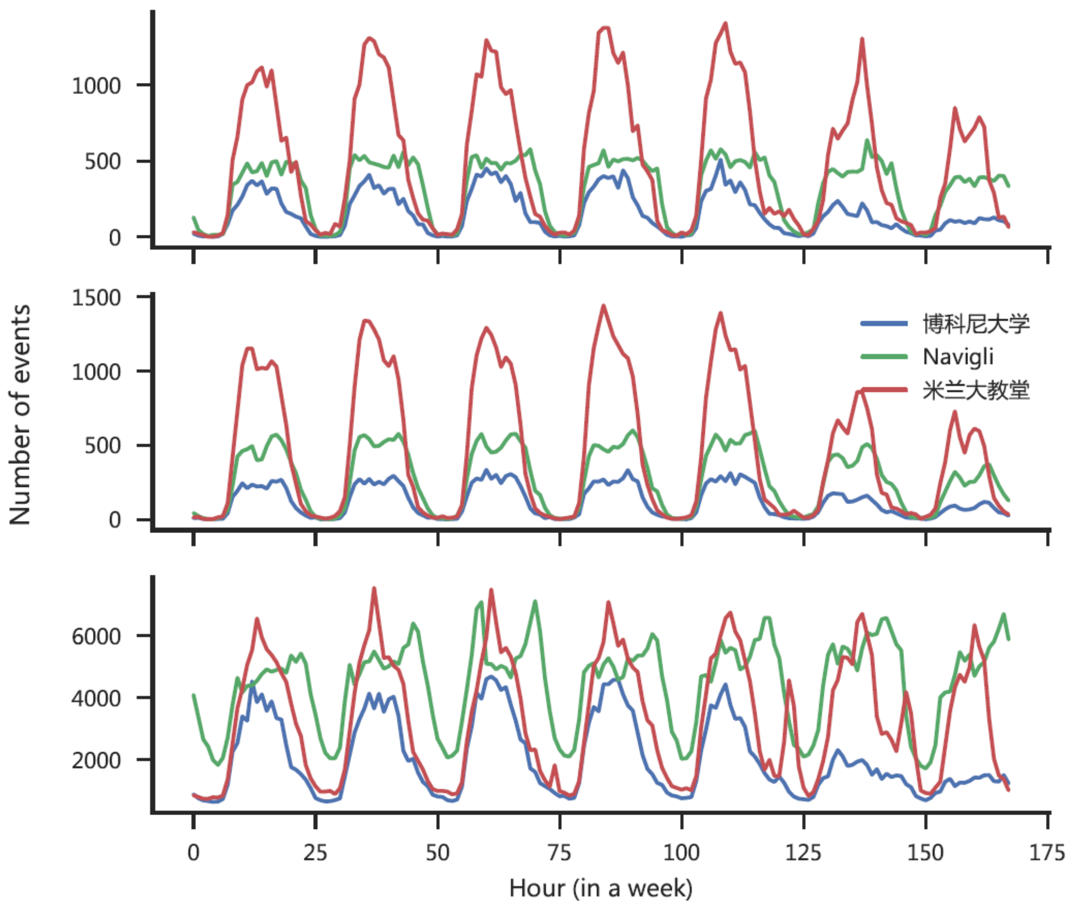
Navigli: a famous place for night life in the city of Milan, lots of bars and entertainment spots

Duomo di Milano: the Duomo is one of Europe’s greatest architectural and cultural landmarks. Italy’s largest church

Bocooni University: a place for study

Data Analysis From the Temporal View

❑Temporal traffic dynamics



In the city center, all three types of traffic are high.

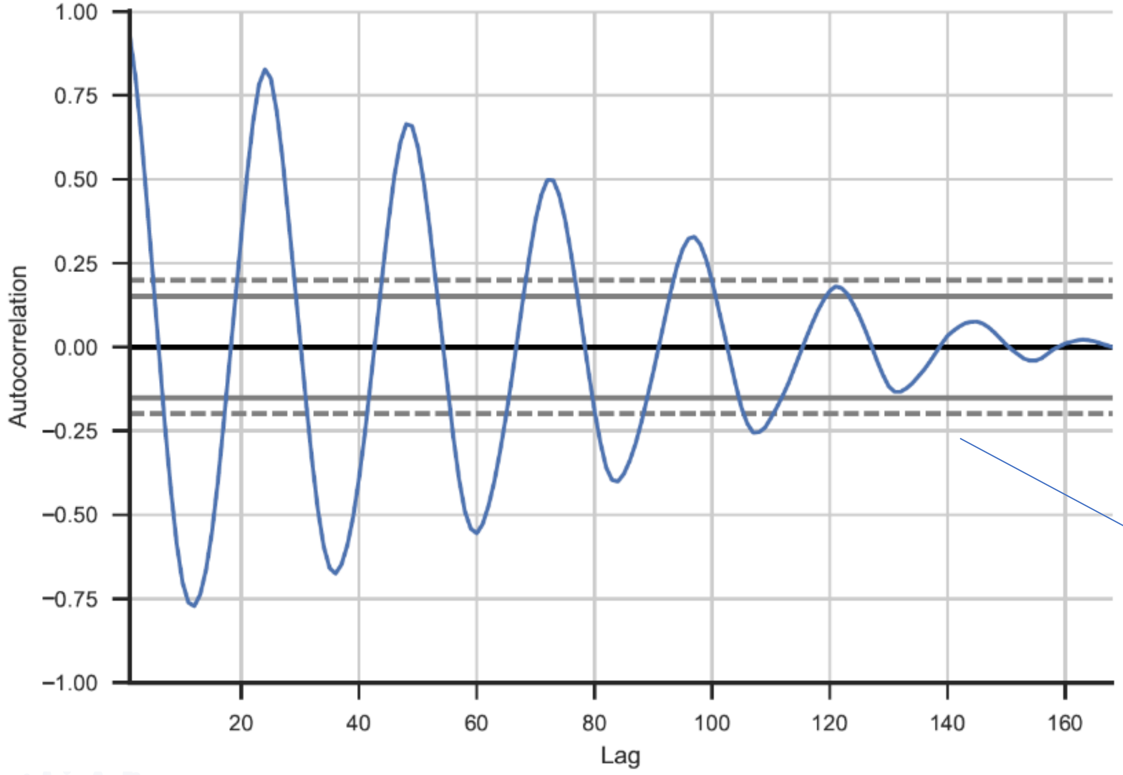
In bustling areas, SMS and CALL traffic decrease on weekends, but internet usage remains unchanged.

In the Navigli area, traffic increases with the arrival of night, showing a significant “delay” during peak hours.

The pattern of internet usage is relatively complex.

Data Analysis From the Temporal View

❑Temporal autocorrelation



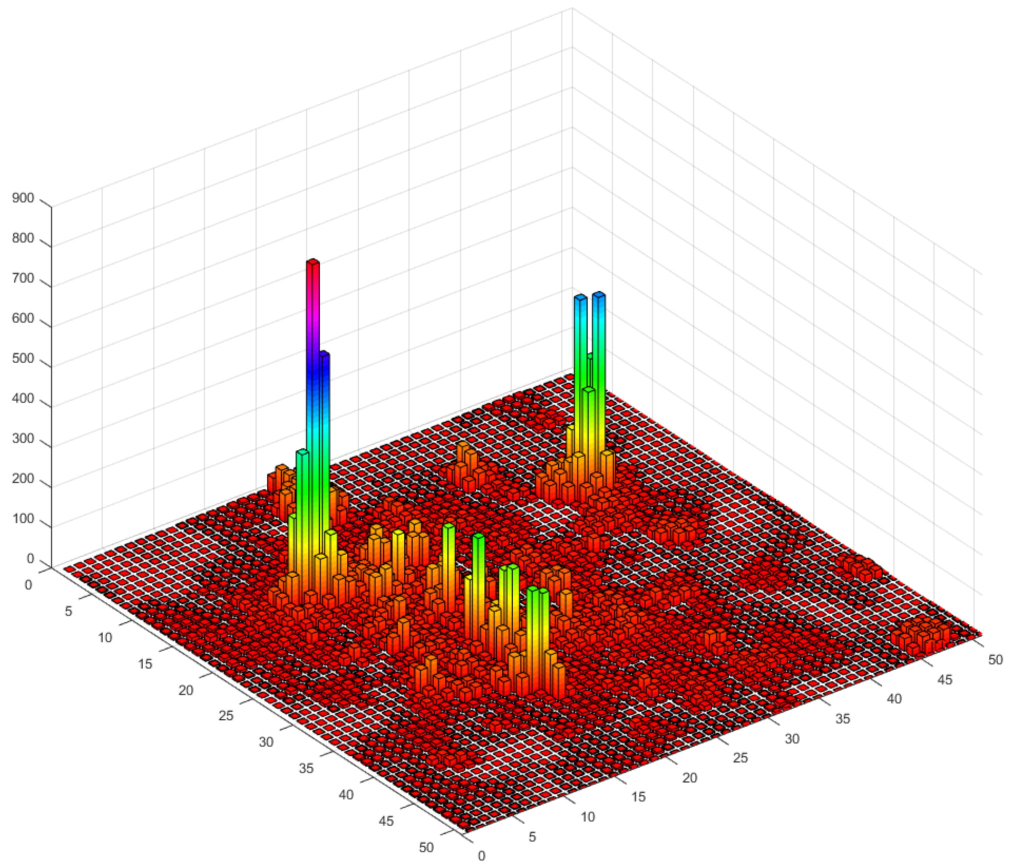
Autocorrelation, also known as serial correlation, is the correlation of a signal with itself at different points in time.

