2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), 25-28 April 2021, Helsinki, Finland
DOI:10.1109/VTC2021-Spring51267.2021.9449040

Forecasting Wireless Demand with Extreme Values using Feature Embedding in Gaussian Processes

Schyler C. Sun[1] and Weisi Guo[1,2*]

Abstract—**Wireless traffic prediction is a fundamental enabler to proactive network optimisation in 5G and beyond. Forecasting extreme demand spikes and troughs is essential to avoiding outages and improving energy efficiency. However, current forecasting methods predominantly focus on overall forecast performance and/or do not offer probabilistic uncertainty quantification. Here, we design a feature embedding (FE) kernel for a Gaussian Process (GP) model to forecast traffic demand. The FE kernel enables us to trade-off overall forecast accuracy against peak-trough accuracy. Using real 4G base station data, we compare its performance against both conventional GPs, ARIMA models, as well as demonstrate the uncertainty quantification output. The advantage over neural network (e.g. CNN, LSTM) models is that the probabilistic forecast uncertainty can directly feed into decision processes in optimisation modules.**

Index Terms—**wireless traffic, forecasting, Gaussian process, machine learning**

I. Introduction

WIRELESS traffic prediction is a key enabler for proactive resource optimisation in 5G and beyond [1], [2], [3]. Proactive optimisation can create user-centric quality-of-service (QoS) and -experience (QoE) improvements across 5G network slices [4], [5]. Direct prediction from historical data [6], [7], [8] and inference from proxy social media data [9] are important inputs to proactive optimisation modules [4] being considered for 5G and beyond applications [10], such as interference management, load balancing, localization [11], and multi-RAT offloading; with implementation on the edge or in CRANs. This is more pertinent for hyper-dense mass autonomy applications where spikes in demand across different quality of service/experience/trust demands require forward prediction [12].

We begin with a review of time-series forecasting algorithms used in wireless traffic prediction and identify a lack of research in both high and low extreme value predictions, which is of critical importance to avoiding network congestion and inefficiencies.

A. State-of-the-Art

Time-series prediction methods can be classified into several types, with training data in high demand.

1) Statistical Models: Statistical time-series modelling using a variety of signal processing and machine learning approaches have been widely applied to predict the wireless traffic. Moving average models with smoothing weights and seasonality works well for univariate forecasting. For example, in [13], seasonal auto-regressive integrated moving average (ARIMA) models were fitted to wireless cellular traffic with two periodicities for prediction. However, this model is insensitive to anomalous values, such as event-driven spike demand. Indeed, predicting and avoiding spike demand is critical to avoiding network outages and improving the consumer experience. Other methods rely on statistical generative functions assuming a quasi-static behaviour, such as the exponential or α-stable model [14], but these do not offer the adpativity of machine learning techniques below.

2) Machine Learning Models: In terms of machine learning approaches, artificial neural networks (ANNs) has been used to predict the self-similar traffic with burstiness in [15]. Although ANNs and deep learning approaches (CNN, LSTM, wavelet, Elman) neural networks [7], [16], [17], [18] performed well in cumulative learning and prediction accuracy, it cannot give a *quantitative uncertainty* due to its intermediate black-box process. Alternatively, Gaussian Processes (GPs) have been used [6] and showed a strong adaptivity to the wireless traffic data. Nevertheless, the usage of traditional kernels are unable to capture long-range period-varying dependent characteristics which limits the efficiency of existing training data.

3) Gaussian Process (GP): Gaussian process (GP) is widely used because of its adaptability to manifold data. As a non-parametric machine learning method, the prior GP model is firstly established with compound kernel functions based on the background of the data. One optimizes the hyper-parameters using the training data to extract its posterior distribution for the predicted outcome. The prediction results given by GPs quantify the statistical significance, which is an important advantage over other black-box machine learning. As such, whilst GPs may not achieve the performance level of ANNs, they are able to quantify risk and that risk can be interpreted back to the features of the data [19].

4) Feature Extraction and Wireless Context: The features of traffic patterns may be correlated if the patterns are driven by the same specified events, i.e. the rush hours, concerts, etc. In these cases, the key point is to find the implied events information from the current flow trend by identifying where its features are close to those in historical data, hence, to predict how will the traffic demand change according to it in the past. [20], [21], [22], [23], [24] addressed the problem of feature selection, in order to determine the most discriminative and relevant features of the classified data.

In the context of wireless traffic forecasting, current literature employ classic kernel functions [6], which cannot memorize the non-periodic data pattern for extended periods. This

[1]Cranfield University, UK. [2]The Alan Turing Institute, UK. *Corresponding Author: Weisi.Guo@cranfield.ac.uk.

means the GP model do not make full use of the training data. Furthermore, wireless traffic forecasting is often interested in predicting extreme events as opposed to the overall pattern of the traffic variation. Extreme demand values are useful in driving proactive network actions (e.g. extreme high demand requires spectrum aggregation and cognitive access, whereas extreme low demand can lead to proactive sleep mode and coverage compensation [25]). Therefore, what is needed is an adaptive kernel in GP models to trade-off prediction accuracy between overall traffic variations and extreme values.

