NovaWireless Representative Generator: A Reproducible Synthetic Call-Center Employee Database

Design, Data Priors, and KPI-Consistent Simulation for Analytics Prototyping

Gina Aulabaugh

February 16, 2026

Abstract

Synthetic datasets are useful for testing analytics pipelines, dashboards, and modeling workflows when production data are unavailable, sensitive, or legally restricted. This paper describes the NovaWireless Representative Generator, a reproducible employee-database generator for a single-queue call center environment. The generator produces a structured roster of customer service representatives (CSRs) with unique identifiers and names, skill-tag strengths, and KPI-consistent performance attributes (e.g., first-call resolution, average handle time, escalation rate). To support realism and reproducibility, the generator uses publicly available datasets as priors for workload pressure, contact-center performance patterns, and workforce distributions. Outputs are designed to integrate directly into a synthetic call generator, enabling repeatable simulations of call outcomes and representative-level dashboards. *All records are synthetic; no real employee or customer data were used.*

Keywords: synthetic data, call centers, KPI simulation, reproducibility, workforce analytics, contact center operations

1 Introduction

Call-center analytics commonly rely on sensitive operational and customer interaction data. In many contexts, analysts need representative datasets for rapid prototyping, experimentation, and portfolio demonstration without exposing personally identifiable information (PII) or proprietary systems. Synthetic data generation provides a practical alternative: it allows the creation of internally consistent data with plausible distributions while keeping the pipeline auditable and reproducible.

The NovaWireless Representative Generator is one component of a broader synthetic environment used to simulate call-center operations. This paper focuses on the employee database: a roster of CSRs who share a single queue and job function but differ in strengths, tendencies, and measurable performance. These representative-level attributes are intentionally structured so they can be used as conditional priors in downstream call simulation.

2 Data Sources and Priors

The generator uses extracted priors derived from multiple public datasets to avoid arbitrary parameter choices. Workforce priors and performance distributions are informed by HR and employee-related datasets (Tank, n.d.; ziya07, n.d.; mexwell, n.d.). Operational pressure and contact-center performance patterns are informed by call-center and service-center datasets (Ahmadzade Jahromi, n.d.; City of San Francisco, n.d.; Metropolitan Transportation Authority, n.d.-a; Metropolitan Transportation Authority, n.d.-b). Telecom-related customer risk segmentation priors are supported by the IBM Telco churn dataset (IBM, n.d.). Issue taxonomy and specialization priors use complaint categories from the FCC CGB Consumer Complaints dataset (Federal Communications Commission, n.d.).

2.1 Why Priors Matter

A common failure mode in synthetic datasets is *independent randomization*: assigning KPIs independently produces unrealistic combinations (e.g., high quality with extremely high escalation rate) and destroys the correlation structure that makes dashboards feel authentic. Priors provide grounded distributions and enable correlated sampling so that “bad weeks cluster” and “high pressure degrades performance” in plausible ways.

3 System Design

3.1 Repository Structure and Reproducibility

The project uses a repo-root structure to ensure portability:

- `data/employee_generation_inputs/`: extracted priors (CSV/JSON)

- `src/`: scripts (generation logic)
- `output/`: generated artifacts (non-destructive, non-overwriting)

Scripts auto-detect the repository root and read from `data/employee_generation_inputs/`. Outputs are written only to `output/` with stable, non-overwriting filenames (including run identifiers derived from the seed and input file fingerprints). This design supports repeated regeneration, auditing, and versioned publication.

3.2 Employee Database Schema

The canonical output artifact (`novawireless_employee_database`) contains one record per CSR. The design assumes a single queue and identical job function for all employees; differentiation occurs through skill tags and performance tendencies.

3.3 Artifact Output and Intended Use

The generator produces `novawireless_employee_database` (CSV), which functions as the representative dimension table for downstream synthetic call generation and KPI dashboards. Each simulated call record references `rep_id` from this table, enabling rep-level aggregation, performance trend analysis, and scenario-based stress testing.

3.3.1 Identity and Organizational Fields

- **`rep_id`**: unique employee identifier (primary key for joins).
- **`first_name`, `last_name`, `rep_name`**: display fields; names are generated from ASCII-only pools to avoid encoding artifacts.
- **`site`**: tenant label (e.g., NovaWireless) for multi-tenant extension.
- **`queue_name`**: queue label (single-queue design).
- **`department`, `job_role`**: fixed to call-center CSR to enforce "same job, same queue."
- **`can_transfer_departments`**: fixed false to enforce "no department transfers." Transfers may still occur *within* the queue as call events, but not as organizational movement.
- **`tenure_months`**: experience proxy; influences variance and baseline outcomes in downstream call simulation.

3.3.2 Skill Tags as Strengths (No Routing Restriction)

All CSRs can handle all call types; skill tags represent *comparative strengths* rather than assignment constraints.

- **`primary_skill_tag`, `secondary_skill_tag`**: taxonomy-based strengths. For interpretability in dashboards, `network_service` is treated as tech support.
- **`primary_skill_label`, `secondary_skill_label`**: human-friendly labels derived from tag mappings.
- **`strengths`, `weaknesses`**: pipe-delimited narrative tags used for scenario realism and coaching dashboards.

3.3.3 KPI Fields and Latent Traits

The generator produces KPI-consistent rep profiles that function as conditional priors for call outcomes.

