# UNDERSTANDING MODEL DRIFT IN A LARGE CELLULAR NETWORK

**Shinan Liu** [1] **Francesco Bronzino** [2] **Paul Schmitt** [3] **Arjun Nitin Bhagoji** [1] **Nick Feamster** [1]
**Hector Garcia Crespo** [4] **Timothy Coyle** [4] **Brian Ward** [4]

## ABSTRACT

Operational networks are increasingly using machine learning models for a variety of tasks, including detecting anomalies, inferring application performance, and forecasting demand. Accurate models are important, yet accuracy can degrade over time due to model drift, whereby either the characteristics of the data change over time (data drift) or the relationship between the features and the target change over time (concept drift). Drift occurs in operational networks for a variety of reasons, ranging from software upgrades to seasonality to changes in user behavior. This paper presents an initial exploration into model drift in a large cellular network in the United States for a major metropolitan area in the context of demand forecasting. We taxonomize data drift and decompose concept drift based on frequency. Furthermore, we identify the sources of concept drift for forecasting downlink volume: lower traffic volumes, and higher speeds tend to exhibit more concept drift; and the features that contribute most to concept drift of volume forecasting are User Equipment (UE) downlink packets, UE uplink packets, and Real-time Transport Protocol (RTP) received packets.

## 1 INTRODUCTION

Network operators are increasingly relying on machine learning models to perform a variety of network operations tasks, including anomaly detection (Shon & Moon, 2007; Shon et al., 2005), performance inference (Futuriom, accessed August, 2021) and diagnosis, and forecasting (Mei et al., 2020; Chinchali et al., 2018). Yet, even if machine learning models are accurate for specific network management tasks on a particular set of test data, deploying and maintaining the models can prove challenging in practice (Sommer & Paxson, 2010). A significant operational challenge is *model drift*, whereby a model that is initially accurate at a particular point in time becomes less accurate over time—either due to a sudden change, periodic shifts, or gradual drift over time. Previous work in applying machine learning models to network management tasks has generally trained models on fixed datasets (Sinclair et al., 1999; Dong & Wang, 2016; Shon & Moon, 2007; Shon et al., 2005; Ayoubi et al., 2018; Mei et al., 2020; Chinchali et al., 2018), demonstrating the ability to predict various network properties or instances at fixed points in time. Yet, a model that performs well at a particular point in time does not necessarily perform well on future data.

Models can become less accurate at predicting target variables for many reasons. One cause is a sudden, drastic change to the environment. For example, the installation of new equipment, a software upgrade, or a sudden change in traffic patterns or demands can cause models to suddenly become inaccurate. One notable instance that exhibited such a sudden shift was the COVID-19 pandemic, an exogenous shock that resulted in significant changes to behavior patterns (Lutu et al., 2020) and traffic demands (Liu et al., 2021). Another characteristic is periodic change. For example, one phenomena that we observe is a drift in model accuracy with a seven-day period; diurnal and periodic patterns in network traffic are, of course, both well-documented and relatively well-understood. A new contribution in this paper, however, is to demonstrate that machine learning models also exhibit similar patterns with respect to accuracy.

Model drift is also a relatively well-understood phenomenon in machine learning and has been studied in the context of many other prediction problems (Schlimmer & Granger, 1986; Lu et al., 2018; Gama et al., 2014). To our knowledge, this paper is the first to study drift in the context of cellular networks. We characterize model drift in the context of a large cellular network, exploring drift for a variety of cellular key performance indicators (KPIs) using nearly three and a half years of data from a major metropolitan area in the United States. *Data drift* refers to changes in the distribution of features that are model inputs. *Concept drift* refers to changes of relationships between model inputs (i.e., features) and the output (i.e., target prediction). In cellular networks, concept drift introduces unique challenges: in

[1]Department of Computer Science, University of Chicago, Illinois, USA [2]Department of Computer Science, Université Savoie Mont Blanc, Annecy, France [3]Information Sciences Institute / University of Southern California, California, USA [4]Verizon Inc., USA. Correspondence to: Shinan Liu <shinanliu@uchicago.edu>.

