Machine Learning at the Edge

A Data-Driven Architecture with Applications to 5G Cellular Networks

Michele Polese[*], Rittwik Jana[+], Velin Kounev[+],

Ke Zhang[+], Supratim Deb[+], *Michele Zorzi*[*]

[*]University of Padova, Padova, Italy

[+]AT&T Labs, Bedminster, NJ

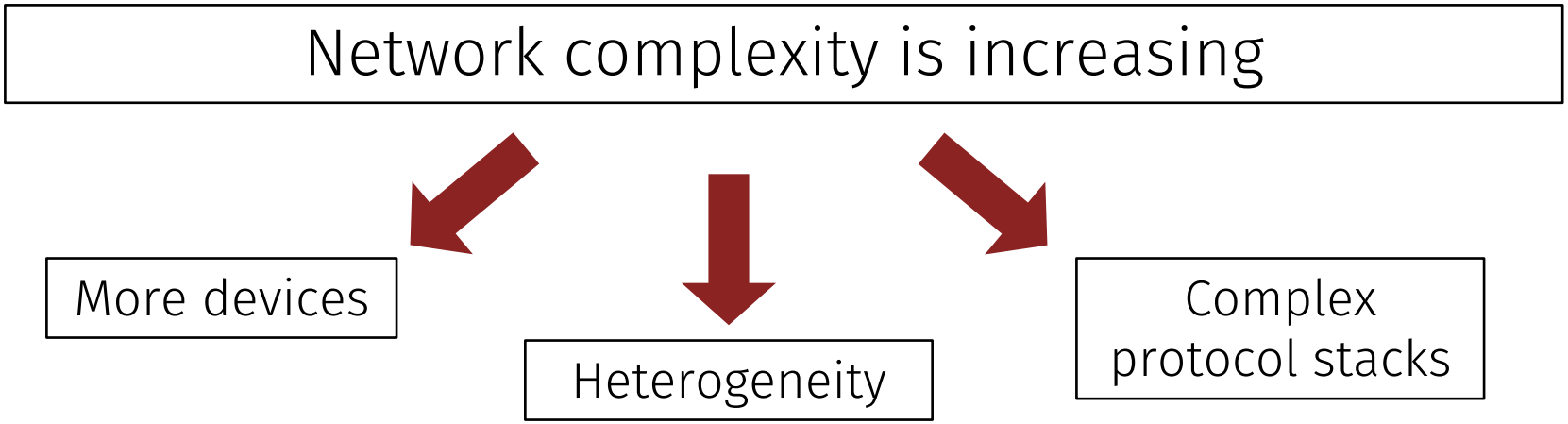
zorzi@dei.unipd.it

ICNC 2019

Outline

- Machine learning in networks: motivation
- Contribution
- The dataset
- 5G data-driven architecture
- Applications
  - Clustering in self-organizing networks
  - Prediction of the number of users in base stations
- Conclusions

Data-driven networks: potentials



- Classic optimization techniques may be infeasible
- Need for autonomous orchestration and configuration
- QoE can improve with context-awareness

Use network data to drive self-optimizing ML algorithms

Data-driven networks: challenges

- Scalability of ML techniques
- Availability of data
- Several open questions to be addressed
  - Which information is needed from the network?
  - How is it possible to efficiently collect this information?
  - How to practically deploy ML/AI algorithms?
  - Which ML techniques perform better?
  - How good is the performance of ML in real networks?

Data-driven networks: our contribution

- Mobile-edge controller-based architecture
  1. Deployable in **5G NR** networks
  2. Capable of handling **data** collection and providing *real-time* analytics and decisions
  3. Better performing than legacy architectures
- Evaluation:
  a. Data-driven dynamic clustering of base stations
  b. Prediction accuracy of the number of UEs per base station

Dataset with hundreds of base stations from major US operator

M. Polese, R. Jana, V. Kounev, K. Zhang, S. Deb, M. Zorzi, *"Machine Learning at the Edge: a Data-Driven Architecture with Applications to 5G Cellular Networks"*, submitted to IEEE JSAC Special Issue on AI and ML for networks

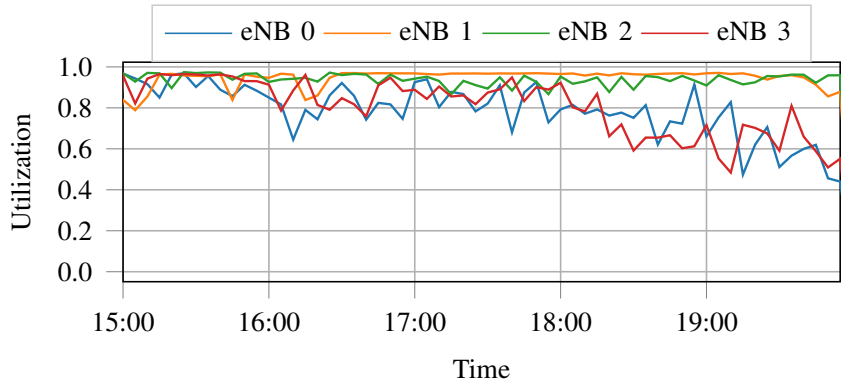
State of the art

- [1] and [2] discuss how big data analytics can be used in networks
- [3] and [4] use network traces to infer human mobility patterns
- [5] models single-user mobility, while we focus on base-station-level behavior

