Generating Realistic Passenger Name Records with Privacy Compliance for Security Analysis

Muhammad Fadlian
University of Sheffield
m.f.fadlian@sheffield.ac.uk

Neil Ireson
University of Sheffield
n.ireson@sheffield.ac.uk

Vitaveska Lanfranchi
University of Sheffield
v.lanfranchi@sheffield.ac.uk

ABSTRACT

Passenger Name Record (PNR) data is essential for transportation analysis and security research, particularly in surveillance and threat detection. However, stringent security and privacy concerns limit access to real PNR data. This study presents a methodology for generating synthetic PNR data that not only replicates statistical properties but also reconstructs passenger social networks, models travel behaviours and preserves individual travel histories for security and mobility analysis. Our approach generates detailed, individual-level data—including passengers, bookings, and flights—while maintaining spatial, temporal, and chronological consistency to ensure realistic movement patterns while upholding privacy. This methodology offers a privacy-preserving alternative for transportation security and behavioural research, expanding access to high-quality data for future studies.

Keywords

Synthetic data, Passenger Name Record (PNR), Agent-based modelling, Privacy-preserving, Security analysis.

INTRODUCTION

In 2024, global passenger traffic is expected to surpass pre-pandemic levels, reaching 9.5 billion passengers, with 43% (4.1 billion) of these being international passengers (ACI, 2025). Such a massive and increasing number of passengers poses significant challenges for border security, necessitating the need to advance automated data processing and analytics to maintain cross-border movements whilst identifying security threats. Passenger Name Record (PNR) data contains detailed information about passengers, including flight itineraries, travel companions, lodging, car rentals, payment information, and other booking details. It has long been a cornerstone of the travel industry, enhancing both operational efficiency and security (Vinod, 2008). Historically, passenger screening systems, such as the Computer Assisted Passenger Pre-screening System (CAPPS) and Secure Flight, have leveraged PNR data to categorize travellers into varying levels of risk. These systems primarily focus on identifying known threats, using predefined rules and government watchlists. Barnett (2004) highlighted the integration of color-coded risk categories supported by data from PNR records, government databases, international partnerships, and flagged individuals ('Privacy Act of 1974: Implementation of Exemptions; Secure Flight Records; Proposed Rule', 2007). While these systems have successfully intercepted passengers already identified as threats, they struggle to detect previously unknown risks and lack algorithmic transparency (Glouftsios & Leese, 2023).

Recent research explores methodologies to analyse travel patterns and personal details for identifying previously unknown high-risk passengers. Zhao et al., (2020) employed neural networks optimized with Particle Swarm Optimization - Back Propagation (PSO-BP) to classify passengers into high, medium, and low-risk categories based on attributes like occupation, credit history, and flight conditions. Their model was trained and validated

using passenger index data collected via airport surveys, with risk levels pre-assessed by civil aviation safety experts based on a standardized evaluation framework. Similarly, Zheng et al., (2017) developed a fuzzy deep learning model capable of handling incomplete data and detecting group-based threats, where collective behaviours may indicate risk even if individual characteristics do not. Their model was trained on a dataset from real-world Air China flight records, supplemented with synthetically generated attacker profiles created by aviation security experts.

Behavioural segmentation, commonly used in marketing, also has security applications for detecting suspicious travel patterns. Mahendru & Singh (2023) segmented travellers using behaviours such as trip span, travel search categories, and time of travel, utilizing datasets containing detailed booking transaction information. Glouftsios & Leese (2023) notes that, under the EU PNR Directive (European Commission, 2020), such profiling can aid in targeting unknown suspicious passengers. Additionally, machine learning models trained on archival PNR data can refine risk-based screening by identifying travel anomalies.

Despite its critical role in security analysis, real PNR data is not publicly available and is primarily restricted to government authorities and airlines (Mottini et al., 2018). Access to such data typically requires formal partnerships. This restriction creates a significant barrier to innovation, as developing and validating security frameworks often depend on real-world datasets (Cavusoglu et al., 2013). Without direct access, researchers must rely on limited, anonymized, or outdated datasets, reducing the effectiveness of experimental security models. These concerns, coupled with strict regulatory restrictions, further limit access to real-world PNR data, particularly within the European Union (European Commission, 2020) making it challenging to develop and validate robust security frameworks. As a result, researchers and organizations without such partnerships face substantial barriers.

To be a reliable substitute for real PNR data in security analysis, synthetic data must replicate statistical properties and capture behavioural and operational dynamics. This includes realistic modelling of passenger identities, travel histories, and booking behaviours, with spatial and temporal consistency in flight sequences, transit times, and routes. Where possible, the model should rely on publicly available datasets to ensure accessibility, privacy, and ethical standards.

This paper addresses the challenges of advancing passenger risk analytics while respecting data privacy by introducing an open-source synthetic PNR framework. We construct a publicly available synthetic dataset derived from demographic, mobility, and aviation data to support transportation security and behavioural research in the absence of real PNR data. Our approach integrates social networks and agent-based modelling to simulate realistic travel behaviours, enabling the identification of group patterns relevant to risk assessment. We evaluate the synthetic data using demographic, statistical, and behavioural metrics to ensure realism and usability.

The remainder of this paper is structured as follows: Section 2 reviews existing approaches to synthetic data generation and agent-based modelling (ABM). Section 3 details our methodology, outlining the steps taken to generate and validate synthetic PNR data. Section 4 presents our results and discussion, comparing the synthetic dataset against real-world benchmarks. Finally, Section 5 concludes the paper, summarizing key findings and potential future directions.

RELATED WORKS

In this section, we explore key methodologies in synthetic data generation, focusing on synthetic reconstruction (SR) and agent-based approaches. SR techniques generate datasets that retain the statistical properties of original data, which is useful when real data is unavailable or restricted. However, SR methods primarily preserve aggregate trends and do not capture individual decision-making processes or behavioural interactions. ABMs address this limitation by simulating individual passenger behaviours, allowing for the study of decision-making and emergent patterns relevant to security research. This overview examines these methods' applications and limitations, especially in synthesising data for airline passenger behaviour studies.