99% confidence interval. If the signal is randomly generated and has no pattern, then the autocorrelation values at any lag approach 0.

$$R(k) = \frac{E[(x_i - \mu_i)(x_{i+k} - \mu_{i+k})]}{\sigma^2}$$

Data Analysis From the Spatial Perspective

❑Spatial traffic dynamics and its correlations

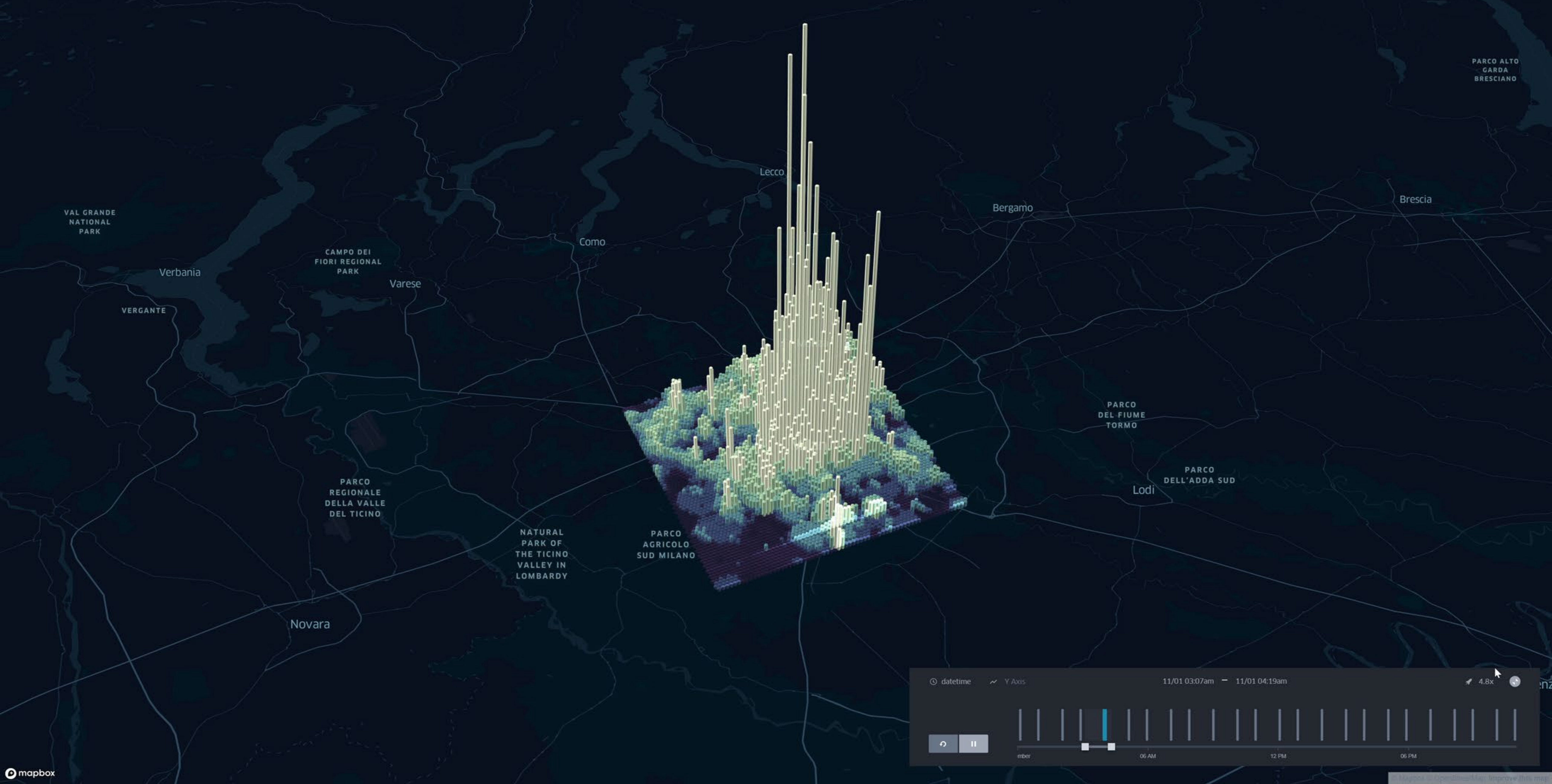


	1	2	3	4	5	6	7	8	9
1	0.62	0.75	0.7	0.56	0.46	0.46	0.69	0.75	0.83
2	0.53	0.65	0.75	0.71	0.77	0.67	0.63	0.8	0.78
3	0.61	0.62	0.72	0.71	0.82	0.73	0.63	0.7	0.72
4	0.69	0.67	0.71	0.71	0.86	0.87	0.81	0.8	0.75
5	0.67	0.7	0.74	0.94	1	0.85	0.84	0.79	0.79
6	0.64	0.78	0.75	0.95	0.96	0.82	0.84	0.75	0.82
7	0.66	0.8	0.76	0.82	0.89	0.87	0.83	0.79	0.79
8	0.58	0.71	0.73	0.74	0.84	0.84	0.81	0.6	0.81
9	0.61	0.65	0.73	0.77	0.85	0.83	0.81	0.81	0.8

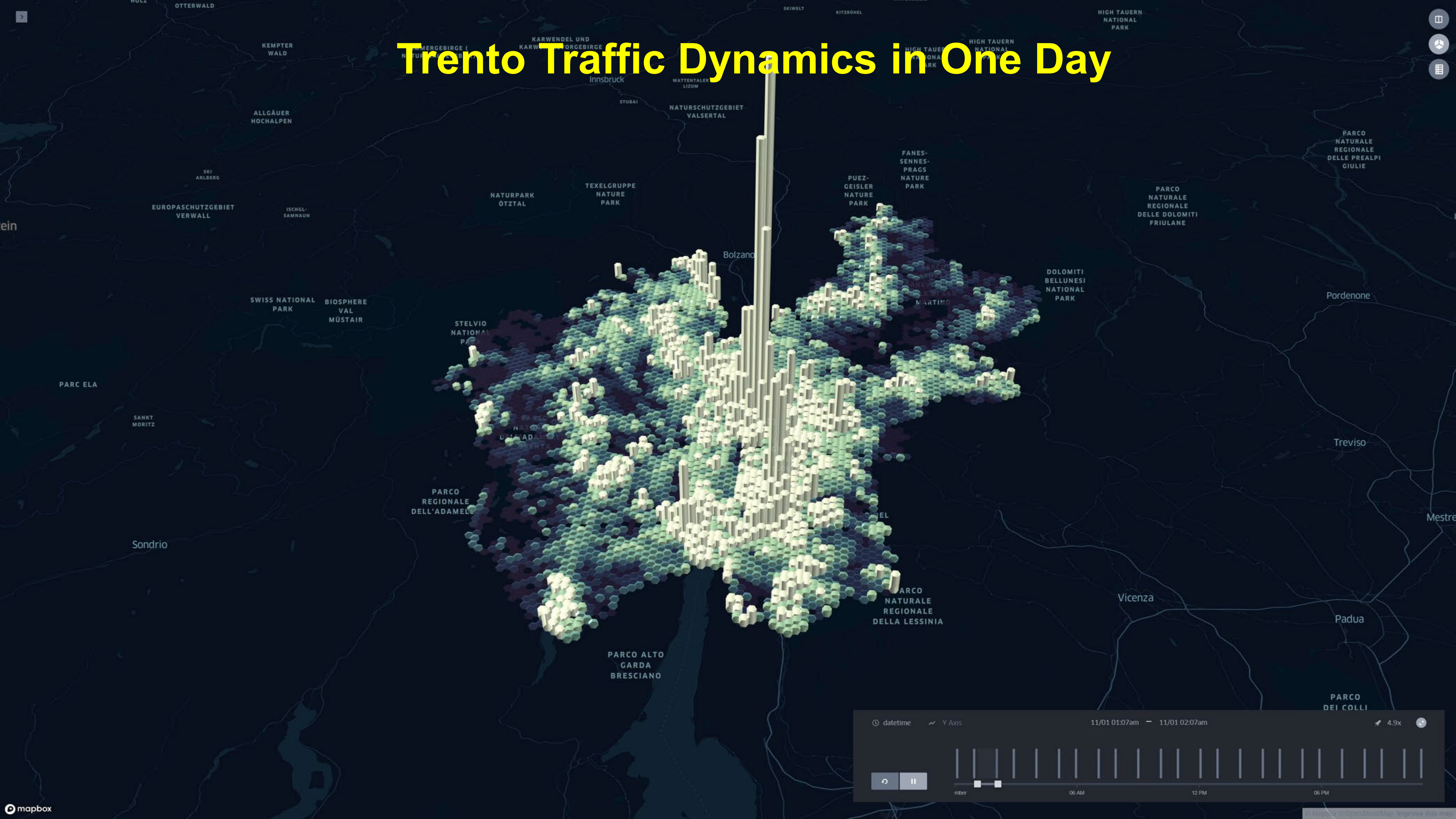
$$\rho_{\mathbf{x},\mathbf{y}} = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} = \frac{E((\mathbf{x}-\mu_{\mathbf{x}})(\mathbf{y}-\mu_{\mathbf{y}}))}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}}$$

The spatial distribution is uneven, but small intervals have correlation, with the strength of correlation related to distance.