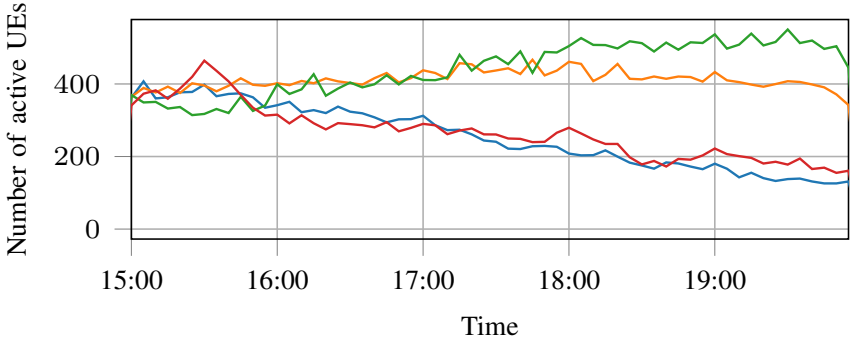
[1] Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao, and R. C. Qiu, “Big data analytics in mobile cellular networks,” IEEE Access, vol. 4, pp. 1985– 1996, March 2016.
[2] A. Imran, A. Zoha, and A. Abu-Dayya, “Challenges in 5G: how to empower SON with big data for enabling 5G,” IEEE Network, vol. 28, no. 6, pp. 27–33, Nov 2014.
[3] R. Becker, R. Caceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky, “Human mobility characterization from cellular network data,” Communications of the ACM, vol. 56, no. 1, pp. 74–82, Jan 2013.
[4] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, “A tale of one city: Using cellular network data for urban planning,” IEEE Pervasive Computing, vol. 10, no. 4, pp. 18–26, April 2011.
[5] W. Dong, N. Duffield, Z. Ge, S. Lee, and J. Pang, “Modeling cellular user mobility using a leap graph,” in International Conference on Passive and Active Network Measurement. Springer, 2013, pp. 53–62.

The dataset

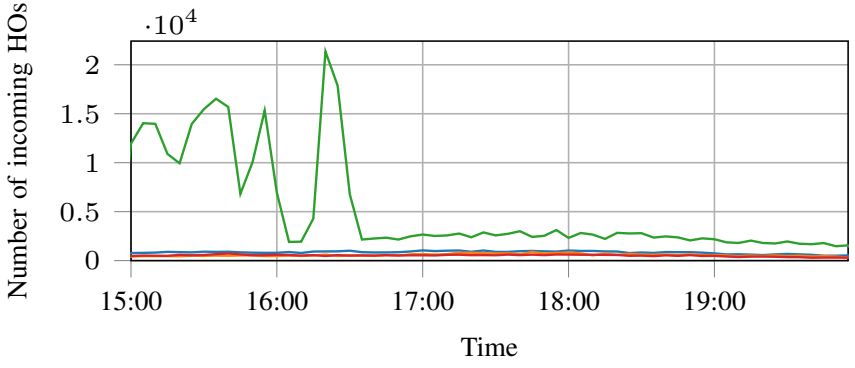
- 472 eNBs in San Francisco
  - February 2017
  - Every day, 3 P.M. to 8 P.M.
- 178 eNBs in Palo Alto
  - June-July 2018
  - Whole day
- 4G LTE deployment
- Data collected:
  - Resource utilization
  - Number of incoming and outgoing handovers
  - Number of active UEs



(a) Utilization (averaged over a 15-minute interval).



(b) Number of active UEs (summed over a 15-minute interval).



(c) Number of incoming handovers (summed over a 15-minute interval).

