Lossy Compression of Cellular Network KPIs

Andrea Pimpinella, Fabio Palmese, Alessandro E. C. Redondi

***Abstract*—Network Key Performance Indicators (KPIs) are a fundamental component of mobile cellular network monitoring and optimization. Their massive volume, resulting from fine-grained measurements collected across many cells over long time horizons, poses significant challenges for storage, transport, and large-scale analysis. In this letter, we show that common cellular KPIs can be efficiently compressed using standard lossy compression schemes based on prediction, quantization, and entropy coding, achieving substantial reductions in reporting overhead. Focusing on traffic volume KPIs, we first characterize their intrinsic compressibility through a rate–distortion analysis, showing that signal-to-noise ratios around 30 dB can be achieved using only 3–4 bits per sample, corresponding to an 8–10× reduction with respect to 32-bit floating-point representations. We then assess the impact of KPI compression on representative downstream analytics tasks. Our results show that aggregation across cells mitigates quantization errors and that prediction accuracy is unaffected beyond a moderate reporting rate. These findings indicate that KPI compression is feasible and transparent to network-level analytics in cellular systems.**

***Index Terms*—Network KPIs, Lossy compression, Network analytics**

I. Introduction

Mobile cellular networks are continuously monitored through a large set of Key Performance Indicators (KPIs), collected from the Radio Access Network (RAN) and the core network. These KPIs are typically represented as floating-point time series and serve as the primary input to operational analytics such as forecasting, anomaly detection, and capacity planning. In modern 4G and 5G networks, KPIs are collected across thousands of cells at fine temporal granularity and over long time horizons, leading to rapidly increasing data volumes and significant challenges for storage, transport, and large-scale analysis.

Despite their scale, KPI time series are often handled using lossless or highly conservative lossy compression, driven by the operational criticality of KPIs and the lack of clear guidelines linking compression to downstream analytics reliability. Yet KPIs are typically consumed through aggregated statistics or predictive models rather than at the raw sample level, suggesting that high pointwise reconstruction fidelity may not be a strict requirement.

In parallel, lossy compression of floating-point time series has been widely studied, albeit in different application domains. Prediction–quantization–entropy coding frameworks represent a dominant approach, as exemplified by LFZip [1], which achieves state-of-the-art rate–distortion performance under explicit error bounds. More recent works, such as Machete [2], focus on time-series databases. While these methods share the same prediction-quantization-entropy coding paradigm, they rely on sample-domain predictors and are designed and evaluated on generic sensor or system-monitoring data, with optimization targets — such as pointwise error bounds and decompression speed — that differ from the rate–distortion and task-centric perspective adopted in this work.

This domain gap is further highlighted by recent surveys [3], which show that existing benchmarks rely on generic datasets and focus on reconstruction fidelity or system efficiency. Network KPIs and mobile cellular workloads are largely absent, and the impact of compression on downstream analytics remains underexplored.

This limitation is particularly relevant for network KPIs, which are inherently aggregated across space and time and used in analytics tasks such as traffic forecasting and capacity planning. However, the relationship between compression rate, aggregation, and downstream performance has not been systematically studied. Existing approaches, such as the Quality of Monitoring (QoM) framework [4], focus on controlling reconstruction error but do not explicitly evaluate the impact on aggregated KPIs or forecasting tasks.

In this paper, we adopt a task-centric perspective and study lossy compression of network KPIs through rate–distortion analysis. Rather than focusing solely on pointwise reconstruction error, we quantify how bitrate reduction affects downstream analytics, including spatial aggregation across cells and short-term traffic forecasting. We consider three representative KPIs—downlink traffic volume, physical resource block (PRB) occupancy, and number of active users—and provide a unified view of compression efficiency and analytical performance.

II. Methodology

A. Dataset

Our analysis is based on operational measurements collected from 2820 LTE (4G) cells deployed in a mid-size European city. The dataset spans one month (October 2023) and consists of hourly sampled time series. For each cell, we consider three widely used radio access network KPIs: downlink traffic volume, physical resource block (PRB) occupancy, and number of active users. These KPIs capture complementary aspects of network behavior related to traffic demand, resource utilization, and user activity, and are routinely monitored in operational cellular networks.

Andrea Pimpinella (andrea.pimpinella@unibg.it) is with the Department of Management, Information and Production Engineering (DIGIP), University of Bergamo, Bergamo, Italy. F. Palmese (fabio.palmese@polimi.it) and A. E. C. Redondi (alessandroenrico.redondi@polimi.it), are with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy. This study was carried out within the PRIN project COMPACT and received funding from Next Generation EU, Mission 4 Component 1, CUP: D53D23001340006.