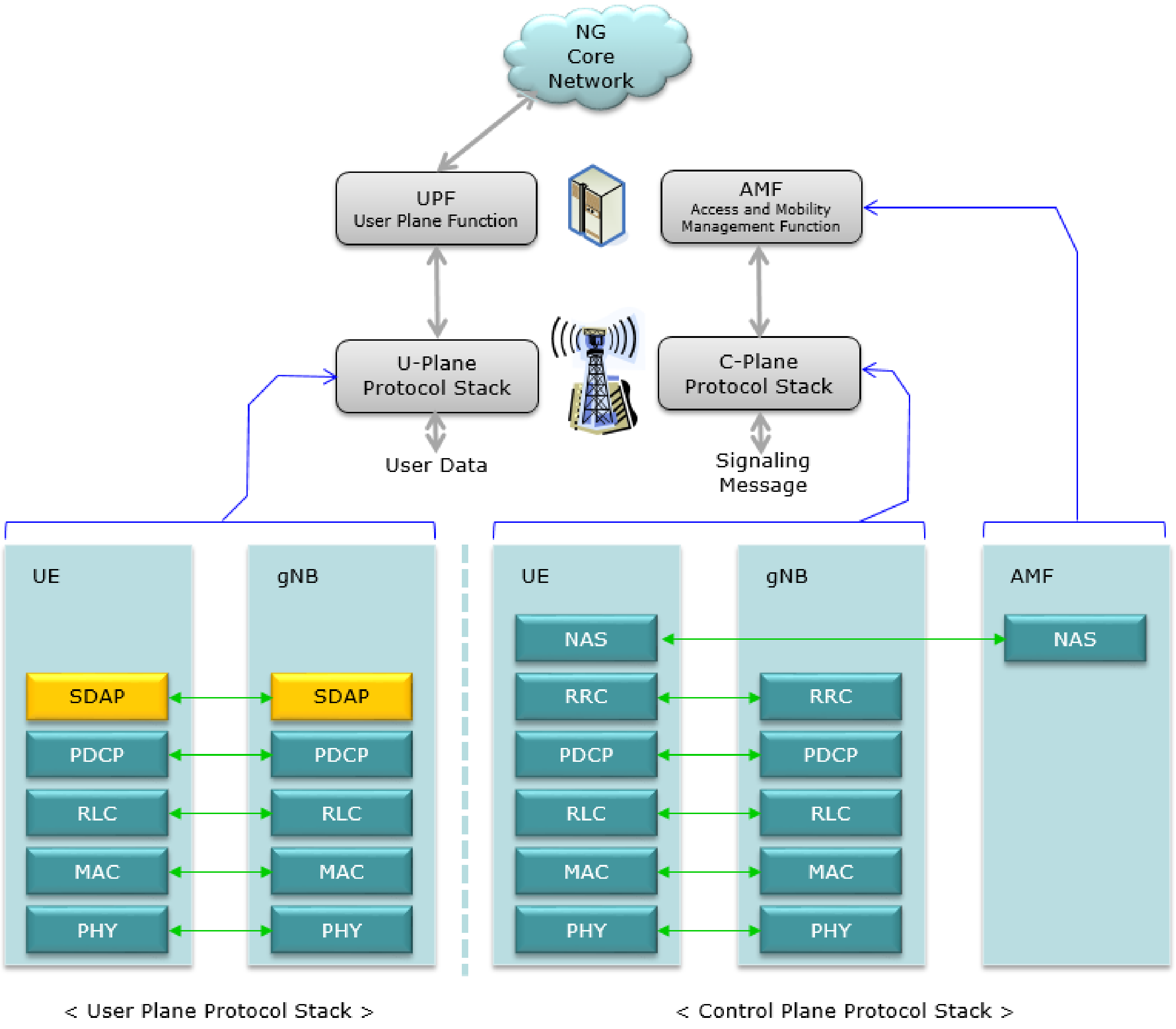
Milan Traffic Dynamics

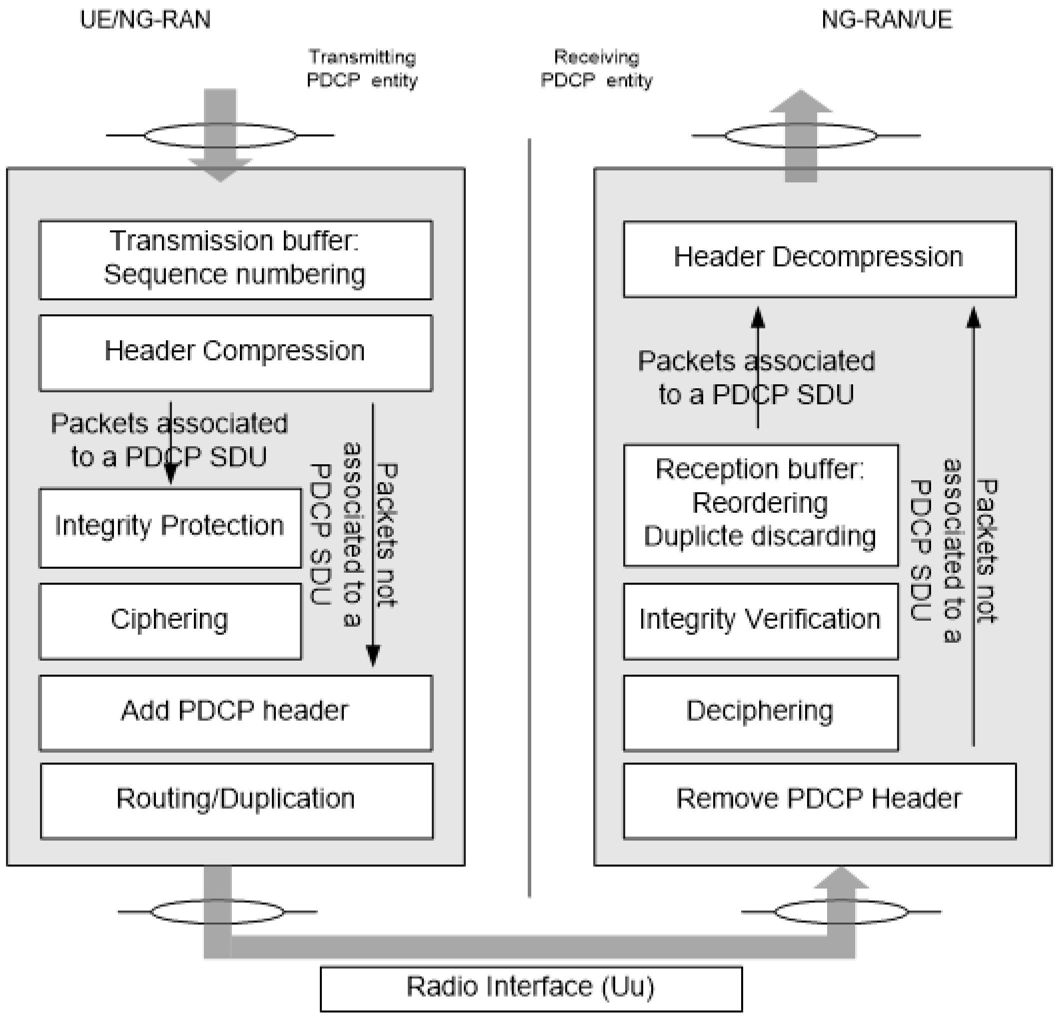


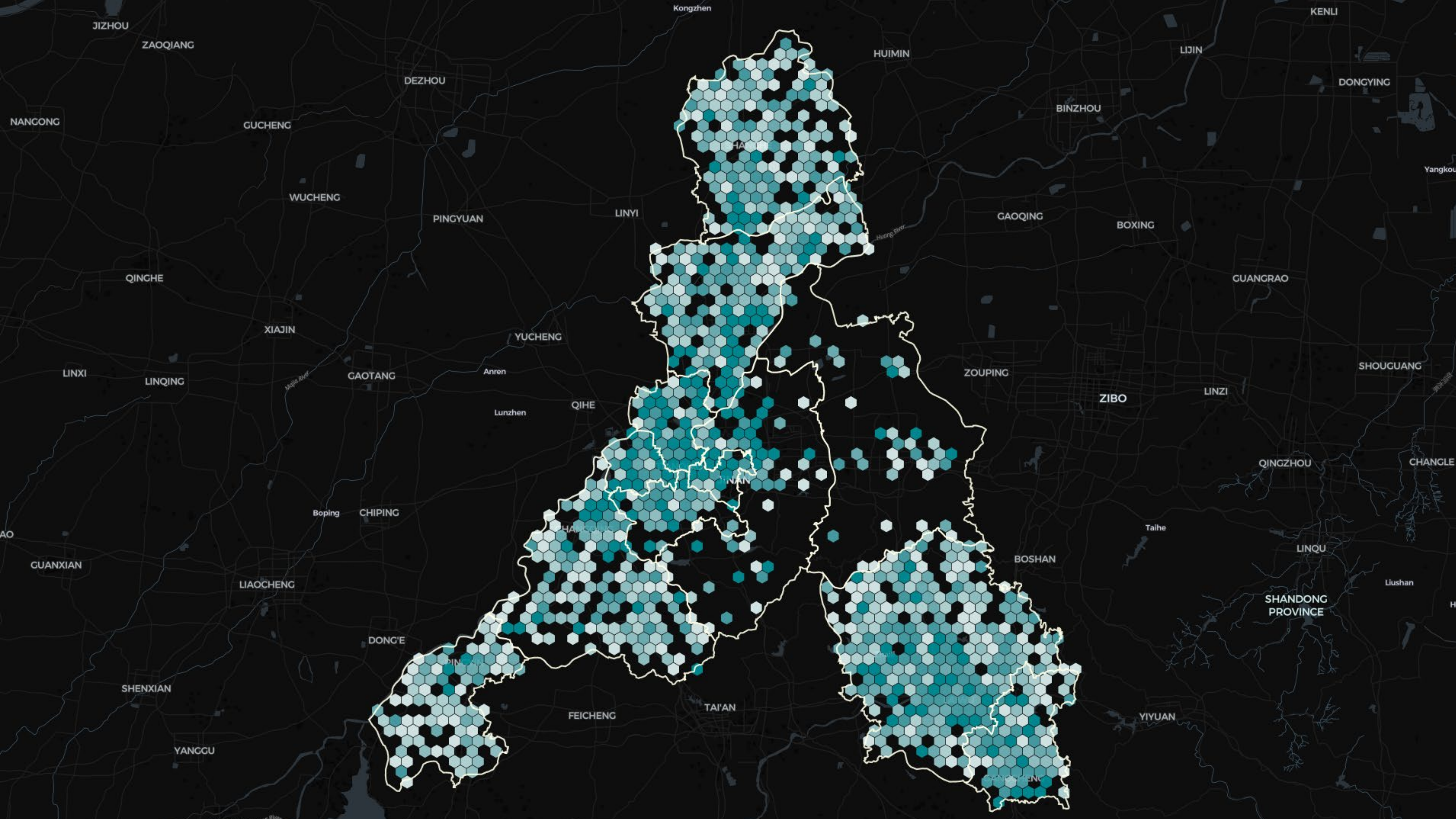