Synthetic Data Generation

Synthetic data generation methods typically extract statistical properties from microdata (i.e., individual-level records) to create representative synthetic datasets (Figueira & Vaz, 2022; Mottini et al., 2018; Patki et al., 2016; Templ et al., 2017). Synthetic Data Vault (SDV) provides various synthesisers, such as Gaussian Copula and Generative Adversarial Networks (GANs), to model data distributions. These methods effectively preserve statistical patterns, making them valuable when real data is restricted. Mottini et al. (2018) applied GAN-based

techniques to generate synthetic PNR data, capturing broad aggregate patterns such as business-to-leisure ratios and passenger volumes. However, the approach was limited in modelling the diversity of individual travel histories and social relationships, which are crucial for security research. In addition, like most synthetic data methods, it relies on existing microdata, meaning it primarily serves to anonymize or obscured identities.

Downscaling techniques provide an alternative to microdata-dependent methods by generating high-resolution microdata from multiple low-resolution sources by modelling the interrelationships between sources. Initially developed to generate high-resolution climate data, it employs various statistical and more recently machine learning techniques (Vandal et al., 2019). The technique has also been applied in the generation of synthetic populations. For example, SynC (Li et al., 2020), which reconstructs detailed individual data, while maintaining aggregate consistency, by fitting Gaussian copula models to socioeconomic and demographic datasets. However, it primarily generates static demographic datasets rather than dynamic, behaviour-driven travel patterns.

Agent-based Trip Planning and Social Networking

Agent-based models (ABMs) provide an alternative approach, focusing on simulating individual passenger behaviour rather than purely reconstructing statistical properties. The AirMarkets Simulator (Parker, 2017), demonstrates how ABMs can integrate discrete choice models to replicate itinerary preferences and global air travel behaviours. One approach that integrates synthetic data with behavioural modelling is persona-based modelling, where synthetic travel behaviour is generated across multiple layers (individual, household, regional level) and integrated with urban transport data (Hörl & Balac, 2021; Stevenson & Mattson, 2019; Vallet et al., 2022) This technique effectively captures demographic diversity and population flows. While originally developed for urban mobility contexts, this approach serves as an inspiration for modelling synthetic airline passenger behaviours.

PROPOSED FRAMEWORK

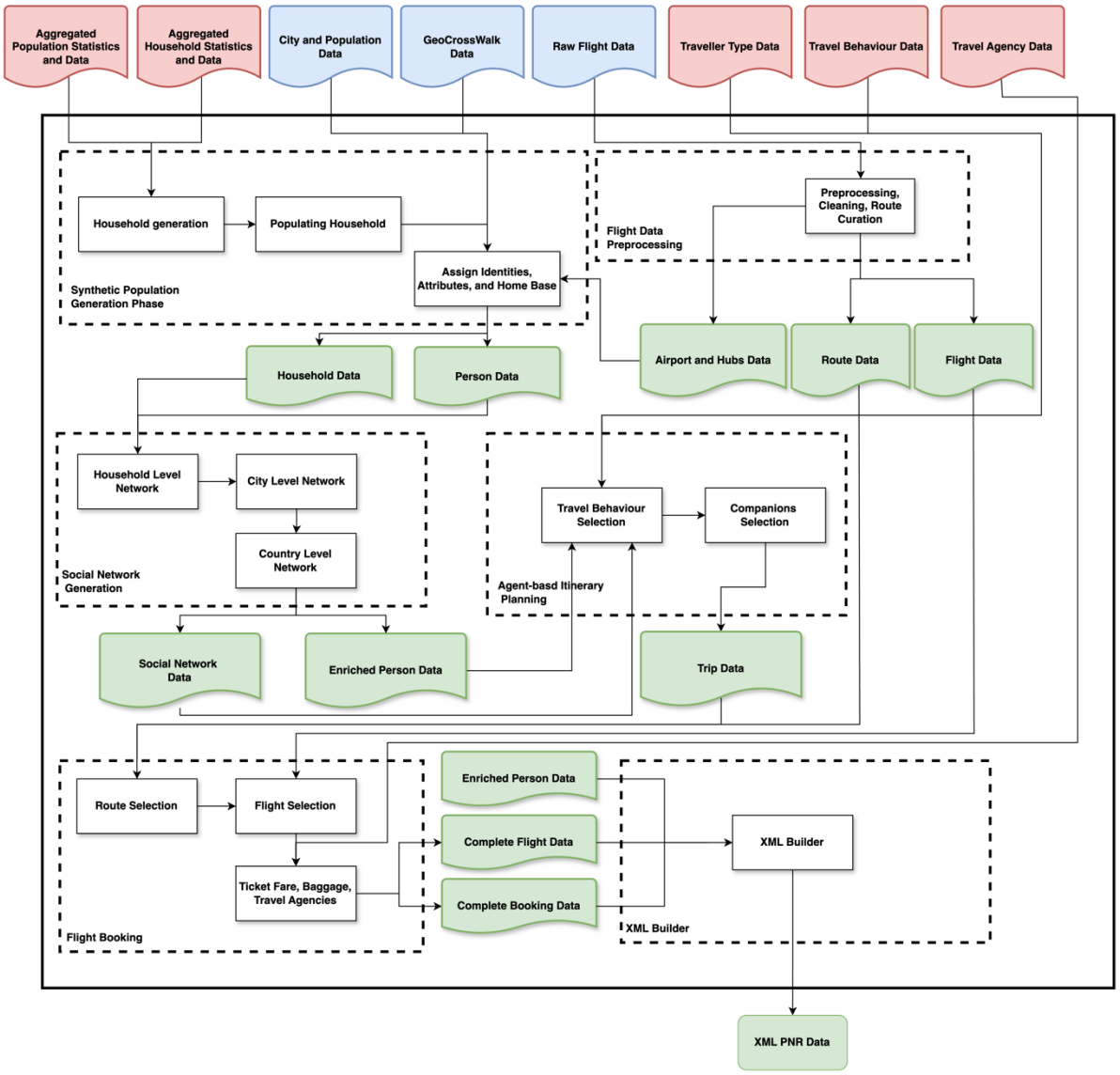
Steps Overview

To generate realistic and security-relevant synthetic PNR data, our framework addresses key challenges: ensuring detailed passenger identities, modelling realistic social connections, capturing travel histories, and incorporating flight booking behaviour. The process follows five key stages:

1. Synthetic Population Generation: Creates passenger base with structured household relationships to enable identity verification and watch-listing.
2. Social Network Construction: Basis of travel companionships based on geographical proximity for association detection.
3. Agent-based Trip Modelling: Simulates travel histories with spatial-temporal consistency, ensuring realistic trip planning and movement patterns.
4. Flight Booking: Assigns flights, tickets, baggage, and payment behaviour to reflect passenger decision-making and passenger travel record.
5. PNR File Generation: Produces standardized .XML PNR files using the International Air Transport Association (IATA) specification, ensuring compatibility with real-world airline systems.