B. KPI Time-Series Model

Let $\mathcal{C}$ denote the set of cells and consider a given KPI. For each cell $c \in \mathcal{C}$, we denote by $x_c[n]$ the corresponding discrete-time KPI time series, sampled at uniform hourly intervals. Time series containing missing samples (NaNs) are discarded before analysis. No missing values are present for the VOL and PRB KPIs. For the RRC KPI, 41 out of 2820 cells (1.5%) are discarded due to missing samples. Given the small fraction of removed cells, this preprocessing step is not expected to introduce significant bias.

C. Quantization Model

All compression schemes considered in this work rely on uniform scalar quantization. We denote by $Q_\Delta(\cdot)$ a uniform scalar quantizer with step size Δ, defined as

$$Q_\Delta(u) = \Delta \cdot \mathrm{round}\left(\frac{u}{\Delta}\right). \tag{1}$$

In our experiments, Δ is swept over a predefined grid of values, shared across all compression schemes, in order to trace their rate–distortion characteristics under consistent quantization conditions.

D. Prediction and Transform-Domain Representations

To characterize the intrinsic rate–distortion behavior of network KPIs, we consider a set of classical sample-domain and transform-domain compression schemes, applied independently to each cell time series.

- *Pulse-code modulation (PCM)*: PCM serves as a baseline and applies no prediction or transform. Quantization is performed directly in the sample domain:
$$\hat{x}_c[n] = Q_\Delta(x_c[n]). \tag{2}$$
- *Differential PCM (DPCM):* DPCM exploits temporal correlation through first-order prediction in a closed-loop configuration. The first sample is directly quantized to initialize the reconstruction, i.e.,
$$\hat{x}_c[0] = Q_\Delta(x_c[0]).$$
For subsequent samples, the prediction residual is computed with respect to the previously reconstructed value and uniformly quantized:
$$\hat{x}_c[n] = \hat{x}_c[n-1] + Q_\Delta\big(x_c[n] - \hat{x}_c[n-1]\big), \quad n \geq 1.$$
The predictor operates sequentially over the entire time series without resetting. As a consequence, quantization errors may accumulate over time, which is inherent to DPCM and is fully reflected in the reported rate–distortion performance. This corresponds to the standard closed-loop implementation of DPCM.
- *Block discrete cosine transform (DCT)*: DCT is applied to weekly KPI vectors. Specifically, non-overlapping blocks of 168 samples (corresponding to one week of hourly measurements) are transformed. The DCT is a fixed, closed-form orthonormal transform that does not depend on the data distribution. Let $\mathbf{D} \in \mathbb{R}^{168\times168}$ denote the DCT matrix. Transform coefficients are quantized independently and reconstruction is obtained via the inverse transform:
$$\hat{\mathbf{x}}_c = \mathbf{D}^\top Q_\Delta(\mathbf{D}\mathbf{x}_c). \tag{3}$$
- *Karhunen–Loève transform (KLT)*: we finally consider the KLT, which optimally decorrelates the KPI samples [5]. Unlike the DCT, the KLT is data-driven and must be estimated from a set of training time series. Specifically, the transform is derived from the empirical covariance matrix
$$\mathbf{R} = \mathbf{X}_{\mathrm{tr}}\mathbf{X}_{\mathrm{tr}}^\top, \tag{4}$$
where $\mathbf{X}_{\mathrm{tr}}$ collects weekly KPI vectors from a subset of cells randomly selected from C. In our experiments, this subset corresponds to the first week of data from 20% of the available cells, sampled uniformly at random. Such data is not used in subsequent analysis to prevent temporal leakage. Let $\mathbf{R} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ denote its eigendecomposition. Compression is performed by quantizing the decorrelated coefficients:
$$\hat{\mathbf{x}}_c = \mathbf{V}Q_\Delta\big(\mathbf{V}^\top\mathbf{x}_c\big). \tag{5}$$

E. Rate–Distortion Metrics

Reconstruction error is measured over all cells and time samples using the mean squared error (MSE) and reported in terms of signal-to-noise ratio (SNR), defined as