*Practical Adoption Challenges of ML for Systems in Industry (PACMI'22), workshop co-located with MLSys 2022* , Santa Clara, CA, USA. 

contrast to tasks such as image or text classification, where the semantics of prediction occurs on a fixed object and characteristics of the features change relatively slowly over time (Fdez-Riverola et al., 2007; Žliobaitė, 2010; Gama et al., 2014), predictions of network characteristics occur continuously over time, and predictions are occurring within the context of a system that changes over time due to both gradual evolution and exogenous shocks. Moreover, dynamic signal interference due to environment changes often occur in wireless networks (Zheng et al., 2014), a phenomenon that adds another layer of complexity and novelty beyond previous studies of drift.

Both data drift and concept drift are important to characterize. While understanding concept drift has obvious importance since it relates to model accuracy, understanding data drift is important because it can help explain concept drift. Although this paper does not go so far as to build a mechanisms to detect or mitigate model drift, our analysis lays the groundwork for techniques that could help to develop detecting and mitigation strategies in the future. We make contributions in three areas.

First, we characterize *data drift* from the input KPIs in a large cellular network. [1] Second, we decompose *concept drift* in a large cellular network based on frequecny components of time-series signals of accuracy. Third, we apply machine learning-based analysis, including contributing factor explanations and feature importance movements to illuminate the causes of model drift.

## 2 Context

In this section, we focus on the dataset and the problem setting. We start with a description of our dataset and then discuss our modelling goal—forecasting target KPIs 180 days in the future.

### 2.1 Dataset

Our analysis of model drift is based on daily eNodeB-level LTE network measurements from a major wireless carrier in the United States. This dataset, collected from a metropolitan area, includes 224 daily Key Performance Indicators (KPIs) for each eNodeB deployed. KPIs are statistics collected in the network and used by the operator of the network to monitor and assess network performance, resource utilization, and user experience, to further support planning, optimization, and troubleshooting. Further, KPIs can relate to either voice or data connections, and some of them have separate directional measurements: the downlink where the network is transmitting the data down to the UEs and the uplink where the network is receiving data from the UEs.

[1] Due to space constraints, we summarize these findings in the main body of the paper; more detailed exploration is in Appendix A.

| | |
|---|---|
| **Collection period** | Jan. 1st 2018 – May 3rd 2021 |
| **Number of KPIs** | 224 |
| **Target KPIs** | Downlink volume<br>Throughput<br>Peak active users<br>RRC est. success<br>S1-U call drop rate |
| **Number of eNBs** | 898 |
| **Number of logs** | 737,577 |

*Table 1.* Summary of dataset.

Finally, the dataset contains features used to uniquely identify eNodeBs (Evolved NodeBs, or the "base station" in the LTE architecture) and their characteristics (*e.g.*, the density of their deployment location).

KPIs are typically calculated using eNodeB counters, measurements, and events generated by user equipment (UE) and radio access network (RAN) equipment. Counters are incremented based on network events that are generated or received by the eNodeB. Some metrics are established to measure the network based on the LTE standard released by 3GPP (3GPP, accessed July, 2021), the body that determines cellular standards. These "raw" measurements can be collected from the eNodeB using external tools that store data for longer periods of time or build formulas (which resemble KPIs but are generated from one or many underlying raw KPIs). KPIs are stored as statistical representations (*e.g.*, min/max, average, percentiles).

Table 1 summarizes the characteristics of the dataset. The dataset contains information collected from 898 eNodeBs in the metropolitan area. For each day, the dataset holds the 224 KPIs collected for the base station and divided mainly across the aforementioned four categories, *i.e.*, traffic, data speeds, retainability, and accessibility. The dataset spans more than three years, from January 1st, 2018 to May 3rd, 2021, totalling more than 737,577 daily KPI logs.