B. Novelty and Contribution

In this paper, we propose to embed the relevant data features in a flexible kernel functions, which enable the GP model to achieve this trade-off. We make three major contributions:

1) A novel feature embedding (FE) kernel GP model is proposed for forecasting wireless traffic. Specifically, fewer hyper-parameters are required in this model, which reduce the computation burden compared with that uses classic hybrid kernels. Meanwhile, the learning rate is improved significantly for irregular training data;

2) The predicted results are quantified into probability density function (PDF), which are more useful to plug into optimisation modules than the mean prediction value. Precisely, the predicted traffic is described to follow a weighted superposition distribution of mixed Gaussian distributions instead of the sum of those in traditional GP;

3) Demonstrating our forecast model on real wireless traffic data, the cumulative error curve of our model is compared against state-of-the-art algorithms used in literature (seasonal ARIMA [13] and traditional GP model [6]). Our model shows the best adaptivity and prediction accuracy trade-off between overall accuracy and extreme value accuracy.

The remainder of this paper is organised as follows. In Section II, we build a system model step by step from preprocessing to prediction. In Section III, we apply the model to the real wireless traffic data and evaluate the performance of it. Section IV concludes this paper and proposes the ideas for future work.

II. System Model

In this paper, we use a sliding window of historical traffic data to predict future traffic demand. In this paper, we focus on wireless downlink (DL) traffic data demanded by end-users at 15m intervals over a two week period - see Fig. 1.

A. Data Decomposition

The raw data is considered to be composed of a daily periodic and an aperiodic pattern from our observation and existing literature. By using a band-pass filter, the raw data can be decomposed into the aforementioned two components, as shown in Fig.1. In order to set the model free from the domination of large-scale periodic patterns, we fix the daily periodic pattern which is derived from the historical data as the established baseline [1] and only make prediction on the rest aperiodic pattern.

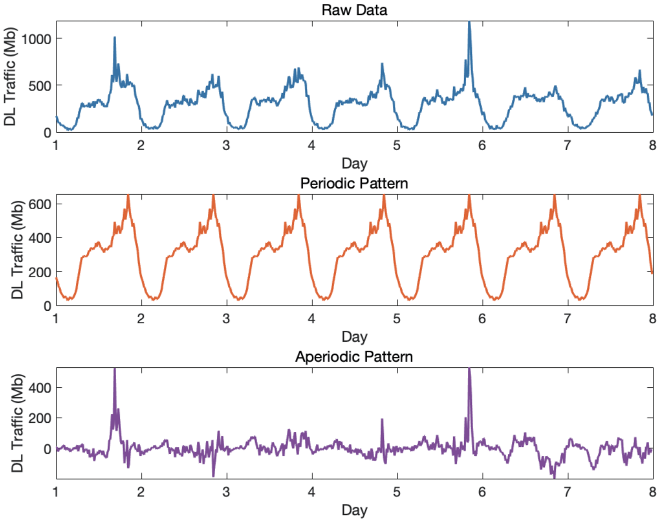

Fig. 1. The traffic demand data is decomposed into daily periodic and aperiodic components.

We assume that the aperiodic traffic consists of a noise flow and a event-driven flow which has an implicit intrinsic correlation. The latter is predictable if we can identify the features relevance in this kind of flow from the noise.

B. Priori Gaussian Process Model

The DL traffic value at time point t is assumed to be a latent GP plus noise as

$$y(t) = f(t) + \epsilon(t), \tag{1}$$

where $f(t)$ is the random variable (RV) which follows a distribution given by GP, and ϵ is the additive Gaussian noise with zero mean and variance σ_n^2. From the continuous time domain, finite number of time points taken as $\boldsymbol{t} = [t_1, t_2, ..., t_n]^T$, the RVs, $\boldsymbol{f(t)} = [f(t_1), f(t_2), ..., f(t_n)]^T$, can be assumed to follow the multivariate Gaussian as [19]

$$\boldsymbol{f(t)} \sim \mathcal{N}(\boldsymbol{M(t)}, \boldsymbol{K(t,t)}) \tag{2}$$

where $\boldsymbol{M(t)}$ is the mean function and $\boldsymbol{K(t,t)}$ is the covariance matrix given by

$$\boldsymbol{K(t,t)} = \begin{bmatrix} k(t_1,t_1) & k(t_1,t_2) & \cdots & k(t_1,t_n) \\ k(t_2,t_1) & k(t_2,t_2) & \cdots & k(t_2,t_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(t_n,t_1) & k(t_n,t_2) & \cdots & k(t_n,t_n) \end{bmatrix} \tag{3}$$

where $k(t_i, t_j)$ is the covariance between $f(t_i)$ and $f(t_j)$ represented by the kernel function.