Figure 1 below illustrate how the flow of the framework.

Figure 1. Framework Overview



Synthetic Population Generation

This stage follows a hierarchical data approach to ensure specific-country data is used as the primary resource for generating the synthetic population. When country-level data is unavailable, the model defaults back to regional or global data.

Input Data Preparation

The data required for this stage are aggregated population statistics and aggregated household statistics which we derived from United Nations (2022 and World Bank (2023). This data pre-processing will be used as the basis to generate head of household (HOH), and fundamental household types (HH Type). Household types are categorized into five groups: one-person households, couples without children, couples with children, single parents with children, and a residual category for other or unidentified types.

Household Generation

The synthetic population generation process starts with calculating the number of households (HH) for each region given the input of the number of households based on the percentage number of households of a country in a region. The data of the number of households in a country is derived from (United Nations, 2022). Once the number of households per country is determined, household structures are generated using synthetic reconstruction methods, specifically Simulated Annealing (Murata & Harada, 2017). This method ensures that

the generated households align with the marginal distributions of four key parameters: the gender of the HOH, the age group of the HOH, household size, and household type. These distributions are derived from aggregated data provided by the United Nations (2022) and World Bank (2023)

Populating Household

Upon establishing household structures, we populate the household by assigning age and sex attributes to the member of household using iterative proportional fitting (Choupani & Mamdoohi, 2016). The gender of the HOH's partner is assigned probabilistically based on marriage and civil partnership data in the UK (Office for National Statistics UK, 2024), The partner's age is assigned within a range of ±5 years relative to the HOH's age (Wilson & Smallwood, 2008). For households with children, the ages of the children are assigned between 20 and 40 years younger than the parent (United Nations, 2024).

Person-level attributes, including first name, surname, date of birth, email addresses, and residential addresses, are generated using the Faker library (Barseghyan, 2022). To reflect immigration, a small proportion of households is assigned nationalities and country of birth that differ from their country of residence (Eurostat, 2022). Locale-specific tools are used to ensure that names and addresses align with the cultural and linguistic characteristics of each country, including transliteration for non-alphabetical scripts. While the Faker library generates valid phone number, addresses, and postcodes, its implementation does not ensure that a given address correctly corresponds to a specific postcode or city.

The synthetic population is then integrated with geographical and flight databases to assign households to specific cities using data from *Geonames* (Rowlingson, 2025) and Airports. Each household is assigned to a city based on the city's population size, with larger cities having a higher likelihood of selection. Home airports are assigned based on proximity to the city of residence and airport traffic volume, ensuring that households in larger cities and near busier airports are more likely to be assigned to those locations. The resulting completed HH and person attributes can be seen in Table 1.

Table 1. Person and Household Attributes

Person Attributes	**Remarks**
Person ID	Unique identification of a person, also used as traveling document ID
Surname & First Name	All lowercase and transliterate to alphabet
DOB & Age	Date of Birth with dd/mm/yyyy format
Gender/Sex	Sex of the person
Payment card information	Vendor of the card, account/card number, and expiry date. Only person above 18 will have this information
Contact (Email & Phone)	
DOCS Information	
Birth location (City & Country) *	Immigration status might be shown with different country of residency and nationality
Community ID	Later used for community/social network group
HH Attributes	
Household ID	Unique identification of each household
Household Size	Size of household, with range from 1 to 8
Country of HH*	
Type of HH	Single, couple, single with children, couple with children, other
Age of HOH	
Sex of HOH	F/M
Nationality*	Nationality of the household
Residency	Residency information of the household that includes address, city, post code, and country
Home Airport	IATA Code for the home airport

*****All country code use alpha-3 country codes

Social Network Generation

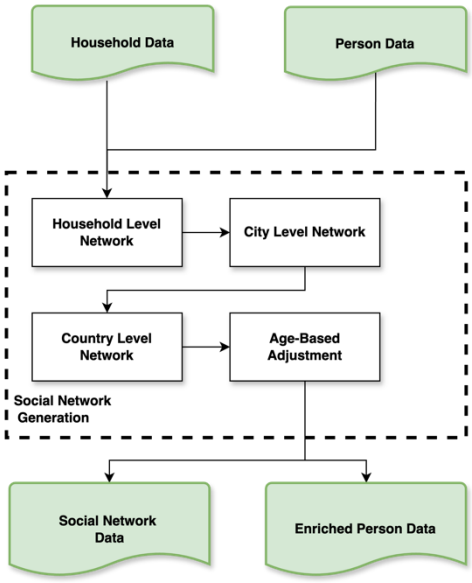
Input Data Preparation

Social networks are constructed to model interpersonal connections within and across households, loosely inspired by methodology of Jiang et al. (2022). Individuals used for this stage obtained from synthetic person data generated on previous stage are represented as nodes, while connections—such as family, friendship, and coworker ties—are represented as edges. The network is built in three stages: household-level, city-level, and country-level connections.

Social Network Generation

The network is built in three stages: First, familial connections are created by linking all members within the same household, assuming nuclear family structure and excluding extended familial ties. Next, city-level networks are established, where individuals within a city group (max 1000 members) can form friendship (20%) and coworker (2%) connections. Based on study by Jugert et al. (2013) and McGill et al. (2012) which highlight the tendency for individuals of the same nationality to form stronger and more stable connections, we assume that shared nationality increases the likelihood of forming connections by 50%, applied as a multiplicative factor. Each city-level network is labelled with a _CI identifier. At the country level, household heads (50% sampled) are linked across cities, but at lower probabilities (50% of city-level values) to reflect reduced interaction over greater distances. These country-wide connections form _CO communities that capture long-distance associations. Next, age-based adjustments are applied to represent homophily in social network (McPherson et al., 2001). Minors (<18) and older adults (>65) have a lower probability (0.01 and 0.05, respectively) of forming external connections, while non-familial edges between individuals with an age gap exceeding 20 years are removed. The resulting social network information (_CI and _CO attributes) is integrated into each person's profile.

Figure 2. Social network generation flowchart



Please note that parameters used to form maximum member, friendship and coworker likelihood, HOH sampling ratio, and age-based adjustment are designed to be configurable, allowing users to integrate their domain-specific expertise for more refined synthetic data.

Agent-based Trip Planning

Input Data Preparation

The stage starts by loading the social network data from previous stage, enriched person data, and a Geocrosswalk data, and flight data. The Geocrosswalk data supports the lookup process for detailed airport information, including IATA codes, country, and country codes. Flight data is used to determine destination probabilities based on airport traffic.