Data-driven 5G architecture

4G systems
- no/limited coordination
- eNBs are self-contained equipment

Proposal

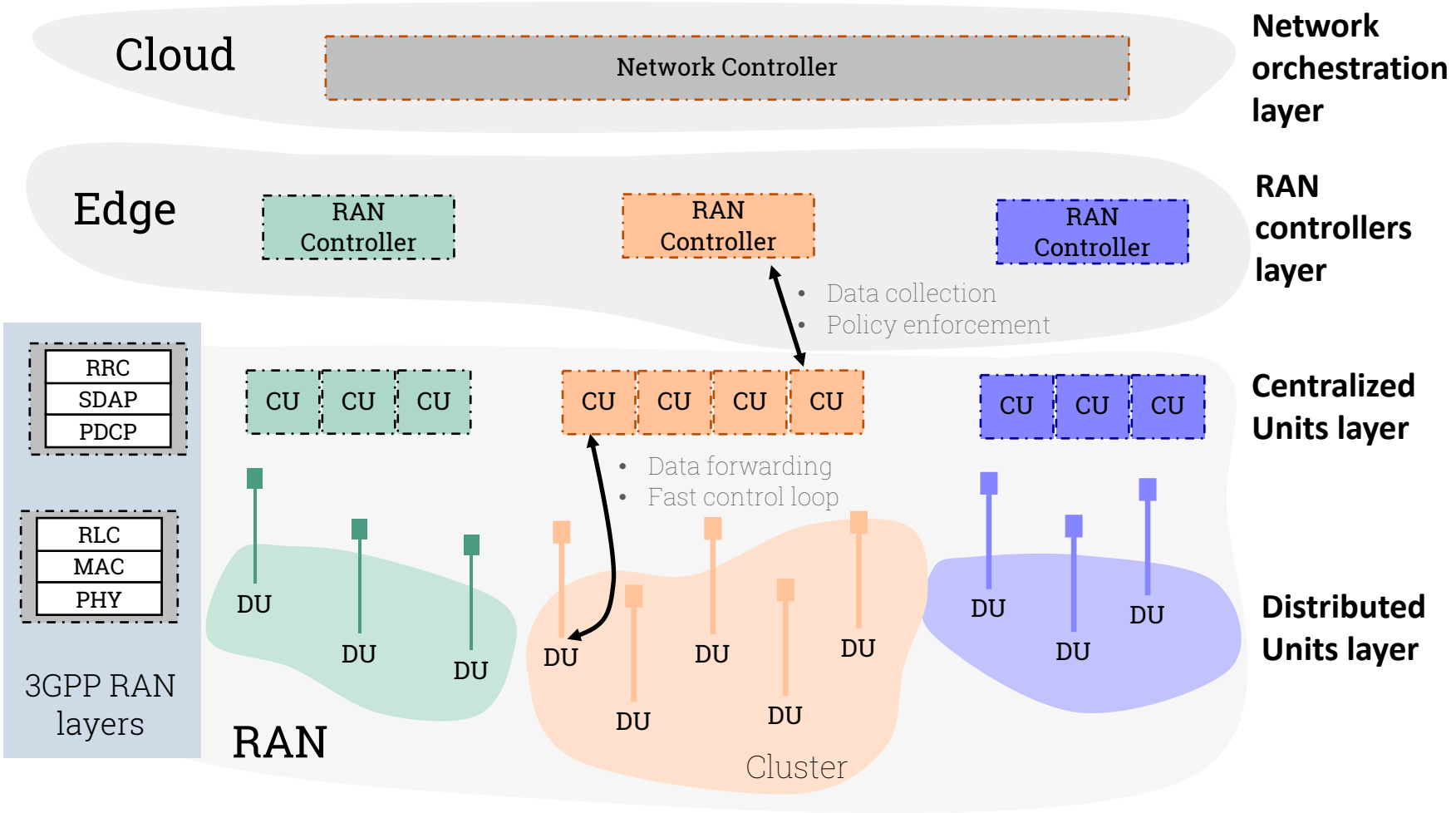
5G systems
- coordination control through xRAN/O-RAN
- CU/DU split

- Learning based on local information/history
- Single-eNB applications

- Learning based on shared information/history
- **Coordinated learning**

Exploit the spatial correlation naturally introduced by user mobility

Multi-layer architecture



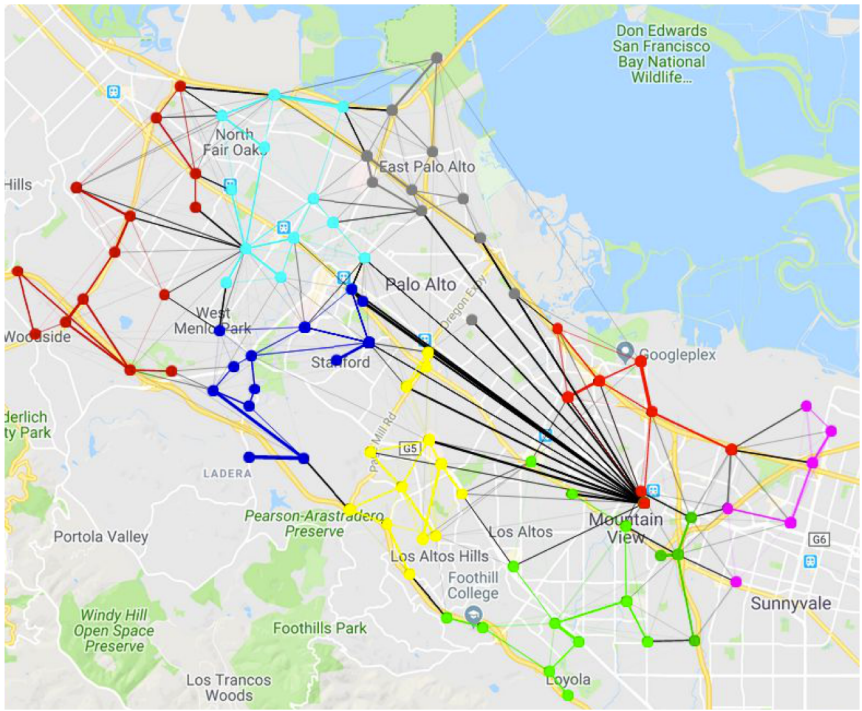
Multi-layer architecture: key components

