**🎯 Top 3 Recommended Target Variable Options**

**1️⃣ Fire Incident Count per Grid per Time Interval (Continuous Variable)**

**Type:** *Regression target*

**Definition:**  
The total **number of fire incidents** that occurred in each **spatial unit (e.g., grid, dissemination area)** during a **specific time interval** (e.g., month).

**Example:**

| **Grid\_ID** | **Year** | **Month** | **Fire\_Incidents\_Count** |
| --- | --- | --- | --- |
| Grid\_001 | 2023 | 01 | 12 |
| Grid\_002 | 2023 | 01 | 0 |

**Why this is a top choice:**

* ✅ Directly measures **fire activity intensity** (the phenomenon you want to predict).
* ✅ Compatible with **time-series regression** or **spatio-temporal models** (e.g., LSTM, Prophet, Random Forest Regressor, XGBoost Regressor).
* ✅ Allows interpretation of trends, seasonality, and spatial clustering.
* ✅ Aggregation naturally matches your constraint (area-level anonymized data).

**Key metric examples:** RMSE, MAE, R².

**Best suited for:** Predicting *how many* fires are likely to occur in a given grid and month.

**2️⃣ Binary Fire Risk Classification (High-Risk vs. Low-Risk Area)**

**Type:** *Classification target*

**Definition:**  
A categorical label indicating whether a grid-area during a time interval is **high risk (1)** or **low risk (0)** for fire occurrence.

**Example:**

| **Grid\_ID** | **Year** | **Month** | **Fire\_Risk\_Class** |
| --- | --- | --- | --- |
| Grid\_001 | 2023 | 01 | 1 |
| Grid\_002 | 2023 | 01 | 0 |

**How to define "high-risk":**

* Option A: Top 10–20% of areas by number of incidents → “high-risk.”
* Option B: Fire count > historical median/threshold for that area.
* Option C: Areas with ≥ 1 fire incident in that month (binary occurrence).

**Why this is a top choice:**

* ✅ Simplifies the problem to **yes/no fire risk**, easy for communication to city departments.
* ✅ Enables **classification algorithms** (Logistic Regression, Random Forest Classifier, Gradient Boosted Trees, Neural Networks).
* ✅ Robust against outliers and data sparsity (many grids will have zero incidents).

**Key metric examples:** Precision, Recall, F1-score, AUC-ROC.

**Best suited for:** Operational decision-making — *“Which areas should we prioritize for inspection next month?”*

**3️⃣ Fire Incident Density (Normalized Count per Area or Population)**

**Type:** *Regression target (continuous, normalized)*

**Definition:**  
The **number of fire incidents normalized by area size (per km²)** or **by population density** within each spatial unit and time window.

**Example:**

| **Grid\_ID** | **Year** | **Month** | **Fire\_Incidents\_per\_km²** |
| --- | --- | --- | --- |
| Grid\_001 | 2023 | 01 | 2.4 |
| Grid\_002 | 2023 | 01 | 0.0 |

**Why this is a top choice:**

* ✅ Controls for the **different sizes** of grid cells or dissemination areas.
* ✅ Allows fair comparison between dense downtown areas and sparse suburbs.
* ✅ Useful for **policy-level insights** and **geographical hotspot detection**.
* ✅ Supports spatial regression or geostatistical modeling (e.g., Geographically Weighted Regression, Spatial Lag Models).

**Key metric examples:** RMSE, MAE, Moran’s I for spatial autocorrelation.

**Best suited for:** Understanding **relative fire intensity** — where risk is higher *per unit area or population*.

**💡 Additional Notes and Clarifications**

| **Aspect** | **Explanation** |
| --- | --- |
| **Data Structure Expected** | One record per (Grid\_ID, Month) pair, with target + all engineered features (building, environmental, socio-economic, etc.) |
| **Target Selection Tip** | Start with **# of fire incidents per month per grid (target #1)** — it’s the most straightforward and scalable baseline. Later, derive target #2 (binary) or #3 (density) for specialized analyses. |
| **Handling Sparse Data** | Include rows with **zero incidents** for months/areas with no fires (important for proper training). |
| **Spatial Division Choice** | Dissemination Areas (DA) or custom grids (e.g., 500m × 500m) are both acceptable; ensure the same spatial unit is used for all features. |
| **Temporal Aggregation Choice** | Monthly aggregation is most balanced — captures seasonality while ensuring enough data points. |

**🧭 Summary Table**

| **Rank** | **Target Variable** | **Type** | **Why It’s Recommended** | **Modeling Examples** |
| --- | --- | --- | --- | --- |
| 1️⃣ | Number of fire incidents per grid-month | Regression | Direct measure of risk, continuous, interpretable | Random Forest Regressor, XGBoost |
| 2️⃣ | High-risk vs. Low-risk grid classification | Classification | Actionable, simple, operational | Logistic Regression, CatBoost, Neural Net |
| 3️⃣ | Fire incident density (per km² or per capita) | Regression | Normalized for area/population; good for fairness | Linear Regression, GWR, GeoML models |

Would you like me to show a **Python example** of how to compute these three target variables (from fire incident data joined with spatial grids using GeoPandas)?  
That would demonstrate how to *create and structure* your target columns correctly for modeling.