Trento Traffic Dynamics in One Day



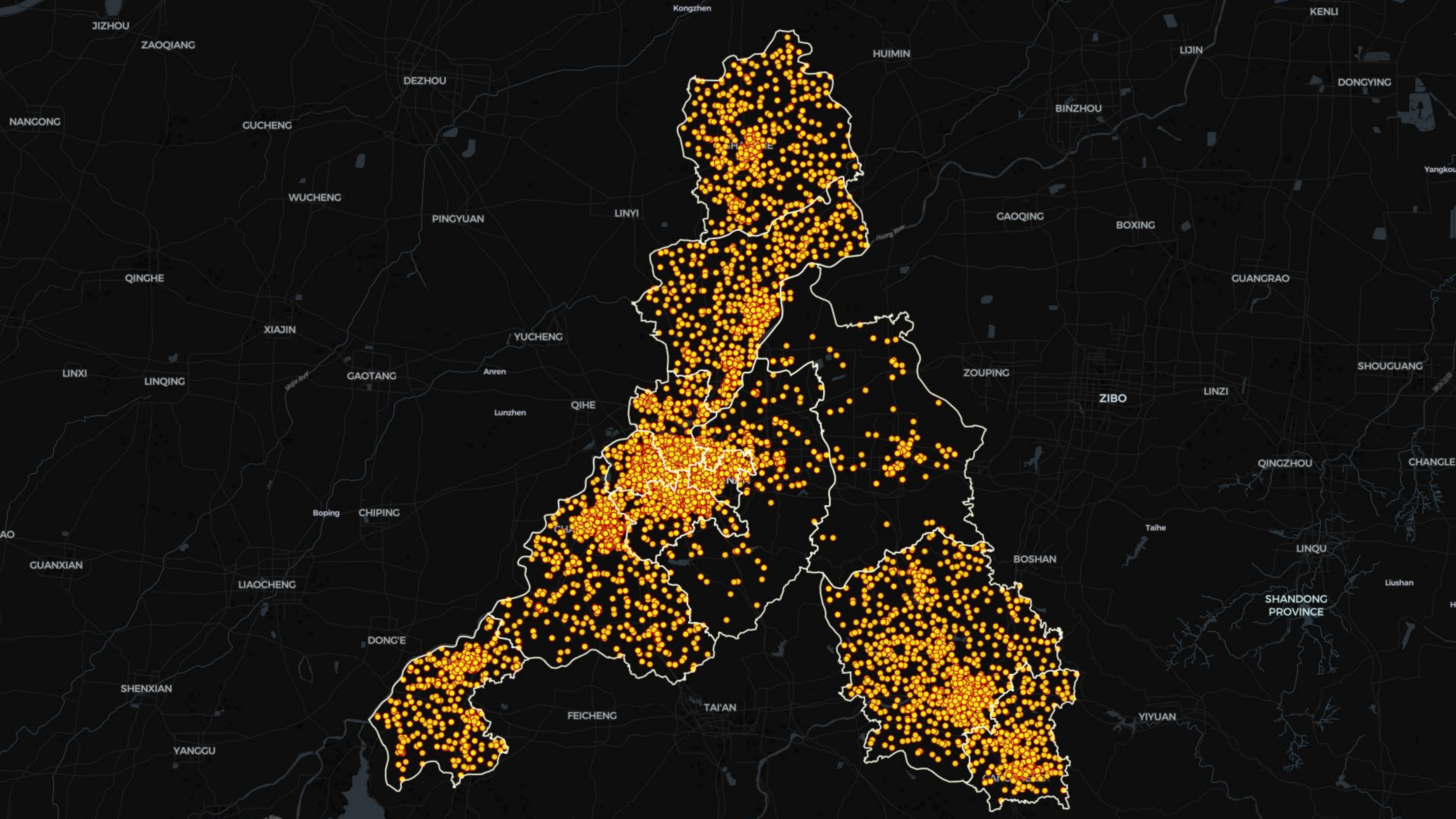
4G & 5G Wireless Data (PDCP SDU分组数据汇聚)