- 3GPP NR **CUs/DUs** for data plane & **local** control decisions
- Mobile Edge Computing facilities
- RAN controllers (xRAN, Open RAN)
  - Deployed at the edge
  - Orchestrate CUs/Dus
  - **Clustered** view on the network
  - Collect data from CUs/DUs to control the network –> use it also to **run ML algorithms**
- Network controller (ONAP)
  - **Centralized** cloud facility
  - RAN controllers orchestration and app-layer services

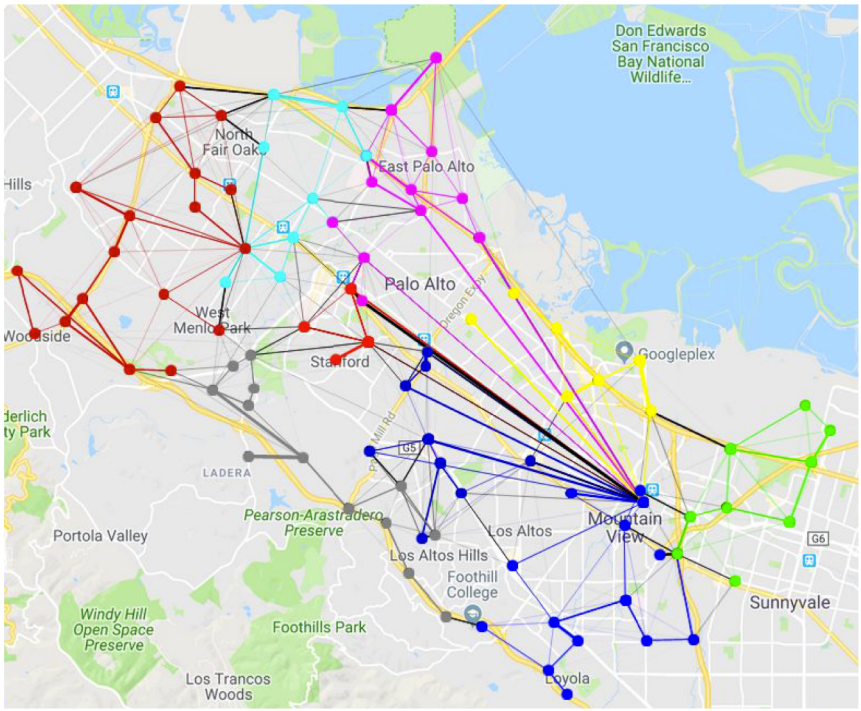
Data-driven operations: RAN clustering

Goal: minimize inter-controller interactions
(impact on control plane latency)

Clustering based on
base station positions
(fixed, no dynamic data)



Clustering based on
handover transitions
(dynamic, based on network data)



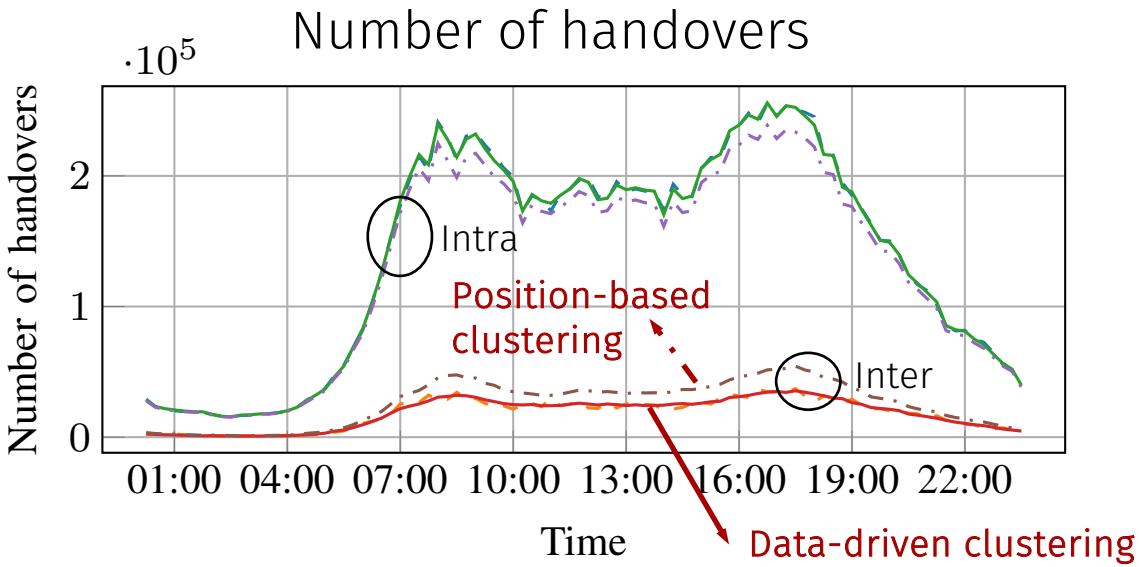
Data-driven operations: RAN clustering

Goal: minimize inter-controller interactions
(impact on control plane latency)