### 2.2 Problem

Network capacity forecasting is important because it guides infrastructure configuration, management, and augmentation. We focus on per-eNodeB level KPI forecasting to provide suggestions for machine learning model deployment, maintenance, and operation in large cellular networks. We use historical data—*i.e.*, all available KPIs up to a given day—to forecast one or more KPIs of interest 180 days in the future. We use a 180-day forecast window because the model outputs we focus on are used *in practice* for network infrastructure provisioning and augmentation over long timescales. We examine the target KPIs (in Table 1) and present the results of downlink volume for brevity.

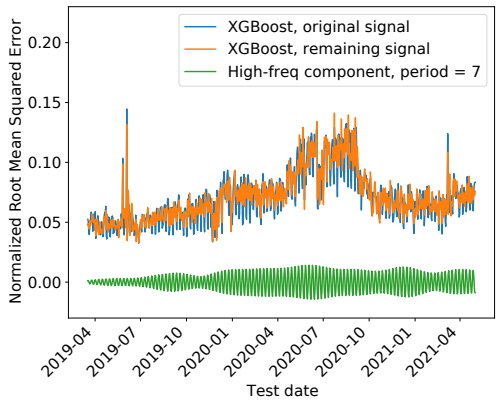



*Figure 1.* Signal decomposition on concept drift of XGBoost model output targeted to downlink volume.

## 3 Drift Components Based on Frequency Decomposition

Given the temporal nature of the target KPIs in the network, we look to apply signal processing techniques to investigate concept drift components. To formalize these patterns and identify the sources, we perform frequency-domain analysis on these signals. We take the output of XGBoost model example to analyze and view the NRMSE (Normalized Root Mean Squared Error) time series output of a forecasting model as a short-time signal that has a sampling rate of one day. The phenomenon on other models and other KPIs are similar, and we omit them for brevity.

**High-frequency drift.** As shown, there is a strong 7-day periodic pattern in Figure 5; this pervasive high-frequency component contributes to the drift on shorter timescales. Therefore, we seek to remove it from the signal using a Butterworth bandpass filter. In Figure 1, we use signal decomposition to remove the weekly component and plot the original signal, the 7-day component, and the remaining signal. As shown, the amplitude of the weekly signal is relatively higher after COVID-19 lockdowns occur around April 2020. This means that a stronger fluctuation of prediction errors happens. Because this model is trained 15 months before COVID-19, it fails to capture more abrupt variance for days of the week during COVID. Over time, the amplitude of the removed signal gradually recovers, except for a slight increase around December 2020. This is possibly because mobility changes during COVID are more difficult to predict and vary more across different eNodeBs than during other time periods.

**Low-frequency drift.** Longer-term evolution in drift are evident in lower frequencies; these components can sometimes be attributed to seasonal (*i.e.*, weather-related) patterns. Figure 5 shows a clear low-frequency pattern around the periodicity of 90 days (roughly a season). The periodicity is discontinued between July 2019 to April 2020 and October 2020 to January 2021, though. In general, seasonality appears to subside during winter periods of each year, which might be due to user behaviors such as the holiday season beginning around Thanksgiving to the new year being more irregular compared with other parts of the year. The low-frequency signal is especially strong during COVID. We believe this could partly be due to people remaining at home, which resulted in more stability in the signals.

**Exogenous shock.** Looking at the remaining signal in Figure 1, we observe a clear indication of an exogenous shock during COVID-19 resulting in sudden drift. The NRMSE drifts from 7% to 12% during that period. This drift corresponds with a range of components, clearly visible as a vertical band in Figure 5 at the same time. Stronger components from 7 to 90 days indicate severe environment, user behavior, or network changes. Unexpected drift often occurs during disasters and other exogenous shocks (Liu et al., 2021), or potentially as a result of a network-wide upgrade (*e.g.*, a wide-scale transition of users from 4G LTE to 5G could result in changes of both network and user behaviors).