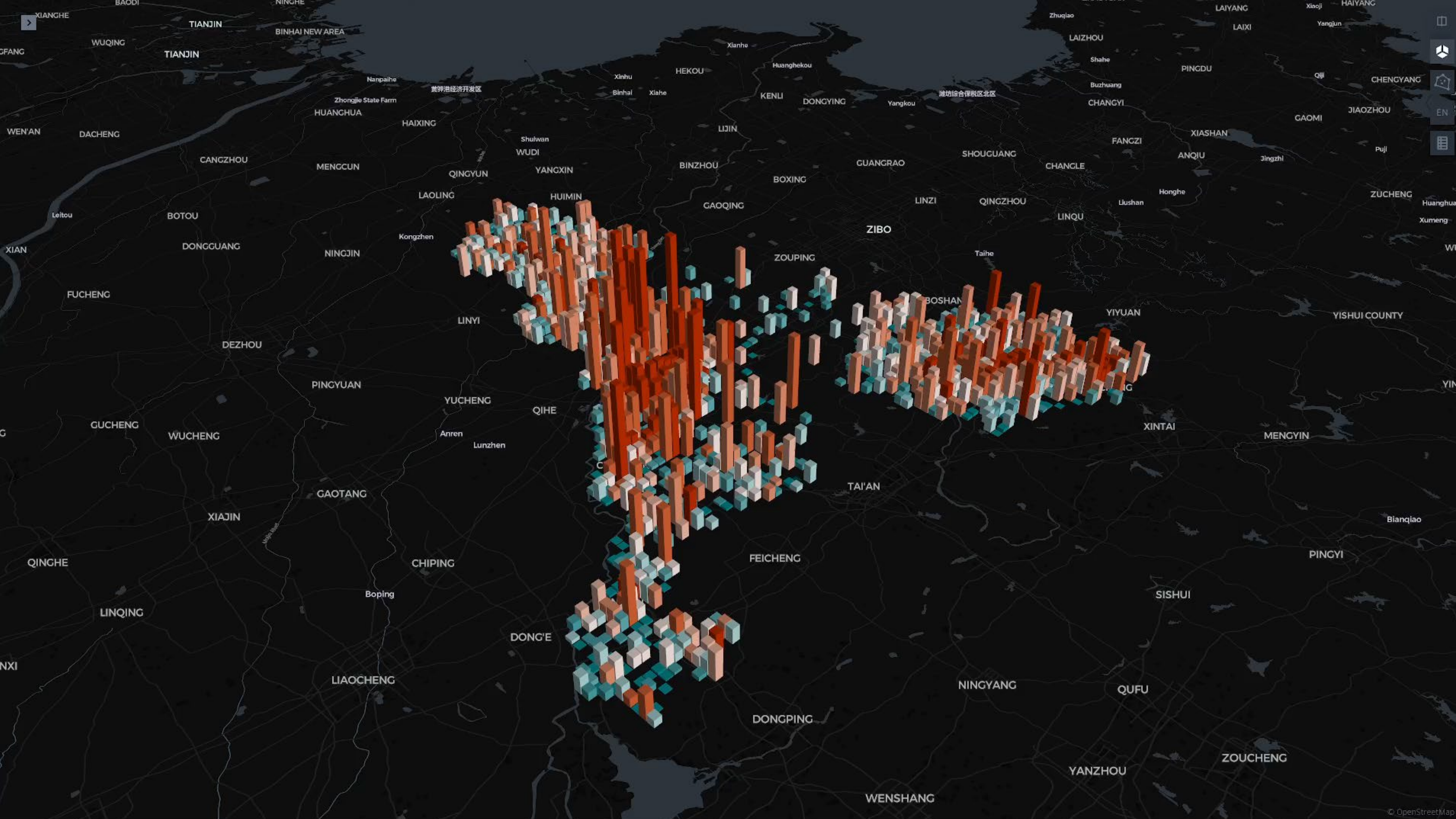




Kongzhen
JIZHOU
ZAOQIANG
KENLI
HUIMIN
LIJIN
DEZHOU
DONGYING
BINZHOU
NANGONG
GUCHENG
Yangkou
WUCHENG
PINGYUAN
LINYI
GAOQING
BOXING
Huang River
QINGHE
GUANGRAO
XIAJIN
YUCHENG
LINXI
LINQING
GAOTANG
Anren
ZOUPING
SHOUGUANG
ZIBO
LINZI
QIHE
Lunzhen
QINGZHOU
CHANGLE
Boping
CHIPING
Taihe
LINQU
GUANXIAN
BOSHAN
LIAOCHENG
Liushan
SHANDONG PROVINCE
DONG'E
SHENXIAN
TAI'AN
FEICHENG
YIYUAN
YANGGU



JIZHOU
ZAOQIANG
Kongzhen
KENLI
HUIMIN
LIJIN
DEZHOU
DONGYING
BINZHOU
NANGONG
GUCHENG
WUCHENG
PINGYUAN
LINYI
GAOQING
BOXING
Huang River
QINGHE
GUANGRAO
XIAJIN
YUCHENG
ZOUPING
SHOUGUANG
LINXI
LINQING
GAOTANG
Anren
ZIBO
LINZI
Majia River
QIHE
Lunzhen
QINGZHOU
CHANGLE
Boping
CHIPING
Taihe
BOSHAN
LINQU
GUANXIAN
LIAOCHENG
Liushan
SHANDONG
PROVINCE
DONG'E
SHENXIAN
TAI'AN
FEICHENG
YIYUAN
YANGGU



XIANGHE
BAODI
NINGHE
TIANJIN
BINHAI NEW AREA
WUQING
TIANJIN
Nanpaihe
Zhongjie State Farm
HUANGHUA
HAIXING
WEN'AN
DACHENG
CANGZHOU
MENGCUN
Shulwan
WUDI
QINGYUN
YANGXIN
Xinhu
Binhai
Xiahe
HEKOU
Xianhe
Huanghekou
KENLI
DONGYING
Yangkou
LIJIN
BINZHOU
BOXING
GUANGRAO
GAOQING
LAOLING
HUIMIN
Leitou
BOTOU
DONGGUANG
XIAN
Kongzhen
NINGJIN
ZIBO
ZOUPING
LINZI
QINGZHOU
LINQU
FUCHENG
LINYI
DEZHOU
PINGYUAN
YUCHENG
QIHE
Anren
Lunzhen
GUCHENG
WUCHENG
GAOTANG
XIAJIN
TAI'AN
FEICHENG
QINGHE
CHIPING
Boping
LINQING
DONG'E
LIAOCHENG
DONGPING
NINGYANG
QUFU
WENSHANG
YANZHOU
ZOUCHENG
SISHUI
PINGYI
MENGYIN
XINTAI
YIYUAN
BOSHAN
Taihe
YISHUI COUNTY
Bianqiao
Zhuqiao
LAIZHOU
Shahe
Buzhuang
CHANGYI
LAIYANG
LAIXI
Xiaoji
Yangjun
HAIYANG
PINGDU
Qiji
CHENGYANG
JIAOZHOU
GAOMI
XIASHAN
FANGZI
ANQIU
Jingzhi
SHOUGUANG
CHANGLE
Liushan
Honghe
Puji
ZUCHENG
Huanghua
Xumeng
EN
© OpenStreetMap

Experiment Settings

- 100 cells are selected for experiments
- Data are scaled to [0,1] using Min-Max normalization
- The first seven weeks data is used for training and the last one weeks data is for test
- Model is a simple three layer LSTM since we care only about FL, each layer has 64 hidden dims
- The lengths of closeness dependence and periodicity dependence are set to 3
- SDG optimizer with learning rate of 0.01 (decay), 100 rounds, 20 local batch size, 0.1 available cells in each round
- The weight of qualsi-global model is selected through a grid search

Experiment Results

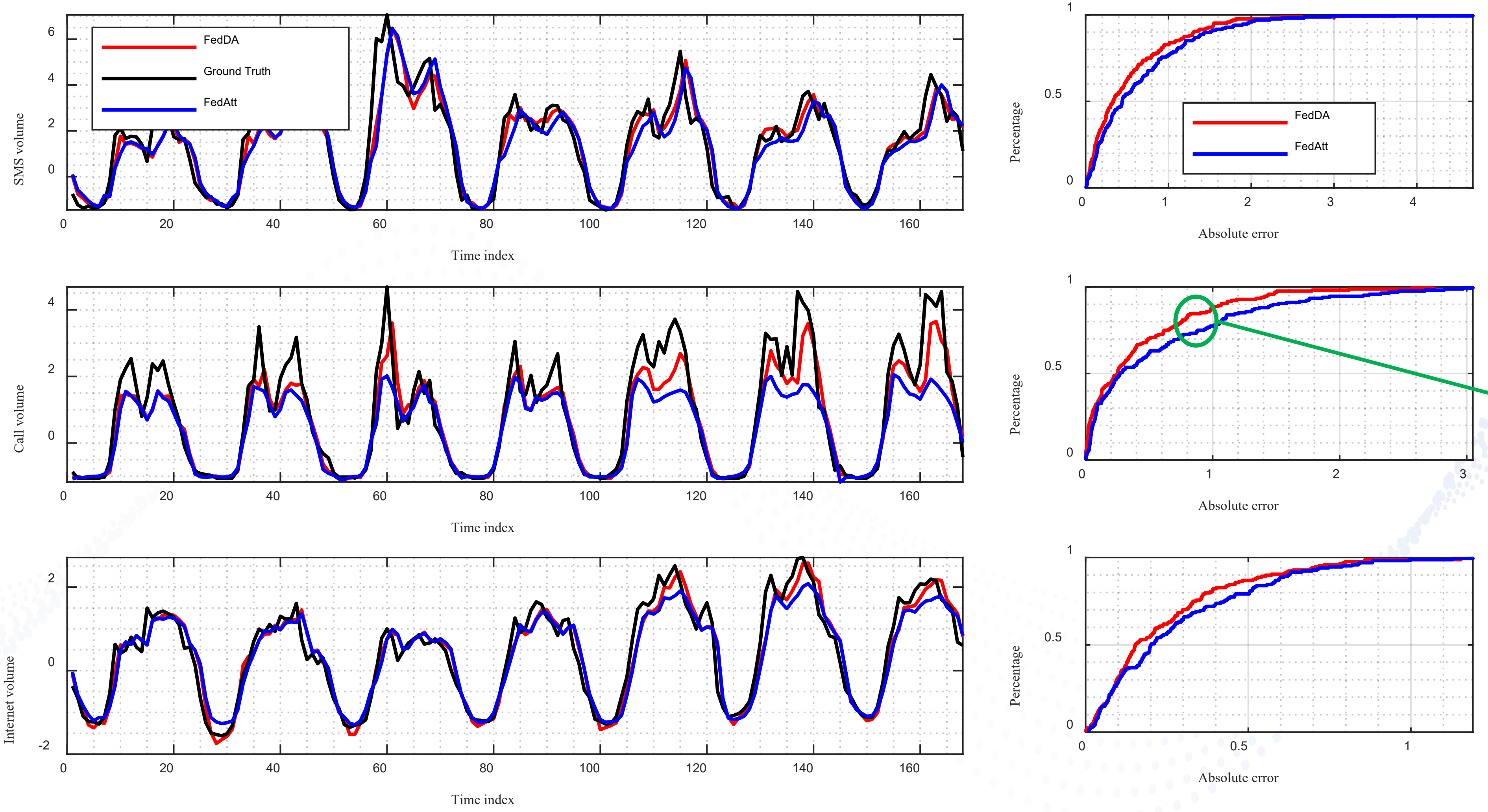
Methods	Milano						Trento					
	MSE			MAE			MSE			MAE		
	SMS	Call	Internet	SMS	Call	Internet	SMS	Call	Internet	SMS	Call	Internet
Lasso	0.7580	0.3003	0.4380	0.6231	0.4684	0.5475	4.7363	1.6277	5.9121	1.3182	0.8258	1.5391
SVR	0.4144	0.0919	*0.1036*	0.3528	0.1852	*0.2220*	5.2285	1.7919	5.9080	1.0390	0.5656	1.0470
LSTM	0.5608	0.1379	0.1697	0.4287	0.2458	0.2936	3.6947	1.1378	4.6976	0.9426	0.5013	1.1193
FedAvg	0.3744	0.0776	0.1096	0.3386	0.1838	0.2319	2.2287	1.6048	4.7988	0.7416	0.5319	1.0668
FedAtt	0.3667	0.0774	0.1096	0.3375	*0.1837*	0.2321	2.1558	1.5967	4.7645	0.7444	0.5306	1.0629
FedDA (φ=1)	0.3559	*0.0752*	0.1118	0.3353	0.1820	0.2367	2.1468	1.4925	4.4335	0.7478	0.5140	1.0212
FedDA (φ=10)	*0.3481*	0.0753	0.1062	*0.3321*	0.1810	0.2275	*2.0719*	*1.1699*	*3.9266*	*0.7320*	*0.4543*	*0.9504*
FedDA (φ=100)	**0.3322**	**0.0659**	**0.1033**	**0.3214**	**0.1741**	**0.2211**	**1.9703**	**1.0592**	**2.4473**	**0.6920**	**0.4281**	**0.7471**
↑ (φ=100)	+9.4%	+14.9%	+5.8%	+4.8%	+5.2%	+4.7%	+8.6%	+33.7%	+48.6%	+7.0%	+19.3%	+29.7%