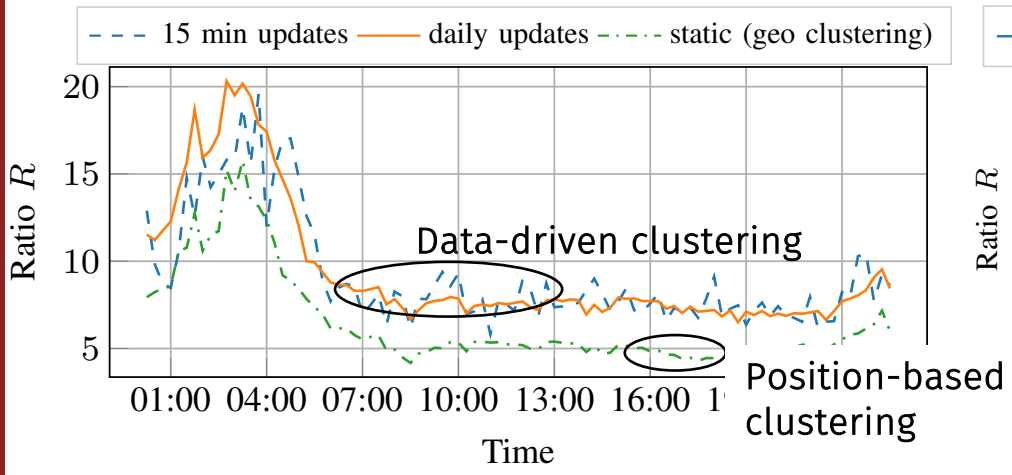
Algorithm 1 Network-data-driven Controller Association Algorithm

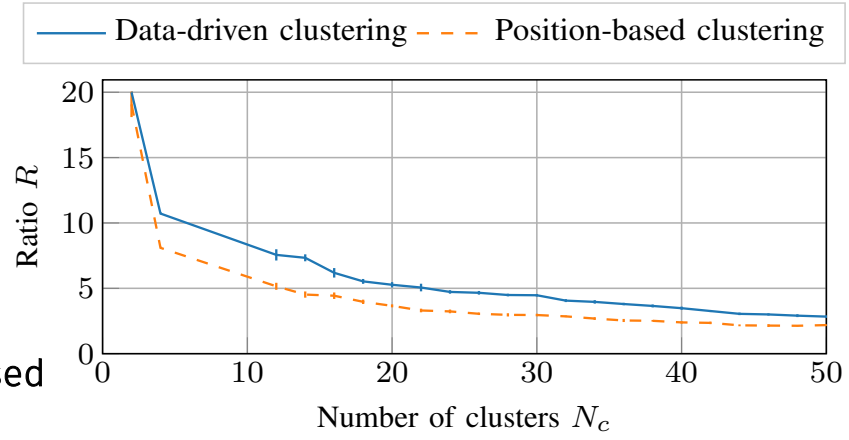
for every time step T_c
  distributed data collection step:
    for every controller $p \in \{0, \ldots, N_c - 1\}$ with associated gNBs set $\mathcal{B}_p$
      for every gNB $i \in \mathcal{B}_p$
        compute the number of handovers $N_{i,j}^{\text{ho}} \forall j \in \mathcal{B}$
      end for
      report the statistics on the number of handovers to the network controller
    end for
  clustering and association step:
    compute the transition probability matrix H based on the handovers between every pair of gNBs
    define weighted graph $G = (V, E)$ with weight $W(G)_{i,j} = H_{i,j} + H_{j,i}$
    perform spectral clustering with constrained K means on G to identify N_c clusters
    apply the new association policy for the next time step
end for

Data-driven operations: RAN clustering



Ratio intra/inter-cluster HOs





Data-driven operations: prediction

Predict the number of active UEs

- Local-based method: train a different model in each BS to predict the number of UEs in each single BS
  - This is what is possible in 4G LTE networks
- Cluster-based method: train a model per cluster, predict a vector with the number of UEs in each BS of the cluster
  - Enabled by our architecture
  - Exploit spatial correlation to improve the prediction

Data preprocessing

- The number of active users is averaged every 5 minutes
- Scaling and log(1+x) applied to the dataset

Local-based prediction

- Target: number of active users in each eNB, with a look-ahead step $L \in \{1, 2 \ldots, 9\}$ 5-minutes steps
- Features:
  - Boolean flag – weekend or weekday
  - Hour of the day
  - Past W samples of the number of active users

Data preprocessing

- The number of active users is averaged every 5 minutes
- Scaling and log(1+x) applied to the dataset

Cluster-based prediction

- Target: vector with the number of active users in each eNB of the cluster, with a look-ahead step $L \in \{1, 2 \ldots, 9\}$ 5-minutes steps
- Features:
  - Boolean flag – weekend or weekday
  - Hour of the day
  - Vector with past W samples of the number of active users in each eNB of the cluster

Data preprocessing

- Sample cluster in San Francisco
- 22 base stations

More details on the clustering process can be found in M. Polese, R. Jana, V. Kounev, K. Zhang, S. Deb, and M. Zorzi, "**Machine learning at the edge: A data-driven architecture with applications to 5G cellular networks**," submitted to IEEE Journal on Selected Areas of Communications, 2018. [Online]. Available: https://arxiv.org/abs/1808.07647