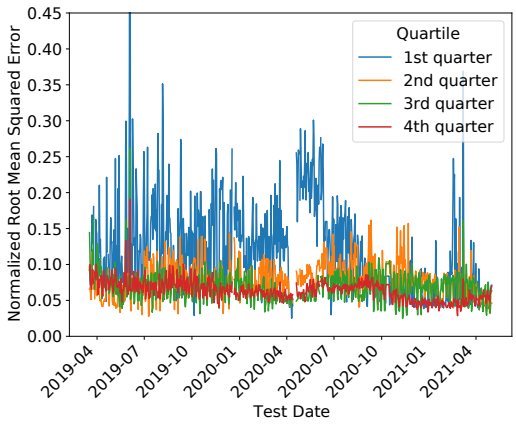
## 4 What Introduces Concept Drift?

For many practical applications, it would be important to unearth more than the presence of drift, but also an understanding of its sources. In this section, we look to provide insight into concept drift from multiple perspectives.

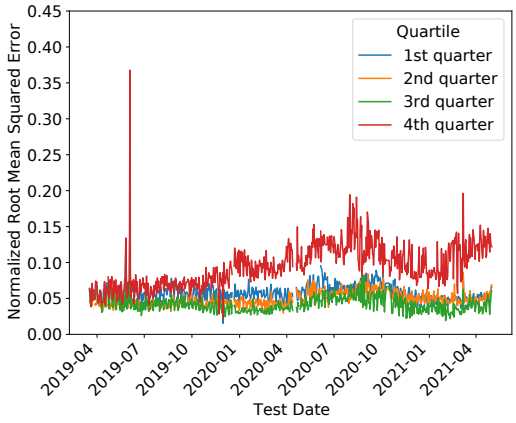
### 4.1 Effects of traffic and performance

To understand drift, we begin by exploring natural classes in the dataset. In this section, we classify the dataset based on traffic tiers, and performance tiers. The goal is still downlink volume 180-day forecast. We keep the model, the training set size and training period the same, *i.e.*, KNeighborsDist and 14 days of historical data with an end date of July 1, 2018. Moreover, we ensure that in each experiment, we have the same number of eNodeBs or number of logs for different classes. Then we retrain and test again based on these classes.

**Traffic.** We classify the eNodeBs in the dataset based on traffic tiers. We split the dataset into four equal-size parts partitioned by 1st quartile, median, and 3rd quartile based on the downlink volume KPI. Then we retrain models and test by date. As shown in Figure 2a, data subsets with higher traffic history have better predictions. The average NMRSE is lower for higher volume eNodeBs compared to low-volume eNodeBs, and the weekly patterns show smaller fluctuations. This may be because that for a volume forecasting problem, if we decompose the dataset based on volume, even though the model performance on raw RMSE could be similar, when normalized the same amount of deviation would result in more concept drift.




**(a)** *Traffic (Volume).*




**(b)** *Performance (Throughput).*

*Figure 2.* Effects of traffic and performance tiers on concept drift. The lowest quarter in traffic and the highest quarter in performance exhibit the highest NRMSE.

**Performance.** We again split the dataset into quartiles, but here we classify based on the throughput. In Figure 2b, again, we see that from 1st to 3rd quartile, the average NRMSE is less and the weekly component has slightly smaller fluctuations. The 4th quartile, though, which is the group of eNodeBs with the highest throughput values, experiences the higher NMRSE values.

### 4.2 Permutation-based feature importance drift

We seek to understand the input features, and their relative importance, that drive model performance. To do so, we discover the most sensitive features through permutation. The higher the score for a feature, the more important it is to the model.

We again use KNeighborsDist and 14 days of historical data with an end date of July 1st, 2018 for the model. We then apply the permutation-based approach to each data subset split by dates. Figure 3 demonstrates the top 6 features that most contribute to model performance using the volume

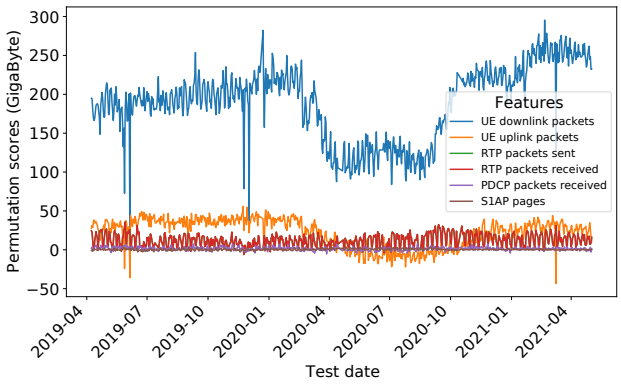



*Figure 3.* Top six features that contribute to concept drift.