Our method achieves the best prediction results

The more data shared, the better prediction performance

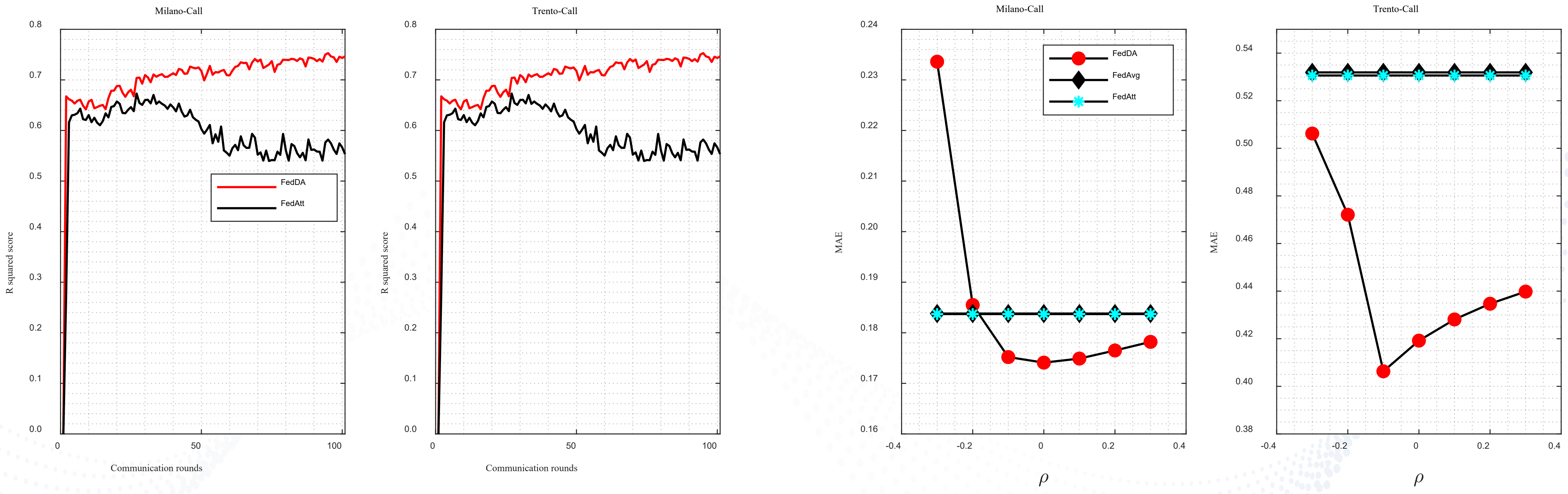
Prediction Versus Ground Truth

FedDA achieved much better performance than baseline, especially when traffic values are large



FedDA has a large portion of low errors

Accuracy Versus Communication Rounds



FedDA can achieve higher prediction accuracy with fewer communications between local client and central server

Quasi-global attention (model) can indeed improve prediction performance

Content

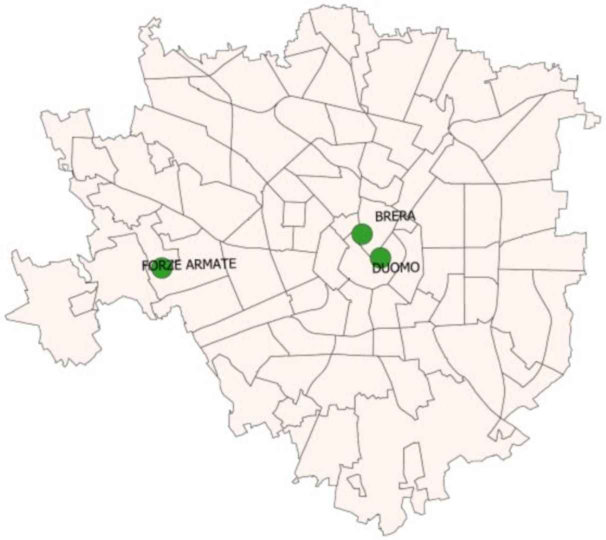
Revisit Wireless Traffic Prediction Under FL

- ❑Spatial modeling under FL relies:
  - BS/Cell/Cloud unit clustering using location information
  - Shared (augmented) data
- ❑Training a model needs frequently communications between local clients (BS/Cell/Cloud unit) and the central server
  - Consumes lots of bandwidth
  - Not works for LLMs

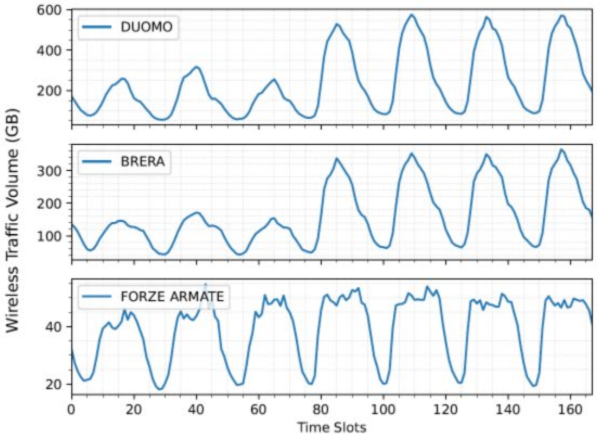
Training wireless traffic prediction model under the scenario of FL with the properties of spatial-temporal modeling and low communications

Evidence on the Deficient of FedAvg for WTP

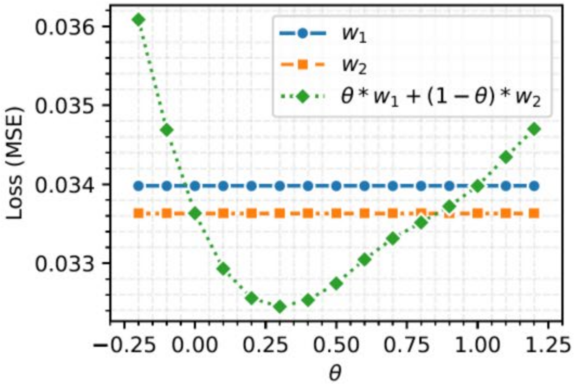
FedAvg works if the traffic of two areas are similar

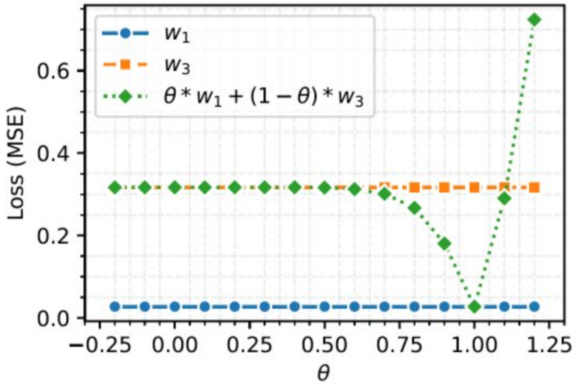


(a)



(b)





The effectiveness of FedAvg does not holds if the traffic of two areas are distinct