Excellent — that’s a *very* clear and meaningful objective, and it puts you on exactly the right track.  
Your ultimate goal isn’t just to predict fires for the sake of accuracy — it’s to **enable preparedness and preventive decision-making** for public authorities.

Let’s unpack this carefully and logically — step by step.

**🎯 1. Clarify the Ultimate Purpose of the Target Variable**

You want your model’s output (the **target variable**) to help **authorities, fire stations, and policy teams**:

* **Anticipate** where and when fires are more likely to occur
* **Allocate** resources (personnel, trucks, inspections, education) more effectively
* **Prioritize** preventive measures (safety inspections, hydrant maintenance, public awareness)
* **Adapt** to changing conditions (weather, urban density, etc.)

So, the **target variable should express “fire risk readiness”**, not just “number of past fires.”

That means the target must be:

* **Predictive** (forward-looking)
* **Actionable** (clear for operational decisions)
* **Interpretable** (understandable to non-technical users like planners or responders)

**🧩 2. Recommended Target Variable Formulations**

Let’s consider the two directions you mentioned — classification and forecasting — and show how they connect to your ultimate objective.

| **Target Type** | **Description** | **Example Definition** | **Why It’s Useful for Readiness / Prevention** |
| --- | --- | --- | --- |
| **A. Fire Risk Classification (3 classes: Low, Medium, High)** | Each grid cell is labeled according to expected fire likelihood for a given period (e.g., next month or next quarter) | High risk = top 20% of predicted fire probability; Medium = middle 40%; Low = bottom 40% | Easy for government dashboards and planners to visualize and act on (e.g., focus inspections on “high” zones) |
| **B. Fire Occurrence Forecasting (Regression)** | Predict the *expected count* or *probability* of fire incidents per cell for the next time window | Predict number of incidents next month per 500 m × 500 m grid cell | Useful for operational planning (e.g., how many incidents to expect per district, allocate trucks/resources) |
| **C. Composite Readiness Risk Index (Hybrid)** | Combines fire likelihood + mitigation capacity (e.g., hydrant density, station distance) to produce an **adjusted risk score** | Risk = PredictedFireRate × (1 - MitigationScore) | This expresses “where risk remains high *despite* infrastructure,” which directly supports preventive prioritization |

✅ **Recommended:** Start with (A) *Risk Classification* (Low/Med/High) — because it’s interpretable — and evolve into (C) the *Composite Readiness Risk Index* once your features are richer.

**🔗 3. How the 4–5 Supplementary Datasets Feed Into This Goal**

| **Dataset** | **Role in Predictive Model** | **Influence on Target Definition** |
| --- | --- | --- |
| **Building data** | Defines exposure (number and vulnerability of structures) | Used for normalization (e.g., incidents per 100 buildings) and for risk features (old buildings = higher baseline risk) |
| **Census / Socioeconomic data** | Captures human activity, demographics, and housing type | Helps explain “why” fires cluster (social factors); informs community-targeted prevention |
| **Weather data** | Introduces time-varying hazard conditions | Enables temporal forecasting — essential for next-month predictions |
| **Fire station / hydrant data** | Represents mitigation / response capacity | Used to adjust predicted risk — *readiness gap* = “high risk but low coverage” |
| **Land use / zoning data** | Distinguishes context of fire risk (industrial, residential, etc.) | Can drive class-specific targets — e.g., “industrial high risk,” “residential medium risk” |

**🧠 4. Step-by-Step Logical Flow for Target Definition**

Here’s how you can systematically define your target variable, aligned with the goal of **actionable fire readiness**:

1. **Grid aggregation:**  
   Aggregate historical SIM interventions into 500 m × 500 m grid cells, with counts per month.
2. **Baseline rate:**  
   Compute historical fire rate per cell:  
   [  
   \text{rate} = \frac{\text{fire count}}{\text{number of buildings or population}}  
   ]
3. **Temporal feature creation:**  
   Create lag features — fire rate of last 1–3 months, seasonal trends, weather variables.
4. **Predictive modeling target (for training):**  
   Define target for next month:
   * Regression: fire count at t+1
   * Classification: risk level based on predicted probability thresholds
5. **Policy-friendly output:**  
   Post-model, translate predictions into clear categories or maps:
   * *“High Risk — limited hydrant coverage”*
   * *“Medium Risk — dense residential area”*
   * *“Low Risk — new buildings, close to station”*

**🧭 5. Summary — The Essence**

Your **target variable** is not just “fire counts.”  
It is a **decision-support signal** that answers this question for the fire authority:

“Where and when should we focus our prevention and preparedness efforts next?”

So the best definition is:

🔹 *A forward-looking classification of fire risk (low, medium, high), derived from historical fire counts and contextual features, designed to guide preventive actions and readiness planning.*

If you’d like, I can now outline **a practical step-by-step plan** (with pseudocode and logic) for:

* how to engineer this target variable from the raw SIM data,
* how to align the 4–5 datasets spatially and temporally, and
* how to produce the “readiness risk classification.”

Would you like me to do that next?

==