Synthetic Fine-Grained Traffic Generation

Engineering Adaptation of the KTH Framework for Sparse Datasets

Ahcene LOUBAR & Maryam BACHTTI

Laboratoire des Signaux et Systèmes (L2S) – CentraleSupélec

February 19, 2026

Abstract

This report documents the implementation of a traffic augmentation pipeline designed to synthesize 1-second resolution traffic loads from 5-minute aggregate measurements. Based on the Interrupted Poisson Process (IPP) framework from KTH Section II-C, we address a critical data-sparsity constraint: the original framework requires multi-day variance statistics which are unavailable in the Bouygues dataset. We propose a *Hybrid Calibration Strategy* using analytical mean preservation (in expectation) and grid-search burstiness tuning. Validation shows high fidelity to coarse envelopes and realistic micro-burst dynamics.

1 INTRODUCTION AND OBJECTIVES

Modern network forecasting models, such as LSTMs or Chronos, require high-resolution data to capture congestion risks. However, industrial datasets often provide only coarse averages (e.g., 5-minute slots). Our objective is to generate synthetic 1-second or 5-second traces that:

1. **Preserve the Mean:** The average of the synthetic signal must equal the observed coarse value.
2. **Capture Burstiness:** The signal must reproduce the ON/OFF stochastic behavior of real-world LTE traffic.

2 THEORETICAL FRAMEWORK: THE KTH IPP MODEL

Following Section II-C of the KTH reference paper, we model traffic as a marked Interrupted Poisson Process (IPP).

2.1 CTMC Activity Model

The system activity is modeled as a two-state Continuous-Time Markov Chain (CTMC):

$$\text{OFF} \xrightarrow{\tau} \text{ON}, \quad \text{ON} \xrightarrow{\zeta} \text{OFF}$$

The stationary probability of the system being in the ON state (active) is given by:

$$P_{\text{ON}} = \frac{\tau}{\tau + \zeta} \tag{1}$$

2.2 Arrivals and Packet Sizes

Conditioned on the ON state, arrivals follow a Poisson process with intensity λ (arrivals/sec). Each arrival carries a mark ψ (packet size in Mbits) drawn from an exponential distribution with mean $E[\psi]$.

The total traffic volume Ψ in a slot of duration T is:

$$E[\Psi] = (\lambda P_{\text{ON}} T) \cdot E[\psi] \tag{2}$$

3 ENGINEERING ADAPTATION: HYBRID CALIBRATION

3.1 The Variance Estimation Problem

The original KTH framework solves for parameters using **Moment Matching** (Eq. 11-12 in the paper), requiring the variance of the traffic volume over the same slot across different days ($Var(O_i)$).
Constraint: The Bouygues dataset is a single time-series; we cannot calculate variance per time-index. Consequently, the second-moment analytical equations are underdetermined.

3.2 Stable Analytical Mean Calibration

To ensure the synthetic data respects the coarse observed rate y_t (in Mbps), we derive a slot-specific $E[\psi_t]$. Setting the synthetic average rate to equal y_t:

$$\frac{E[\Psi_t]}{T} = y_t \implies \frac{\lambda P_{\mathrm{ON}} T E[\psi_t]}{T} = y_t$$

Solving for the mean mark size:

$$E[\psi_t] = \frac{y_t}{\lambda P_{\mathrm{ON}}} \tag{3}$$

This derivation is the core of our `kth_ipp.py` implementation, guaranteeing trend preservation regardless of the burstiness parameters.

3.3 Grid-Search for Burstiness Hyperparameters (Implemented in `test.ipynb`)

Because the Bouygues dataset provides a single time series per sector, the KTH moment-matching system (which relies on per-slot multi-day variance) cannot be solved. We therefore treat (τ, ζ, λ) as *hyperparameters* and tune them via a grid-search on a small set of representative sectors.

Representative sectors. We select three sectors whose mean traffic lies approximately at the 10%, 50% and 90% quantiles across sectors. This provides a low/medium/high load coverage while keeping the tuning computationally light.

