Insights from Field Data

Number of Event X per Exposure Time < Threshold?

Sandro Wiedmer, 2025-10-02

CAS Applied Data Science, Module 2 Project

Agenda

1. Introduction
2. Data Description
3. Question to be answered
4. Descriptive Statistics
5. Statistical Test
6. Conclusion
7. Questions and Discussion

Introduction: Use Case

- **Product intro: X-ray generator + X-ray tube:**
  - Unipolar (160 and 225 kV) and bipolar (320 and 450 kV) generators (this specific product line)
  - Several tube types
  - Arcs (electrical discharges) in the tube as one of the most stressful situations for the generator

- **Why field data matters:**
  - Knowledge is hidden in operational data
  - Field performance = real-world learning opportunity
  - Lab data is limited and not the ‘real world’
  - Helps validate assumptions, improve designs, prevent repeat issues
  - Field data closes critical (R&D) feedback loop

Introduction: databricks

- Unified cloud platform for efficient data analysis + machine learning
- Handles large data volumes (e.g. log files / diagnostic reports)
- Integrates with established tools (Python notebooks, SQL)
- PySpark DataFrame vs Pandas DF (distribution in compute cluster)

Data Flow: From Generator to Analysis in Databricks

Simplified Overview:

- **Data Source System (X-Ray Generator)**
  - The X-ray generator writes diagnostic reports in a specific format
  - Executable script automatically uploads the files to web application (hosted on AWS)
- **Cloud Storage & Notifications (AWS S3 + SNS):** Simple Storage Service, Simple Notification Service
- **Ingestion Layer (Databricks Autoloader):** Subscribes to SNS topic and polls for new notifications
- **Processing:** A scheduled Databricks job triggers a notebook that executes the event extraction script
- **Consumption Layer:** Databricks notebooks (and dashboards) are built here, **e.g. for Module 2 project**
- **Alerts:** Databricks alerts are configured (e-mail notifications to relevant stakeholders)

Data Description

- Number of rows x columns: 25’762’956 x 36 (raw) resp. 25’596’346 x 50 (pre-processed)
- Each line has a timestamp, device ID, event type (e.g. arc), counts, working point (voltages, currents, …), duration, tube type, …
- Number of generator devices: 800 (where field data is available, biased - ev. “pessimistic”)

Question to be answered

“Is the accumulated number of tube arcs (type of event) per accumulated exposure time (in hours) per generator device below a certain threshold, e.g. 4’000 arcs per 10’000 hours* (rate 0.4 arcs per hour)?”

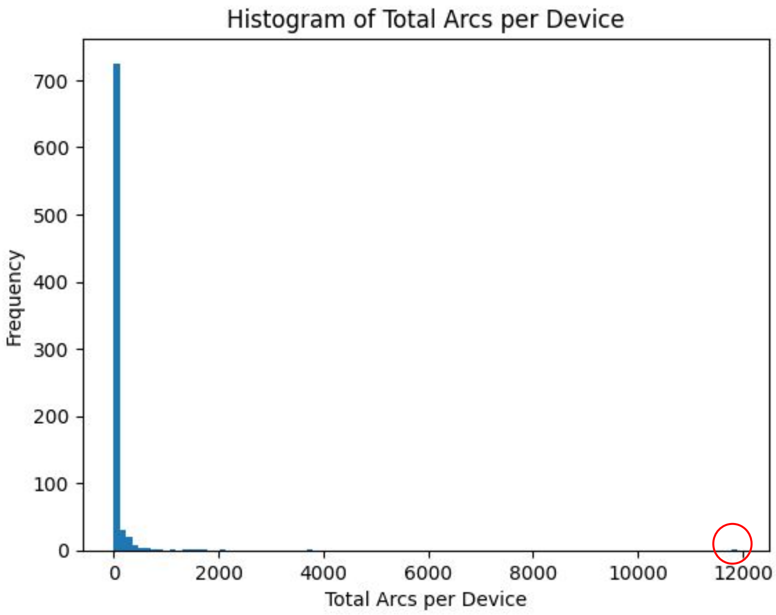
*: from lifetime related requirements, validated in lab for each generator type

=> Accumulated number of arcs per accumulated exposure time per device has to be extracted

Descriptive Statistics: Arcs per Device

- Total number of arcs: 54’626
- Indication for not having normal distribution: Skewness/overdispersion
- Summary:

