Useful Data, Not Just More Data: Cost-Aware Floating Car Data Generation in Large-Scale Traffic Simulation

Paulo Silva, Pavlína Smolková, Kateřina Slaninová, Jan Martinovič
IT4Innovations, VSB – Technical University of Ostrava
Ostrava, Czech Republic
Email (corresponding author): paulo.silva@vsb.cz

I. Introduction

Large-scale traffic simulations generate Floating Car Data (FCD): time-dependent vehicle positions, speeds, and movements used in mobility analysis [1]. More records do not necessarily provide more useful information, but they require storage and processing. Using Ruth, a deterministic mesoscopic traffic simulator [2], we examine how to retain useful traffic information without unnecessary output. The findings can inform workflows such as Ruth's LEXIS-based interface [3].

II. Methodology

Experiments ran on LUMI-C using Chicago 2019 AM/PM demand from the Chicago Metropolitan Agency for Planning (CMAP) c24q4 2019 Base dataset [4]. The input contains 6,897,778 vehicle trips, not simultaneous vehicles, on a graph with 748,509 nodes and 1,867,225 directed edges. Comparable runs reuse the same demand and cached map.

F and S denote the configured vehicle-update and FCD-sampling periods in seconds. F20/S5, F20/S10, and F20/S20 keep F fixed; the remaining configurations in Table I change both periods. Unlike varying S alone, changing F or rerouting participation can alter simulated traffic.

Comparisons used 4 nodes, 8 tasks per node, 16 CPUs per task, 8-hour Slurm limits, and asynchronous output in Hierarchical Data Format version 5 (HDF5) [5]. Scaling tests used 1–16 nodes with 1-hour limits, from cold starts and saved checkpoints. Following earlier routing work [6], we compared 50% and 80% rerouting participation for F20/S5, F60/S60, and F120/S60. Participation denotes the configured share using Ruth's route-alternative mechanism.

III. Metrics

Usefulness is assessed against the F20/S5 simulation, not independently measured traffic speeds. Coverage is

$$\text{Coverage} = \frac{\text{reference elements also observed}}{\text{all reference elements}} \times 100. \quad (1)$$

For road-edge coverage, elements are directed road segments; for temporal coverage, they are road segments in particular 60-second intervals. Mean absolute error (MAE) and Pearson correlation compare mean speeds on intervals observed in both outputs. Missing intervals reduce coverage but are excluded from these speed statistics.

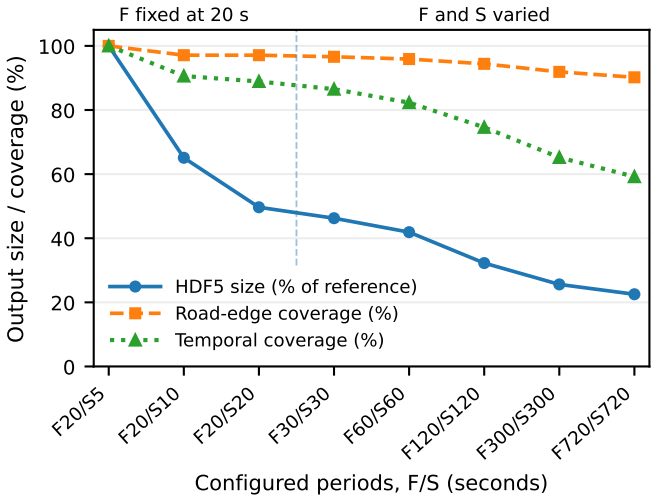


Fig. 1. Output size and coverage relative to F20/S5. The divider separates fixed-F20 sampling tests from runs changing both periods. Record counts and speed statistics remain in Table I.

Measurements include output size, record count, memory, and node-level CPU-busy percentages in scaling runs. FCD temporal span is the interval between the earliest and latest saved FCD timestamps, an output-based progress measure rather than exact simulator progress. Wall s/sim s divides wall-clock seconds by this span in seconds; lower is better. Runs can cover different simulated intervals.

IV. Results and Discussion

Figure 1 and Table I show the trade-off. F20/S20 approximately halves output while retaining 97.1% road-edge coverage, 88.9% temporal coverage, and correlation 0.897. F60/S60 reduces output further with lower temporal coverage. F720/S720 retains 90.2% road-edge coverage but only 59.2% temporal coverage: observing roads does not guarantee observing their changing traffic.

MAE and correlation need not worsen monotonically: each setting is evaluated on its shared road–time intervals, and

TABLE I
ABSOLUTE OUTPUT SIZES AND RECORD COUNTS; COVERAGE AND SPEED METRICS RELATIVE TO F20/S5.