According to (1) and (2), the priori GP probability model of DL traffic can be expressed as

$$\boldsymbol{y(t)} \sim \mathcal{N}(\boldsymbol{M(t)}, \boldsymbol{K(t,t)} + \sigma_n^2 \boldsymbol{I_n}). \tag{4}$$

[1] We acknowledge that there are baseline variations between each day of the week, but we focus on the aperiodic prediction, which is the main challenge.

C. Feature Embedding Kernel Function

In GP, the covariance between every two RVs is quantified by the kernel function which interprets the potential correlation of RVs in a high dimensional space. Here we use the Gaussian radial basis function (RBF) kernel with a feature embedding (FE) norm:

$$\begin{aligned} k_G(t_i,t_j) &= \sigma^2\exp(-\frac{\|\boldsymbol{\Lambda_i^\Theta}-\boldsymbol{\Lambda_j^\Theta}\|_2^2}{2\beta^2}) \\ \mathbb{R}\times\mathbb{R} &\xrightarrow{\text{FE}} \mathbb{R}^\Theta\times\mathbb{R}^\Theta \xrightarrow{k_G} \mathbb{R}, \end{aligned} \tag{5}$$

where $\boldsymbol{\Lambda_l^\Theta}$ is defined as the Θ dimensions weighted feature matrix of the RV at time point t_l:

$$\boldsymbol{\Lambda_l^\Theta}=[w_1\lambda_l^1,w_2\lambda_l^2,...,w_\Theta\lambda_l^\Theta]^T, \tag{6}$$

where the θ^{th} feature of RV $f(t_l)$ in the matrix is from a feature generator function $h_\theta(.)$ which can either be homogeneous or non-homogeneous of former values ($L<<l$):

$$\lambda_l^\theta=h_\theta[y(t_{l-1}),y(t_{l-2}),...,y(t_{l-L})]. \tag{7}$$

Due to the symmetry, it can be easily proved that our new kernel function still meets the conditions of Mercer's theorem.

The feature generator $h_\theta(.)$ is the key to capture the events. Here we describe an event from three perspectives. Firstly, the baseline of the events, i.e. $\Sigma_1^i y(t_{l-i})$ with $(i<<l)$. This feature helps to measure the base size of traffic and works for events traffic that are proportional to the former values. Secondly, the differences in time series values, i.e. $y(t_{l-i})-y(t_{l-j})$ with $(i,j<<l)$. This allows the absolute change, which is associated with an event, being captured. Thirdly, the fluctuation degree, i.e. $\frac{y(t_{l-a})-y(t_{l-b})}{y(t_{l-c})-y(t_{l-d})}$ with $(a,b,c,d<<l)$ or the standard deviation of former values. This feature enables the prediction giving a better result with the evaluation of the current traffic volatility.

In BS (coordinated) control systems (e.g. radio resource management or beamforming), understanding sharp changes in traffic demand (especially when above the cell capacity or significantly below economic profitability thresholds) is more important than average demand trends. As such, the proposed feature weighting process in this paper focus on building a trade-off between general prediction accuracy and the aforementioned *extreme demand values*. Another advantage of feature embedding in this way is that all the features are explainable (see Section III-B), which is critical from an explainable AI perspective [10].

To achieve this, we set a threshold ξ of traffic varying value Δy at each sample time point based on historical data as shown in Fig.2. If Δy_p at t_p is outside the $\xi\times 100\%$ confidence interval in the distribution of Δy, the associated feature $\boldsymbol{\Lambda_p^\Theta}$ will be tagged as an outlier and assigned to category $\mathcal{A}$; otherwise it is assigned to category $\mathcal{B}$. The *Relief* idea in [26] is utilized, whereby the feature weights are optimized by maximizing the sum of margin from each $\boldsymbol{\Lambda_{\mathcal{A}n}^\Theta}$ to the nearest point with a different category $N_{\mathcal{A}n}(\boldsymbol{\Lambda_{\mathcal{B}}^\Theta})$. This process is expressed as:

$$\max_w\sum_n^{|\mathcal{B}|}(M_w(\boldsymbol{\Lambda_{\mathcal{A}n}^\Theta},N_{\mathcal{A}n}(\boldsymbol{\Lambda_{\mathcal{B}}^\Theta}))\ s.t.\ \|w\|_2^2=1,w\geq 0 \tag{8}$$

where $M_w(\boldsymbol{\Lambda_p^\Theta},\boldsymbol{\Lambda_q^\Theta})=\sum_{\theta=1}^{\Theta}w_\theta\left|\boldsymbol{\Lambda_p^\theta}-\boldsymbol{\Lambda_q^\theta}\right|$, which projects the high dimensional feature vectors' norm onto one dimension, and w_θ is the weight of the θ^{th} feature.