Please note that parameters used to determine stay duration, travel propensity, travel decision, trip type, travel mode, travel purpose, are designed to be configurable.

Agent Initialisation

Agents are initialized from population data, with each individual aged 18 and above acting as an autonomous agent. Agents are using assigned demographic attributes such as age, gender, household information, frequent flyer status, and network attributes (City_community_ID and Country_community_ID). Each agent maintains a global tracker containing:

- Last trip date
- Last known airport
- Last known country
- A travel history that stores details of past trips

The process of generating an agent's travel event involves five interdependent sub-processes:

1. Travel propensity - each agent has a likelihood of travelling a given number of times per year, this is highly country specific and dependent on an individual's economic status. It is important to note that in practice travel propensity does not dictate that an agent must travel within the specified period. Instead, it represents the minimum cooldown period before they can consider booking their next trip.

2. Trip Type and Travel Mode - Agents select the trip type and travel mode based on predefined probabilities. Trip type is a choice between return trips or one-way trips, travel mode determines if the type is solo or in a group, and travel purpose determines if the trip is for business or leisure.

3. Companion Selection - given the choice to travel in a group the agents identify companions from their social network based on the travel purpose: selecting coworkers for business travel and family and friends for leisure travel. Companions must be available for the same period and share the agent's location, as tracked by the global tracker. The global tracker is updated every time an agent travels or is selected as a companion for another agent's trip. If no companions are found, the agent defaults to solo travel. The maximum group size, including the main traveller, is capped at seven.

4. Stay Duration - The stay duration is also based on the trip purpose, with business trips being shorter than leisure.

5. Destination Selection - The destination selection process filters and selects destinations for each agent based on flight data, origin airport and country of residence. The algorithm follows these steps:

    - Filtering Destinations: Destinations within the same country as the origin and the origin airport itself are excluded.

    - Probability Calculation: The probability of selecting a destination is based on its popularity, represented by the total sum of flight capacity to that destination. Destination capacities are normalized to create a probability distribution.

    - A destination is selected through weighted random sampling. Destinations with higher capacities have a greater probability of being chosen.

    - The selected destination airport and country code are returned. If no valid destination is found, the algorithm returns None.

The trip planning process is conducted every day. Agents with a scheduled trip on the current day proceed to plan trips and generate travel bookings. The result of each trip plan itinerary is described in Table .

Table 2. Trip Attributes

Trip Attributes	Remarks
Booking ID/Trip ID	Unique id for all trips. Also used as booking ID
Main Passengers	Person ID of the main passenger. The person's email, phone, payment information, address will be used as the primary contact information of the booking
Companions / List of Passengers	List of Person ID for the companions of this trip, including the main passenger
Number of Passengers	Number of persons in this trip
Solo/Group	Solo or Group mark
Travel Purpose	The purpose of the travel - Business, Holiday, Visiting Family and Relatives (VFR), Others
Return/One-way	Return or One-way information of the trip
Travel Date	Start date and the end date of the trip. End date is the return date of the trip. If the trip is one way, then the end date is empty
Stay Duration	Duration (days) in the destination. The value is empty of one-way trip. Also consist of start and end date of the trip.
Airport (Destination and Origin)	IATA Code of airport

Flight Selection

Input Data Preparation

The stage starts by loading trip data from the previous stage also flight data and route information derived from Strohmeier et al. (2021). The flight data is pre-processed and filtered to include only commercial passenger aircraft. Trip data includes trip-level details such as origin, destination, travel dates, and group size. Flight data contain information on origin and destination airports, flight date and time, capacity, and occupancy. Routes define possible connections between airports, with only routes that operate for a minimum of 210 days per year being considered.

Please note that parameters used to determine ticket purchase timing, baggage weight, ticket pricing, and payment method are designed to be configurable.

Route Selection

For each trip, the system identifies possible routes between the origin and destination airports. Routes are divided into legs, with each leg representing a direct flight or connection between two airports. For example, a trip between Changi Singapore and London Heathrow can be done with several route options:

- One direct leg (SIN-LHR)
- Two legs via Dubai (SIN-DXB-LHR).

If no valid routes are found, the booking is marked as incomplete.

Flight Search and Booking

The flight search process is responsible for allocating flights to bookings based on availability, date constraints, and layover windows. The process starts by retrieving all available flights for a given origin-destination pair (referred to as a leg pair). Flights that do not fall within the ±2-day search window are removed. For connecting flight, additional time constraints are applied to ensure that the selected flight departs within 1 to 8 hours.

Flights that meet the date and time constraints are then filtered based on seat availability. The system verifies that the flight has enough remaining capacity to accommodate the booking party size. If multiple flights satisfy the conditions, the system selects one using weighted probability, where flights with higher capacity have a greater chance of being assigned.

Once a flight is selected, the flight's occupancy counts and the number of associated PNRs is adjusted for each flight. The booking ID is added to the reservation list. If no valid flights are found, the booking is marked as "no flight", allowing for post-processing adjustments.

Post-Processing

The post-processing phase enriches flight booking data with attributes critical for modelling behavioural patterns and identifying security risks. Key enhancements include:

1. Travel Agency can link bookings to country-specific or international agencies enables monitoring of high-risk entities (e.g., agencies on watchlists) and detection of suspicious traveller-agency associations (e.g., frequent use of agencies in conflict zones).

2. Ticket Purchase Timing Before Flight data is derived from Groves & Gini, (2015) and Office for National Statistics UK (2023). Rokita, (2016) asserts passenger behaviour and decision making affect their timing on booking flight. In addition based on National Academies of Sciences, Engineering, and Medicine (2015) and (Palmer & Boissy, 2009) the price of the flight also drives when the passengers book the flight.

3. Anomalies in luggage weight paired with destination/return-date analysis can flag undeclared cargo, hazardous materials, or potential smuggling risks.

4. Cash usage, while not inherently criminal, is disproportionately linked to illicit activities(Riccardi & Levi, 2018) Simulating payment preferences helps identify transactions requiring enhanced scrutiny (e.g., cash payments for high-value leisure trips).

The Output of this process is the list of the flight data completed with the occupancy and the list of bookings of the flight and the list of the bookings complete with the list of passengers, trip information, and list of flights used by the booking.

XML PNR Data Generation

The XML-formatted PNR provides a structured representation of the synthetic PNR data. Each .XML file corresponds to a flight and contains multiple PNRs, with each PNR including passenger details.