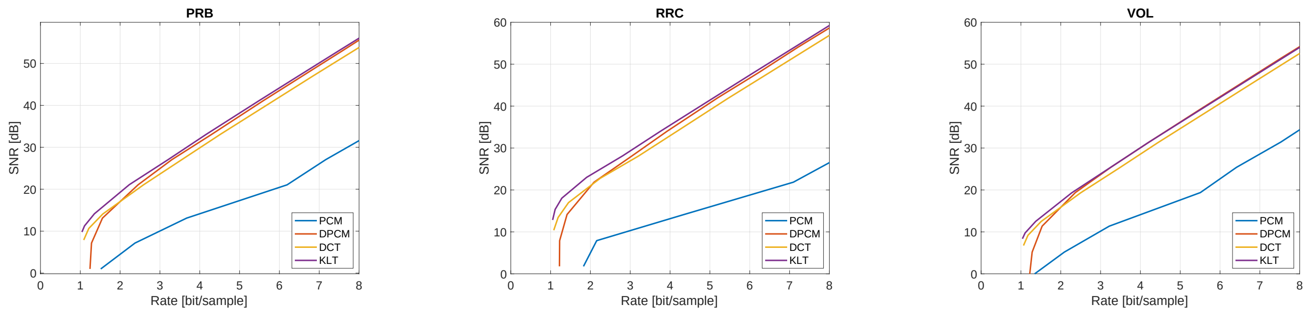
$$\mathrm{SNR} = 10\log_{10}\left(\frac{\mathrm{Var}(x_c[n])}{\mathbb{E}[(x_c[n] - \hat{x}_c[n])^2]}\right), \tag{6}$$

where variance and expectation are computed over all cells $c \in \mathcal{C}$ and time indices n.

To estimate the coding rate, we adopt a practical entropy-coding model based on Huffman coding. For each compression scheme and quantization step, the quantization indices obtained on a training subset of the data (the same used for KLT estimation) are used to estimate their empirical probability mass function and to build the corresponding Huffman dictionary. The bitrate is then computed from the average codeword length induced by this dictionary on the encoded symbols, and is reported in bits per sample. By sweeping Δ, this procedure yields the rate–distortion characteristics of each representation under consistent coding conditions.

III. Rate–Distortion Analysis

Fig. 1 reports the rate–distortion characteristics of the considered compression schemes for the three KPIs under study: PRB occupancy, number of active users (RRC), and downlink traffic volume. Across all KPIs, transform-based schemes substantially outperform sample-domain methods, confirming strong temporal correlation. PCM shows the lowest efficiency, while DPCM provides moderate gains by exploiting short-term dependencies. DCT and KLT achieve significantly higher performance, with KLT consistently attaining the highest SNR at a given bitrate. In the practically relevant range of 2–4 bit/sample, transform-based schemes achieve approximately 20–30 dB SNR, corresponding to compression factors of about



(a) PRB occupancy (b) Active users (RRC) (c) Downlink traffic volume

Fig. 1. Rate–distortion characteristics for the three considered KPIs. SNR is reported as a function of the rate [bit/sample] for PCM, DPCM, DCT, and KLT.

TABLE I
PER-CELL SNR STATISTICS AT REPRESENTATIVE KLT OPERATING POINTS (CLOSEST TO ∼30 DB GLOBAL SNR).

KPI	Average Rate [bit/sample]	Global [dB]	Median [dB]	P10 [dB]	P90 [dB]	Worst [dB]
PRB	3.21	27.06	22.63	19.00	24.98	8.03
VOL	4.19	31.28	27.10	23.87	29.63	16.32
RRC	2.81	28.13	22.30	18.29	24.80	11.91

8×–16×. The gap between KLT and DCT remains limited, suggesting that closed-form transforms may offer a favorable complexity–performance trade-off.

While the rate–distortion curves in Fig. 1 report aggregate SNR values across all cells and time samples, they may hide variability across individual cells. To better characterize this aspect, we report in Table I the distribution of per-cell SNR values at representative operating points. The results show that the median per-cell SNR is consistently lower than the global SNR, with non-negligible variability across cells. In particular, lower percentiles and worst-case values indicate that a subset of cells experiences reduced reconstruction quality, although the typical performance remains satisfactory. Such cells, which are typically characterized by irregular or non-periodic traffic patterns, may benefit from rate-adaptive compression strategies, where the bitrate is tuned per cell based on local signal statistics.

IV. Impact on Downstream Analytics

In this section, we evaluate the impact of transform-based compression on representative downstream analytics tasks commonly employed in cellular network monitoring and optimization.