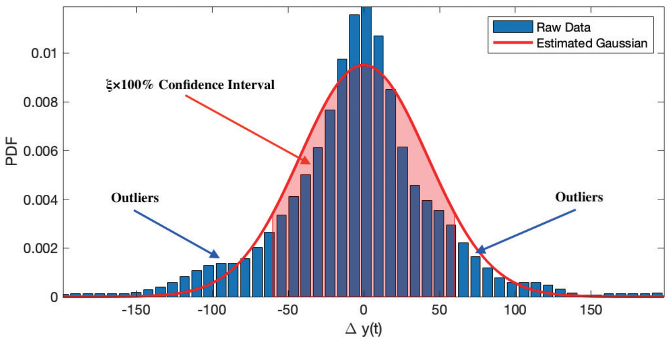

Fig. 2. Historical time points are collected into two categories according to their estimated Gaussian distribution.

In the Gaussian RBF kernel (5), the feature space can be mapped to an infinite dimensional kernel space ($e^x=\sum_{n=0}^{\infty}\frac{x^n}{n!}$). The hyper-parameter β controls the higher dimensional attenuation rate and has amplitude σ. Hyper-parameters are tuned by maximizing the corresponding log marginal likelihood function which is equivalent to minimizing the cost function $l(\beta,\sigma)$ [6]:

$$\arg\min_{\beta,\sigma} l(\beta,\sigma)=\boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y}+\log|\boldsymbol{C}|\,, \tag{9}$$

where $\boldsymbol{C}=\boldsymbol{K(t,t)}+\sigma_n^2\boldsymbol{I_n}$ and $\boldsymbol{y}$ is the matrix of known values $[y(t_1),y(t_2),...,y(t_n)]^T$. The quasi-Newton and gradient descent methods can be used in this optimization problem.

D. Posteriori Prediction

After the hyper-parameters are determined, the covariance of every two RVs in the training set can be quantified by $\boldsymbol{C(t,t}|\hat{\beta},\hat{\sigma})$, where $\hat{\beta},\hat{\sigma}$ are the optimized parameters. Let us assume that at a future time point t_f, the RV $y(t_f)$ follows the same model as the $y(t_1)\sim y(t_n)$ training set. Therefore, $\boldsymbol{K}(t_f,\boldsymbol{t}|\hat{\beta},\hat{\sigma})$ yields the covariance of $y(t_f)$ with historical RVs. The multivariate distribution for any $t_i(i\in\{1,2,...,n\})$ and t_f is

$$\begin{bmatrix} y(t_i) \\ y(t_f) \end{bmatrix}\sim\mathcal{N}\left(\boldsymbol{M}_{i,f},\boldsymbol{\Sigma}_{i,f}^2\right) \tag{10}$$

with mean $\boldsymbol{M}_{i,f}=\begin{bmatrix} M(t_i) \\ M(t_f) \end{bmatrix}$, and covariance matrix $\boldsymbol{\Sigma}_{i,f}^2=\begin{bmatrix} k(t_i,t_i)+\sigma_n^2 & k(t_i,t_f) \\ k(t_f,t_i) & k(t_f,t_f)+\sigma_n^2 \end{bmatrix}$.

So $\boldsymbol{\mathcal{Y}_i}=[y(t_i),M(t_i),\hat{\sigma},\hat{\beta}]$ given, the posterior distribution of $y(t_f)$ can be derived as

$$y_i(t_f)|\boldsymbol{\mathcal{Y}_i}\sim\mathcal{N}\left(\hat{\mu}_{i,f},\hat{\sigma}_{i,f}^2\right) \tag{11}$$

with

$$\begin{aligned} \hat{\mu}_{i,f} &= M(t_f)+\frac{k(t_f,t_i)}{k(t_i,t_i)+\sigma_n^2}(y(t_i)-M(t_i)) \\ \hat{\sigma}_{i,f}^2 &= \sigma_n^2+k(t_f,t_f)-\frac{k(t_f,t_i)k(t_i,t_f)}{k(t_i,t_i)+\sigma_n^2}. \end{aligned} \tag{12}$$

For each previous time point t_i in this model, a posterior distribution component of $y_i(t_f|\boldsymbol{\mathcal{Y}_i})$ can be generated. In naive GP, the predicted distribution $y(t_f)$ is also a Gaussian distribution which sums the influence of each previous point on its mean and variance [19]. In our proposed FE-GP forecasting model, the predicted distribution uses a Gaussian mixed model (GMM). Consider the GMM resultant PDF of $y(t_f)$ is the superposition of every individual distribution components from each $y(t_i)$ and $y(t_f)$ with a normalization coefficient as

$$P(y(t_f)|\boldsymbol{\mathcal{Y}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi\hat{\sigma}_{i,f}^2}}\exp(-\frac{(y_f-\hat{\mu}_{i,f})^2}{2\hat{\sigma}_{i,f}^2}) \qquad (13)$$