The XML generation process follows these steps:

- For each flight, create a <Flight> element.
- Insert all PNRs (booking data) associated with the in an <PNR> element.
- For each PNR, insert passenger details (<Passenger> elements) using the corresponding person data.

RESULTS AND DISCUSSION

In this section, we present a validation of our framework by conducting a comparative analysis between the aggregated synthetic data and the original dataset. This comparison focuses on assessing the fidelity of the synthetic data in replicating key statistical properties of the source.

To assess the result was obtained by generating 1,792,293 households of 36 countries with 177 different nationalities with 4,675,477 individuals. The synthetic population then combined with European flight data from 1st January 2019 to 31 December 2019 from Strohmeier et al. (2021). For this purpose we include only those routes with almost daily flights and generate trips that either end or start in France or Greece, as several of the main partners of our project consortium is based in Greece, while France is included to provide a comparative case without specific selection bias.

Table 3 shows the stats specifically for flights between France and Greece.

Table 3. Flight Stats

Flight Stats	Value
Number of Flights	4250
Total Passenger Capacity	664,346
Total Occupancy (%)	378,257
Average Occupancy (%)	56.87
Unique Routes	4
Number of Countries	2
Number of Cities	2
Number of Airports	3

Population Comparison

To validate the demographic accuracy of the synthetic population, we compare the age-gender distribution of the generated data against real-world census data for France and Greece. The comparison is quantitatively assessed using the following metrics:

- Kullback-Leibler Divergence (KLD): Measures the similarity between two asymmetric probability distributions. The smaller the value, the more similar the probability
- Wasserstein Distance (WD): Evaluates the difference in cumulative distributions.
- Root Mean Square Error (RMSE): Highlights deviations.

As shown in ***Figure 2*** and ***Table*** 4, The KLD values indicate that the synthetic population maintains a high degree of similarity to real-world demographic distributions, with lower divergence values signifying better alignment. The Wasserstein distances suggest that the synthetic and real-world distributions are closely matched, particularly for female populations, which has WD, KLD, and RMSE lower compared to the Male population. Male Population's distribution is "protruded" around the age of 35-49 probably due to more male being assigned as the HOH in the first stage.

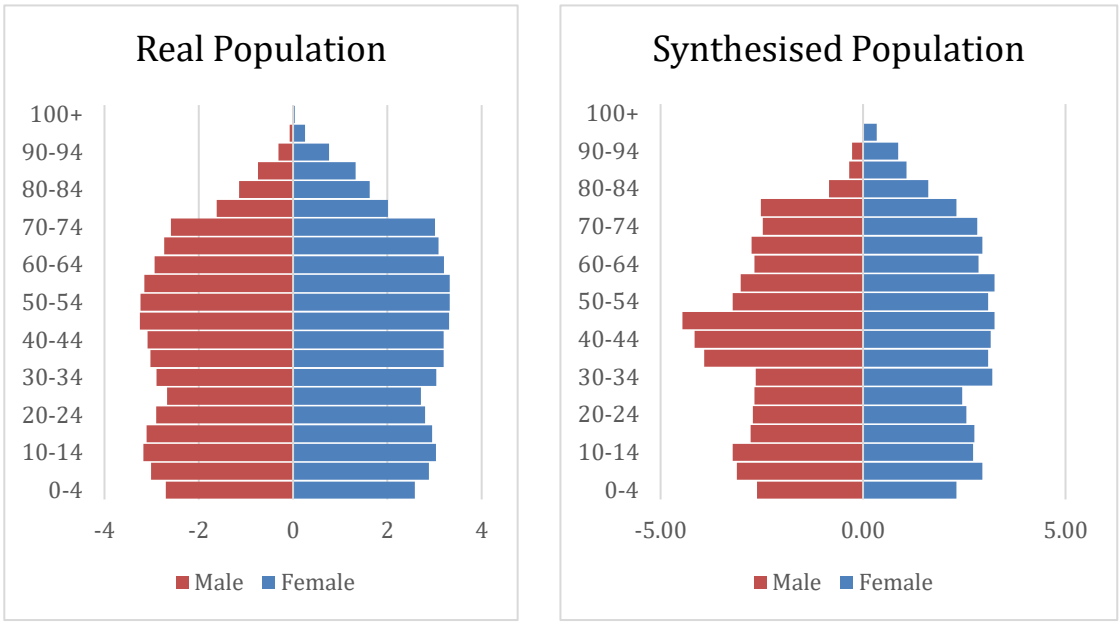


Figure 3. Population Comparison

Table 4. Population Comparison Test Result

Test	Value
KLD (Male)	0.048
KLD (Female)	0.015
Wesserstein Distance (Male)	0.277
Wesserstein Distance (Female)	0.146
RMSE (Male)	0.478
RMSE (Female)	0.195

Trip Comparison

The generated synthetic travel data was evaluated by comparing key travel attributes against real-world data. The comparison focuses on trip type (one-way/return), travel purpose, solo/group travel preference, average stay duration, and ticket purchase timing before flight. Based on Table 2 generally the synthesised data follows the real-life data trends, this suggests that the synthetic framework captures the general trend but slightly overestimates or underestimate the attributes.

Notable difference is in solo vs. group travel. The proportion of solo travellers is higher in the synthetic dataset (83.43%) compared to real-world data (70%), while group travel is significantly lower (16.57% vs. 30%). This suggests that the agent-based simulation may be generating more solo trips than expected. A likely reason is that finding companions during the trip-planning stage is more difficult, and when an agent fails to find a companion, the fallback mechanism defaults to solo travel.

Synthesised Baggage weight also produce larger average compared to reference. This might be caused by our algorithm that allows for very large baggage (50-100Kg) with 0.05% chance.