- **`qa_score`**: quality/procedural correctness proxy.
- **`fcr_30d`**: first-call resolution propensity.
- **`repeat_contact_rate`**: repeat-contact propensity (inverse-related to FCR).
- **`aht_secs`**: average handle time (seconds), used as a center for call duration sampling.
- **`csat_proxy`**: satisfaction proxy conditioned on resolution and interaction friction.
- **`transfer_rate`**: likelihood of transferring to another agent within the queue.
- **`escalation_rate`**: likelihood of supervisor escalation.
- **`compliance_risk`**: probability-like proxy for policy/behavioral risk.
- **`productivity_index`**: composite performance summary (balanced blend of FCR/QA/AHT proxies).

Latent traits drive variability and stress response:

- **`burnout_index`**: chronic fatigue/stress proxy.
- **`resilience_index`**: stability/recovery proxy (inverse-related to burnout).
- **`volatility_index`**: per-call variability proxy (affects noise in simulated outcomes).

- **`strain_tier`**: categorical tier (`low`, `medium`, `high`, `very_high`) derived from baseline strain and burnout.
- **`pressure_index_baseline`**: global operating pressure used for generation; downstream call simulation may replace this with time-varying pressure curves.

4 Generation Method

4.1 Uniqueness Constraints

The generator enforces uniqueness on:

- **Employee IDs**: deterministic unique IDs (e.g., `REP00001`).
- **Names**: unique `(first_name, last_name)` pairs to avoid duplicates in dashboards and transcript references.

4.2 KPI-Consistent Synthesis

Rep KPIs are synthesized as correlated proxies rather than independent draws. Intuitively:

- Higher burnout increases average handle time, escalation likelihood, repeat contact probability, and compliance risk.
- Higher quality increases resolution probability and reduces repeat contact probability and escalation likelihood.
- Pressure increases handle time and worsens outcomes, with larger effects for high-burnout profiles.

Skill tags provide small outcome tilts when a call issue matches a rep's primary or secondary strength, without restricting who can receive the call.

5 Integration with Call and Scenario Generation

The employee database is designed to condition the call generator. For each simulated call:

1. Assign `rep_id` to the call record (enabling rep-level dashboards).
2. Use `primary_skill_tag`/`secondary_skill_tag` to compute a *skill-match modifier* on resolution probability and handle time.

3. Sample call duration around `aht_secs`, with variance scaled by `volatility_index`.
4. Sample escalation and transfer events using `escalation_rate` and `transfer_rate`, adjusted by pressure and burnout.
5. Compute CSAT using `csat_proxy` as a baseline, modified by resolution, escalation, and time.

This structure yields call-level realism (e.g., peaks cause friction) and supports dashboards that behave like real operations (e.g., some reps are consistently steadier; some show high variance).

6 Ethical and Practical Considerations

All outputs are synthetic and generated for analysis and demonstration. No real employee identities, customer identities, or proprietary interaction records are used. Public datasets are used only to parameterize priors and distributions, and the approach emphasizes reproducibility and auditability (seeded generation and non-destructive artifacts) to support transparent experimentation.

7 Limitations and Future Work

Current outputs represent a single-queue call center with a single job role. Future work may extend the framework to:

- Time-series KPI panels per rep (monthly drift under time-varying pressure).
- Explicit issue-to-skill mapping tables for consistent skill-match modifiers.
- Scenario-generation layers (policy changes, outages, promotion misapplication spikes).
- A call-type generator that consumes telecom issue taxonomies and seasonality priors.

8 Conclusion

The NovaWireless Representative Generator provides a reproducible method for synthesizing a call-center employee roster with KPI-consistent rep profiles. By grounding distributions in public datasets and encoding correlation structure through latent traits and pressure modifiers, the generator produces realistic artifacts suitable for dashboards, analytics prototyping, and synthetic scenario testing.

References

Ahmadzade Jahromi, M. (n.d.). *Call center data (call center daily performance)* [Data set]. Kaggle. Retrieved February 15, 2026, from https://www.kaggle.com/datasets/satvicoder/call-center-data

City of San Francisco. (n.d.). *Call center metrics for the Health Service System* [Data set]. Data.gov. Retrieved February 15, 2026, from https://catalog.data.gov/dataset/call-center-metrics-for-the-health-service-system

Federal Communications Commission. (n.d.). *CGB Consumer Complaints Data* [Data set]. FCC Open Data. Retrieved February 16, 2026, from https://opendata.fcc.gov/Consumer/CGB-Consumer-Complaints-Data/

IBM. (n.d.). *Telco Customer Churn* [Data set]. Retrieved February 15, 2026, from https://raw.githubusercontent.com/IBM/telco-customer-churn-on-icp4d/master/data/Telco-Customer-Churn.csv

Metropolitan Transportation Authority. (n.d.-a). *MTA NYCT customer engagement statistics: 2017–2022* [Data set]. Data.gov. Retrieved February 15, 2026, from https://catalog.data.gov/dataset/mta-customer-engagement-statistics-beginning-may-2017

Metropolitan Transportation Authority. (n.d.-b). *MTA NYCT paratransit call center performance: Beginning 2016* [Data set]. Data.gov. Retrieved February 15, 2026, from https://catalog.data.gov/dataset/mta-access-a-ride-call-center-performance-beginning-2016

mexwell. (n.d.). *Employee performance and productivity data* [Data set]. Kaggle. Retrieved February 15, 2026, from https://www.kaggle.com/datasets/mexwell/employee-performance-and-productivity-data

Tank, P. (n.d.). *Employee HR dataset* [Data set]. Kaggle. Retrieved February 15, 2026, from https://www.kaggle.com/datasets/prishatank/employee-hr-dataset

ziya07. (n.d.). *Employee churn data* [Data set]. Kaggle. Retrieved February 15, 2026, from https://www.kaggle.com/datasets/ziya07/employee-churn-data