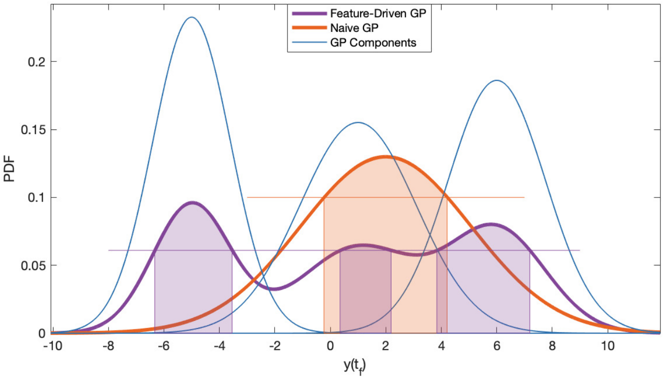

Fig. 3. The purple and orange shadows have the same area representing the same probability.

An example is shown in Fig.3. Blue lines are distribution components, derived by the covariance matrix of three previous points with the future point. Naive GP gives the average prediction result of this future point, i.e. also a Gaussian distribution, under integrated impacts from all components. While in FE-GP, the GMM prediction result of this future point is assumed to have an equal probability to follow one of these three distribution components. The purple line gives the overall PDF of FE-GP.

III. Experimental Results and Discussion

A. Data Source

The data we use for training comes from base stations (BSs) in a 4G metropolitan area. The anonymous data is given by our industrial collaborator. It consists of aggregated downlink (DL) and uplink (UL) traffic demand volume per 15 minute interval over several weeks. We have selected a few example BSs at random to demonstrate our forecasting algorithm's performance.

B. Explainable Feature Matrix for FE-GP

When applying FE-GP to wireless DL traffic forecasting, the first to be considered is what does the feature matrix consist of. In our experiment, the features are set to be:

$$\begin{gathered}\lambda_l^1 = y(t_{l-1});\;\; \lambda_l^2 = y(t_{l-2});\\ \lambda_l^3 = y(t_{l-3});\;\; \lambda_l^4 = y(t_{l-4});\\ \lambda_l^5 = y(t_{l-2}) - y(t_{l-5});\;\; \lambda_l^6 = y(t_{l-3}) + y(t_{l-4});\\ \lambda_l^7 = \frac{y(t_{l-2}) - y(t_{l-1})}{y(t_{l-3}) - y(t_{l-2})};\;\; \lambda_l^8 = \frac{y(t_{l-2}) - y(t_{l-3})}{y(t_{l-3}) - y(t_{l-4})};\\ \lambda_l^9 = std.[y(t_{l-1}), y(t_{l-2}), y(t_{l-3}), y(t_{l-4}), y(t_{l-5})];\end{gathered} \qquad (14)$$

where $std.(\boldsymbol{y})$ is the standard deviation of elements of $\boldsymbol{y}$.

C. Performance Metrics

We use the absolute cumulative error (ACE) as the performance metric:

$$\text{ACE} = \sum_{n=i}^{j}\left|\hat{y}(n) - y(n)\right|, \qquad (15)$$

where $\hat{y}$ is the predicted DL traffic and $y(n)$ is the real data. For a fixed value forecast (one-step-ahead forecasting of the DL traffic), we assign $\hat{y}$ to be the value that has the maximum posterior probability.

D. Results Analysis

Fig.4 shows a comparison of forecasting algorithms over a week (672 points): (1) proposed FE-GP, (2) classical Naive-GP, (3) Seasonal ARIMA, against real 4G DL data. The cumulative error is shown for 2 different representative parts of the data: (left) average demand shows similar performance between FE-GP and Naive-GP; and (right) extreme spike demand shows superior performance by FE-GP against both Naive-GP and S-ARIMA. From the GP models perspective, in the average part (left), both FE-GP and Naive-GP consider most of the traffic demands as noise flow, i.e. $\epsilon(t)$ in the initial model, thus they perform similarly; In the extreme spike part (right), FE-GP can correctly recognize a potential event-driven flow, which has happened before, using features from the last few points, yet Naive-GP cannot, hence FE-GP gives a better prediction. In Fig. 6, we demonstrate that the proposed FE-GP perform the best overall due to its adaptive to the extreme values, even though it might be a little worse on average demand in few specific time stamps.

E. Uncertainty Quantification

Posterior distribution of both models at a few representative points are given in Fig.5. Different from single peak Gaussian distribution predicted by Naive-GP, the GMM in FE-GP gives more general distributions for prediction. In the absence of a known periodicity, Naive-GP sums the effect of the last few time points, while FE-GP consider the effect from all time points according to their similarity in features with the predicted point. Consequently, there may be several peaks scattered over the forecast, which will inform proactive optimisation modules.