Table 2. Summary of Trip Data Comparison

One-way/Return	Real Percentage	Synthesised Percentage
One-way	40%	44.056%
Return	60%	55.940%
Travel Purpose	**Real Percentage**	**Synthesised Percentage**
Business	12.3%	8.937%
Holiday	64%	74.760%
VFR	20.9%	13.413%
Other	2.8%	2.889%
Solo or Group	**Real Percentage**	**Synthesised Percentage**
Solo	70%	83.428%
Group	30%	16.571%
Travel Purpose	**Real Average Stay (Day)**	**Synthesised Average Stay (Day)**
Business	4	3.664
Holiday	10	9.799
VFR	15	14.97
Other	7	6.626
Travel Purpose	**Real Average Ticket Purchase Timing Before Flight (Day)**	**Synthesised Average Ticket Purchase Timing Before Flight (Day)**
Business	7	6.5992
Holiday	30	29.545
VFR	20	19.294
Other	15	14.6
Travel Purpose	**Real Average Baggage Weight per Person (Kg)**	**Synthesised Average Baggage Weight per Person (Kg)**
Business	22.5	24.71
Holiday	27.5 or 35	32.110
VFR	30	47.601
Other	25	41.94

Flight Comparison

To evaluate the generated flight booking data, we compare the hourly, daily, and monthly flight occupancy patterns between the synthetic dataset and real-world flight records. The KLD score is used to measure the statistical similarity between these distributions. As shown in Figure 4, the generated flight data closely matches the real-world distribution in terms of hourly flight occupancy, with a KLD of 0.015 indicating the synthetic model effectively captures the hourly passenger flow dynamics.

However, discrepancies emerge when comparing daily occupancy trends. The generated flight data on Figure 6 deviates from real-world patterns over the long term, particularly in the later months of the simulation. This divergence is likely due to insufficient population size and imbalance passenger distribution over time.

KLD of month March – May in Figure 5 (0.071) is better compared to the later months as Figure 6, with the KLD for the full year trend is 0.160. This suggests a need for adjustments in trip generation algorithms, such as improving seasonality modelling and better constraints on travel frequency to maintain similarity throughout the entire year.

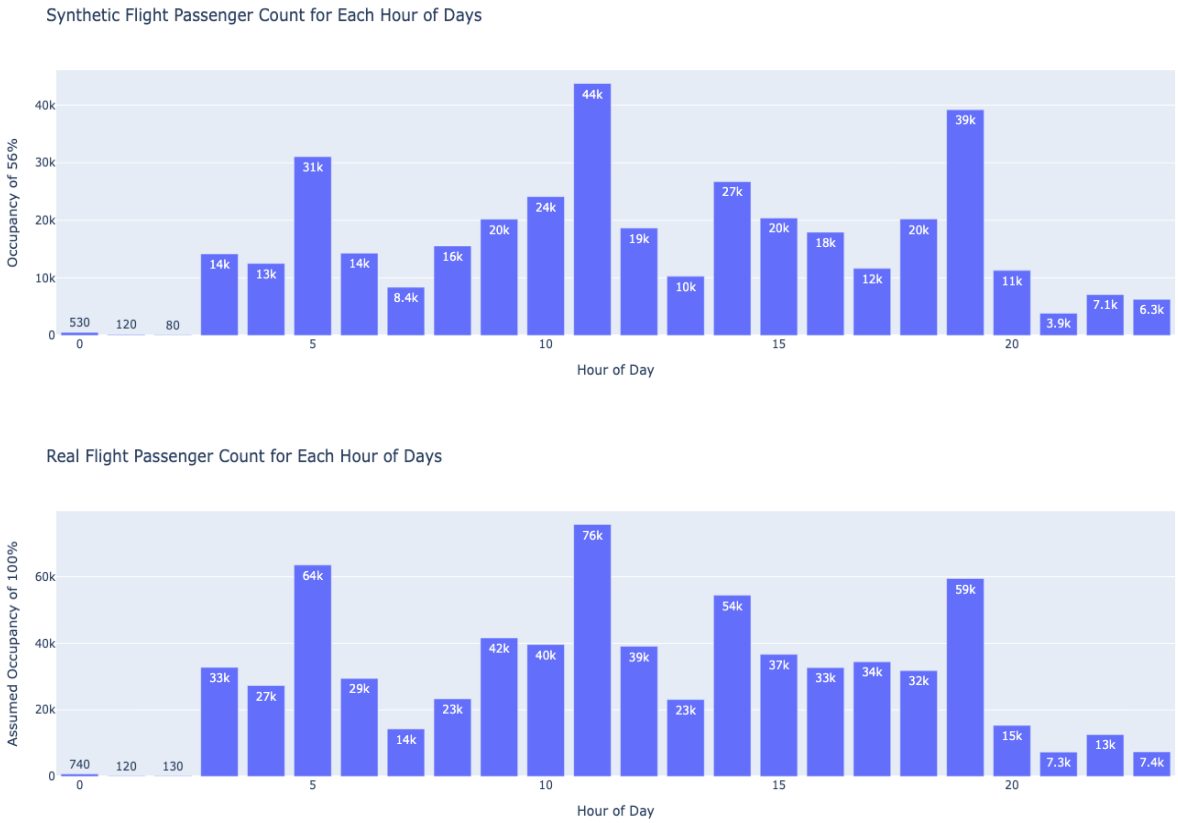


Figure 4. Hourly Flight Comparison

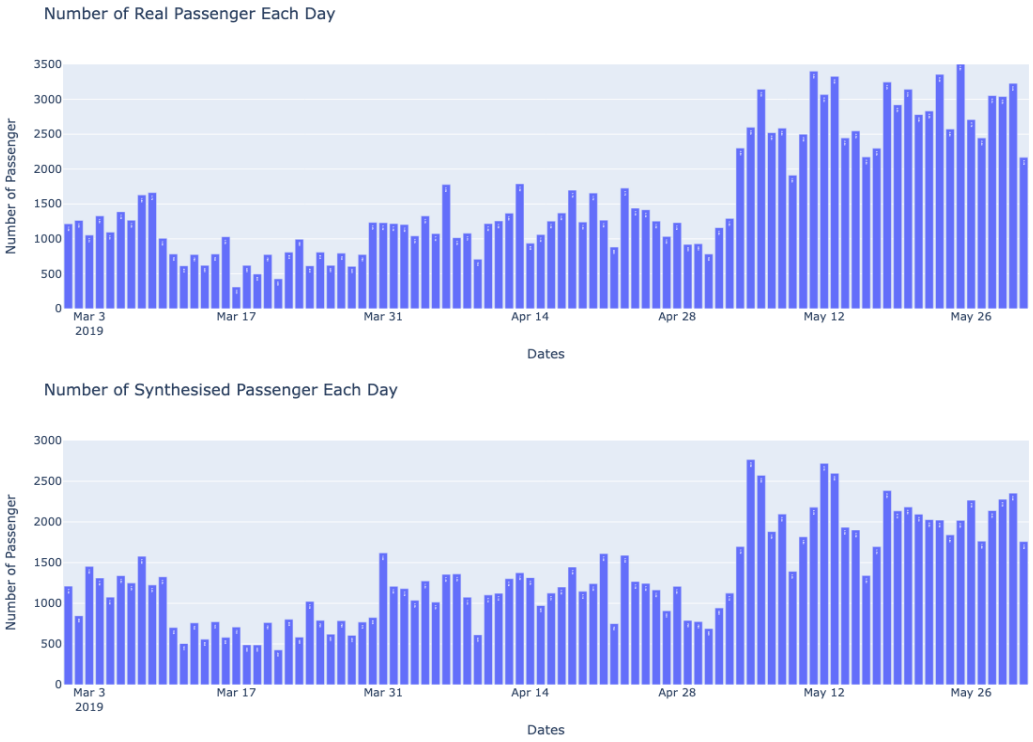


Figure 5. Daily Passengers (March-May)