KPI, meaning the permutation score related to the change in the number of MB predicted. As shown, the top three features dominate the majority of the predicted volume changes using this technique. Before the COVID-19 lockdown, permutations on User Equipment (*i.e.*UE) downlink packets can cause the predicted downlink volume to deviate over 200k MB. At the same time, the UE uplink packets have a permutation score of 48k MB on average. Then beginning in early 2020, effect related to COVID-19 impact the network, and these two features both experience significant drops. UE downlink packets drops to roughly 120k MB, while UE uplink packets even drops below 0, which indicates that the feature has begun to harm the prediction performance. After October 2020, the permutation score of UE downlink packets returns to its prior range and even increases, reaching up to 250k MB. Over the entire period, the third most important feature using the permutation technique is the Real-time Transport Protocol (*i.e.*RTP), which tends to maintain a relatively stable permutation score.

We find that decreases of the top two features negatively correlate with the increase in NRMSE, indicating that some features were less important during COVID-19. The magnitude of permutation score decreases may be a good indicator for the sensitivity of a particular feature to drift.

## 5 Conclusion

This paper characterized the sources of model drift in a large cellular network in the context of KPI forecasting. We explored how data drift can contribute to concept drift. Using dataset decomposition and permutation-based importance explanations, we explored the sources of the concept drift when predicting KPIs, focusing on downlink volume forecast. We find that low traffic volume, and higher network throughput or speed also correspond to more concept drift. We also find that UE downlink data packets, UE uplink data packets, and total RTP received packets are the top three features that drift in ways that affect model performance over time.

## REFERENCES


3GPP. *LTE standard, 3GPP TS 32.425 version 9.5.0 Release 9*, accessed July, 2021. https://www.etsi.org/deliver/etsi_ts/132400_132499/132425/09.05.00_60/ts_132425v090500p.pdf.

Ayoubi, S., Limam, N., Salahuddin, M. A., Shahriar, N., Boutaba, R., Estrada-Solano, F., and Caicedo, O. M. Machine learning for cognitive network management. *IEEE Communications Magazine*, 56(1):158–165, 2018.

Chinchali, S., Hu, P., Chu, T., Sharma, M., Bansal, M., Misra, R., Pavone, M., and Katti, S. Cellular network traffic scheduling with deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Dong, B. and Wang, X. Comparison deep learning method to traditional methods using for network intrusion detection. In *2016 8th IEEE International Conference on Communication Software and Networks (ICCSN)*, pp. 581–585. IEEE, 2016.

Fdez-Riverola, F., Iglesias, E. L., Díaz, F., Méndez, J. R., and Corchado, J. M. Applying lazy learning algorithms to tackle concept drift in spam filtering. *Expert Systems with Applications*, 33(1):36–48, 2007.

Futuriom. *Verizon Applies Machine Learning to Operations*, accessed August, 2021. https://www.futuriom.com/articles/news/verizon-applies-machine-learning-to-operations/2018/08.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

Liu, S., Schmitt, P., Bronzino, F., and Feamster, N. Characterizing service provider response to the covid-19 pandemic in the united states. In *PAM 2021-Passive and Active Measurement Conference*, 2021.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12): 2346–2363, 2018.

Lutu, A., Perino, D., Bagnulo, M., Frias-Martinez, E., and Khangosstar, J. A characterization of the covid-19 pandemic impact on a mobile network operator traffic. In *Proceedings of the ACM Internet Measurement Conference*, pp. 19–33, 2020.

Mei, L., Hu, R., Cao, H., Liu, Y., Han, Z., Li, F., and Li, J. Realtime mobile bandwidth prediction using lstm neural network and bayesian fusion. *Computer Networks*, 182: 107515, 2020.