In data-driven wireless resource proactive optimization system [2], we ought to focus on not only the benefits brought by the system decision, but also the potential risks that drive regret

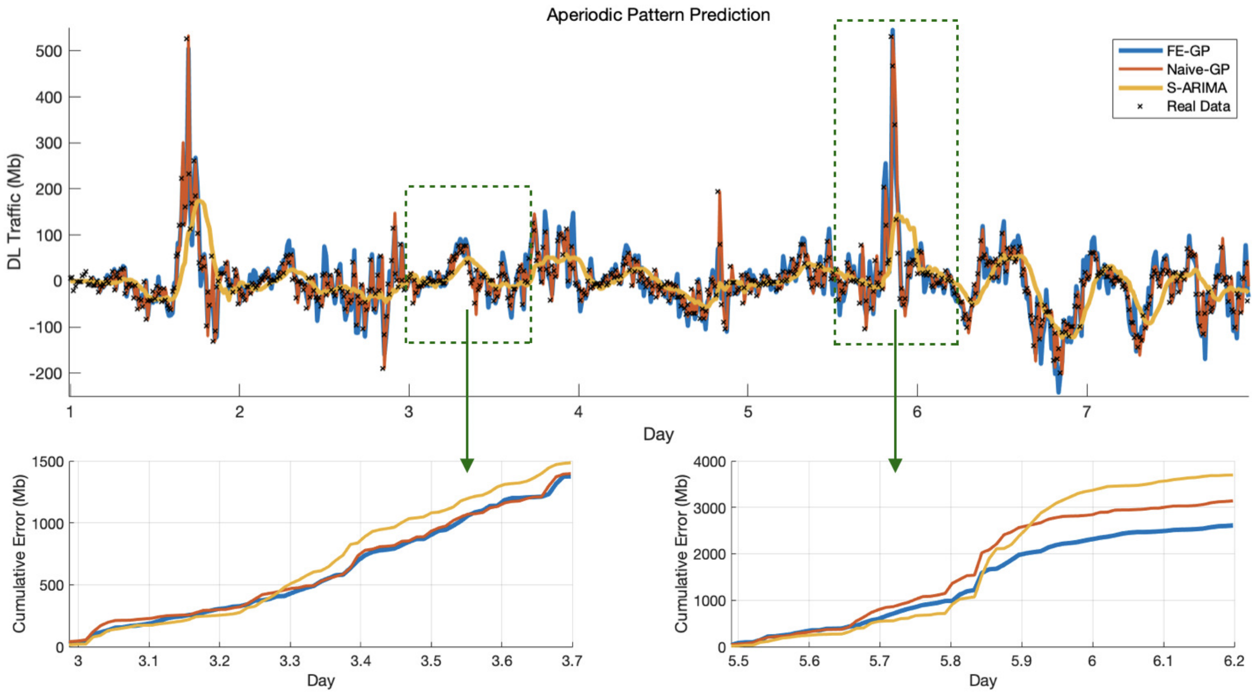

Fig. 4. The purple and orange shadows have the same area representing the same probability.

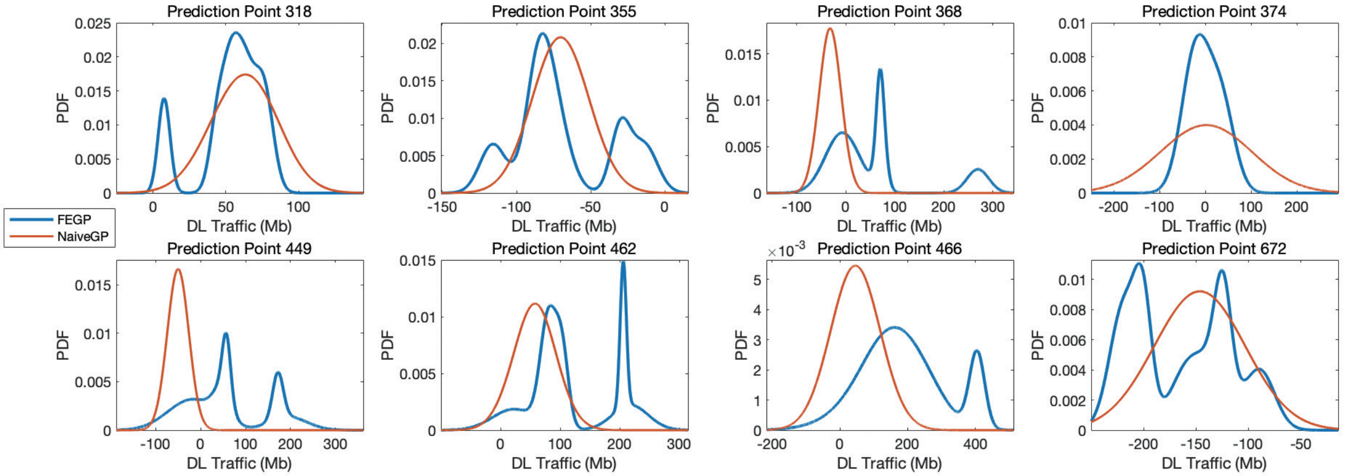

Fig. 5. Comparison of forecasts against 4G DL data. The cumulative error for 2 representative parts: (left) average demand; and (right) spike demand.