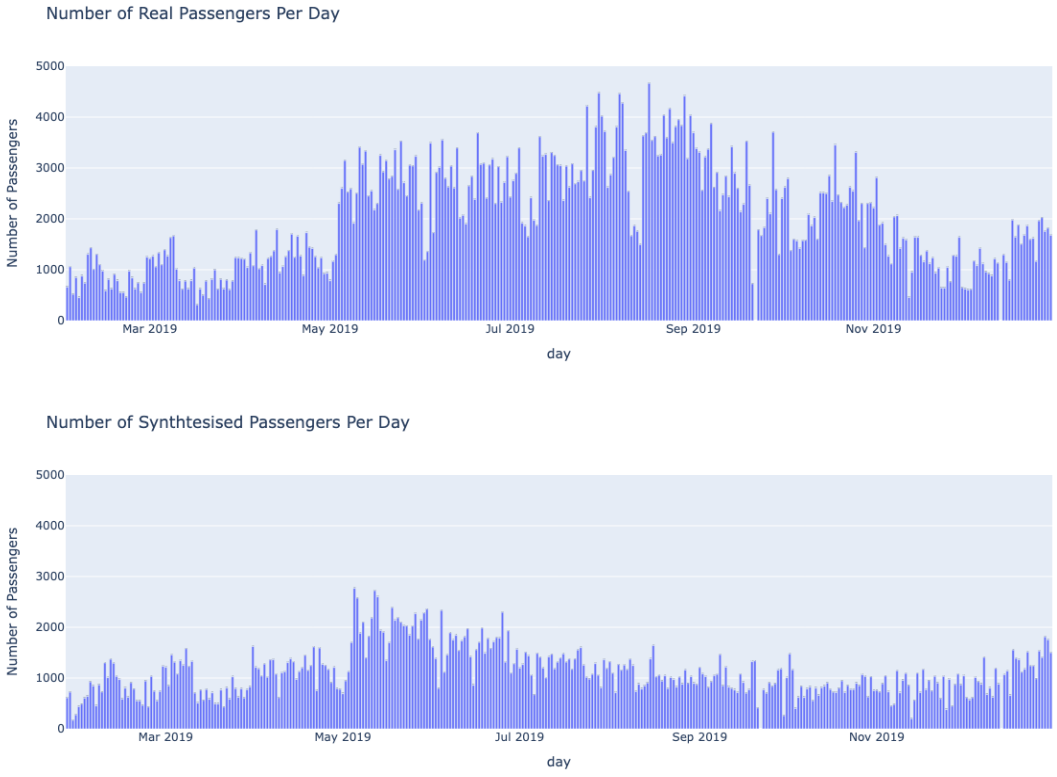


Figure 6. Daily Passengers (Full)

Limitations

Two main limitations affect the current approach. First, external validation by domain experts is required to assess the realism of the agent-based model. Second, computational constraints limit the population size and social network density, reducing the likelihood of group travel formation and affecting flight occupancy outcomes.

CONCLUSION AND FUTUREWORKS

In this work, we propose a framework to generate individual-level PNR data from aggregated data using synthetic reconstruction and agent-based methods. The framework is evaluated by comparing the synthetic data with real aggregated data across population characteristics, trips, bookings, and flights. The results show that the framework replicates key patterns in flight occupancy at the hourly level, with some deviations in daily and monthly trends, especially in later simulation periods. These are likely due to population size limits and uneven passenger distribution.

Future work includes increasing population size and social network density, refining behavioural rules, incorporating expert validation, improving seasonality modelling, and adding constraints to trip generation. Enhancements to temporal consistency and integration of external mobility data are also planned.

The code used in this paper is available online (Fadlian, 2024) and the data is available on request and is regularly maintained and updated.

ACKNOWLEDGMENTS

This work is part of the TENACITy project and has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101074048.

REFERENCES

Barnett, A. (2004). CAPPS II: The Foundation of Aviation Security? *Risk Analysis*, *24*(4), 909–916. https://doi.org/10.1111/j.0272-4332.2004.

Barseghyan, A. (2022). Faker-file: Create files with fake data. In many formats. With no efforts. In *GitHub repository* [Computer software]. GitHub. https://github.com/barseghyanartur/faker-file

Cavusoglu, H., Kwark, Y., Mai, B., & Raghunathan, S. (2013). Passenger Profiling and Screening for Aviation Security in the Presence of Strategic Attackers. *Decision Analysis*, *10*(1), 63–81. https://doi.org/10.1287/deca.1120.0258

Choupani, A.-A., & Mamdoohi, A. R. (2016). Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research. *Transportation Research Procedia*, *17*, 223–233. https://doi.org/10.1016/j.trpro.2016.11.078

European Commission. (2020). *Report from the Commission to the European Parliament and the Council on the review of Directive 2016/681 on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime* (Report No. COM(2020) 305 final). European Commission. https://home-affairs.ec.europa.eu/policies/law-enforcement-cooperation/passenger-data_en

Fadlian, M. (2024). *Synthetic PNR Generation* [Computer software]. https://github.com/fafadlian/Synthetic-PNR-Generation

Figueira, A., & Vaz, B. (2022). Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics*, *10*(15), 2733. https://doi.org/10.3390/math10152733

Glouftsios, G., & Leese, M. (2023). Epistemic fusion: Passenger Information Units and the making of international security. *Review of International Studies*, *49*(1), 125–142. https://doi.org/10.1017/S0260210522000365

Groves, W., & Gini, M. (2015). On Optimizing Airline Ticket Purchase Timing. *ACM Trans. Intell. Syst. Technol.*, *7*(1). https://doi.org/10.1145/2733384