System Model

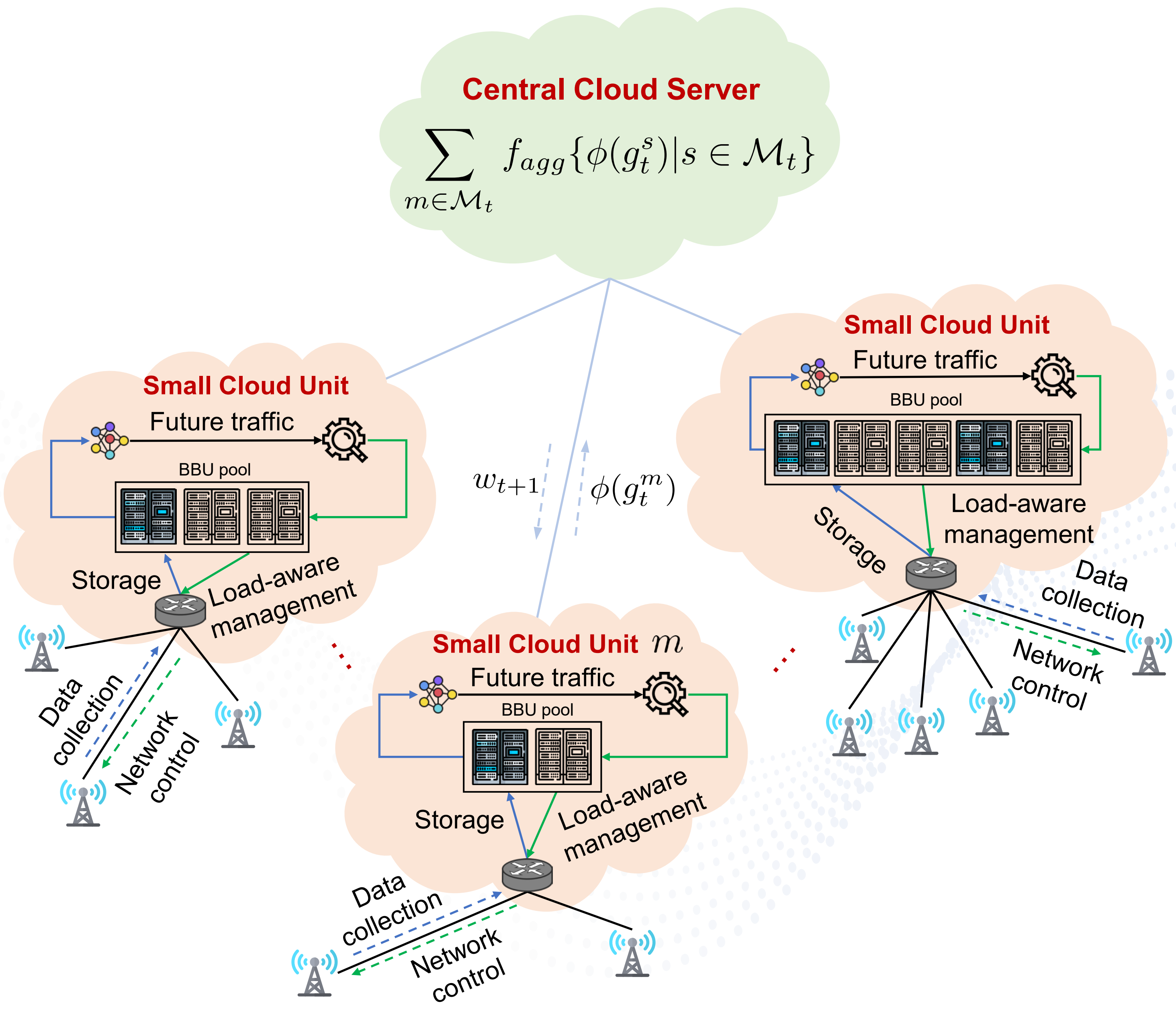
Low communications solution

Gradient Compression

Spatial dependence modeling solution

Gradient Correlation

Federated Learning with **G**radient **C**ompression and **C**orrelation for Wireless Traffic Prediction



FedGCC Algorithm

❑Global model optimization

Traditional global model optimization

$$w_{t+1} = w_t - \frac{\eta}{|\mathcal{M}_t|} \sum_{m \in \mathcal{M}_t} g_t^m$$

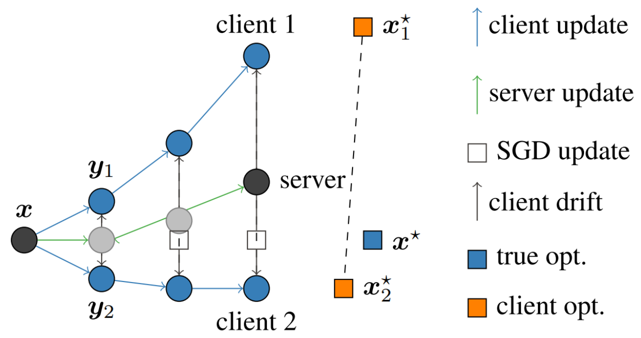
Our global model optimization

$$w_{t+1} = w_t - \frac{\eta}{|\mathcal{M}_t|} \sum_{m \in \mathcal{M}_t} f_{agg}\{\phi(g_t^s)|s \in \mathcal{M}_t\}$$

Heuristic personalized gradient aggregation strategy

Gradient compression scheme at the local client side

FedGCC Algorithm



Example of gradient tracking

❑Local model optimization with error feedback

$$g_{t,c}^m = e_t^m + \nabla f_m(w_{t,c}^m; \mathcal{B}_{t,c}^m) - h_t^m$$

round

step

client

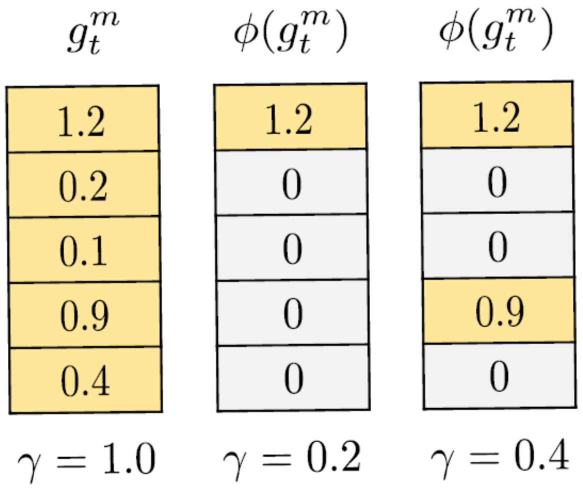
Error feedback: accum. of gradient errors

Gradient of current step on batch data B with parameter w

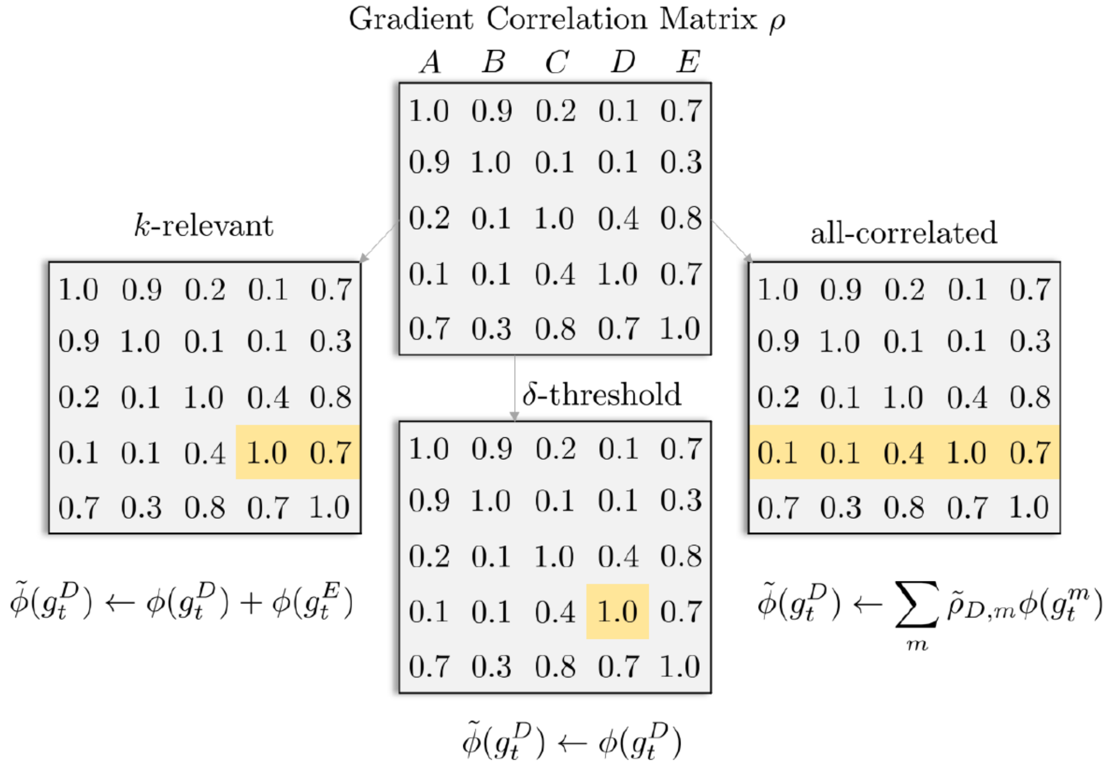
Gradient tracking: correct the gradient direction