Schlimmer, J. C. and Granger, R. H. Incremental learning from noisy data. *Machine learning*, 1(3):317–354, 1986.

Shon, T. and Moon, J. A hybrid machine learning approach to network anomaly detection. *Information Sciences*, 177 (18):3799–3821, 2007.

Shon, T., Kim, Y., Lee, C., and Moon, J. A machine learning framework for network anomaly detection using svm and ga. In *Proceedings from the sixth annual IEEE SMC information assurance workshop*, pp. 176–183. IEEE, 2005.

Sinclair, C., Pierce, L., and Matzner, S. An application of machine learning to network intrusion detection. In *Proceedings 15th Annual Computer Security Applications Conference (ACSAC'99)*, pp. 371–377. IEEE, 1999.

Sommer, R. and Paxson, V. Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE symposium on security and privacy*, pp. 305–316. IEEE, 2010.

Zheng, G., Krikidis, I., Masouros, C., Timotheou, S., Toumpakaris, D.-A., and Ding, Z. Rethinking the role of interference in wireless networks. *IEEE Communications Magazine*, 52(11):152–158, 2014.

Žliobaitė, I. Adaptive training set formation. 2010.

# A Data Drift Taxonomy

Data drift is caused by distributional changes in features that are used as input in a machine learning pipeline. Such changes can be predictable or unpredictable; sudden or gradual; planned or unplanned; recurring or one-time. In this section, we provide a brief taxonomy of data drift using examples from our dataset. Figure 4 demonstrates long-term evolvements of downlink volume, with 3-week inset figures that provide a closer look at weekly patterns.

**Gradual drift.** Gradual drift occurs when the distribution of a variable changes over a period of time. Figure 4a shows an example of gradual drift using the data volume KPI for an eNodeB in our dataset. As shown, over the roughly 3.5 years in the plot, the volume increases for the eNodeB from 0.4 normalized data volume to a somewhat steady point around 0.8 in the latter half of 2020[2]. Zooming in, the 3-week inset reveals a periodic pattern three times, which indicates a weekly component. Gradual drift can be caused by a number of factors in real-world networks. For example, cellular carriers continually augment and improve their infrastructure. Upgraded infrastructure could then allow customers to send and receive more data on the network, causing the KPI to increase over time for a given eNodeB.

**Incremental drift.** Incremental drift is similar to gradual drift in that the variable shifts to a different distribution; however, incremental drift includes more distributions in the interim of a shift (*i.e.*, phase shifting from one steady state to another with a period of transition in-between). Figure 4d illustrates incremental drift for an eNodeB in our dataset, again using the data volume KPI. As the plot shows, volume on this eNodeB steadily increases over the observation period from roughly 0.2 normalized volume to 0.7. This incremental drift could be attributable to network changes such as the introduction of a new service plan and promotion. For example, carriers may add a plan that offers customers unlimited data. As users increasingly adopt such plans over metered plans, data volume increases for network eNodeBs. Note that even if long-term evolution is incremental, we still observe weekly patterns on shorter timescales.

**Sudden drift.** Sudden drifts correspond to abrupt shifts in the dataset distribution. In the context of cellular networks, many events can be the cause for a possibly very noticeable, sudden drift. We can categorize these sources of sudden drift into *operator events*, (*e.g.*, network-wide upgrades(, *scheduled events* (*e.g.*, a special event such as the Super Bowl), or *unexpected events* (*e.g.*, the COVID-19 pandemic or natural disasters).

Figure 4c shows a clear instance of sudden drift extracted from our dataset concerning total downlink volume for a given eNodeB. We observe that the normalized total drastically decreases starting in early 2020, then stabilizes to a consistently low average from March 2020 onwards (almost 60% less experience volume). These dates suggest that these sudden changes could be due to the events caused by the COVID-19 pandemic where people's mobility was heavily affected, and cellular networks with it (Lutu et al., 2020).