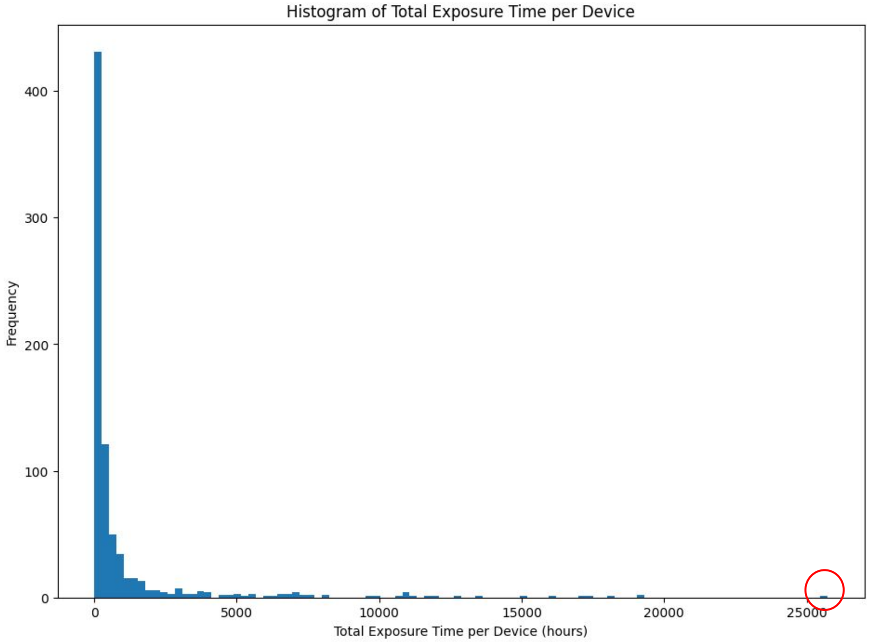
```
min: 0
max: 11894
mean: 68.28
median: 4.0
std: 468.496
```



Descriptive Statistics: Exposure Time per Device

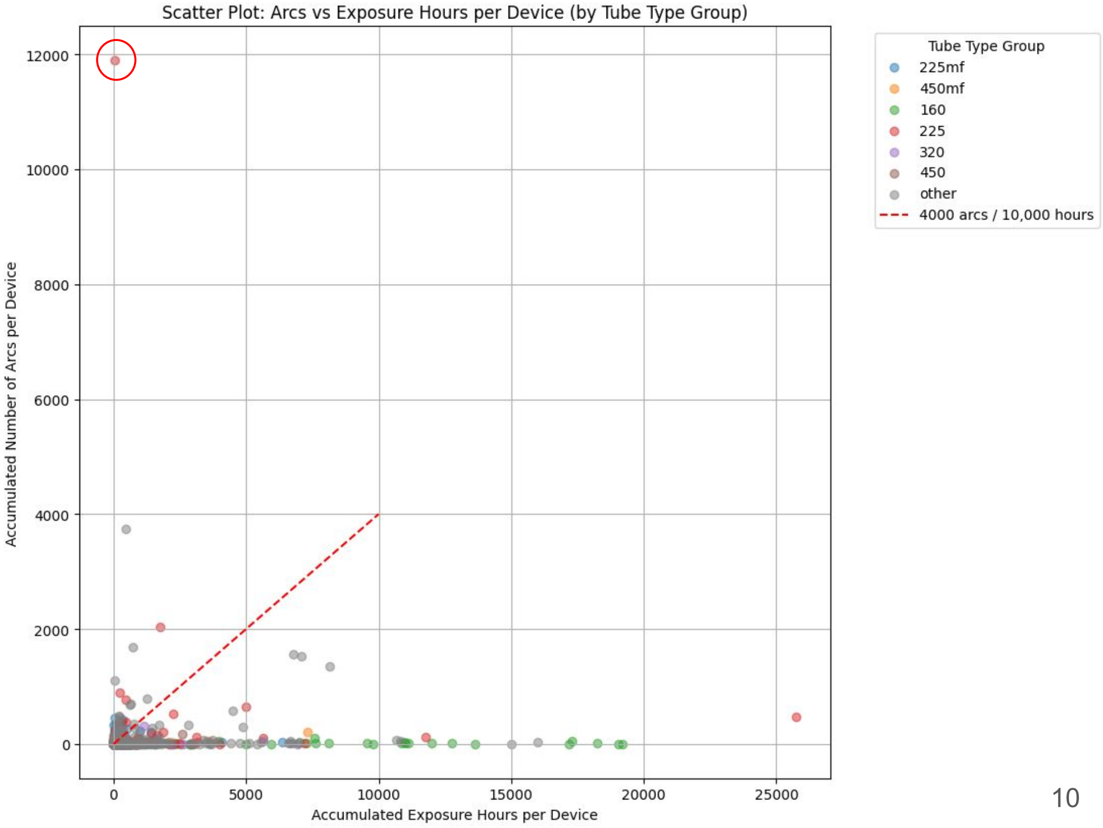
- Total exposure time: 780’570 hours
- Indication for not having normal distribution: Skewness/overdispersion
- Summary:

```
sum: 780570.5
min: 0.000167
max: 25735.39
mean: 1015.046
median: 181.878
std: 2611.528
```