Hörl, S., & Balac, M. (2021). Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transportation Research Part C: Emerging Technologies*, *130*, 103291. https://doi.org/10.1016/j.trc.2021.103291

Jiang, N., Crooks, A. T., Kavak, H., Burger, A., & Kennedy, W. G. (2022). A method to create a synthetic population with social networks for geographically-explicit agent-based models. *Computational Urban Science*, *2*(1), 7. https://doi.org/10.1007/s43762-022-00034-1

Joint ACI World-ICAO. (2025, January 28). *Passenger Traffic Report, Trends, and Outlook, Advisory Bulletins*. https://aci.aero/2025/01/28/joint-aci-world-icao-passenger-traffic-report-trends-and-outlook/

Jugert, P., Noack, P., & Rutland, A. (2013). Children's cross-ethnic friendships: Why are they less stable than same-ethnic friendships? *European Journal of Developmental Psychology*, *10*, 649–662. https://doi.org/10.1080/17405629.2012.734136

Li, Z., Zhao, Y., & Fu, J. (2020). *SYNC: A Copula based Framework for Generating Synthetic Data from Aggregated Sources* (No. arXiv:2009.09471). arXiv. https://doi.org/10.48550/arXiv.2009.09471

Mahendru, M., & Singh, A. (2023). Airline Customer Segmentation based on Complex Behavioral Approach using K-Mode and XG-Boost Algorithm. *2023 International Conference on Disruptive Technologies (ICDT)*, 685–689. https://doi.org/10.1109/ICDT57929.2023.10151011

McGill, R. K., Way, N., & Hughes, D. (2012). Intra- and interracial best friendships during middle school: Links to social and emotional well-being. *Journal of Research on Adolescence*, *22*, 722–738. https://doi.org/10.1111/j.1532-7795.2012.00826.x

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*, 415–444.

Mottini, A., Lheritier, A., & Acuna-Agost, R. (2018). *Airline Passenger Name Record Generation using Generative Adversarial Networks* (No. arXiv:1807.06657). arXiv. https://doi.org/10.48550/arXiv.1807.06657

Murata, T., & Harada, T. (2017). Nation-wide synthetic reconstruction method. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–6. https://doi.org/10.1109/SSCI.2017.8285394

National Academies of Sciences, Engineering, and Medicine. (2015). *Passenger Value of Time, Benefit-Cost Analysis and Airport Capital Investment Decisions, Volume 1: Guidebook for Valuing User Time Savings in Airport Capital Investment Decision Analysis*. The National Academies Press. https://doi.org/10.17226/22162

Office for National Statistics UK. (2023, May). *Travel trends estimates: UK residents' visits abroad—Office for National Statistics*. https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/datasets/ukresidentsvisitsabroad

Palmer, A., & Boissy, S. (2009). The Effects of Airline Price Presentations on Buyers' Choice. *Journal of Vacation Marketing*, *15*(1), 39–52. https://doi.org/10.1177/1356766708098170

Parker, R. A. (2017). An Agent-Based Simulation of Air Travel Itinerary Choice. *Procedia Computer Science*, *109*, 905–910. https://doi.org/10.1016/j.procs.2017.05.419

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410. https://doi.org/10.1109/DSAA.2016.49

Privacy Act of 1974: Implementation of Exemptions; Secure Flight Records; Proposed Rule. (2007). *Federal Register*, *72*(163), 48356.

Riccardi, M., & Levi, M. (2018). Cash, Crime and Anti-Money Laundering. In C. King, C. Walker, & J. Gurulé (Eds.), *The Palgrave Handbook of Criminal and Terrorism Financing Law* (pp. 135–163). Springer International Publishing. https://doi.org/10.1007/978-3-319-64498-1_7

Rokita, M. Z. (2016). *Decision Factors Affecting Online Flight Ticket Booking: Can Loyalty Programme Effects Be Observed?* [Master's Thesis]. Copenhagen Business School.

Rowlingson, B. (2025). *geonames: Interface to the 'Geonames' Spatial Query Web Service*.

Stevenson, P. D., & Mattson, C. A. (2019). The personification of big data. *Proceedings of the International Conference on Engineering Design, ICED*, *2019-August*, 4019–4028. https://doi.org/10.1017/dsi.2019.409

Strohmeier, M., Olive, X., Lübbe, J., Schäfer, M., & Lenders, V. (2021). Crowdsourced air traffic data from the OpenSky Network 2019–2020. *Earth System Science Data*, *13*(2), 357–366. https://doi.org/10.5194/essd-13-357-2021

Templ, M., Meindl, B., Kowarik, A., & Dupriez, O. (2017). Simulation of Synthetic Complex Data: The R Package simPop. *Journal of Statistical Software*, *79*(10). https://doi.org/10.18637/jss.v079.i10

United Nations. (2022). *Household Size and Composition | Population division*. https://www.un.org/development/desa/pd/data/household-size-and-composition

United Nations. (2024). *World Population Prospects (2024) – Processed by Our World in Data*. https://ourworldindata.org

Vallet, F., Hörl, S., & Gall, T. (2022). Matching Synthetic Populations with Personas: A Test Application for Urban Mobility. *Proceedings of the Design Society*, *2*, 1795–1804. https://doi.org/10.1017/pds.2022.182

Vandal, T., Kodra, E., & Ganguly, A. R. (2019). Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation. *Theoretical and Applied Climatology*, *137*, 557–570. https://doi.org/10.1007/s00704-018-2613-3

Vinod, B. (2008). The continuing evolution: Customer-centric revenue management. *Journal of Revenue and Pricing Management*, *7*(1), 27–39. https://doi.org/10.1057/palgrave.rpm.5160117

Wilson, B., & Smallwood, S. (2008). Age differences at marriage and divorce. *Population Trends*, *132*, 17–25.

World Bank. (2023). *Heads of households, female (% of households with a female head) | World Bank Gender Data Portal*. https://genderdata.worldbank.org/en/indicator/sp-hou-fema-zs

Zhao, Z., Zhang, C., & Guo, D. (2020). Analysis of Risk-Based Airport Passenger Classification with PSO-BP Neural Network. *2020 39th Chinese Control Conference (CCC)*, 7344–7349. https://doi.org/10.23919/CCC50068.2020.9188750

Zheng, Y.-J., Sheng, W.-G., Sun, X.-M., & Chen, S.-Y. (2017). Airline Passenger Profiling Based on Fuzzy Deep Machine Learning. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(12), 2911–2923. https://doi.org/10.1109/TNNLS.2016.2609437