functions, i.e., the occurrence of extreme demands. In our proposed FE-GP prediction model, the risks can be quantified from posterior distribution. For example, in Fig.5:

(1) **Low-traffic triggers proactive sleep mode and coverage compensation:** Our FE-GP prediction points 318 and 672 demonstrates clear non-negligible probability of low traffic whilst the mean prediction is similar to that of the naive-GP. That is to say, we may need to proactive sleep selected cells to achieve more energy efficient operations, whilst using other neighbouring cells across RATs to compensate [25]. The risk of doing so is quantified by the posterior distribution (e.g., there is a small risk that there is actually high demand and compensated coverage is not enough).

(1) **Spike-traffic triggers proactive spectrum aggregation and offloading:** prediction point 368, there is a non-negligible high probability density area appearing at extreme value, which is far away from the predicted mean value. This can be used to inform proactive spectrum aggregation and off-loading of non-vital traffic to delay-tolerant RATs. The risk of doing so is quantified by the posterior distribution (e.g., there is a small risk that there is actually no demand for high capacity).

F. Training Process

As the training set increases over time, the FE-GP model becomes more sensitive to spikes due to its adaptivity to features. Nevertheless, the cost of computing goes up with the size of training set as well, thus we have to set a size threshold to the training set. In Naive-GP, we can discard data in reverse chronological order without affecting the performance of the model. However, in FE-GP, we must make a trade-off between

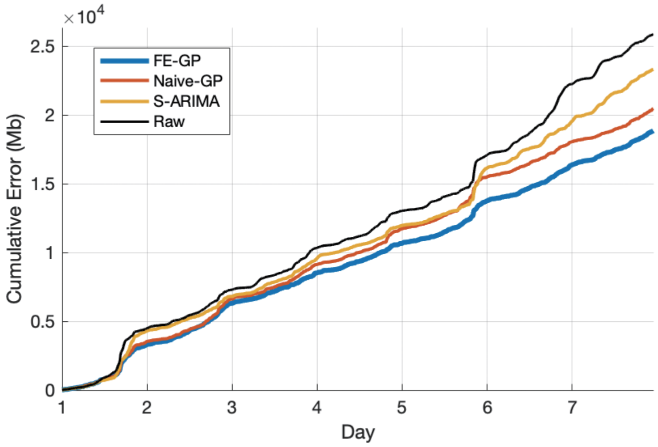

Fig. 6. Cumulative error comparison between forecasting algorithms.

the sensitivity of spikes and overall prediction accuracy, i.e., keeping more extreme value time points means the model is more sensitive to spikes prediction but may reduced overall performance. This need to be done case by case with each pre-exiting resource proactive optimization system.

IV. Conclusion and Future Work

Forecasting extreme demand spikes and troughs is essential to avoiding outages and improving energy efficiency. Whilst significant research into traffic forecasting using ARIMA, GPs, and ANNs have been conducted, current methods predominantly focus on overall performance and/or do not offer probabilistic uncertainty quantification. Here, we designed a feature embedding (FE) kernel for a GP model to forecast traffic demand. The FE kernel enabled us to trade-off overall forecast accuracy against peak-trough accuracy. We compared its performance against both conventional GPs, ARIMA models, as well as demonstrate the uncertainty quantification output. The advantage over neural network (e.g. CNN, LSTM) models is that the probabilistic forecast uncertainty can directly feed into decision processes in self-organizing-network (SON) modules in the form of both predicted average KPI benefit and regret functions using methods such as probabilistic numerics. Our future work will focus on expanding to spatial-temporal dimension via Gaussian random fields integration, consider multi-variate forecasting across different service slices, as well as employing Bayesian training in Deep Gaussian Process (DGP) models [27] to avoid catastrophic forgetting and to combat the dynamiticity of the traffic process.

References

[1] B. Ma, W. Guo, and J. Zhang, “A survey of online data-driven proactive 5g network optimisation using machine learning,” *IEEE Access*, vol. 8, pp. 35 606–35 637, 2020.

[2] Z. Du, Y. Sun, W. Guo, Y. Xu, Q. Wu, and J. Zhang, “Data-driven deployment and cooperative self-organization in ultra-dense small cell networks,” *IEEE Access*, vol. 6, pp. 22 839–22 848, 2018.

[3] S. O. Somuyiwa, A. Gyorgy, and D. Gundz, “A reinforcement-learning approach to proactive caching in wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 36, June 2018.

[4] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs, “Mobile traffic forecasting for maximizing 5G network slicing resource utilization,” in *IEEE Conference on Computer Communications (INFOCOM)*, May 2017, pp. 1–9.