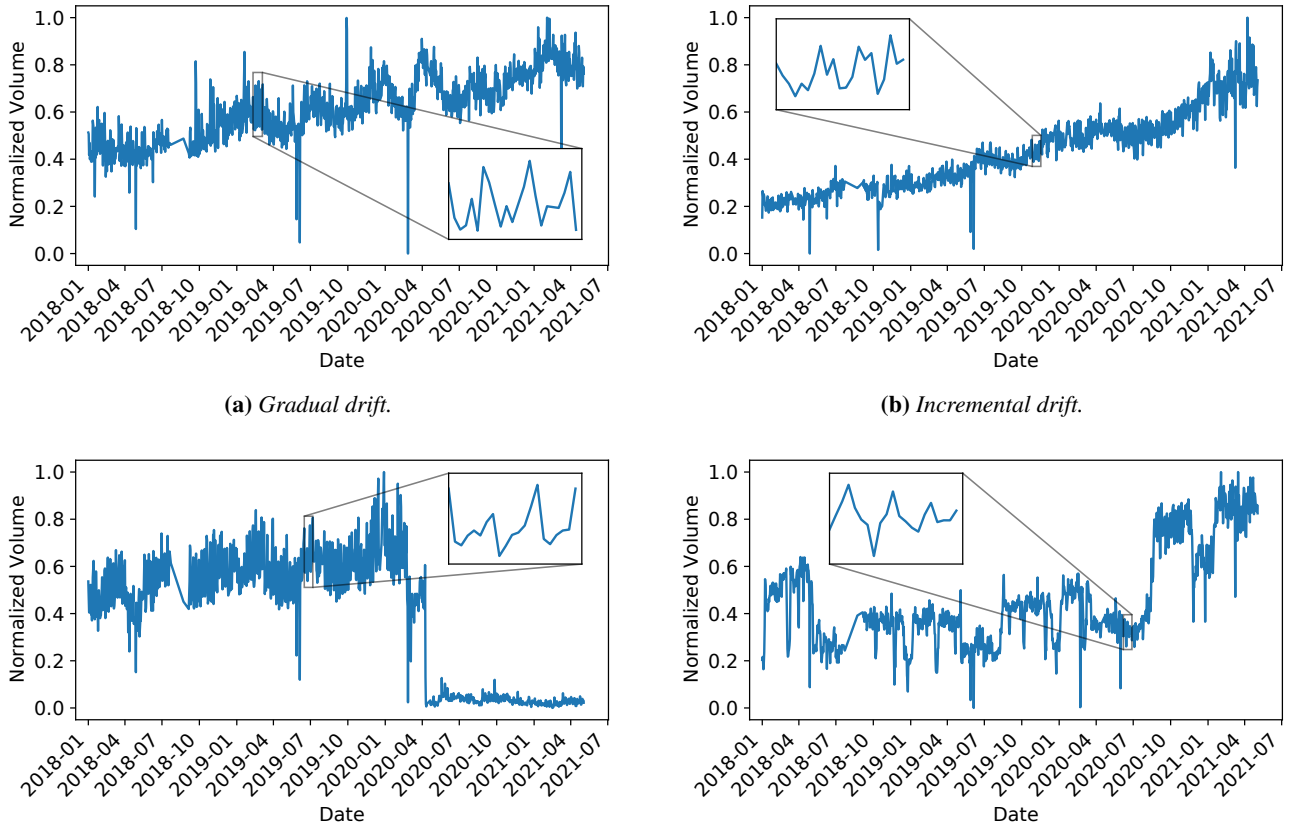
In contrast, previous research into concept drift adaptation has primarily focused on how to minimize the reduction in accuracy by recovering using new data samples. Although this approach can help in some cases where the sudden changes are permanent (*e.g.*, network updates), it can be counter-effective when applied to temporary and unpredictable shifts that cause one-off, ephemeral deviations in the data distribution (*e.g.*, trends shown during the pandemic lockdowns).