Algorithms

- Bayesian Ridge Regressor (BRR)
  - Local-based only
- Random Forest Regressor (RFR)
  - Local- and cluster-based
- Gaussian Process Regressor (GPR)
  - Local- and cluster-based
  - Combined kernel with
    - Dot product kernel (non stationary behavior)
    - Rational quadratic kernel (mixture of stationary behaviors)
    - White kernel (noisy input)

Hyperparameters	
Bayesian Ridge Regressor [18], [19]	
α	$\{10^{-6}, 10^{-3}, 1, 10, 100\}$
λ	$\{10^{-6}, 10^{-3}, 1, 10, 100\}$
Random Forest Regressor [20], [21]	
Number of trees N_{rf}	$\{1000, 5000, 10000\}$
Gaussian Process Regressor [22]	
α	$\{10^{-6}, 10^{-4}, 10^{-2}, 0.1\}$
σ_k	$\{0.001, 0.01\}$

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

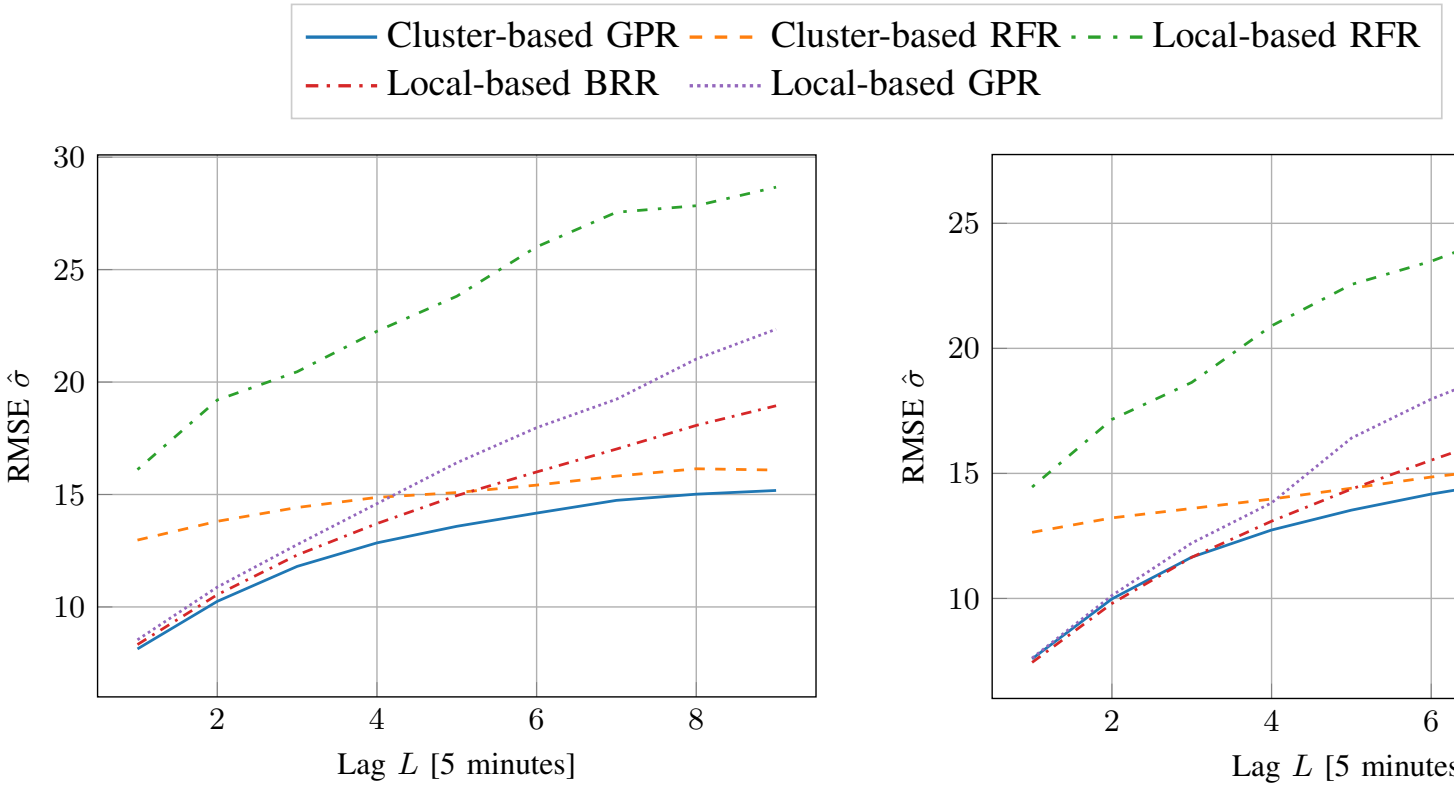
Training and testing

- 3-fold cross validation to select hyperparameters (with time-consistent split)
- RMSE considered for the prediction error:

$$\sigma_i = \sqrt{\frac{1}{N_{te}} \sum_{t=1}^{N_{te}} (y_i(t) - \hat{y}_i(t))^2}$$