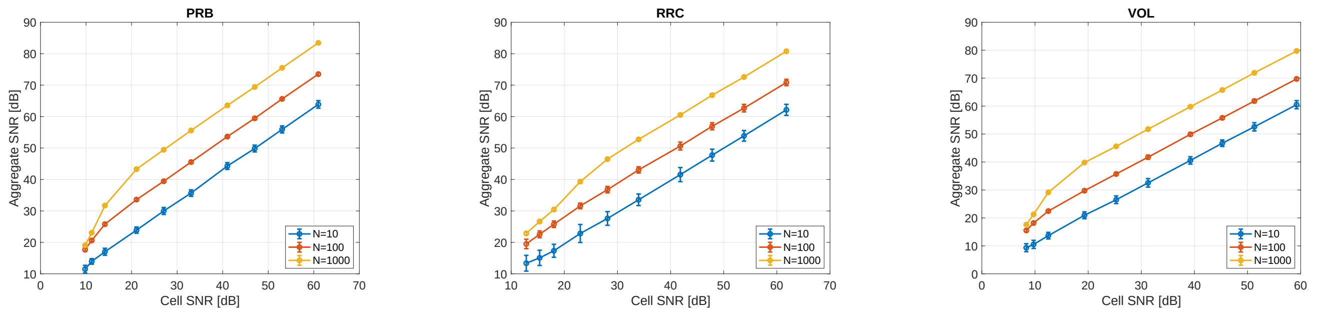
A. Aggregated KPI Accuracy

We first assess the impact of compression on aggregated KPIs that are directly relevant at the core network level. Let $\mathbf{x}_c \in \mathbb{R}^T$ and $\hat{\mathbf{x}}_c$ denote the original and reconstructed volume KPI time series for cell $c \in \mathcal{C}$, respectively. In our case study, aggregation is performed by summation across cells, which corresponds to the total traffic observed at the core network and represents a key quantity for network monitoring and capacity planning. Formally, the aggregated traffic obtained from the original KPIs is given by $\mathbf{s} = \sum_{c\in\mathcal{C}} \mathbf{x}_c$, while aggregation of the reconstructed KPIs yields $\hat{\mathbf{s}} = \sum_{c\in\mathcal{C}} \hat{\mathbf{x}}_c$.

We quantify the effect of compression by measuring the signal-to-noise ratio (SNR) between s and $\hat{s}$. Fig. 2 reports the aggregate SNR as a function of the average per-cell SNR for different numbers of aggregated cells ($N = 10, 100, 1000$). For each value of N, we repeat the aggregation experiment over 50 randomly selected groups of cells and report the average performance together with the corresponding standard deviation. The standard deviations are small across all operating points, with the largest variability observed for small N; for $N = 10$, standard deviations are at most a few dB, and become negligible as N increases. The results show that aggregation yields a substantial SNR gain with respect to the cell-level reconstruction quality. In particular, when aggregating $N = 1000$ cells, an aggregate SNR of approximately 30 dB is achieved even when the average per-cell SNR is only about 15 dB. This corresponds to operating points where transform-based compression schemes such as KLT or DCT require only 1–2 bit/sample to achieve the necessary cell-level reconstruction quality. As N increases, the aggregate SNR improves further. This behavior can be understood analytically: quantization errors accumulate incoherently across cells, so the aggregate error power grows as N, while the dominant traffic components—which share common daily and weekly periodicity—combine coherently, so the aggregate signal power grows approximately as N^2. The aggregate SNR therefore grows linearly with N. As a result, compression operating points that appear moderate at the cell level translate into high-fidelity representations of aggregated traffic at the core network.

