DataGenesis

# **Synthetic population**
# powered by AI

Synthetic population data for Disaster
Response and Climate Resilience

June 2024

**In an era marked by the escalating threat of climate change, the specter of natural disasters and extreme weather, increasingly frequent and severe, looms larger than ever before. Amidst this reality, it is crucial to recognize that not all communities are equally equipped to weather the storm.**

Certain populations, particularly vulnerable demographics and those exposed to chronic environmental hazards, poverty, and insufficient access to healthcare, face disproportionately higher impacts when a catastrophe strikes. Moreover, as climate change spurs population shifts and migratory pressures in the medium to long term, the challenge of planning and preparing for future disasters becomes increasingly complex.

In the face of such challenges, **access to reliable and up-to-date information becomes critical for effective relief efforts.** Understanding the dynamics between vulnerable populations, uncertain disasters, and transportation infrastructure becomes paramount to make better decisions in emergency evacuation and supply chain resilience. Yet, working with population data faces significant barriers. Privacy concerns surrounding personal data hinder the flow of critical information, while the absence of accurate models for population projection complicates planning efforts. Nevertheless, the urgency of the situation demands action. "What-if" scenario analysis is not just desirable but imperative to quantify risks and chart a course towards climate resilience.

## Miami, FL: the Perfect Storm

The First Street Foundation reports that Florida, the fastest growing state in the U.S., is also the state with the most increased number of days of extreme heat due to climate change. By 2053, Miami is estimated to rise from 50 to 91 days/year with heat index above 100 degrees. Miami-Dade County, FL is also at risk of extreme winds from hurricanes, tornados, and storms, which dramatically increase the risk for rapid spreading of wildfires and smoke plumes from nearby forests, and most importantly floods. 51% of Miami properties face flooding risks, which will increase to 77% over the next 30 years, when most of the city will be under sea level.

**What local communities are most exposed to the risks of heatwaves, decreased air quality, and flooding? Where are vulnerable populations located, and how are they expected to evolve in the future as climate change effects become visible?**

Download and explore the distribution of demographics, socioeconomics and health factors in the DataGenesis synthetic population for the city of Miami, Miami-Dade County, FL, for the year 2021, generated at scale at https://skymantics.com/datagenesis/. We want to know your use cases and needs, please share at datagenesis@skymantics.com

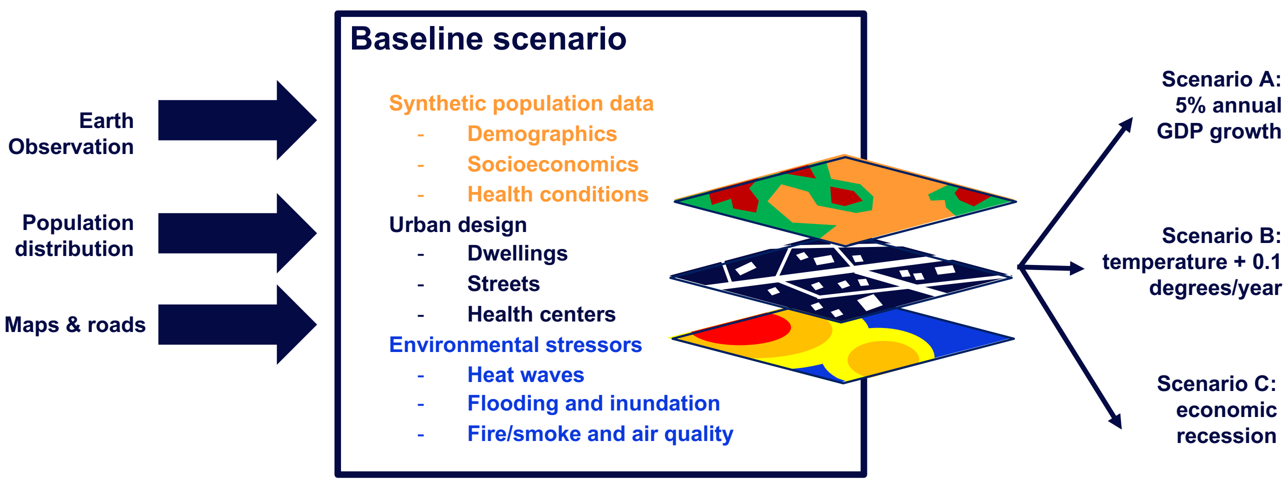
**To successfully tackle climate resiliency and disaster relief, emergency management organizations, law enforcement, and infrastructure planners should adopt AI-generated synthetic populations to overcome current data limitations and inform proactive strategies for more resilient communities. Let us heed the call for action, recognizing that when we face climate-induced disasters, preparedness and foresight are our greatest assets.**