Config.	HDF5 (GiB)	Records (10^6)	Road (%)	Temporal (%)	MAE (m/s)	Corr.
F20/S5	1.172	27.951	100.0	100.0	0.000	1.000
F20/S10	0.763	18.191	97.1	90.6	0.719	0.902
F20/S20	0.582	13.871	97.1	88.9	0.759	0.897
F30/S30	0.542	12.928	96.6	86.5	0.991	0.861
F60/S60	0.491	11.710	95.9	82.3	1.332	0.770
F120/S120	0.378	9.009	94.4	74.6	1.625	0.679
F300/S300	0.300	7.159	91.9	65.2	1.374	0.738
F720/S720	0.264	6.290	90.2	59.2	1.081	0.837

TABLE II
CHANGES WITH 50% VERSUS 80% PARTICIPATION, AT IDENTICAL F/S SETTINGS. TIME NORMALIZATION USES FCD SPAN.

Config.	FCD span change	Wall s/sim s change	HDF5 change	HDF5/sim h change
F20/S5	+22.0%	-18.0%	+48.8%	+22.0%
F60/S60	+19.6%	-16.4%	+40.0%	+17.1%
F120/S60	+18.4%	-15.5%	+21.4%	+2.5%

changing F can alter traffic. Higher correlation at coarser settings does not establish better fidelity when coverage falls; the aggregate results do not isolate the rebound's cause.

The fixed-F20 runs reach about 12.5 simulated minutes in eight hours according to simulation logs. Reduced sampling lowers output volume without materially improving runtime progress; post-processing savings were not directly measured.

Cold-start F60/S60 produces 3.50 minutes of FCD span on 2 nodes and 6.48 on 16 nodes within one hour. CPU busy decreases across these cold runs. Within one hour, cold single-node runs emit no FCD; logs show initial route computation. These results show startup cost and sublinear output-based gains, not the individual contributions of synchronization or load imbalance.

Reducing participation from 80% to 50% increases FCD span by 18.4–22.0% and lowers wall s/sim s by 15.5–18.0% (Table II). Total output rises; for F120/S60, output per hour of FCD span rises by only 2.5%. These fixed-budget results do not establish equivalent completed-trip travel times or traffic fidelity.

V. CONCLUSION

This work examined how FCD settings, rerouting participation, and node count affect data generation in large-scale traffic simulation. Selected coarser sampling settings reduced output while retaining much of the reference information, but did not noticeably improve runtime progress. Reducing rerouting participation improved output-based progress, while additional nodes provided sublinear gains. These findings show why output volume, retained information, and runtime cost need to be considered together.

For practical use, FCD settings should follow the needs of the intended analysis. Keeping vehicle updates fixed allows users to select the smallest output meeting their coverage and speed-error requirements. Within our fixed-F20 comparison, requiring at least 90% temporal coverage and MAE below 1 m/s selects F20/S10; accepting 88% permits F20/S20 and approximately halves output. These examples illustrate the selection process rather than universal defaults. Changes to rerouting participation still require validation of traffic outcomes.

ACKNOWLEDGMENT

This poster is partially based upon work from the project "Floreon WIA", VK02020253, supported by the Ministry of the Interior of the Czech Republic from the OPSEC Programme. This work was also supported by the Ministry of Education, Youth and Sports of the Czech Republic (ID: MC2102) and through e-INFRA CZ (ID:90254).

REFERENCES

[1] J. C. Herrera, D. B. Work, R. Herring, X. J. Ban, Q. Jacobson, and A. M. Bayen, "Evaluation of traffic data obtained via GPS-enabled mobile phones: The mobile century field experiment," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 4, pp. 568–583, 2010.

[2] IT4Innovations, "Ruth - traffic simulator," https://github.com/It4innovations/ruth, 2026.

[3] M. Golasowski, J. Martinovič, J. Křenek, K. Slaninová, M. Levrier, P. Harsh, M. Derquennes, F. Donnat, and O. Terzo, "The LEXIS platform for distributed workflow execution and data management," in *HPC, Big Data, and AI Convergence Towards Exascale*. CRC Press, 2022, pp. 17–35.

[4] Chicago Metropolitan Agency for Planning, "Travel Demand Model Data User Guide (c24q4)," https://datahub.cmap.illinois.gov/documents/a14d641dbe184704ab597479b103986c/explore, 2019 Base dataset: c24q4_100.zip. Accessed: 2026-07-21.

[5] M. Folk, G. Heber, Q. Koziol, E. Pourmal, and D. Robinson, "An overview of the HDF5 technology suite and its applications," in *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases*. ACM, 2011, pp. 36–47.

[6] P. Silva, P. Smolková, S. Michailidu, J. Beránek, R. Macháček, K. Slaninová, J. Martinovič, and R. Cmar, "High-performance computing for distributed route computation in traffic flow models," *Procedia Computer Science*, vol. 255, pp. 83–92, 2025, proceedings of the Second EuroHPC user day. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050925006246