Local gradient is corrected by using a) ***non-transferred gradients*** *in previous rounds; b)* ***current 'true' gradients*** *on local batch data; c) the* ***gradient difference*** *between local client and central server.*

Example of Gradient Compression & Correlation



Gradient compression (sparsification) with different ratios

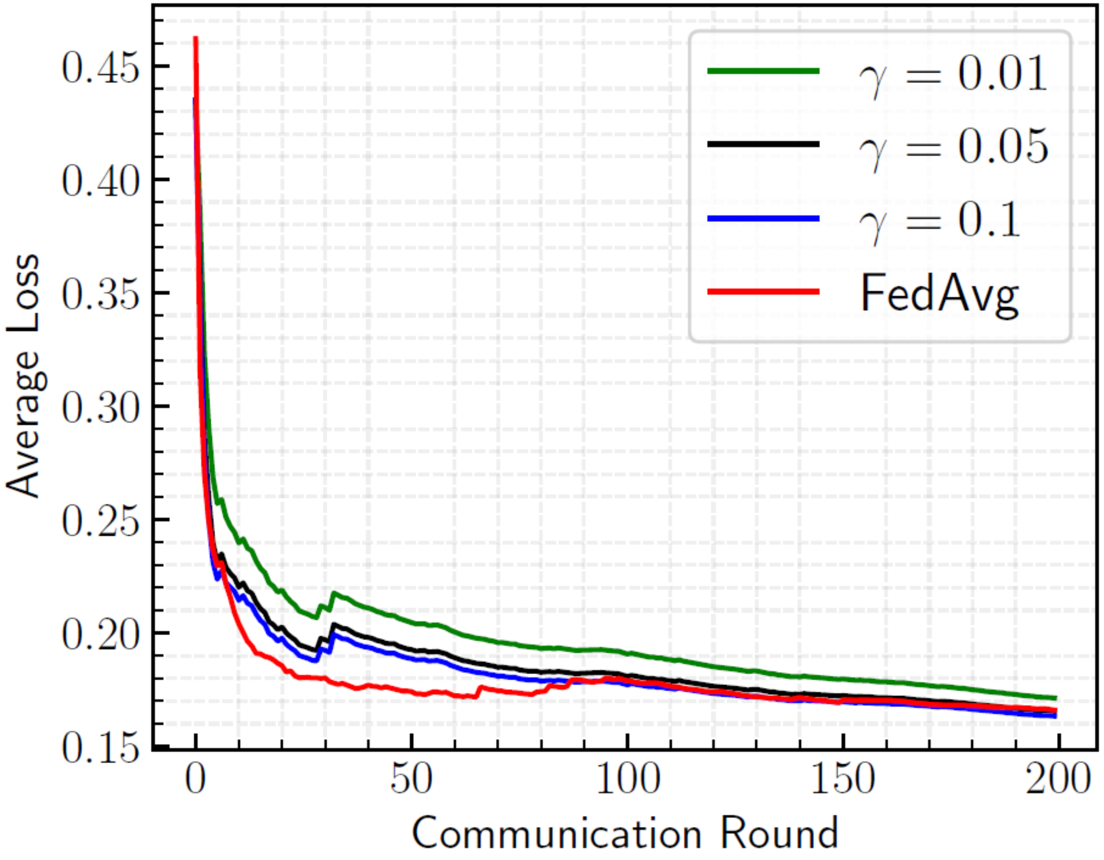


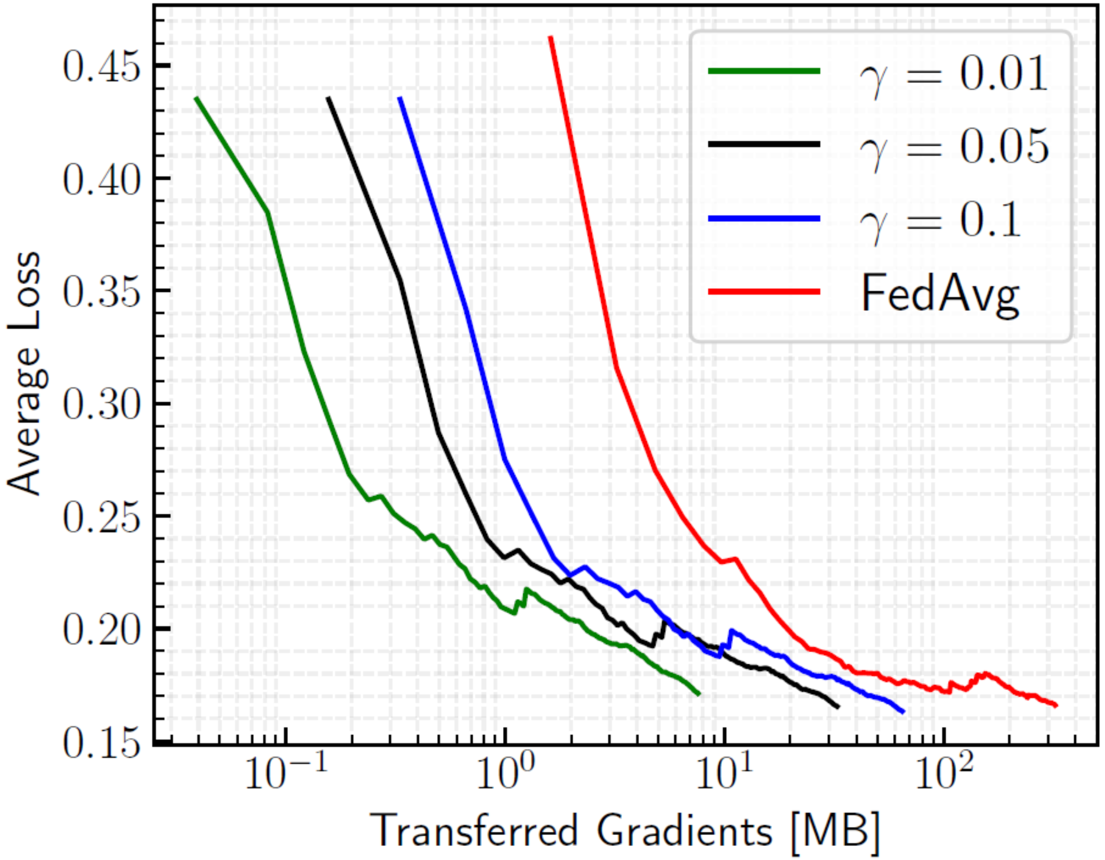
Gradient correlation with different strategies

Experiment Results

Method	Notes	Milan			Trentino			ΔC (MB)
		RMSE	MAE	R^2 Score	RMSE	MAE	R^2 Score	
FedAvg [37]	-	0.1401	0.0965	0.9371	1.0504	0.5834	0.7871	126.27/322.68
FedProx [24]	$\mu = 0.01$	0.1398	0.0960	0.9373	1.0057	0.5581	0.8129	126.27/322.68
	$\mu = 0.1$	0.1400	0.0963	0.9373	1.0402	0.5774	0.7936	
	$\mu = 1$	0.1339	0.0869	0.9446	1.1615	0.6475	0.7114	
FedAtt [38]	-	0.1357	0.0898	0.9431	0.9194	0.5269	0.8637	126.27/322.68
FedDA [36]	$\varphi = 1$	0.1339	0.0816	0.9466	0.7391	0.3933	0.9264	126.30/322.74
	$\varphi = 10$	0.1308	0.0795	0.9493	0.7823	0.4188	0.9143	126.55/323.39
	$\varphi = 100$	0.1301	0.0790	0.9493	0.7711	0.3918	0.9217	129.06/329.81
FedCOMGATE [22]	-	0.1438	0.1027	0.9317	0.7427	0.3849	0.9273	14.373/38.449
Proposed	k-relevant	**0.1299**	**0.0788**	**0.9501**	**0.6935**	**0.3621**	0.9431	**3.1494/7.5843**
	δ-threshold	0.1301	0.0797	0.9486	0.6943	0.3638	0.9423	
	all-correlated	0.1300	0.0795	0.9483	0.7048	0.3805	**0.9499**	

Experiment Results





Content

1 Background

2 FedDA

3 FedGCC

4 Conclusion

Conclusion

- ❑Wireless traffic prediction supports AI native of 6G
- ❑FedDA: Dual attention based wireless traffic prediction
  - Clustering for spatial dependence modeling
  - Augmented data sharing for reducing heterogeneity
  - Dual attention based federated optimization
- ❑FedGCC: Gradient compression and correlation for wireless traffic prediction
  - Gradient compression for reducing communication between local clients and the central server
  - Gradient correlation for spatial dependence modeling

★ 网络天下 ★

数据通信路由技术与验证算法技术论坛

Thanks for your time !

chuanting.zhang@sdu.edu.cn

https://chuanting.github.io