# Towards Synthetic Digital Twins for climate resilience

**Digital Twins, virtual representations of real-world systems, are being widely adopted to simulate and plan for emergency scenarios.** To existing models of physical objects (dwellings, streets, health resources) and processes (weather events, impact risks, itineraries), Skymantics' DataGenesis adds a synthetic population data layer. Compared to current approaches, where aggregate statistics are used to model demographic processes, synthetic data creates rich contextual information of the population distributed per socioeconomic and geospatial segments.

The resulting Synthetic Digital Twin enables discrete event simulation and projection of health baselines into hypothetical scenarios of population growth, shifts, or humanitarian crises driving migrations and displacements. **The innovation of a Synthetic Digital Twin is that it allows, from a baseline reality, to spawn alternative Digital Twins representing different branches of the same scenario.**

By configuring macroscopic parameters such as rate of population growth, urban surface growth, demographic age shift, economic growth, temperature rise, or rate of disease prevalence, a stakeholder can explore how the climate/disaster risk indexes would vary over the years in the advent of a specific trend or event. **The resulting resilience indicators can then be used by climate scientists, healthcare providers, housing authorities, and policymakers to inform resource planning focused on impacted communities as they evolve.**




**Synthetic Digital Twin conceptual architecture**

**A Synthetic Digital Twin acts as an aggregator of a geospatially enabled, multi-layer decision support data pool.** This data pool contains environmental stressors from extreme weather events into population data layers distributed in urban areas and characterized by indicators of vulnerability due to demographic and social determinants (e.g., elderly, poverty level, or chronic health conditions) or to accessibility factors (e.g., limited mobility or low access to resources). **This is the base for a Climate Resilience Information System (CRIS) with adaptation and forecast capabilities.**

## The limitations of using personal data for disaster response

**Emergency organizations generally have access to citizen data such as addresses, demographics, health conditions, and history of past emergencies. However, these pieces of data live in different jurisdictions – an emergency organization, a law enforcement agency, a county/municipal authority, or an EMS provider.** Often, each organization ends up solely working with the limited information they have because they cannot secure adequate tools or agreements to share such complex and sensitive data in compliance with regulations. Such hurdles prevent entities from achieving a common picture for coordinated disaster relief.

**Privacy.** Two types of information attributes allow the identification of an individual:

- **Personally Identifiable Information (PII)** such as name, address, or SSN.
- **Protected health information (PHI)** from health treatment, payment or operations.

Both are protected and strictly regulated by privacy rules such as HIPAA, GDPR and CCPA. Records can be anonymized but are still identifiable if combined with other attributes. As a consequence, PII/PHI datasets require large organizational resources to collect, curate and protect. **Data cycles** for updating and sharing **are long and expensive**, as they require strict security controls and privacy validations in the production environment that severely limit the flow of timely data for analysis.

### Personal data (production environment)

Difficult to share between jurisdictions

Expensive to collect and maintain

Small scale - irrelevant for forecast modeling

Narrow – deterministic test cases

**Coverage.** Personal records represent only population with the surveyed attributes in the jurisdictional area at a specific point of time. This is statistically irrelevant to develop data analytics models for scenario forecasting. As a result, test cases reflecting scenarios of analysis have to be defined manually based on deterministic rules. This approach misses out on complex events that require contextual links between population attributes **(e.g., what if a recession occurs five years before a major natural disaster?).**

## What about insurance?

Over the past decade, U.S. insurers have suffered a staggering three-fold increase in disaster-related yearly losses, leading the industry to scale back coverages and hike premiums. Central to this dilemma is the difficulty in forecasting the evolution of the future disasters and quantifying the impact likelihood and severity. Data-driven insights on the nature of disasters and shifts in population vulnerability distributions hold the key to bringing certainty to insurance planning in an increasingly volatile landscape. However, the utility of insurance customer data is limited by stringent privacy considerations, as safeguarding against misuse remains a priority.