- The RMSE is averaged over the base stations of each cluster
- Training dataset from 01/31 to 02/20
- Testing dataset from 02/21 to 02/26

Performance evaluation



Fixed $W = 1$ 5-minute step

Best W for each L

Look-ahead step L	1	2	3	4	5	6	7	8	9
BRR	6	6	4	4	3	3	3	2	2
cluster-GPR	3	2	2	2	2	1	6	5	4

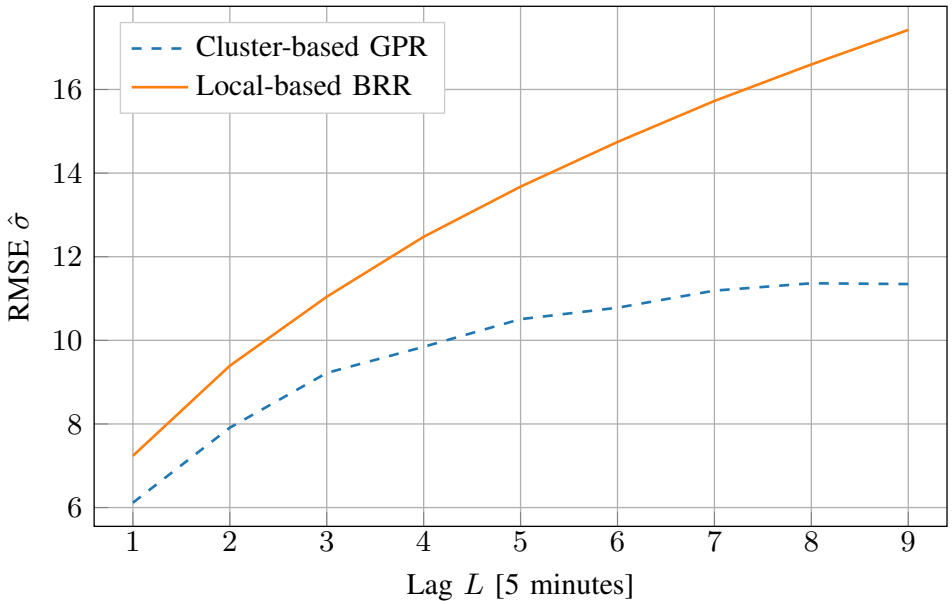
Performance evaluation

- Spatial correlation (cluster- vs local-based) is more impactful than temporal correlation

53% RMSE reduction

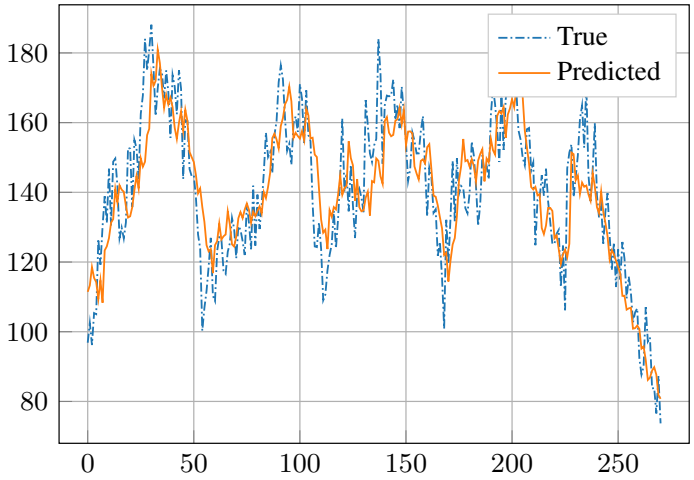
5% RMSE reduction when increasing W

- Exploit geographic constraints on mobility flows
- When considering all the 472 eNBs (in 22 clusters):