[5] L. Le, D. Sinh, L. Tung, and B. P. Lin, “A practical model for traffic forecasting based on big data, machine-learning, and network KPIs,” in *2018 15th IEEE Annual Consumer Communications Networking Conference (CCNC)*, Jan 2018, pp. 1–4.

[6] Y. Xu, W. Xu, F. Yin, J. Lin, and S. Cui, “High-accuracy wireless traffic prediction: A GP-based machine learning approach,” in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2017, pp. 1–6.

[7] K. Zhang, G. Chuai, W. Gao, X. Liu, S. Maimaiti, and Z. Si, “A new method for traffic forecasting in urban wireless communication network,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 66, Mar 2019.

[8] X. Cao, Y. Zhong, Y. Zhou, J. Wang, C. Zhu, and W. Zhang, “Interactive temporal recurrent convolution network for traffic prediction in data centers,” *IEEE Access*, vol. 6, pp. 5276–5289, 2018.

[9] B. Yang, W. Guo, B. Chen, G. Yang, and J. Zhang, “Estimating mobile traffic demand using Twitter,” *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 380–383, Aug 2016.

[10] W. Guo, “Explainable Artificial Intelligence (XAI) for 6G: Improving Trust between Human and Machine,” *IEEE Communications Magazine*, vol. 58, no. 6, 2020.

[11] Z. Wei, B. Li, W. Guo, W. Hu, and C. Zhao, “On the accuracy and efficiency of sensing and localization for robotic,” *IEEE Transactions on Mobile Computing*, pp. 1–1, 2020.

[12] C. Li, W. Guo, S. Sun, S. Al-Rubaye, and A. Tsourdos, “Trustworthy Deep Learning in 6G Enabled Mass Autonomy: from Concept to Quality-of-Trust KPIs,” *IEEE Vehicular Technology Magazine*, 2020.

[13] F. Xu, Y. Lin, J. Huang, D. Wu, H. Shi, J. Song, and Y. Li, “Big data driven mobile traffic understanding and forecasting: A time series approach,” *IEEE Transactions on Services Computing*, vol. 9, no. 5, pp. 796–805, Sep. 2016.

[14] R. Li, Z. Zhao, J. Zheng, C. Mei, Y. Cai, and H. Zhang, “The learning and prediction of application-level traffic data in cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3899–3912, June 2017.

[15] L. Xiang, X. Ge, C. Liu, L. Shu, and C. Wang, “A new hybrid network traffic prediction method,” in *IEEE Global Telecommunications Conference (GLOBECOM)*, Dec 2010, pp. 1–5.

[16] J. Feng, X. Chen, R. Gao, M. Zeng, and Y. Li, “Deeptp: An end-to-end neural network for mobile cellular traffic prediction,” *IEEE Network*, vol. 32, no. 6, pp. 108–115, November 2018.

[17] X. Wang, Z. Zhou, F. Xiao, K. Xing, Z. Yang, Y. Liu, and C. Peng, “Spatio-temporal analysis and prediction of cellular traffic in metropolis,” *IEEE Transactions on Mobile Computing*, pp. 1–1, 2018.

[18] F. Ni, Y. Zang, and Z. Feng, “A study on cellular wireless traffic modeling and prediction using elman neural networks,” in *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, vol. 01, Dec 2015, pp. 490–494.

[19] A. G. Wilson, “Covariance kernels for fast automatic pattern discovery and extrapolation with gaussian processes,” Ph.D. dissertation, University of Cambridge, 2014.

[20] B. Cao, D. Shen, J.-T. Sun, Q. Yang, and Z. Chen, “Feature selection in a kernel space,” in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 121–128.

[21] Z. Wei, W. Guo, B. Li, J. Charmet, and C. Zhao, “High-dimensional metric combining for non-coherent molecular signal detection,” *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1479–1493, 2020.

[22] M. Ramona, G. Richard, and B. David, “Multiclass feature selection with kernel gram-matrix-based criteria,” *IEEE transactions on neural networks and learning systems*, vol. 23, no. 10, pp. 1611–1623, 2012.

[23] Z. Wei, B. Li, C. Sun, and W. Guo, “Sampling and inference of networked dynamics using log-koopman nonlinear graph fourier transform,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 6187–6197, 2020.

[24] K.-P. Wu and S.-D. Wang, “Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space,” *Pattern Recognition*, vol. 42, no. 5, pp. 710–717, 2009.

[25] W. Guo and T. O’Farrell, “Dynamic cell expansion with self-organizing cooperation,” *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 5, pp. 851–860, May 2013.

[26] Y. Sun, “Iterative RELIEF for feature weighting: algorithms, theories, and applications,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1035–1051, 2007.

[27] S. Lee, J. Kim, J. Jun, J. Ha, and B. Zhang, “Overcoming catastrophic forgetting by incremental moment matching,” *Advances in Neural Information Processing Systems (NIPS)*, 2017.