# DataGenesis, AI-powered synthetic population generator

**To address the need for fully dimensionalized synthetic population data, Skymantics created DataGenesis. DataGenesis uses a Generative AI approach to recreate synthetic datasets that mimic actual populations at scale including more than 200 demographic, socioeconomic, and health attributes**. Populations are distributed geographically and respecting statistical accuracy per neighborhood, thus allowing for geospatial analysis of climate and disaster impacts.

## Synthetic data (test environment)

Non-identifiable, safe to share

Agile data cycles

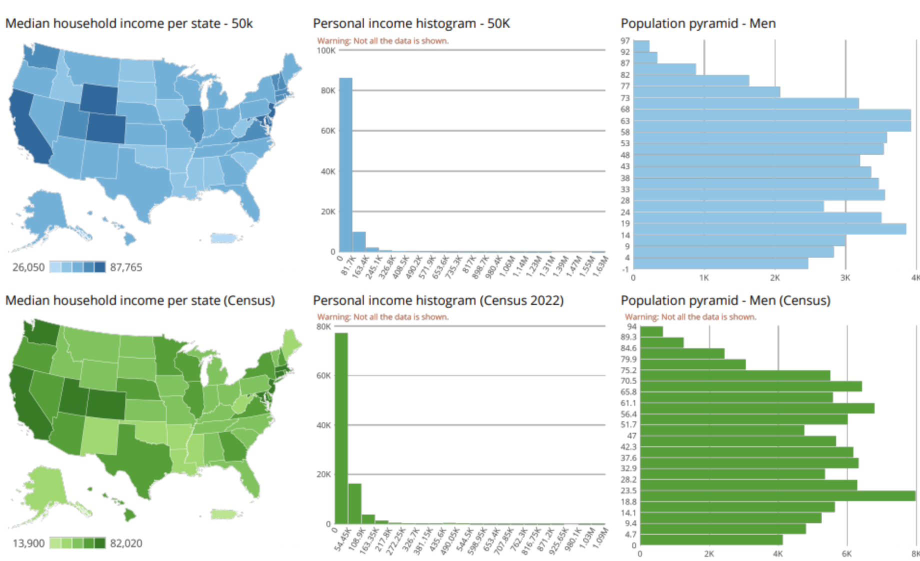
Dimensionalized for AI modeling

Automated, contextual test cases

**DataGenesis is fed from more than 140 public authoritative sources to create built-in AI models for population generation and extraneous events models.** The synthesized populations are untraceable to real personal records and thus safe to share, while maintaining statistically accurate population distributions fit for analysis and model development. DataGenesis incorporates a **data validation suite** to verify accuracy and representativity of the generated population.

The DataGenesis approach to population synthesis is unique. Unlike other synthetic data generators, DataGenesis preserves AI-generated multivariable correlations between variables to ensure the consistency of population groups, family structures, socioeconomic levels, and relationships among individuals, households, and dwellings.

This built-in context allows to simulate interactive events within the population, such as evacuations or aging, accurately. Aging is forecasted by application of natural life events (births, dates, marital status changes, family relocations, migrations, job changes, diseases and disabilities) to create **realistic consistent multiverses for the same population in future years** and support calculation of impact trends in the future.