Candidate grid. We search:

$$\tau \in \left\{ \frac{1}{120}, \frac{1}{60}, \frac{1}{30}, \frac{1}{15} \right\}, \quad \zeta \in \left\{ \frac{1}{60}, \frac{1}{30}, \frac{1}{15}, \frac{1}{8} \right\}, \quad \lambda \in \{0.1, 0.3, 0.5, 1.0, 2.0\}.$$

Evaluation metrics. For each candidate (τ, ζ, λ) and each selected sector, we generate a fine-grained series and compute: (i) reconstruction MAPE between the original coarse series and the re-aggregated fine series, (ii) a coefficient-of-variation proxy (median of std/mean over slots), (iii) median fraction of zero bins (silence periods), (iv) a spike ratio proxy (max/mean within each slot).

Scoring rule. We minimize a composite score dominated by reconstruction MAPE, with additional penalties if burstiness proxies fall outside target ranges. This implements a constraint-like tuning: keep coarse fidelity high while encouraging plausible micro-burst behavior.

Selected parameters. The best candidate under this score was:

$$\tau = \frac{1}{15}, \quad \zeta = \frac{1}{60}, \quad \lambda = 2.0,$$

with median metrics across the three sectors approximately: MAPE ≈ 0.127, CV ≈ 0.655, zero_frac ≈ 0.133, spike_ratio ≈ 2.67.

4 IMPLEMENTATION ARCHITECTURE

We implement an **event-driven** simulation loop. This is more accurate than discrete-time steps as it samples exact exponential holding times for the ON/OFF states.

Algorithm 1: Event-Driven IPP Generation (`kth_ipp.py`)

Input: Coarse rate y_t, Resolution dt, Params $(\tau, \zeta, \lambda, T)$
Output: Fine-grained rate series $\{x_{t,k}\}$
$P_{\text{ON}} \leftarrow \tau/(\tau+\zeta)$
$E[\psi_t] \leftarrow y_t/(\lambda P_{\text{ON}})$ `// Eq. 3 calibration`
Initialize bins B of size T/dt
Sample initial state $\in$ {ON, OFF} from P_{ON}
while $t_{current} < T$ **do**
    Sample duration $D \sim \text{Exp(rate)}$
    $t_{\text{end}} \leftarrow \min(t_{\text{current}} + D, T)$
    if *State is ON* **then**
        Sample $N \sim \text{Poisson}(\lambda \cdot (t_{\text{end}} - t_{\text{current}}))$
        Distribute N arrivals uniformly in $[t_{\text{current}}, t_{\text{end}}]$
        Assign sizes $\psi_j \sim \text{Exp}(E[\psi_t])$
        Add sizes to bins: $B[\lfloor \text{arrival\_time}/dt \rfloor] \leftarrow \sum \psi_j$
    $t_{\text{current}} \leftarrow t_{\text{end}}$; Flip state
return $x \leftarrow B/dt$

5 VALIDATION AND SANITY CHECKS

5.1 Coarse-to-Fine-to-Coarse Consistency

We validate the pipeline by re-aggregating the fine-grained rate series back to 5-minute resolution and comparing it to the original coarse series. Because the mean mark size $E[\psi_t]$ is calibrated analytically, the coarse fidelity is enforced *in expectation.* In practice, due to stochastic sampling and discretization, we observe a non-zero reconstruction error.

Under the selected hyperparameters $(\tau, \zeta, \lambda) = (1/15, 1/60, 2.0)$ with $dt = 5\text{s}$, the median reconstruction MAPE on the representative sectors is approximately 0.127 (i.e., about 12.7%).