**Recurring context.** Recurring context differs from previous categories as it does not refer to a permanent or persistent change in the dataset distribution. In contrast, recurring context identifies changes that occur periodically in the data. Such changes are strongly present in networking data as multiple factors can affect network performance and behavior. For example, users' behavior and mobility patterns vary during the days of the week (weekdays vs weekend). Weather can also cyclically affect network performance; for example, rain and foliage can negatively affect the radio environment, and thus the transmission rates of a base station.

Figure 4d demonstrates the presence of recurring context of one eNodeB in our dataset. From the figure, we observe that summer months (*i.e.*, May to September) periodically show up to 50% lower data volumes for the base station, possibly a product of a lesser users' presence and therefore usage of this particular eNodeB. These recurring patterns are heavily affected in 2020, a byproduct of the pandemic that led to lower volume demand in March, *i.e.*, the moment when stay-home regulations started. Weekly pattern exists during stable months such as February 2020.

In the literature (Gama et al., 2014; Lu et al., 2018), recurring context has normally been tackled by the use of historical data matching the current data distribution. This approach, however, can introduce new challenges for networking data, which typically experiences multiple types of drift at once, such as the contemporaneous sudden and recurring context drifts from the previous example.

[2]Note that our examples of drift in operational networks may not perfectly resemble textbook definitions of data drift, as real-world network KPIs are inherently noisy.




**(a)** *Gradual drift.*

**(b)** *Incremental drift.*

**(c)** *Sudden drift.*

**(d)** *Recurring context.*

*Figure 4.* Data drift taxonomy examples demonstrated by normalized downlink volume. Inset figures exhibit a 3-week view (all starting from Sunday) of normalized downlink volume for the box-selected period. Note that gradual and incremental drifts are different because incremental drift contains more intermediate distributions.

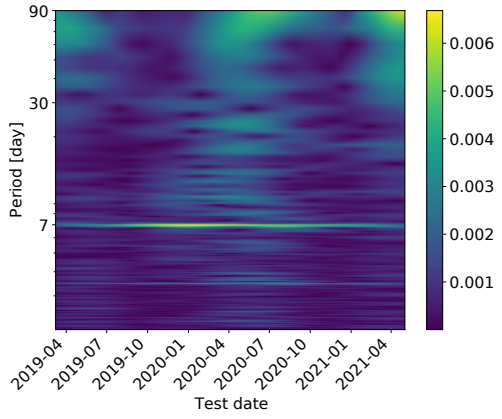



*Figure 5.* STFT Periodigram, whose y-axis is in log scale.

## B DRIFT COMPONENTS BASED ON FREQUENCY DECOMPOSITION

Given the temporal nature of the target KPIs in the network, we next look to apply signal processing techniques to further investigate drift components. To formalize these patterns and identify the sources, we perform frequency-domain analysis on these signals. If we view the NRMSE time series output as a short-time signal that has a sampling rate of one day, then we can apply a Short Time Fourier Transform (STFT) to the time series. STFT can assist us in discovering major frequency components and then we can use a bandpass filter to yield a filtered signal with a specified frequency.

We take the output of XGBoost model to analyze. Since this is a signal with relatively short duration—containing 765 date-dependant NRMSE scores, we use a sliding window of 365 days that slides one day for each sample to compute the Fourier Transform separately. Figure 5 shows the STFT of NRMSE time series output from XGBoost on the downlink volume KPI. We plot the y-axis in log-scale in order to show lower frequency values more clearly. On the x-axis, each test date represent the Fourier Transform executed on this date and the NRMSE in the next whole year, *e.g.*, July 1st, 2019 actually contains all NRMSE from that date to July 1, 2020. Unsurprisingly, as shown in the figure, there is

a very strong component on the period of 7 days in the STFT, illustrating a consistent weekly pattern. Note that there's also an apparent pattern around 3.5, which is due to the harmonic effect of the Fourier Transform. We also observe seasonal effects on the components, and we see a periodicity shift around 90 days. However, from July 2019 to April 2020 and October 2020 to January 2021, appears on the 90-day components. During COVID-19 lockdown, components of many different periodicity appear vertically in Figure 5. We expect that these changes we attributable to a fundamental shift in signals during COVID-19, manifesting in abnormal drift.