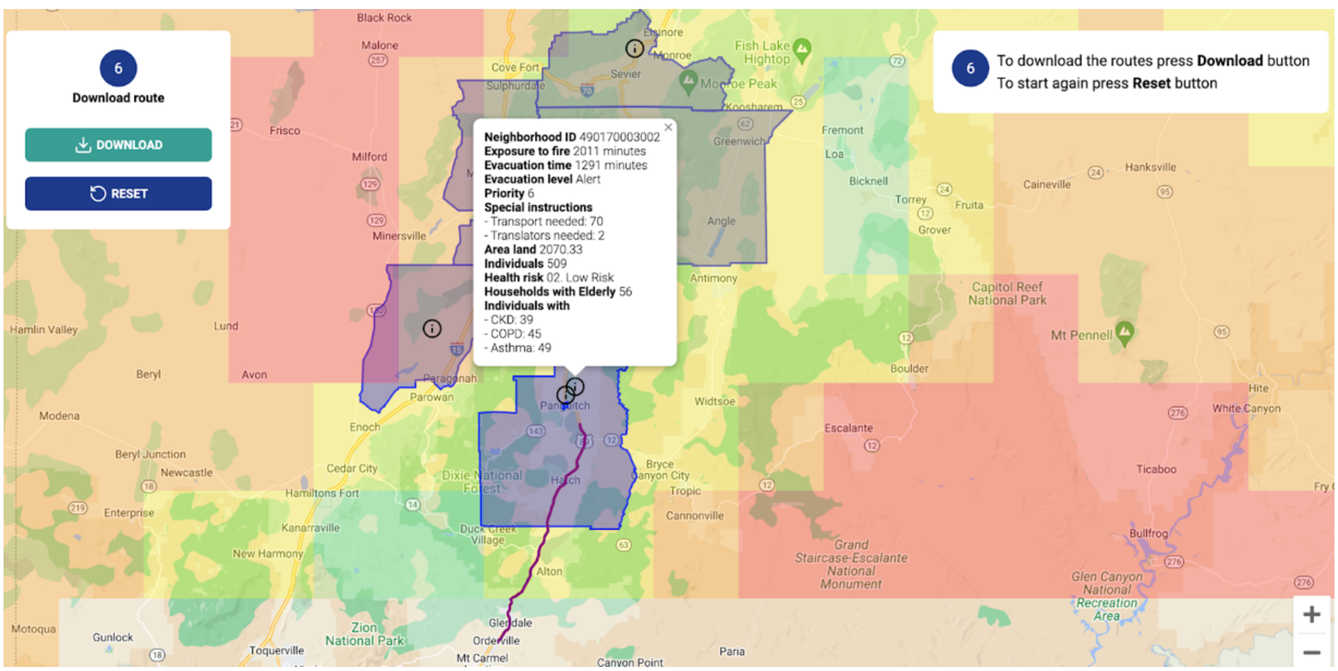

**Comparison of 50,000 households – synthetic vs real (U.S. Census - 2022)**

## Success story

# Evacuation scenario planning and training

**When disaster hits, the protection and evacuation of affected population needs to happen first in a safe and efficient manner to clear the area for disaster suppression.** Evacuation resources are multiple: emergency management, law enforcement, EMTs, and community leaders coordinate to lead citizen groups to designated reception centers in an orderly manner. Currently, each response team rely on data available in their jurisdiction and fixed pre-planned responses, especially in information-poor rural areas. Unfortunately, this approach is too rigid as each disaster scenario is different and contains many unknowns (population behaviors, road blockages, health stressors such as smoke) that require "what-if" scenario planning.

This is where an evacuation planning tool is critical. **Fusing satellite imagery, synthetic data of affected population, socioeconomic/health factors and road information, DataGenesis offers an evacuation scenario application** that has been tested in urban (St Louis, MO) and rural (Fish Lake, UT and Manitoba, Canada) areas. The application allows to declare disasters and blocked areas, and estimates vulnerability indexes from demographics, health conditions, and mobility limitations of affected neighborhoods. Running the scenario indicates the best distribution of people among available evacuation routes, time to evacuate, prioritization of evacuation orders, and special instructions to serve specific vulnerable communities. Evacuation algorithms can be configured to maximize capacity usage, minimize crowding, or other preferences.




**Evacuation scenario planning in Fish Lake, UT (2023)**

This tool is used in **tabletop exercises and training**, to identify **vulnerable neighborhoods**, find bottlenecks in the **road infrastructure,** determine optimal **locations for reception centers**, or forecast **future pressures by migrations or demographic shifts.** Results are safe to share with stakeholders and fit for collaborative analysis.