Algorithm 2: Grid-Search Burstiness Tuning (`test.ipynb`)

foreach (τ, ζ, λ) *in grid* **do**
    foreach *sector in {low, median, high}* **do**
        Generate fine series using Algorithm 1
        Compute metrics: MAPE, CV, zero_frac, spike_ratio
    Aggregate sector metrics by median
    Compute score = $50 \cdot$ MAPE + range penalties
Select parameters minimizing score

5.2 Visual Fidelity

As shown in our experimental plots (Ref: Fig. 6(a) replication), the synthetic signal transforms the "staircase" coarse measurements into a series of high-frequency pulses. These pulses correctly align with the coarse envelope while introducing the stochastic silence periods characteristic of real-world IPP traffic.

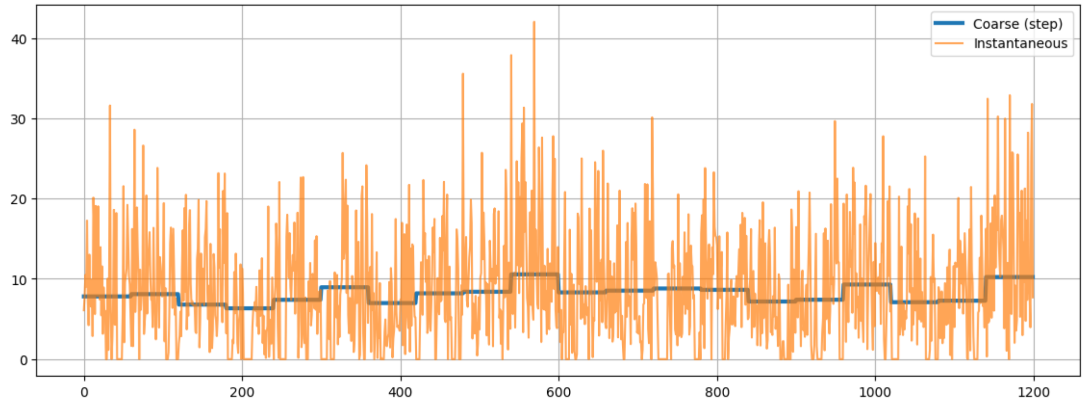

Figure 1: Instantaneous IPP signal versus coarse staircase signal.

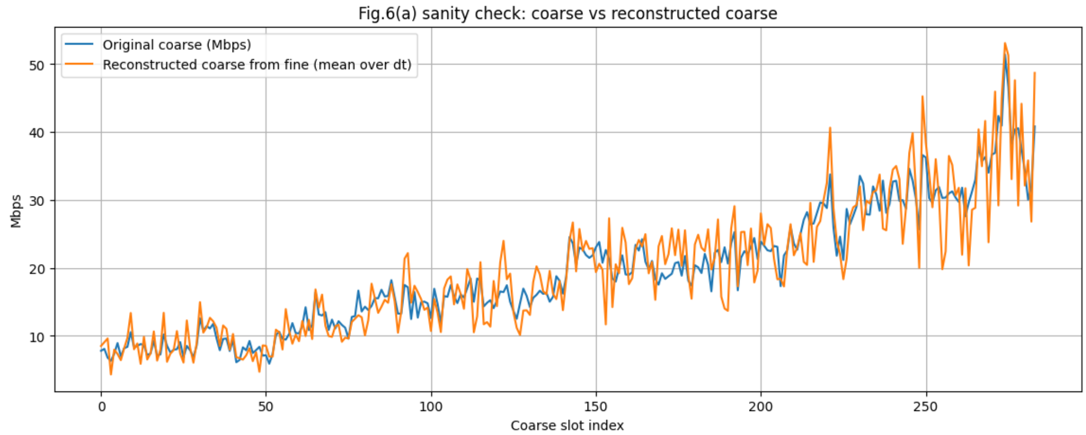

Figure 2: Reconstructed coarse signal from fine-grained data.

6 CONCLUSION

The developed pipeline successfully adapts the KTH IPP framework for the sparse Bouygues dataset. By separating the calibration into an analytical mean-matching step and an empirical burstiness tuning step, we created a robust generator that satisfies both physical constraints (mean preservation) and statistical realism (bursty behavior).