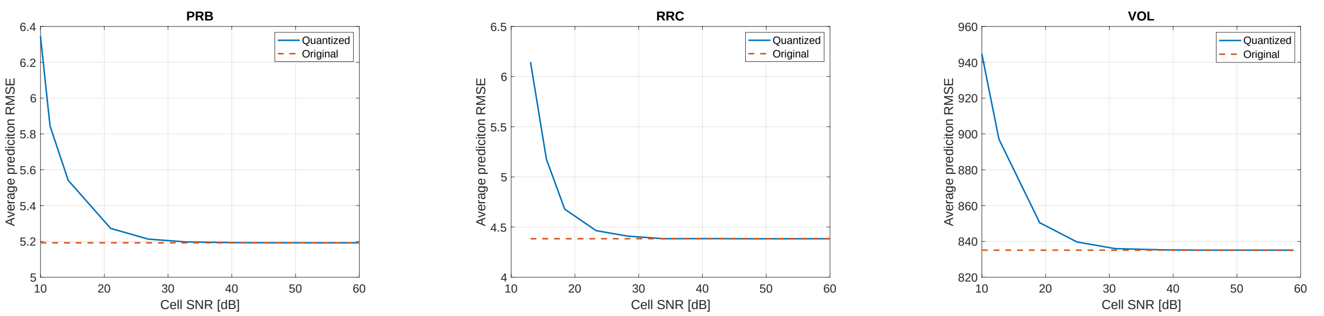
B. Forecasting Performance

We next evaluate the impact of compression on a downstream forecasting task. For this experiment, compressed KPIs are obtained exclusively using KLT-based transform coding, which provides the best rate–distortion performance among the considered schemes. Specifically, we consider the Median Weekly Signature (MWS) predictor [6], which computes the sample-wise median over three consecutive weeks of historical data and uses it to forecast the fourth week, independently for each cell and KPI. Forecasting performance is evaluated by computing the root mean square error (RMSE) for each cell



(a) PRB occupancy (b) Active users (RRC) (c) Downlink traffic volume

Fig. 2. Aggregate SNR vs per-cell SNR for three representative KPIs: (a) PRB, (b) RRC, and (c) VOL. Results are shown for different numbers of aggregated cells. Aggregation by summation yields a substantial SNR gain that increases with N, reflecting the attenuation of incoherent quantization noise across cells.



(a) PRB occupancy (b) Active users (RRC) (c) Downlink traffic volume

Fig. 3. Forecasting RMSE as a function of per-cell SNR for (a) PRB, (b) RRC, and (c) VOL, reported in KPI-native units (percentage, number of users, and MB/hour, respectively). Results are shown for original and KLT-quantized data.

and KPI and averaging the results across all cells. Forecasts are generated starting from either the original KPI time series $\mathbf{x}_c$ or their compressed and reconstructed counterparts $\hat{\mathbf{x}}_c$. Prediction accuracy is evaluated as a function of the per-cell reconstruction quality, measured in terms of SNR. Fig. 3 reports the forecasting RMSE as a function of the cell-level SNR for the three KPIs, comparing original and KLT-quantized data. Fig. 3 shows that, for per-cell SNR values above approximately 30 dB, the forecasting RMSE obtained from KLT-quantized KPIs closely matches that obtained from the original data. In this regime, no measurable degradation in forecasting accuracy is observed. While the MWS predictor is inherently robust to noise due to its median-based formulation, the observed robustness is primarily driven by the operating regime: at 30 dB per-cell SNR, quantization errors are small relative to the signal and do not alter the dominant weekly patterns, indicating that preservation of low-frequency structures is the key factor underlying forecasting performance.

V. Conclusions

This paper investigated lossy compression of cellular network KPIs from a rate–distortion and task-centric perspective. Our results show that operating points around 30 dB per-cell SNR can be achieved at 3–4 bit/sample using transform-based schemes, corresponding to an 8×–10× reduction with respect to 32-bit representations. Aggregation across cells enables reliable representations at even lower rates (1–2 bit/sample), as quantization errors are attenuated through incoherent accumulation. Forecasting performance is largely unaffected at moderate compression levels, indicating that the dominant temporal structures of KPI time series are preserved. The compression schemes considered are general-purpose and do not rely on dataset-specific assumptions, although validation across diverse deployments and network generations including 5G remains an open question. This will be explored in future work, together with compression strategies that exploit spatial redundancy across cells and learning-based approaches.

References

[1] S. Chandak, K. Tatwawadi, C. Wen, L. Wang, J. A. Ojea, and T. Weissman, "Lfzip: Lossy compression of multivariate floating-point time series data via improved prediction," in *2020 Data Compression Conference (DCC)*. IEEE, 2020, pp. 342–351.

[2] Y. Shi, X. Zou, X. Chen, S. Jin, D. Tao, C. Deng, Y. Chen, and W. Xia, "Machete: An efficient lossy floating-point compressor designed for time series databases," in *2024 Data Compression Conference (DCC)*. IEEE, 2024, pp. 532–541.

[3] G. Chiarot and C. Silvestri, "Time series compression survey," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–32, 2023.

[4] N. H. Motlagh, S. Kapoor, R. Alhalaseh, S. Tarkoma, and K. Hätönen, "Quality of monitoring for cellular networks," *IEEE Transactions on Network and Service Management*, vol. 19, no. 1, pp. 381–391, 2021.

[5] R. Dony *et al.*, "Karhunen-loeve transform," *The transform and data compression handbook*, vol. 1, no. 1-34, p. 29, 2001.

[6] A. Furno, M. Fiore, R. Stanica, C. Ziemlicki, and Z. Smoreda, "A tale of ten cities: Characterizing signatures of mobile traffic in urban areas," *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2682–2696, 2016.