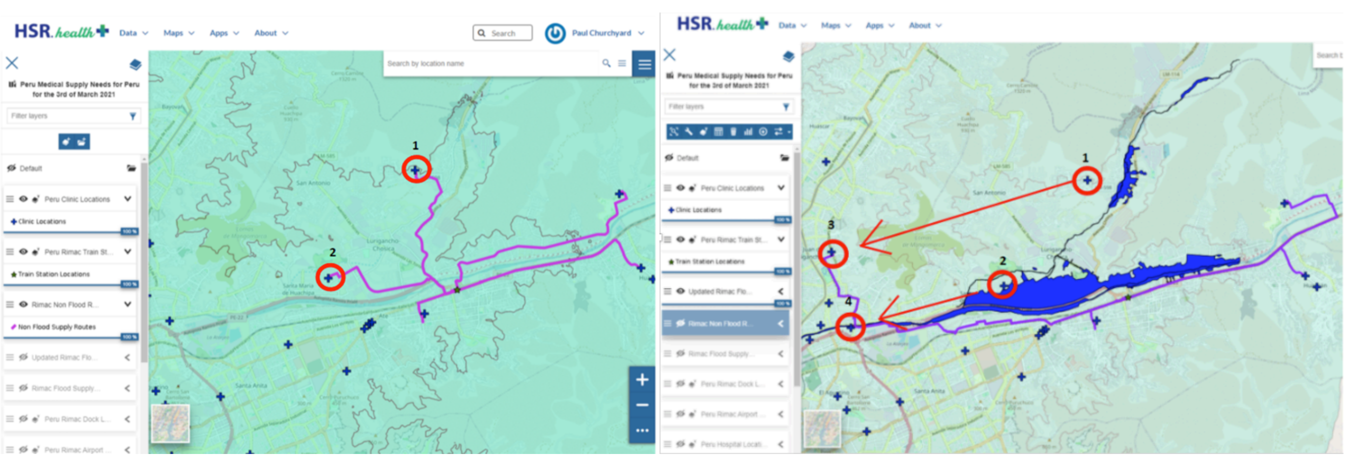
## Success story

# Emergency supply chain route modeling

**Supply chain management and resource allocation for healthcare supplies is especially complicated when the evolution of a disaster or conflict in the future is unpredictable. Demographic changes and disease incidences produce shifts in health coverage demand, which require adjustments in planning of facilities and inventory.** Rapidly growing urban areas are especially dangerous, as new communities may be established in currently underpopulated areas with unknown disaster impact indicators such as heatwaves or flooding.

In this uncertain environment, population segments can be exposed to new vulnerabilities unexpectedly, such as high temperatures, degraded air quality, or direct threat on life and property. Due to how disasters affect infrastructure, entire communities can be suddenly cut off from access to healthcare, food and water, or safe shelter.

**The Skymantics Supply Chain engine, powered by DataGenesis,** is a resilience planning solution developed in collaboration with partners [HSR.*Health*] and [RSS-Hydro], and tested successfully in Open Geospatial Consortium Disasters Pilot 21 in the Rimac River basin (Peru). This engine allows infrastructure planners, health providers, and local authorities to understand disaster trends of disasters and develop probability scenarios of emergency supply chain to communities at risk.




**Flood impact on supply chain routes in Rimac River basin, Peru (visualized in HSR.*Health* GeoMD)**

**The supply chain engine generates a road accessibility and risk calculation that identifies feasible routes for delivery of emergency supplies and the coverage area of critical resources to population. DataGenesis** can in turn generate emergency demand models for the affected population to assign appropriate volume and origin of resources considering specific needs of vulnerable population in areas at risk.

The engine provides accurate assessments of how vehicle characteristics can impact logistics routes, and can consume real-time, up to date external data sources (Earth Observation, sensors, or manual input) to form geolocated hazard levels. The subsequent impact on road usage is then used to update accessibility index and recalculate best routes. **The resulting demand coverage can be used to identify weak points in the emergency logistics network and assess the habitability ratio of the affected areas in long-term environment changes.**

DataGenesis
by skymantics

**SBA 8(a)-certified Disadvantaged Small Business**

We provide AI, data and geospatial solutions in Government, healthcare and financial industries

**www.skymantics.com**

**info@skymantics.com**