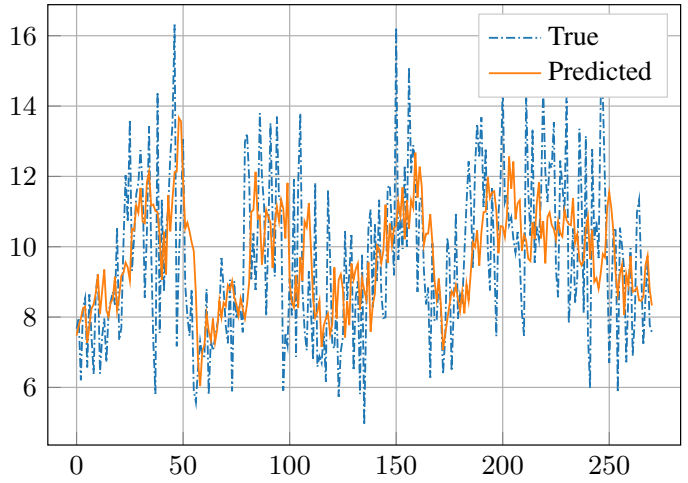


Example of predicted timeseries

- Cluster-based GPR
- High number of users



- Low number of users

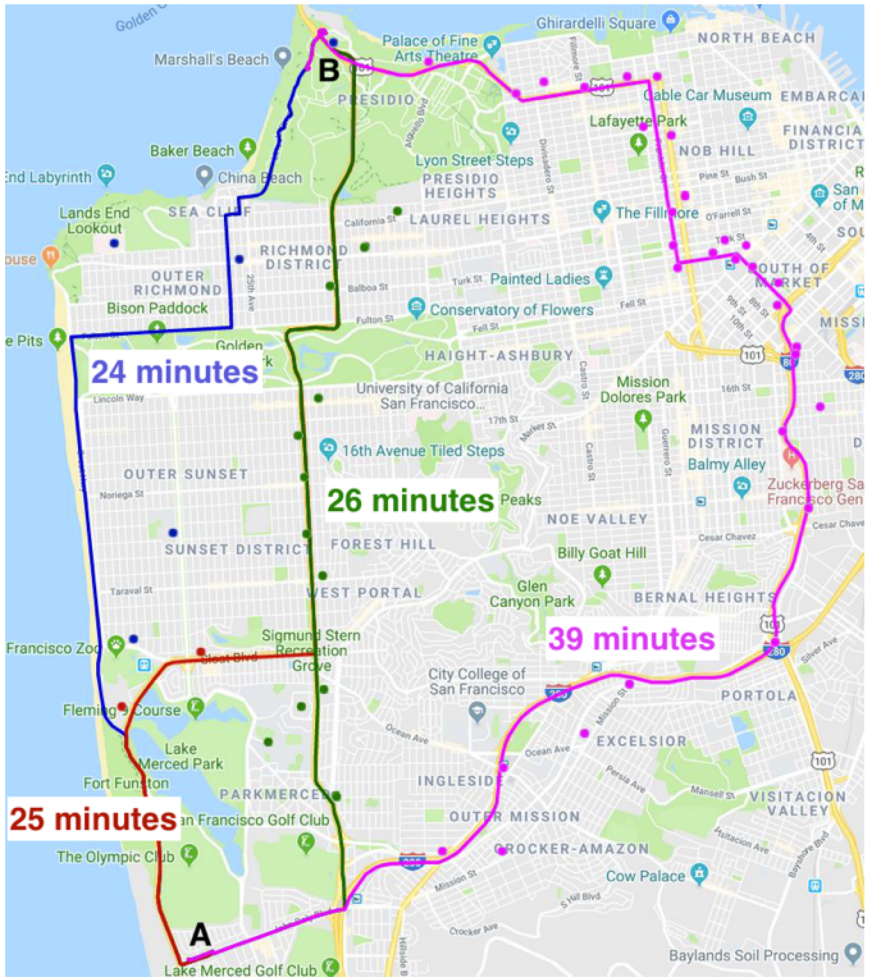


Good tracking of daily patterns
Very noisy traces

Use cases for the prediction

- Medium-timescale horizon (5 – 45 minutes)
- Network management and operations:
  - Predictive load-balancing
  - Bearer pre-configuration for anticipatory mobility
  - Radio resource scaling
- New services to the end-users
  - Vehicular route optimization with network KPIs
    (provide transit directions tailored on the network performance)

Vehicular route optimization with network KPIs



Predicted throughput
(as a function of number of active users)

Route	R1	R2	R3	R4
	Feb. 23rd, 19:00			
$\hat{S}$ [Mbit/s]	1.93	2.51	2.36	2.74
$D_{o,\max}$ [s]	133.47	157.8	172.5	171.2
	Feb. 24th, 19:00			
$\hat{S}$ [Mbit/s]	1.72	2.00	2.28	2.89
$D_{o,\max}$ [s]	152.4	157	148.8	169.1
	Feb. 24th, 19:20			
$\hat{S}$ [Mbit/s]	2.05	2.49	1.98	2.86
$D_{o,\max}$ [s]	152.1	123.7	172.5	116.7

Predicted outage duration
(as a function of number of active users)

Conclusions

- Proposed a data-driven architecture for 5G networks
- Evaluation of learning approaches on a large-scale dataset from a network operator
  - RAN clustering
  - Prediction
- Exploiting spatial correlation is beneficial for medium-term prediction
- Reduction of the prediction error up to 53%
- Enabler of new use cases – both for RAN control and innovative user services

More details can be found in M. Polese, R. Jana, V. Kounev, K. Zhang, S. Deb, and M. Zorzi, "**Machine learning at the edge: A data-driven architecture with applications to 5G cellular networks**," submitted to IEEE Journal on Selected Areas of Communications, 2018. [Online]. Available: https://arxiv.org/abs/1808.07647

signet.dei.unipd.it
mmwave.dei.unipd.it

Machine Learning at the Edge

A Data-Driven Architecture with Applications to 5G Cellular Networks

Michele Polese[*], Rittwik Jana[+], Velin Kounev[+],

Ke Zhang[+], Supratim Deb[+], *Michele Zorzi*[*]

[*]University of Padova, Padova, Italy

[+]AT&T Labs, Bedminster, NJ

zorzi@dei.unipd.it

ICNC 2019