Descriptive Statistics: Arcs vs Exposure per Device

1. Each point represents a generator device
2. Tube type related pattern?
3. Above/below red line



Descriptive Statistics: Arcs vs Exposure per Device

Summary

```
min: 0.0
```

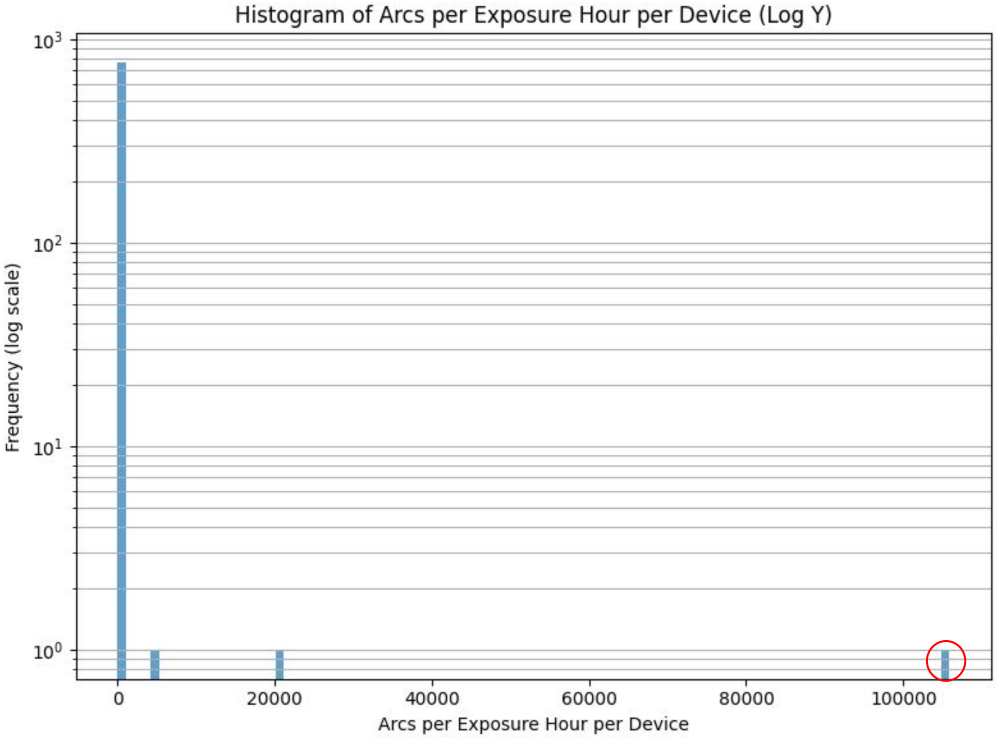
```
max: 105882.35294117646
```

```
mean: 172.23575505036175
```

```
median: 0.013743855947591622
```

```
std: 3891.3803374426057
```

Again, indication for not having normal distribution: Skewness/overdispersion



Statistical Test

- Hypotheses
  - λ = true mean rate of arcs/hour
  - Threshold y = 0.4 arcs/hour (4’000 arcs / 10’000 hours)
  - H0: $\lambda \geq y$ (arc rate is at or above threshold, i.e. not acceptable)
  - H1: $\lambda < y$ (arc rate is significantly below threshold, i.e. acceptable)
- Chosen test type:
  - 1.) Fleet-level: Exact Poisson test, sanity check 54’626 arcs / 780’570 exp. hours (0.07 arcs per hour)
  - 2.) Device-level: Negative binomial regression w. exposure offset (generalization of Poisson regression)
- Implementation:
  - 1.) scipy.stats Poisson CDF (cumulative density function)
  - 2.) statsmodels GLM (generalized linear models, negative binomial with offset)

- Result 1.) H0 rejected

```
One-sided p-value = 0.0000
```

Result 2.) Fail to reject H0

```
one-sided p-value = 1.0000
```

Conclusion

1. Fleet-level: Poisson
    a. Rejected H0 → Fleet overall arc rate significantly below threshold
    b. Homogenous fleet assumed
    c. Evidence of compliance, “fleet level is safe”
2. Device-level: Negative Binomial Regression
    a. Did not reject H0 → Some devices may exceed threshold
    b. Cannot guarantee all devices are below threshold
    c. Reveals important variability risk that should be managed
3. Implication
    a. Fleet is compliant on average, but a subset of devices drives variability
    b. Tail of risk - some devices are problematic
    c. Investigate high arc rate devices (e.g. 't3-1716055-1936' with 11’894 arcs: Comet internal)

Discussion Starter

- Which type of test would you have chosen (counts per exposure, not normally distributed, overdispersed)?
- What could be changed or added to the procedure?

Questions?